Workshop 1
From a collection of documents to a published edition : how to use an end-to-end publication pipeline

Floriane Chiffoleau, PhD candidate at Le Mans Université (3.LAM) and Inria (ALMAnaCH), and Hugo Scheithauer, Research and Development Engineer at Inria (ALMAnaCH)
TEI 2022
September 12th, 2022

ALMAnaCH project-team

Inria

Le Mans Université

3L.AM
Langues, Littératures, Linguistique
Le Mans Université
Université d'Angers

INTRODUCTION

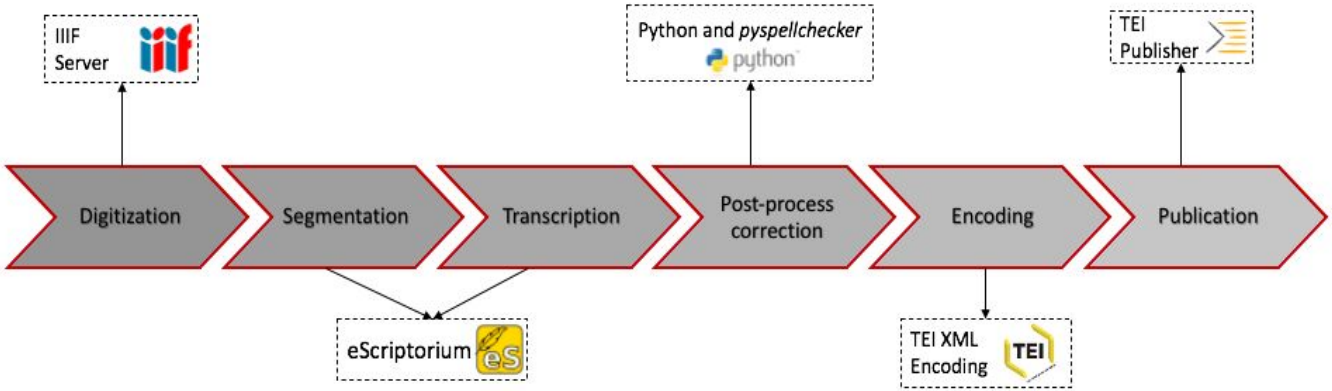
Pipeline for digital scholarly edition of historical documents

- Why did we create this pipeline ?

  → To facilitate the digitization of data extracted from archival collections, and their dissemination to the public in the form of digital documents in various formats and/or as an online edition

  → To make available an open system with open source tools, interoperable and easy-to-use

- What is it made of ?



A bit of background information about you

- ❏ Who already used a IIIF server/fetched a IIIF link ?
- ❏ Who already did automatic transcription (OCR, HTR, etc.) ?
- ❏ Who knows how to do transformation with XSLT ?
- ❏ Who knows how to do transformation with Python ?
- ❏ Who already encoded texts in XML TEI ?
- ❏ Who already used TEI Publisher ?

DIGITIZATION

Displaying a facsimile with IIIF

- What is IIIF ?
  - Stands for *International Image Interoperability Framework*
  - Standardized method of describing and delivering images over the web
- Why IIIF ?
  - The images are not directly in the publication platform, which alleviates its weights and gives leeway for more content
  - Putting them in a specific server ensures the sustainability of high-quality images

Displaying a facsimile: example of IIIF servers and links

- Gallica (Digital Library of the BNF): https://gallica.bnf.fr/accueil/en/content/accueil-en
  - DDHC, 1789: https://gallica.bnf.fr/iiif/ark:/12148/btv1b69480451/f1/full/full/0/default.jpg
- Internet Archive (American Digital Library): https://archive.org/
  - Versailles Treaty, 1919: https://iiif.archivelab.org/iiif/treatypeacewith00goog#6/full/full/0/default.jpg
- NAKALA (HumaNum project): https://nakala.fr/
  - Holocaust Testimony, 1945: https://api.nakala.fr/iiif/10.34847/nkl.bffevbx7/7e089763b1f519027d9868cf941b09ce926d1f0e

Adding images on your local Cantaloupe server: setting up the folder

Link to instruction

1. Cantaloupe must be downloaded on your computer and in an easily accessible place

2. Create a folder for images or choose an already existing folder with images and copy the absolute path of this folder

3. Go to your Cantaloupe folder, make a copy of the file *cantaloupe.properties.sample*, rename it by removing ".sample"

4. Open the file, search for `*FilesystemSource.BasicLookupStrategy.path_prefix*` and change the path existing to the path leading to your folder of images (step 2)

Adding images on your local Cantaloupe server: launching the server

Link to instruction

5. Open your command line interface (terminal), go to the place where Cantaloupe is installed (using *cd*) and activate Cantaloupe with the following command

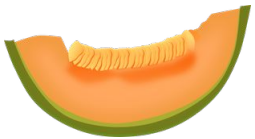
```
$ java -Dcantaloupe.config=cantaloupe.properties
-Xmx2g -jar cantaloupe-5.0.5.jar
```

6. Once this is done, open your browser and enter the following URL
http://localhost:8182/iiif/3/{name_of_the_file.extension}/full/max/0/default.jpg

7. Your image should be displayed on the browser

Different ways to see an image with IIIF

- Metadata of the images: http://localhost:8182/iiif/3/jane_austen.jpg/info.json
- Full view: http://localhost:8182/iiif/3/jane_austen.jpg/full/max/0/default.jpg
- Gray version: http://localhost:8182/iiif/3/jane_austen.jpg/full/max/0/gray.jpg
- Zoom in: http://localhost:8182/iiif/3/jane_austen.jpg/160,80,200,300/max/0/default.jpg
- Rotated: http://localhost:8182/iiif/3/jane_austen.jpg/full/max/68/default.jpg
- Smaller size: http://localhost:8182/iiif/3/jane_austen.jpg/full/,250/0/default.jpg



SEGMENTATION/TRANSCRIPTION/ POST-OCR CORRECTION

What is OCR/HTR and how to do it ?

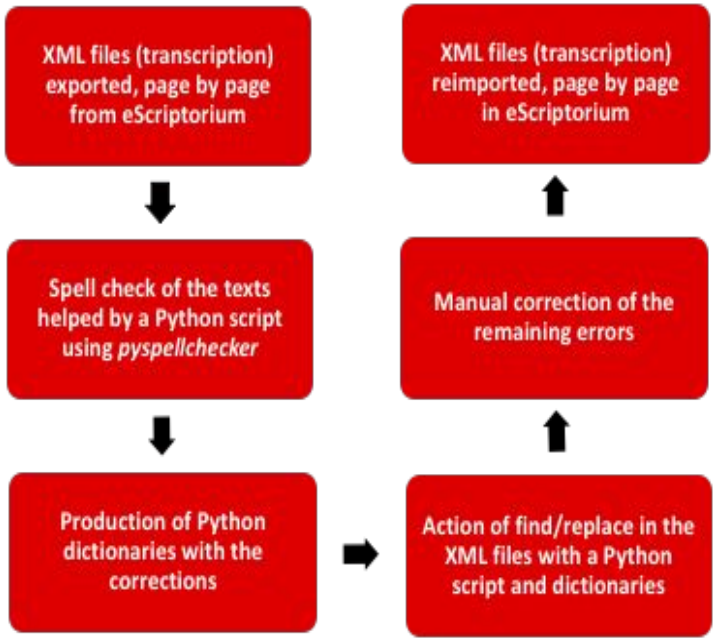
- What is OCR/HTR ?
  - OCR = Optical Character Recognition
  - HTR = Handwritten Text Recognition
- How is it composed ?
  - Segmentation
  - Text Recognition
  - Model training
- Different tools
  - Proprietary software OCR → Abbyy FineReader
  - Open source OCR → Tesseract
  - For HTR → Transkribus, eScriptorium

Working with eScriptorium

- Web interface for collaborative and automatic transcription projects, relying on the OCR software Kraken
- How to do OCR on your corpus ?
  - Create a project
  - Segment
  - Transcribe
  - Export

Post-OCR correction

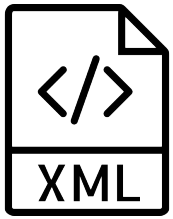
- Automatic transcription → errors to correct → manual correction?
- No fixed way to do it and no definite tool created for it
- Personal solution:
  - Python and *pyspellchecker*
  - Generation of dictionaries of errors and corrections
  - Diminish the time used for the manual correction



ENCODING

The Text Encoding Initiative (TEI)

- Consortium which develops and maintains a standard for the representation of texts in digital form
- Set of guidelines that specify elements and attributes to use to encode every part of your transcription



From TXT/ALTO/PAGE to TEI

- Transformation of the TXT version
  - Python script: *text_to_tei.py*
  - Instruction: https://github.com/FloChiff/workshop-discholed-tei2022/blob/main/instructions/transforming_transcriptions_into_tei.md#transforming-text-files
- Transformation of the ALTO/PAGE version
  - XSLT transformation
    - From ALTO to TEI: alto_to_tei.xsl
    - From PAGE to TEI: xmlpage_to_tei.xsl
    - From TEI (multiple files) to TEI (single file): tei_to_tei.xsl or tei_to_tei.py
  - Instruction: https://github.com/FloChiff/workshop-discholed-tei2022/blob/main/instructions/transforming_transcriptions_into_tei.md#transforming-xml

Providing the document metadata: the <teiHeader>

- <fileDesc>
  - <titleStmt> → title with a xml:lang, principal with their name on it, <respStmt>
  - <publicationStmt> → mention of their institution if they have one, the licence, and the date
  - <seriesStmt> → if the text is part of a whole, mention the collection, mention of the genre and topic possible too
  - <sourceDesc → <msDesc if the source is kept somewhere and to have physical description, (create a <listPlace> and <listPerson for the NE)
- <encodingDesc> → gives quick information on why it is done (Here for example, “encoding done during a workshop”)
- <profileDesc> → information about the author, place of writing, etc.
- <revisionDesc> → register the changes

From flat transcription to diplomatic version

- The TEI Guidelines offers the possibility to encode a text very precisely, from the layout to the faded words, from the additions/deletions to the named entities
- Possible additions to the <body>:
  - @rend to <p> (paragraph), <title>, <note>, etc.
  - <add>, <del>, <unclear>, etc. to writing specificities in the text
  - <persName>, <placeName>, <orgName>, <rs>, etc. to the named entities
  - <note @resp="#..."> to add comments about part of the text

PUBLICATION

What is TEI Publisher ?

- TEI Publisher is …
  - an easy-to-use tool to publish your TEI XML files
  - a *prêt-à-porter* application customizable with few tweaks
- Few platforms created with TEI Publisher

When the Wall Came Down

Van Gogh Letters

Shakespeare's Plays

Early English Books

Digital Scholarly Editions
(Made by me)

Discovering TEI Publisher

- Demo Collection:
  - Various XML files (letters, plays, novels, etc.) with each their ODD and templates, to showcase what TEI Publisher has to offer by displaying specific features
  - Link: http://localhost:8080/exist/apps/tei-publisher/index.html?tab=0&collection=test
- Playground:
  - Area to upload your own documents and to test them with every ODD and templates to evaluate your need in terms of features
  - Playground: http://localhost:8080/exist/apps/tei-publisher/index.html?tab=0&collection=playground
  - Instruction: https://github.com/FloChiff/workshop-discholed-tei2022/blob/main/instructions/working_with_tei_publisher.md#discovering-the-playground

Working with your own production

- Developing your own ODD:
  - The ODD is the file where we will add the elements we want to display specifically (rendition, predicate, behaviour, template and/or parameters)
  - Instruction: https://github.com/FloChiff/workshop-dischole d-tei2022/blob/main/instructions/working_with _tei_publisher.md#developing-your-own-odd
- Generating your own application
  - With the ODD and a chosen template, you can generate an application with all the basics elements needed and then, custom-made it
  - Instruction: https://github.com/FloChiff/workshop-dischole d-tei2022/blob/main/instructions/working_with _tei_publisher.md#generating-your-own-applicat ion

Modifying the application as you want

- Displaying the facsimile: Visualize next to each other the transcription and the image it's coming from
- Creating and displaying modes: Showing the same element in various ways
- Working with the index: Having access on the same page to the information about the named entities of the text
- Displaying the sourceDoc: Exhibiting next to each other the flat transcription and its diplomatic version
- Creating a collection: Offering the possibility of various corpus into one application

Feedback session:
Any questions?
Any remarks?

Contact
Floriane Chiffoleau: floriane.chiffoleau[at]inria.fr
Hugo Scheithaeur: hugo.scheithaeur[at]inria.fr

RESOURCES

Tools introduced during this workshop

- Cantaloupe (Open-source dynamic image server for on-demand generation of derivatives of high-resolution source images): https://cantaloupe-project.github.io/
- eScriptorium (A Digital Text Production Pipeline for Printed and Handwritten Texts using machine learning): https://escriptorium.paris.inria.fr/
- Oxygen XML (off-the-shelf XML editing software, providing must-have tools, and covering most XML standards): https://www.oxygenxml.com/
- TEI Guidelines (Guidelines for Electronic Text Encoding and Interchange): https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
- TEI Publisher (Instant Publishing Toolbox, developed by e-editiones:): https://teipublisher.com/index.html

BIBLIOGRAPHY

Previous presentations of this pipeline

- Chagué, Alix, and Floriane Chiffoleau. *An accessible and transparent pipeline for publishing historical ego documents*. 2021. ⟨hal-03180669⟩
- Chiffoleau, Floriane, Anne Baillot, Manon Ovide. “A TEI-based publication pipeline for historical ego documents -the DAHN project.” *Next Gen TEI, 2021 – TEI Conference and Members’ Meeting*, Oct 2021, Virtual, United States. ⟨hal-03451421⟩
- Chiffoleau, Floriane, Anne Baillot. *Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives*. 2022. ⟨hal-03628094⟩

Articles about the pipeline

Development of the pipeline:

- Chiffoleau, Floriane, DAHN Project, *Digital Intellectuals*, 2020-2021: https://digitalintellectuals.hypotheses.org/category/dahn

Steps from the pipeline:

- Chagué, Alix, and Hugo Scheithauer. 2021. *page2tei*, an XSL Transformation to transform PAGE XML into TEI XML (Version 1.0.0) [Computer software]
- Kiessling, Benjamin et al. 2019. “eScriptorium: An Open Source Platform for Historical Document Analysis”. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2, pp. 19–19. DOI:10.1109/ICDARW.2019.10032
- Pierazzo, Elena. 2019. What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter. *International Journal for Digital Humanities*, Springer, 1, pp.1-12.⟨10.1007/s42803-019-00019-3⟩. ⟨hal-02117714⟩