



**HAL**  
open science

## An accessible and transparent pipeline for publishing historical egodocuments

Alix Chagué, Floriane Chiffolleau

### ► To cite this version:

Alix Chagué, Floriane Chiffolleau. An accessible and transparent pipeline for publishing historical egodocuments. WPIP21 - What's Past is Prologue: The NewsEye International Conference, Mar 2021, Virtual, Austria. hal-03173038

**HAL Id: hal-03173038**

**<https://hal.science/hal-03173038>**

Submitted on 18 Mar 2021

**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

# An accessible and transparent pipeline for publishing historical egodocuments

What's Past is Prologue : The NewsEye International Conference  
17 March 2021

DAHN Project  
Alix Chagué - Floriane Chiffolleau  
Research and Development Engineers at Inria



# WHY A PIPELINE ?

## ❖ Black box:

- does everything we want; but no interception of the data, no customization;



## ❖ Scattered toolbox:

- risk of obsolescence and discontinuity; portability of data not guaranteed;



## ❖ Pipeline:

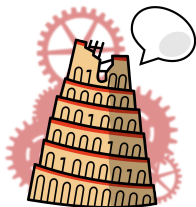
- clear steps, compatible softwares, extensive documentation
- data integrity, user in control



# THE DAHN PROJECT

*“Dispositif de soutien à l'Archivistique et aux Humanités Numériques”*

- ❖ Members: Inria + EHESS + University of Le Mans | MESRI;
- ❖ Scientific digital edition program for archival corpus;
- ❖ Goal: facilitate the digitization of data extracted from archival collections and their dissemination to the public in the form of digital documents and/or as online editions
- ❖ ALMANACH in charge of developing and/or enhancing the tools that are part of the pipeline.



*Inria*



**Le Mans  
Université**

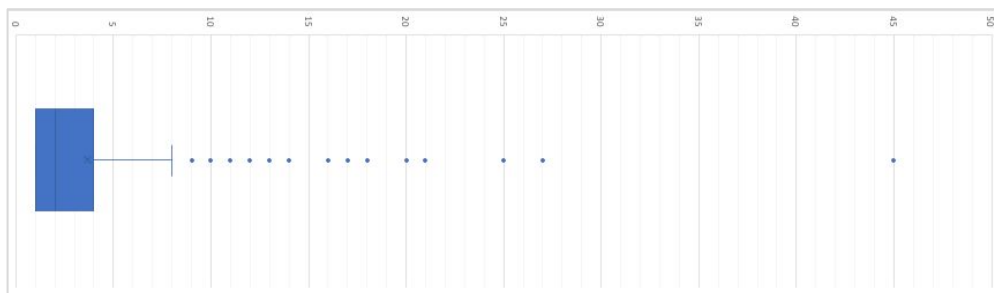


# THE CORPUS: PAUL D'ESTOURNELLES DE CONSTANT CORRESPONDENCE

- ❖ The context:
  - War reports
  - Exchange of opinions between two pacifists
- ❖ The corpus, a perfect test subject:
  - Typewritten documents: easier OCR, no HTR, faster work
  - Voluminous corpus: 1500 letters, 430 from the war period alone (from April 15, 1914 to November 19, 1918), one-to-several tens pages



The sender:  
Paul d'Estournelles  
de Constant



Distribution of pages per letter

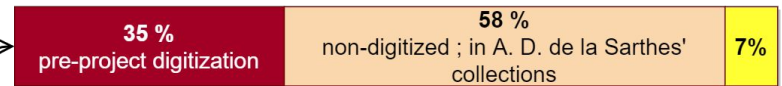
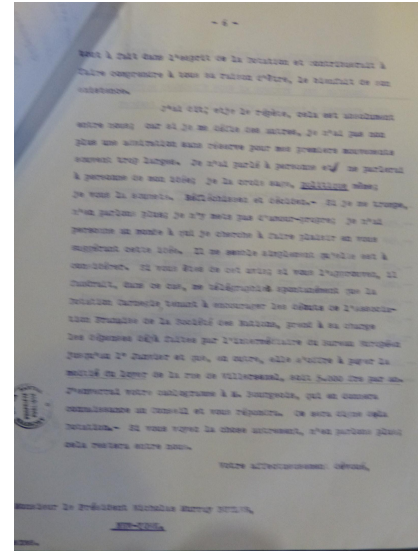
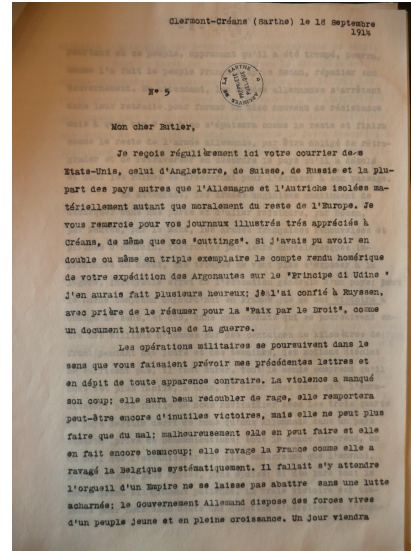


The recipient:  
Nicholas Murray  
Butler

# THE CORPUS: DIGITIZATION OF THE CORRESPONDENCE

An heterogeneous corpus of digitizations:

- ❖ Manual photo shooting of the corpus, by a researcher of the project, prior to its creation: medium quality images, case of blurred lines, cropped images and/or incomplete pages
- ❖ Institutional and exhaustive digitization campaign with proper material by the institution where the corpus is kept: high quality images

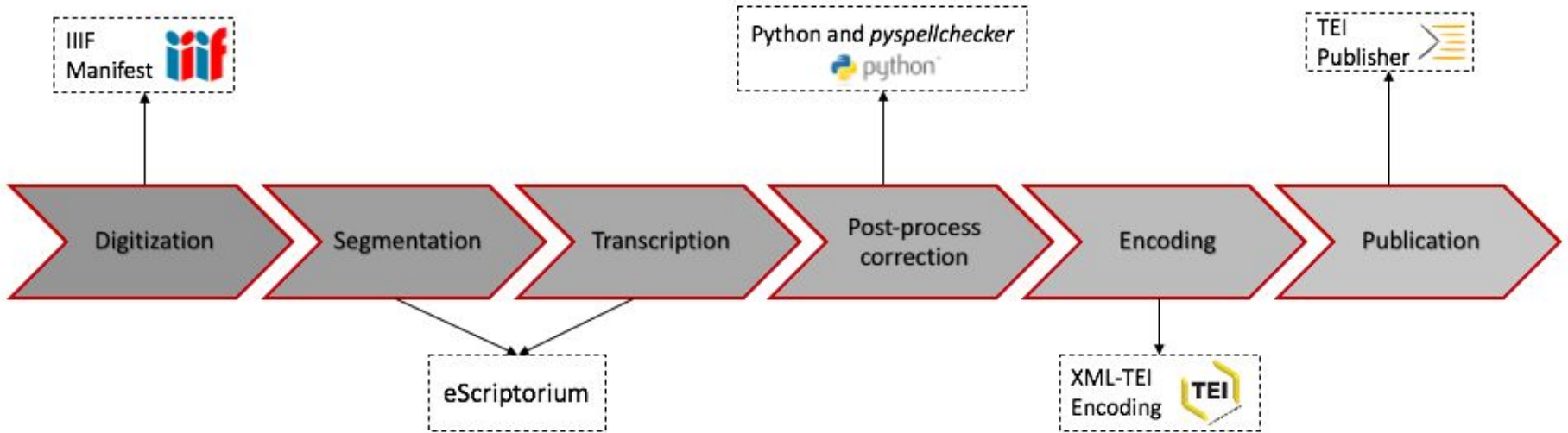


includes a few pages requiring a new digitization

only partially included in first wave of institutional digitization

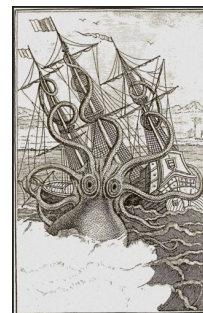
portion of the logical corpus held in another institution

# THE PIPELINE

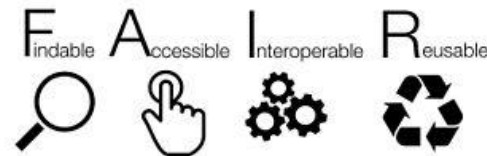


# THE OCR ENGINE: A CAPSTONE

- ❖ Powered by eScriptorium and Kraken
  - eScriptorium: web interface for collaborative and automatic transcription projects;
  - Historical ties between ALMAAnCH and SCRIPTA PSL (project team behind eScriptorium);
  - Intuitive platform and user-friendly on multiple features;
  - Entirely open-source and up-to-date formats;



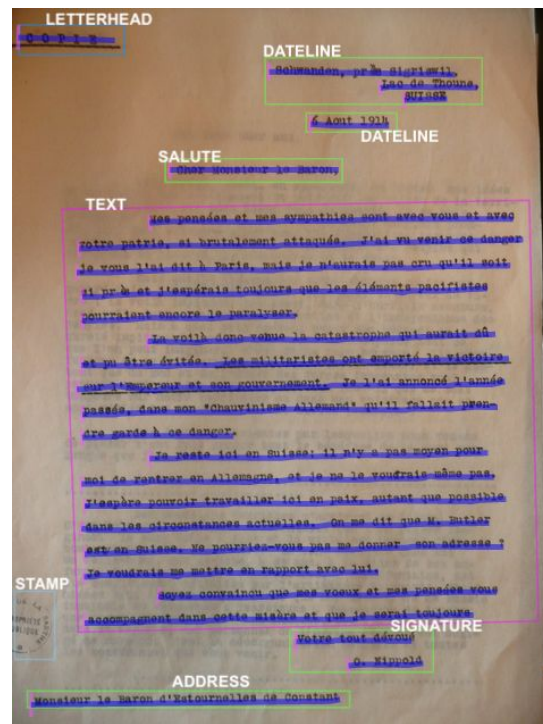
- ❖ Guiding principles:
  - Every step of the pipeline relies on open-source software or services;
  - Compliance with the FAIR principles.





# THE PIPELINE STEP BY STEP: SEGMENTATION

- ❖ Segmentation and layout annotation with a system of lines and zones tagging coupled with an ontology;
- ❖ Possible to build your own modelization and pass it on to a model;
- ❖ Integration of the TEI framework early-on;
- ❖ Participation of few members of the DAHN project in [SegmOnto](https://github.com/SegmOnto), a working group aiming at creating a general TEI-based ontology for HTR projects.



Example of an image annotated for Segmonto

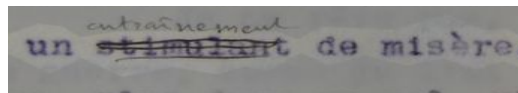
# THE PIPELINE STEP BY STEP: TRANSCRIPTION

## ❖ Training of a transcription model

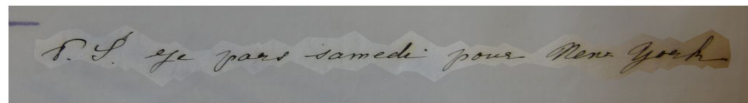
- Ground truth for the first model: 84 pages (10 letters) → not efficient enough
- Construction of a dataset more representative: focus on specific parts of text (capital letters, numbers, narrow text) - about 100 of additional pages → 92,74% accuracy
- Fine-tuning a pre-existing model (“tapuscorpus”, trained on varied typewritten docs): resulting model potentially more capable of generalization → 93,90% accuracy

## ❖ A model capable to render meaningful formatting of the text:

- Handwritten annotation → no recognition required, a double ‘££’ for every handwritten word
- Deletions → a ‘€’ at the start of the deletion and at the end, even if when spanning over several words



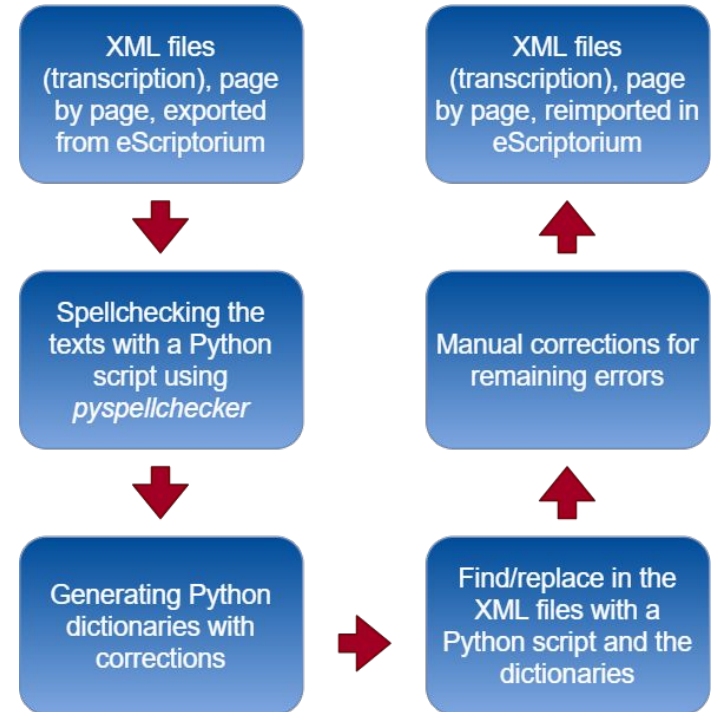
un €stimulant€ de misè



££ ££ ££ ££ ££ ££ ££

# THE PIPELINE STEP BY STEP: POST-PROCESS CORRECTIONS

- ❖ Export/import with eScriptorium
- ❖ Spell-checking:
  - Python script and *pyspellchecker* module
  - Generation of Python dictionaries for correction
- ❖ Correction of texts
  - Python script and find/replace
  - Manual correction of remaining errors
- ❖ Goal → implementation of this method in the eScriptorium API



# THE PIPELINE STEP BY STEP: ENCODING

## ❖ Header

- General encoding of the header, with manuscript description for the corpus
- Correspondence data in the <profileDesc>

## ❖ Body

- Default text structure (titles, paragraphs), changes in text (addition and deletion), difficulties of the corpus (unclear, gaps) and named entities (person, place, organization)
- Correspondence part:
  - Easy tags (opener, closer, address, signatures, etc.)
  - Assistance available for difficult tags: TEI mailing-list, [Correspondence Special Interest Group](#)

## ❖ Helping tools for the encoding

- Script for the encoding of metadata in the header, using an inventory with essential information
- Script for the encoding of the body, using regular expressions

# THE PIPELINE STEP BY STEP: PUBLICATION

## ❖ TEI Publisher

- From exist-db, an open source software project for databases built on XML technology
- Transformation files (ODD) and templates (HTML)
- Presentation by collections

## ❖ Application for ego documents

- Text and image opposed
- Metadata available with a specific button
- ODD containing all the tags related to ego documents
- One corpus = one collection



## Collection de corpus d'égodocuments

A screenshot of the TEI Publisher application interface. The page has a light blue header with the word "Documents" on the left. Below the header, there is a search bar with the text "rechercher..." and a magnifying glass icon. To the left of the search bar, there are options for "Trier par" (set to "Titre") and "Filtrer selon" (set to "Titre"). Below the search bar, there is a "SUPPRIMER" button with a trash icon. The main content area displays a list of document collections. The first collection is titled "Correspondance de d'Estournelles de Constant" and includes a portrait of Paul d'Estournelles de Constant. The second collection is titled "Correspondance des Intellectuels Berlinois" and includes a portrait of a man. The text below each title describes the content of the collection.

# FAVORING THE USE OF THE PIPELINE: DOCUMENTATION

- ❖ Initial documentation provided with the software, standards and framework
- ❖ Self-produced documentation
  - Comprehensive series of guidelines for the encoding of ego documents (ODD)
  - Jupyter notebooks and markdown files commenting the use of Python scripts (encoding and correction)
  - Provision of the data from the steps of the pipeline on the dedicated repository

# GENERALIZATION AND MODULARITY

- ❖ Corpus of the [Berlin intellectuals](#), correspondence of the start of the 19th century
  - Modularity: Plug data in without following the pipeline from the beginning
  - Generalization: No limit of space and time for the encoding and publication
- ❖ Sustainability
  - Dissemination of the data in useful formats
  - Ground truth available in [HTR-United](#)



## TO CONCLUDE: KEY TAKEAWAYS

- ❖ Scenario fully documented and based on open-source and standards to avoid blackbox and/or scattered toolbox
- ❖ Adaptability of the input and output formats of softwares
- ❖ Prerequisites to use the pipeline:
  - Digitized images
  - A transcription system such as eScriptorium/Kraken
  - A TEI modelisation of the structure of your documents
  - A server to deploy a TEI Publisher application



# THANK YOU FOR YOUR ATTENTION

## CONTACTS

[alix.chague@inria.fr](mailto:alix.chague@inria.fr)

[floriane.chiffolleau@inria.fr](mailto:floriane.chiffolleau@inria.fr)

## LINKS

Repository of the project: <https://github.com/FloChiff/DAHNPProject>

eScriptorium: <http://traces6.paris.inria.fr/>