Leveraging EHRI Online Editions for training automated edition tools

Floriane Chiffoleau, Hugo Scheithauer

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

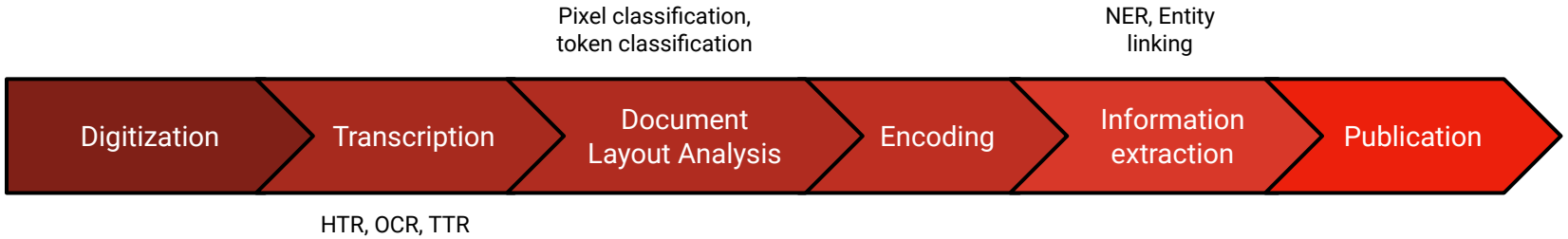
Hugo Scheithauer

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 ALMAnaCH 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 ?

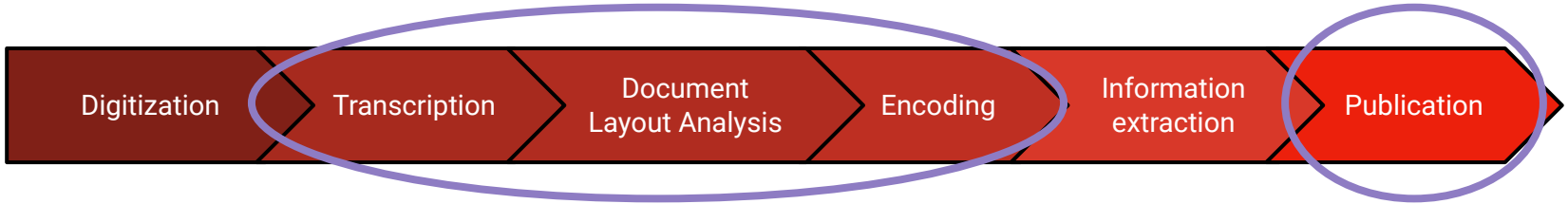
The EHRI consortium has supported the development and publication of six Holocaust-related online editions.

- **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 ?

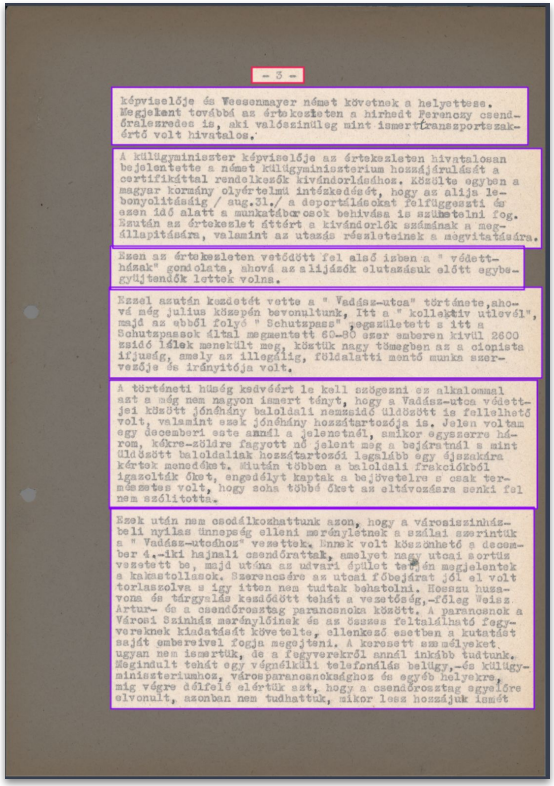
For today’s presentation, we will talk about:



SEMI-AUTOMATIC ENCODING WITH LAYOUT ANALYSIS

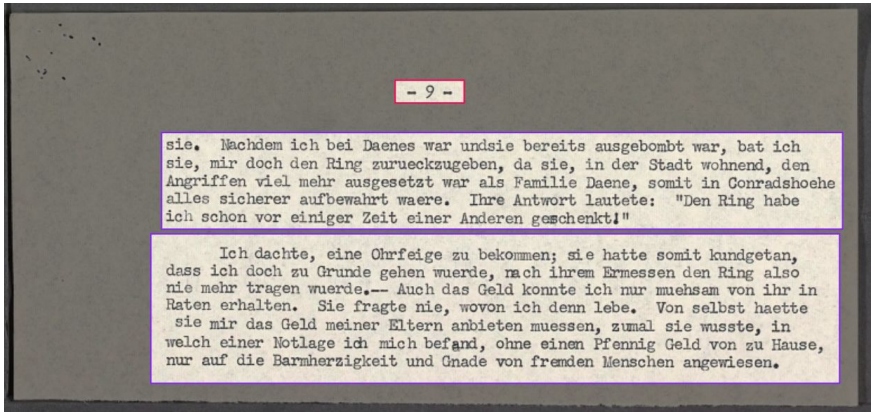
Semi-automatic encoding with layout analysis

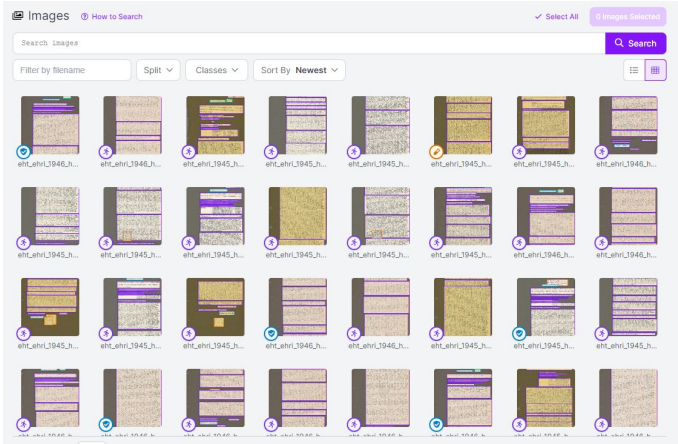
- Document Layout Analysis (DLA): a task that aims at detecting the layout components and hierarchy of a document.
- SOTA DLA systems are based on two approaches: **visual features** (YOLOv8), and a combination of **textual and visual features** (LayoutLM).



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).

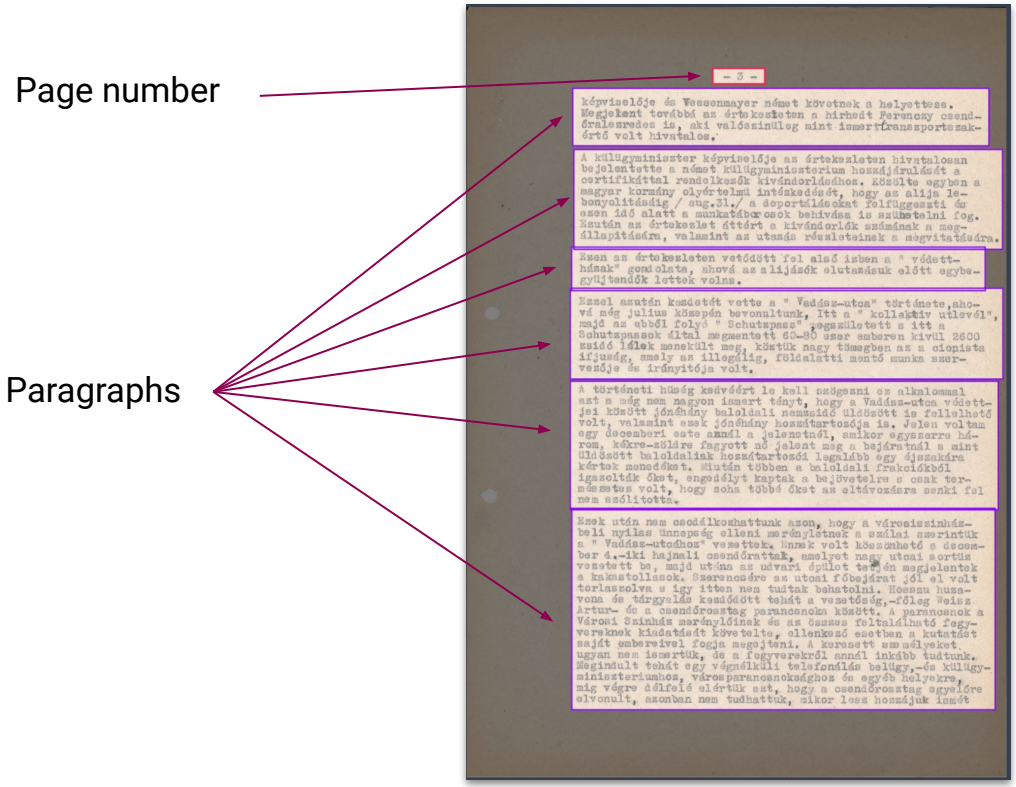




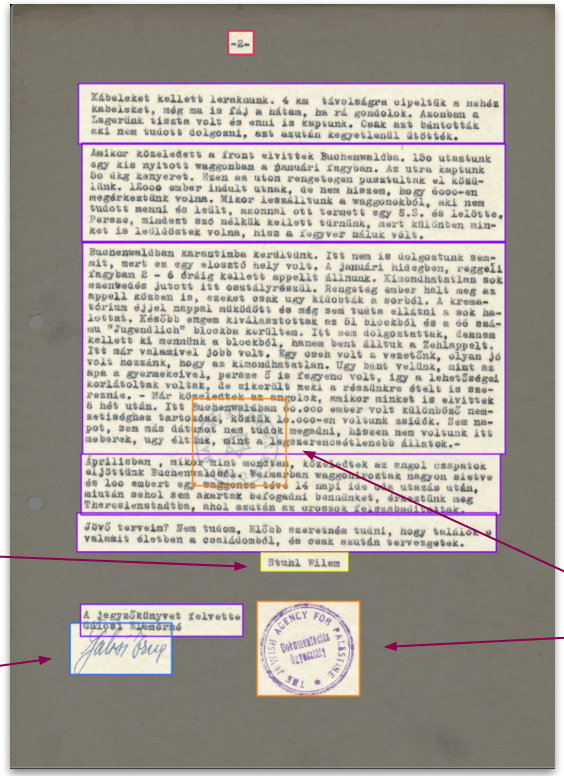
Thibault Clérice - You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. Journal of Data Mining & Digital Humanities, 26 décembre 2023, Documents historiques et reconnaissance automatique de texte - https://doi.org/10.46298/jdmdh.9806

Thibault Clérice, Juliette Janes, Hugo Scheithauer, Sarah Bénière, 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

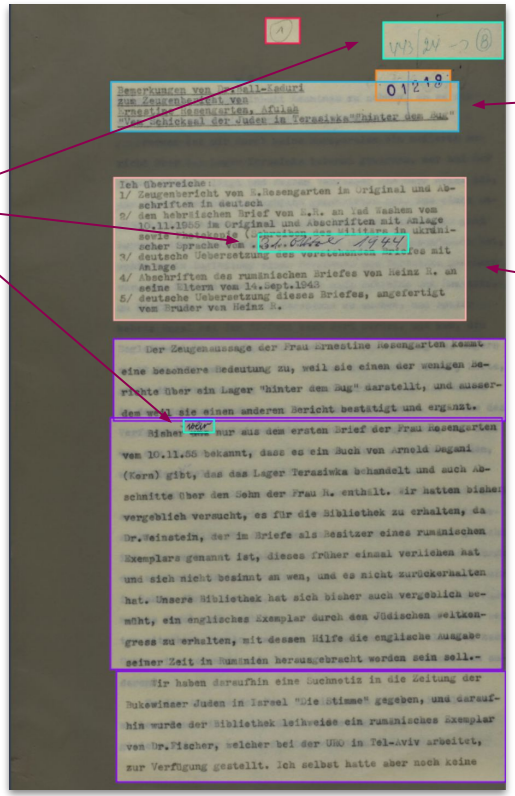


Semi-automatic encoding with layout analysis



Semi-automatic encoding with layout analysis

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



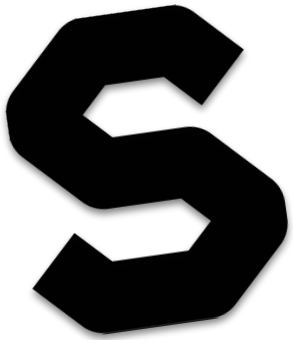
Semi-automatic encoding with layout analysis

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.

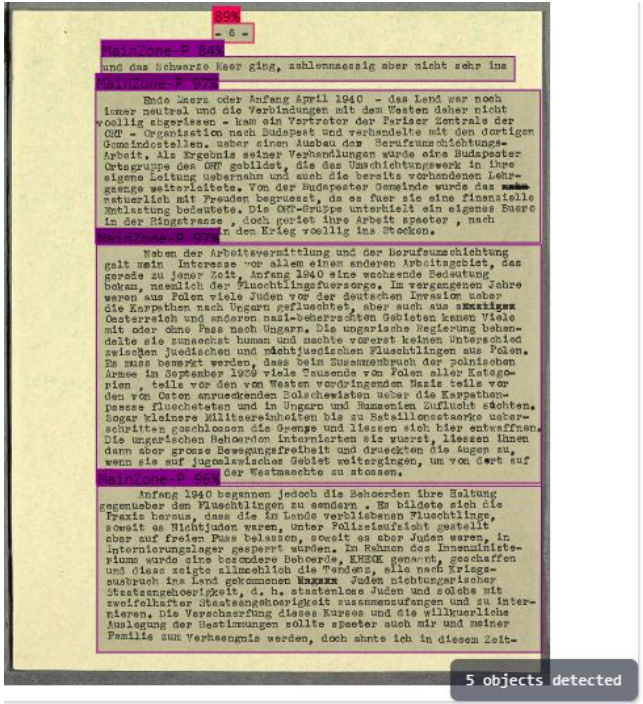
By using SegmOnto, the annotated dataset is easily interoperable and reusable.

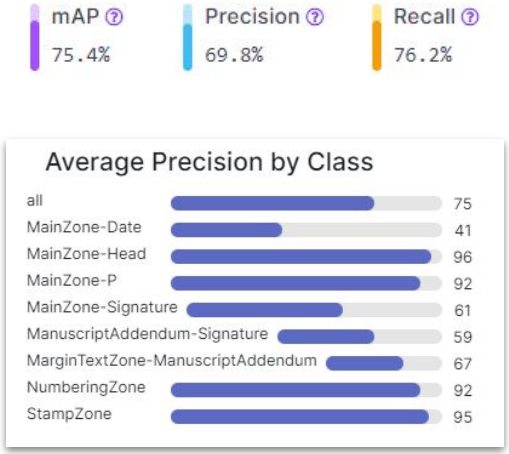
See SegmOnto's website: https://segmonto.github.io/ and Simon Gabay, Ariane Pinche, Kelly Christensen, Jean-Baptiste Camps. SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles. 2023. ⟨hal-04343404⟩.

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

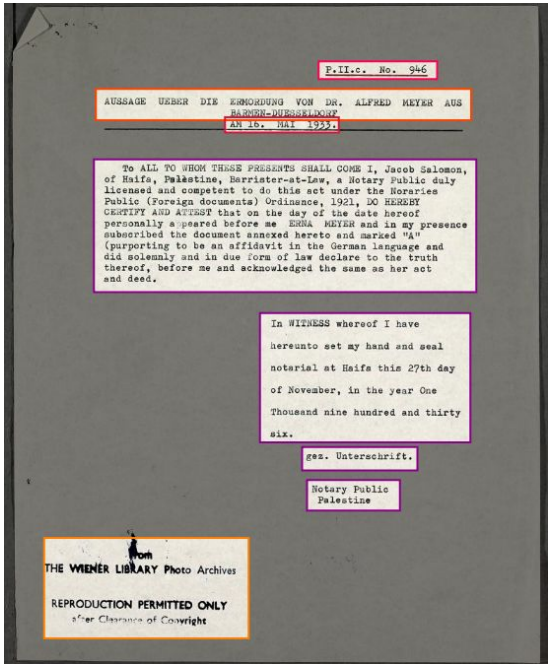
Semi-automatic encoding with layout analysis



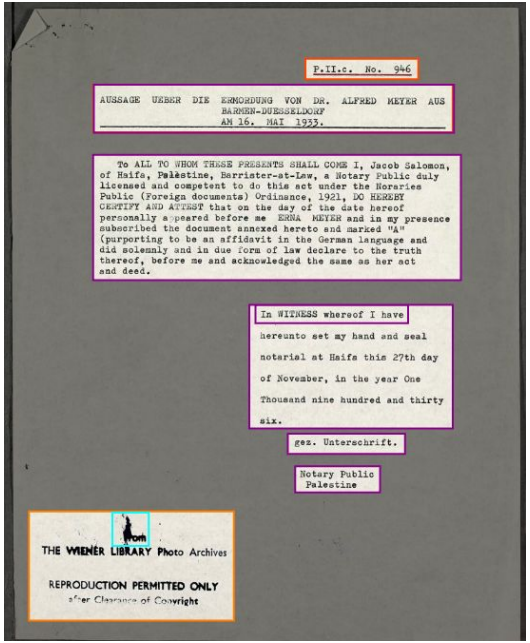


Preliminary experiment DLA model is available at:
https://app.roboflow.com/ehri/ehri-ladas/visualize/3

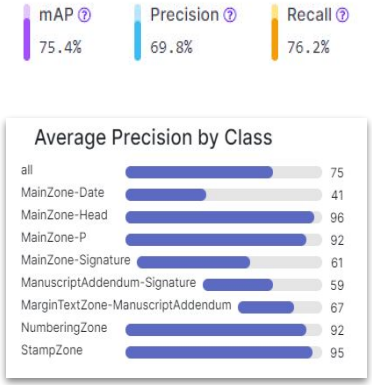
Semi-automatic encoding with layout analysis



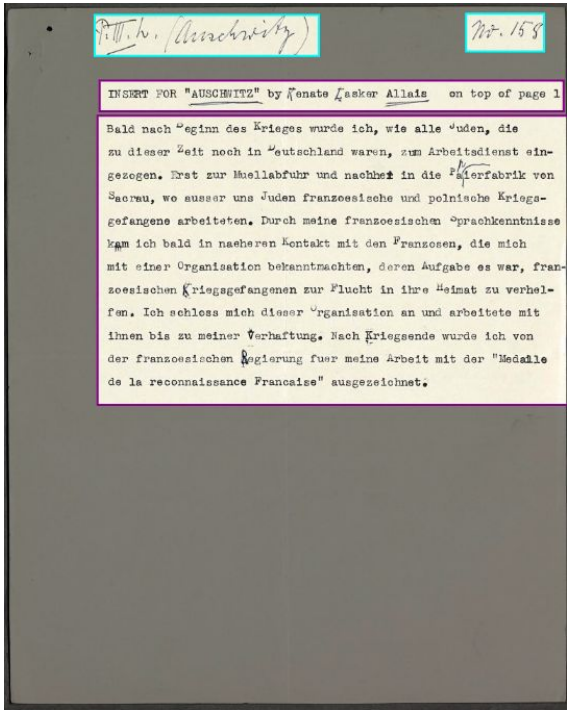
Ground truth



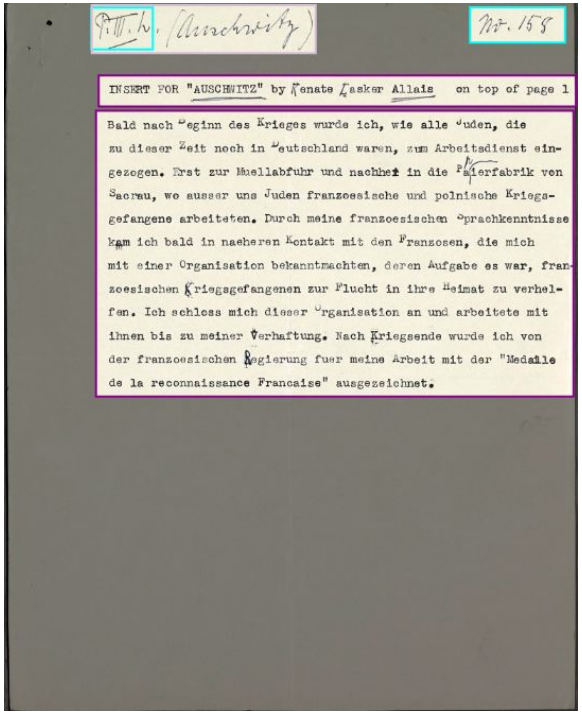
Prediction



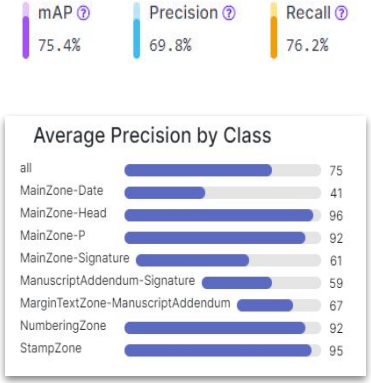
Semi-automatic encoding with layout analysis



Ground truth

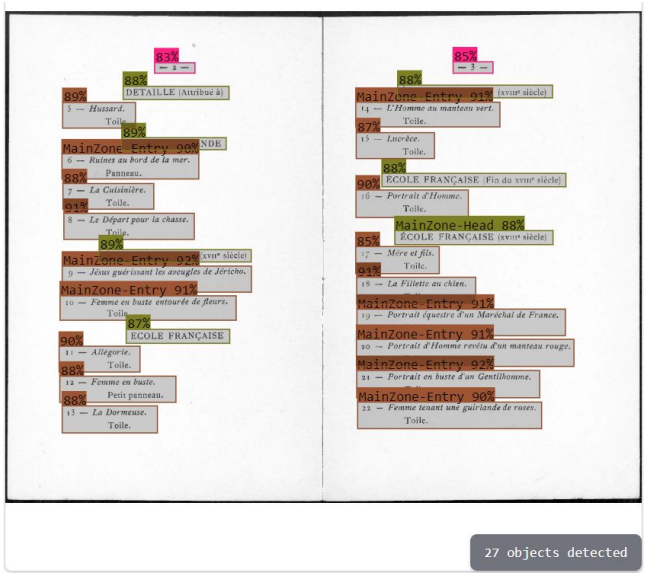


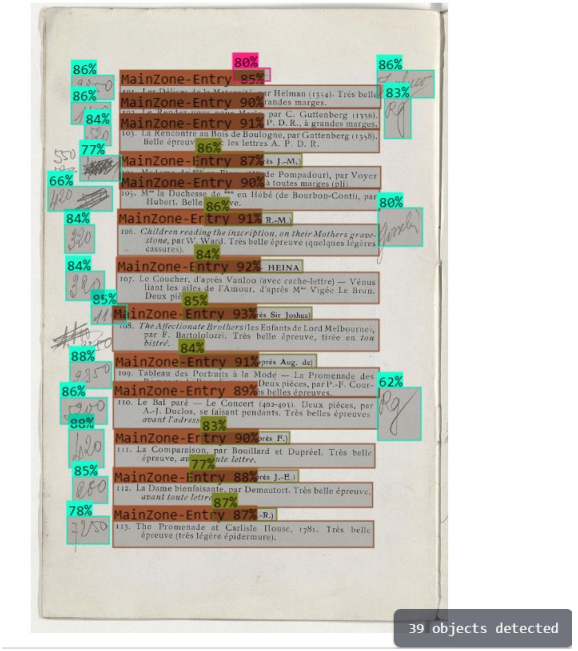
Prediction



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.

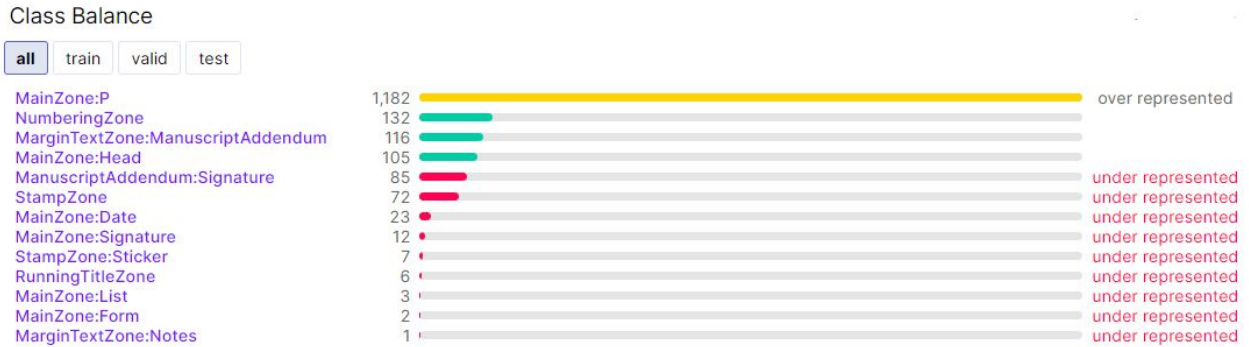




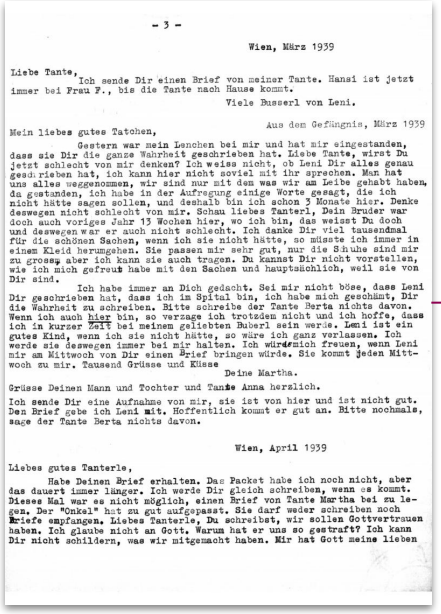
Semi-automatic encoding with layout analysis

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 Busserl von Leni.

Aus dem Gefängnis, März 1939

Mein liebes gutes Tatchen,

Gestern war mein Lenchen 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 geschrieben hat, ich kann hier nicht soviel mit ihr sprechen. Man hat uns alles weggenommen, wir sind nur mit dem was wir am Leibe 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. Denke deswegen nicht schlecht von mir. Schau liebes Tanterl, Dein Bruder war doch auch voriges Jahr 13 Wochen hier, wo ich bin, das weisst 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 verzage ich trotzdem nicht und ich hoffe, dass ich in kurzer Zeit bei meinem geliebten Buberl 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 würde mich freuen, wenn Leni mir am Mittwoch von Dir einen Brief bringen würde. Sie kommt jeden Mittwoch zu mir. Tausend Grüsse und Küsse

Deine Martha.

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.

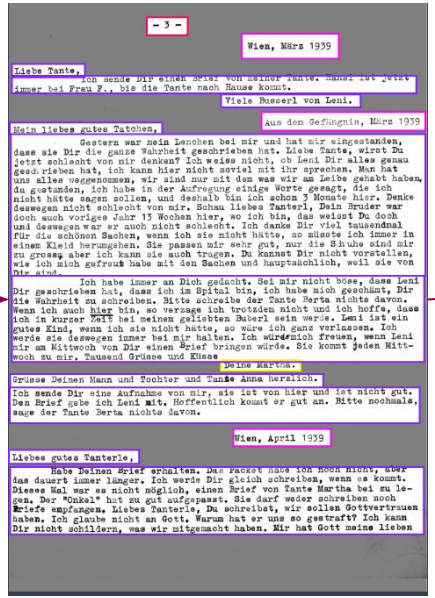
Wien, April 1939

Liebes gutes Tanterle,

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 legen. Der "Onkel" hat zu gut aufgepasst. Sie darf weder schreiben noch Briefe empfangen. Liebes Tanterle, Du schreibst, wir sollen Gottvertrauen haben. Ich glaube nicht an Gott. Warum hat er uns so gestraft? Ich kann Dir nicht schildern, was wir mitgemacht haben. Mir hat Gott meine lieben

EHRI-ET-WL1375B310

DLA



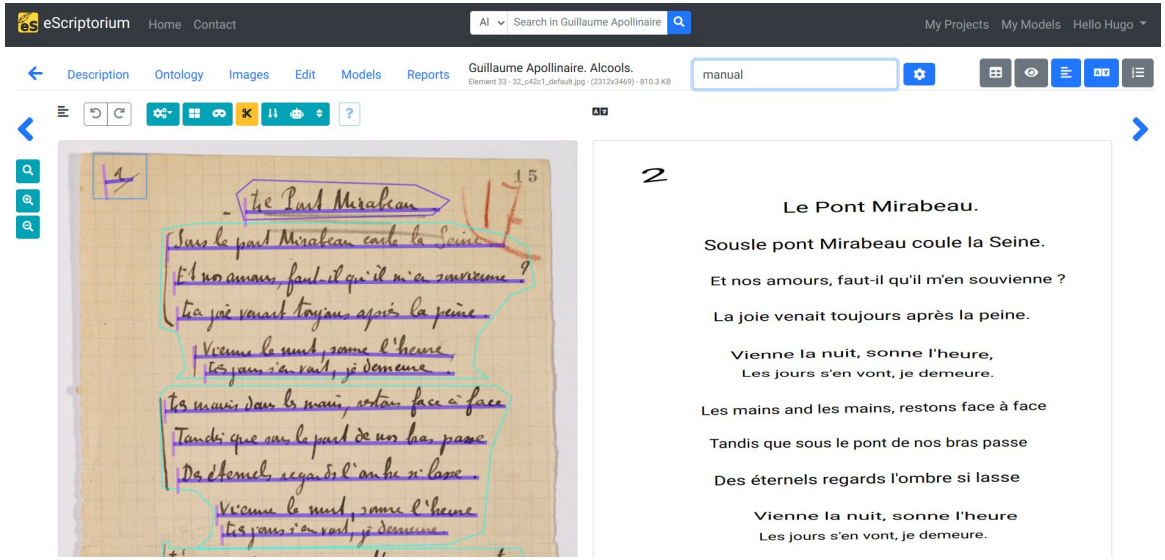
Automatic text recognition on segmented components

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!).



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

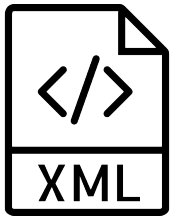
Creation of a multilingual recognition model

- 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**

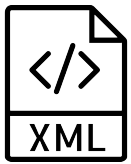


Text Encoding Initiative

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

Home Collections About History rechercher... Langue Français Se connecter

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

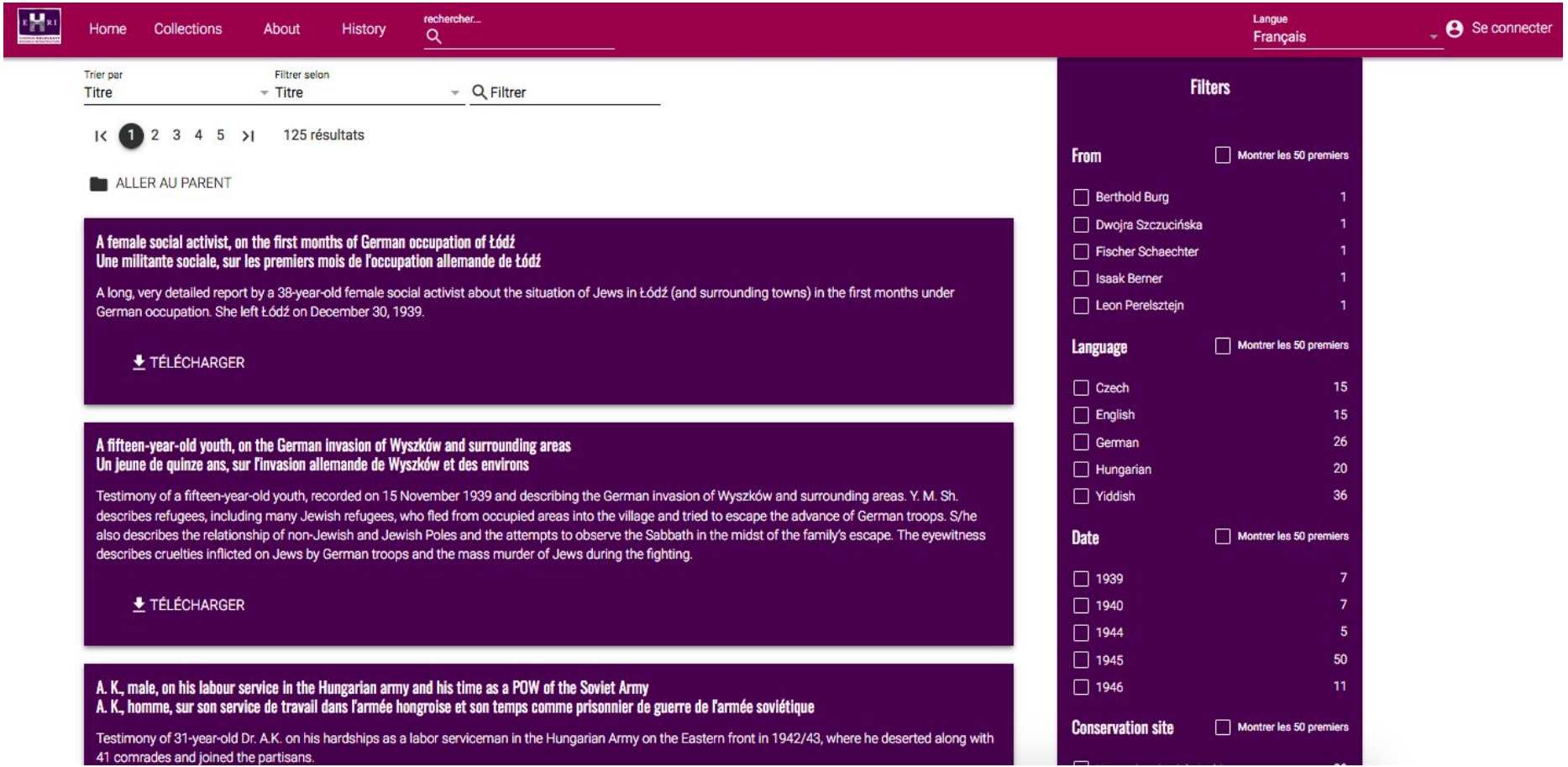
This edition gathers documents on the deportation of thousands of Jews to Nisko am San (Poland).

TEI Publisher **0.1** / web components **2.19.0** / API **1.0.0**

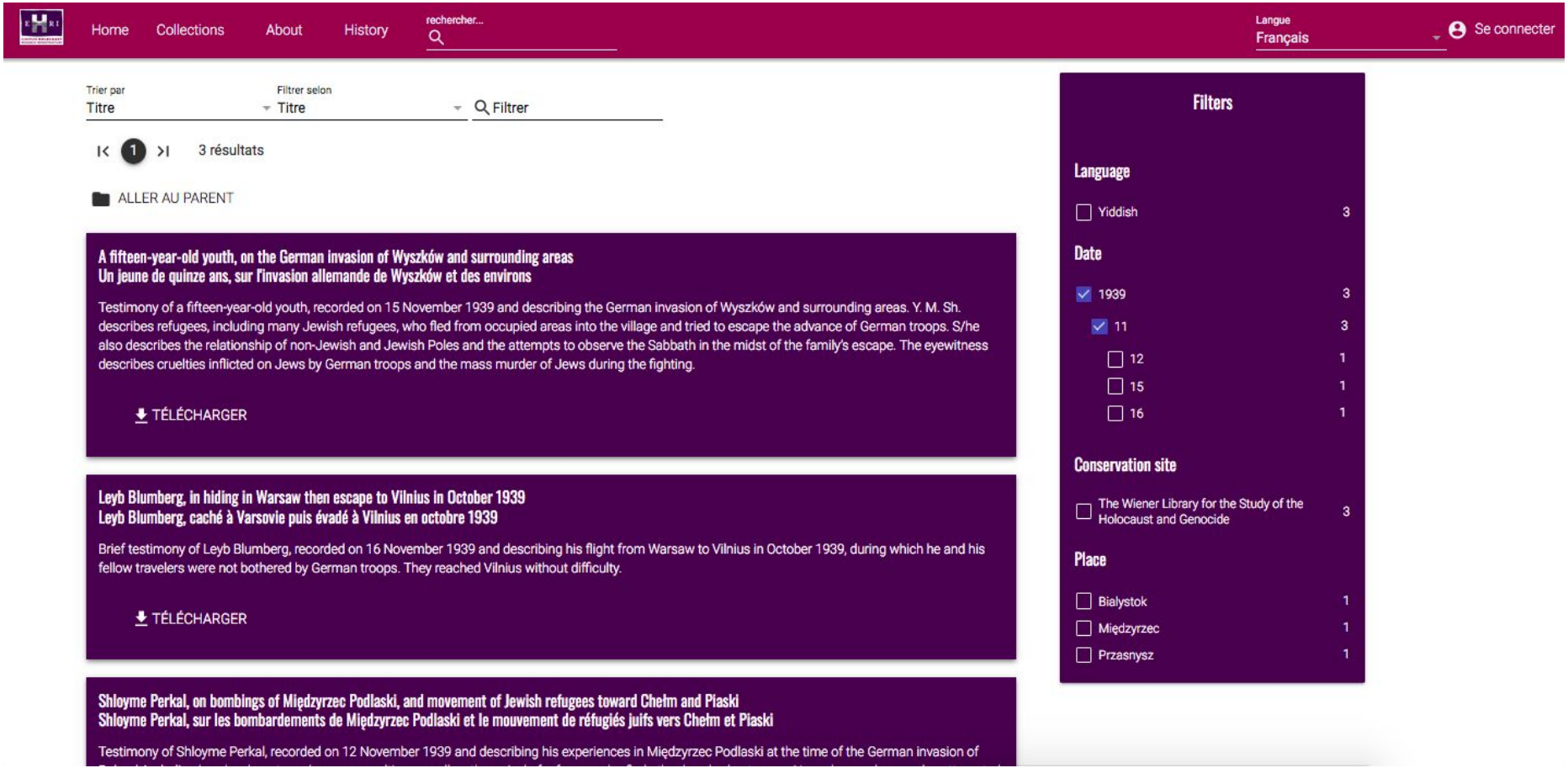
This application was developed as part of the EHRI project with the help of TEI Publisher.

Powered by editiones

Homepage of the TEI Publisher EHRI application



Early Holocaust Testimonies collection



Filter options for a collection

View
Translation

Committee to Collect Material

about the Destruction of **Polish** Jewry 1939

Y. M. Sh.

Wyszków

Age: 15 years; living with his parents

Testimony number 3

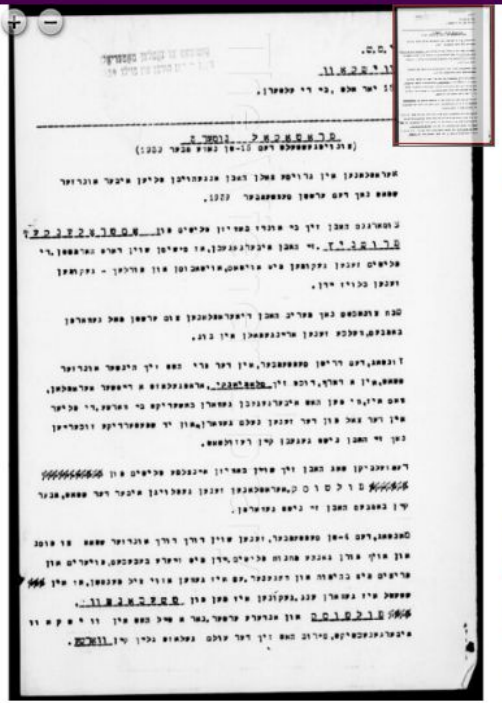
(recorded on the 15th of November, 1939)

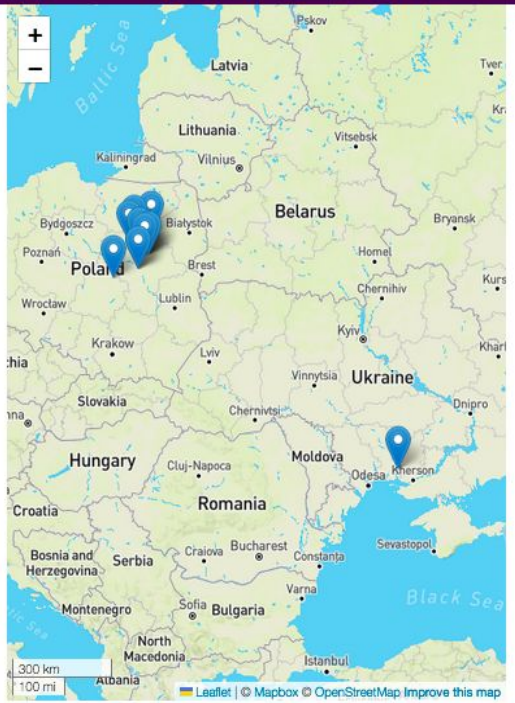
A lot of airplanes started to fly over our town after the 1st of September 1939.

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

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

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





Translated text with the facsimile and the map

Home Collections About History rechercher... Langue Français Se connecter

TABLE OF CONTENTS METADATA DISPLAY CITATION LICENCE NEXT PAGE

View
Original version

קאָמיטעט צו זאַמלען מאַטעריאַלן

וועגן ייִדישן חורבן אין **פוילן** 1939

ש.מ.י.

ווישקאָוו

18 יאָר אַלט, בייַ די עלטערן

פּראָטאָקאָל נומער 3

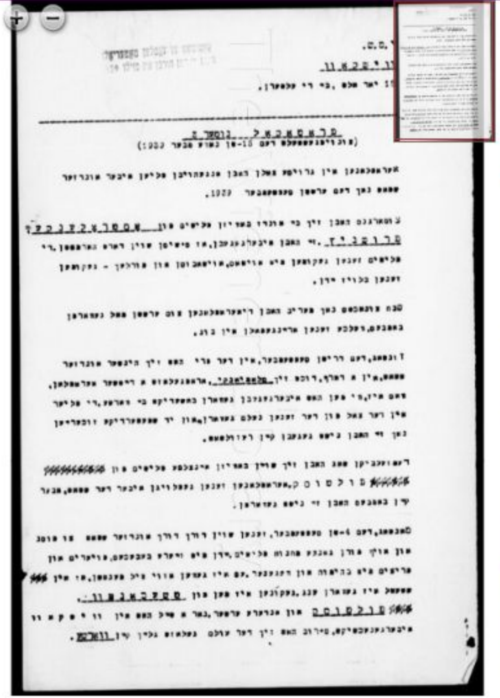
(צונויפגעשטעלט דעם 15טן נאָוועמבער 1939)

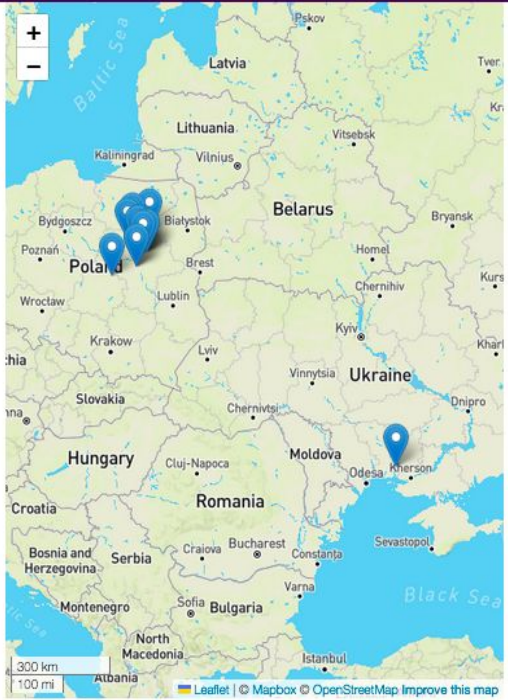
אַעראָפּלאַנען אין גרויסע צאָלן האָבן אָנגעהויבן פּליִען איבער אונדזער שטאָט נאָך דעם ערשטן סעפּטעמבער 1939.

צומאָרגנס האָבן זיך בייַ אונדז באַוויזן פּליטים פֿון **אָסטראָלענקע, פּרושניץ**. זיי האָבן איבערגעגעבן, אַז ס'שיסן שוין דאָרט האַרמאַטן. די פּליטים זענען געקומען מיט אויטאָס, אויטאָבוסן און פֿורלעך – געקומען זענען בלויז ייִדן.

שבת צו נאַכטס נאָך מעריבֿ האָבן די אַעראָפּלאַנען צום ערשטן מאָל געוואָרפֿן באָמבעס, וועלכע זענען אַרײַנגעפֿאַלן אין **בוג**.

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





Original text with the facsimile and the map

Home Collections About History rechercher... Langue Français Se connecter

PREVIOUS PAGE | TABLE OF CONTENTS | METADATA | DISPLAY | CITATION | LICENCE | NEXT PAGE

Metadata

Title

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

Time and place of writing

15.11.1939

Collection history

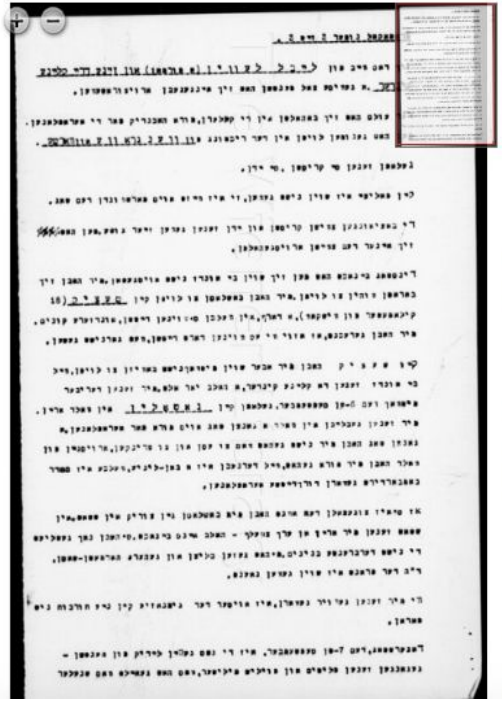
- Collection : Persecution of Jews in Poland: reports and statements, microfilm (coll. 532)
- Institution : The Wiener Library for the Study of the Holocaust and Genocide

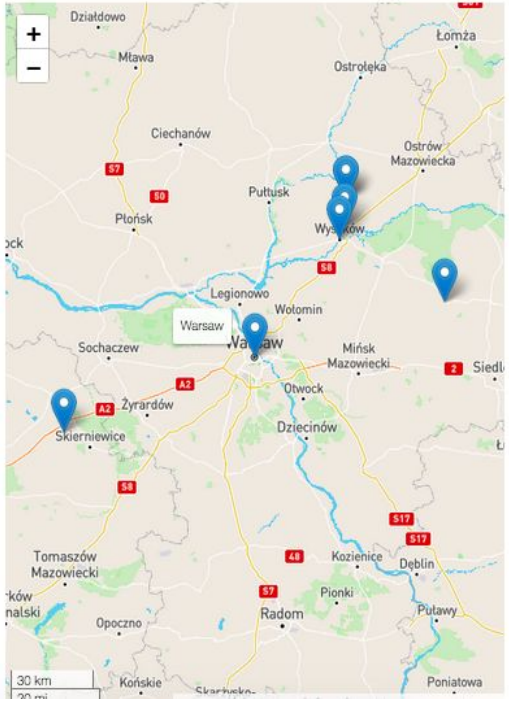
Encoding

- Project :

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

d 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 **Sieczychy**, because in





Display of the metadata on the TEI Publisher application

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

View
Translation

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

Levin, Leybl
Coachman from Wyszków.

People [illegible] of the airplanes. Other escaped in the directions of **Węgrow** and **Warsaw**.

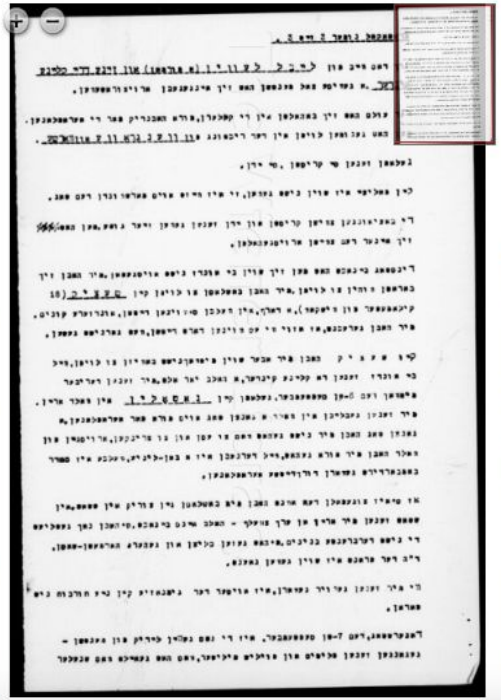
Both Christians and Jews were running away.

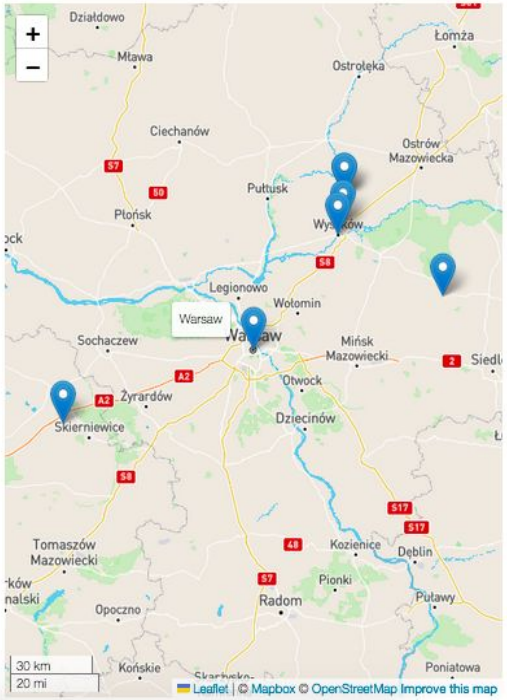
There were no policemen any more, apparently they disappeared the same day.

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

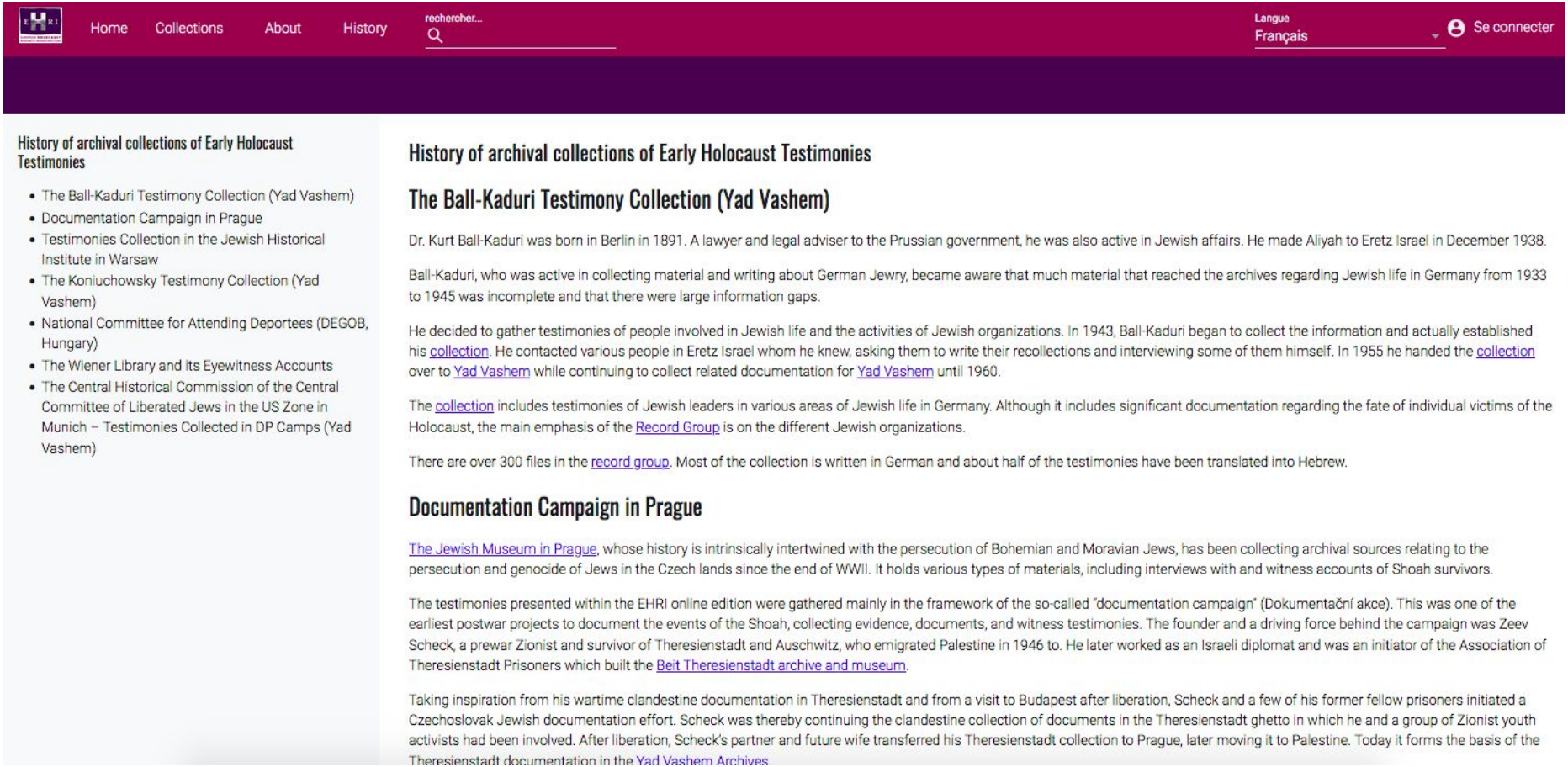
Tuesday in the evening we did not get undressed for the night. We discussed to where to escape. We decided to run to the village of **Sieczychy** (18 km from **Wyszków**), where some Germans were living, our customers. We believed that nothing would happen in a place where Germans were living.

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





Display of a named entity on the TEI Publisher application



HTML page of the "History of archival collections of Early Holocaust Testimonies"

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