



Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition

Alix Chagué, Hugo Scheithauer, Lucas Terriel, Floriane Chiffolleau, Yves Tadjó-Takianpi

► To cite this version:

Alix Chagué, Hugo Scheithauer, Lucas Terriel, Floriane Chiffolleau, Yves Tadjó-Takianpi. Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition. Digital Humanities 2022: Responding to Asian Diversity, ADHO; University of Tokyo, Jul 2022, Tokyo, Japan. hal-03739767

HAL Id: hal-03739767

<https://inria.hal.science/hal-03739767>

Submitted on 28 Jul 2022

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



TAKE A SIP OF TEI AND RELAX

A proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition

Alix Chagué, Hugo Scheithauer, Lucas Terriel,
Floriane Chiffolleau, Yves Tadj-Takianpi



Inria

DH2022

DIGITAL HUMANITIES 2022

HTR IS A GAME CHANGER TO ACCESS TEXTUAL DATA

- HTR : handwritten text recognition
- Accessible to non-specialists via software like Transkribus or eScriptorium (and others)
- Takes a digitized document and create digital text equivalent after analyzing the layout
- Let the computer work for you and get thousands of text files!
- It works for many types of languages and writing systems



eScriptorium



<https://gitlab.com/scripta/escriptorium>



<https://transkribus.eu/>



Digital
image

PIPELINE(S) FOR DIGITAL EDITIONS

STEP 1

Text acquisition
(HTR)



STEP 2

Post
processing

STEP 3

Information extraction
(NER, etc)

STEP 4

Publication



index & references



Digital
image

PIPELINE(S) FOR DIGITAL EDITIONS

STEP 1

Text acquisition
(HTR)



STEP 2

Post
processing

Values?
Conceptual
frameworks?



STEP 3

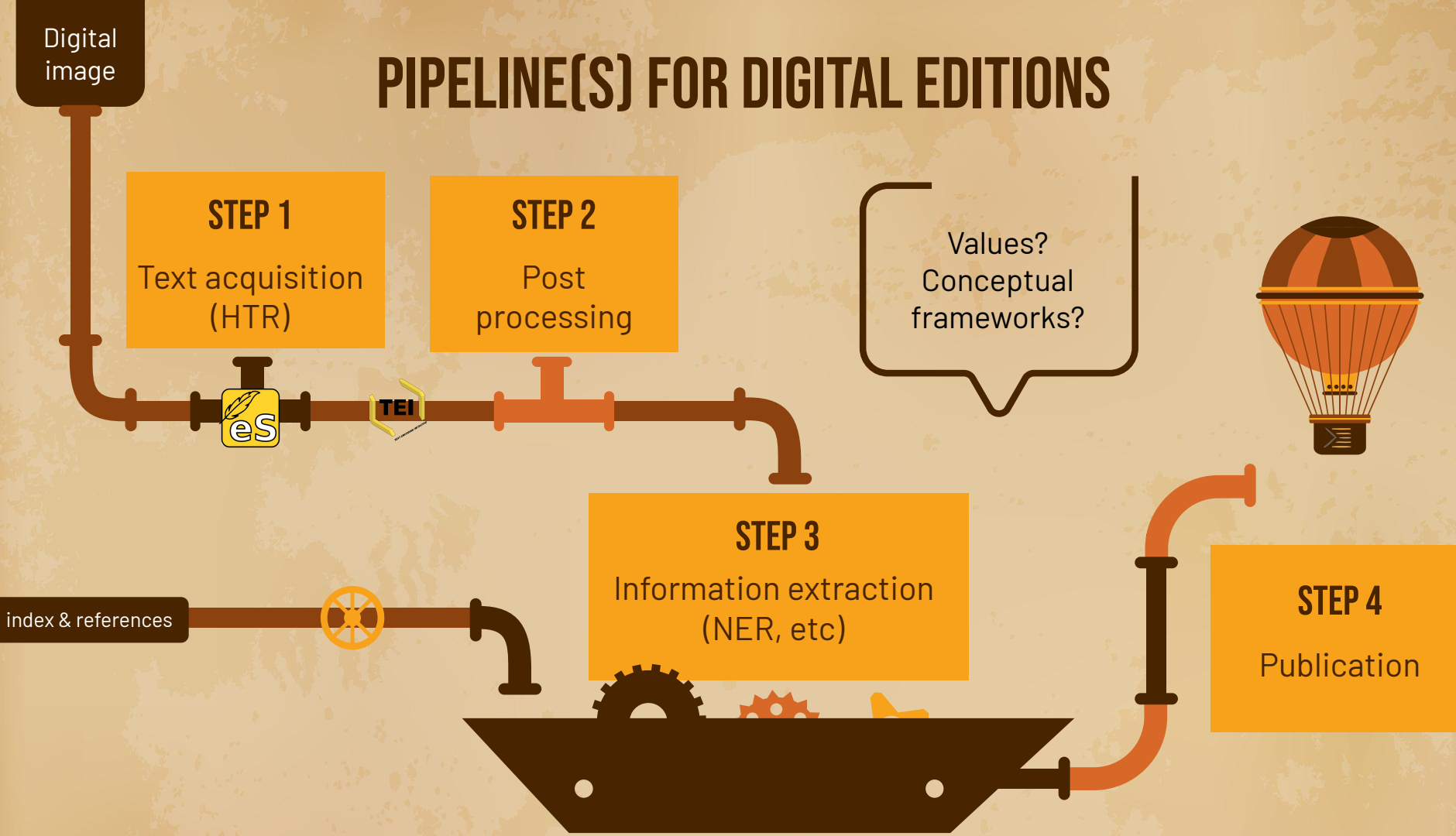
Information extraction
(NER, etc)

index & references



STEP 4

Publication



THE CORNERSTONE: TEI XML



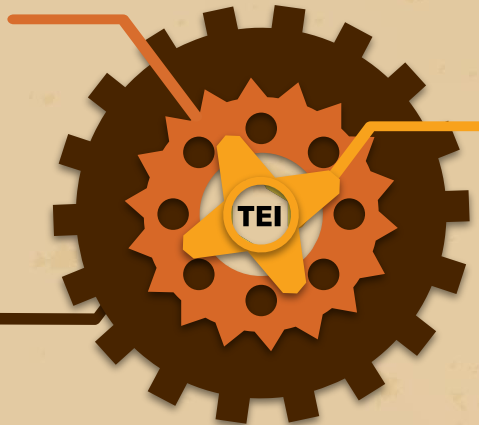
- Frequent loss of information during switch from ALTO/PAGE to TEI
- TEI is a stable and documented standard for the representation of texts in digital form
- Extended markup to register every information needed for the source document :

<SOURCEDOC>

Data about the coordinates of the lines and regions on the facsimile and raw transcription

<BODY>

Text and annotations



<TEIHEADER>

Metadata about the manuscript description, the members of the project, specificities in the content, etc.

ENRICHING TEI WITH INFORMATION EXTRACTION

- "Named-entity recognition is **a subtask of information extraction**
- **Locate and classify named entities** (person names, organizations, locations, etc.)

Edward Teach **PER**, better known as "Blackbeard **MISC**", was born at Bristol **LOC** and was an English **MISC** pirate. On 22 November 1718 following a ferocious battle Teach **PER** and several of his crew were killed by a small force of sailors led by Lieutenant Robert Maynard **PER** in Ocracoke **LOC**, North Carolina **LOC**.

Many possible approaches:

- **rule-based matching**: regular expression, gazetteers.
- **feature engineering-based machine learning**: statistical model trained on word features (wordcase, type of prefix/suffix, word length).
- **neural network**: train an algorithm on pre-annotated entities in dataset.

Typical use cases:

- Information retrieval / faceted search,
- Quantitative or statistical studies based on named entity,
- Entity linking to explore relation with semantic web,
- Build a Knowledge graph to infer new potential sources, etc.

SEMANTIC@

- Handling TEI XML and Named Entity Recognition is a difficult task.
- Semantic@:
 - is a proof of concept, open source and fully customizable,
 - offers a graphical user-friendly interface,
 - gathers in one place many tools for semantic annotation.
- It relies on existing libraries:
 - David Lassner's Standoff converter (<https://github.com/standoff-nlp/standoffconverter>)
 - The Natural Language Processing library SpaCy,
 - Flask and SQLite (Front end).



SEMANTIC@

SpaCy

Named Entity
Recognition (NER)

standoff-converter
/ XSLT

Re-injection of NE
in source XML



import

01

XML parsing and
text extraction

standoff-converter

02

Manual correction
of NER predictions

RecogitoJS

03

04



export



<https://github.com/Lucaterre/semanticat>

Semantic@ Menu

332 annotations

Search for annotations... (clear)?

LOC: 273

PER: 13

ORG: 25

Document: Lettre569_3octobre1919.xml

de notre pays". La pluie battante et glacée ne m'a pas empêché d'arriver le Dimanche matin au du canton de **Comice Agricole** de **St-Paterne**, où ma présence était d'autant plus attendue des populations que le Préfet craignait de se compromettre en y assistant et refusait même de s'y faire représenter ; - parce que là nous étions dans l'arrondissement de M. **Caillaux**. - et que, si M. **Caillaux** n'est toujours pas jugé, il n'en est pas moins toujours en prison et présumé coupable par le Gouvernement qui l'y a fait mettre. Voyez, entre mille, les petits effets de cette réclusion de M. **Caillaux** ; le pays qui l'a élu est mis à l'index; ses **Comices Agricoles** ignorés du Gouvernement : quiconque ose trouver que le procès se fait bien attendre est mal noté ; et je ne parle pas de la façon dont furent traités pendant la guerre les soldats natifs de l'arrondissement de **Mamers**, qu'il fallait punir, sinon supprimer : les électeurs à **Caillaux** ! Une fanfare de cors de chasse accueillit ma descente de voiture ; je visitai rapidement, sous des parapluies, l'exposition des bestiaux, peu importante, et je me rendis à la **mairie** où le banquet était pré-paré, exactement comme avant la guerre, plus convives, malgré le temps et les circonstances si défavorables. Des allocations furent échangées et je m'excusai de ne pouvoir prendre place à table. On m'attendait à l'arrondissement, à **Mamers**, pour présider une fête semblable, mais plus importante d'un vin d'honneur offert par la Municipalité aux soldats démobilisés rentrés à leurs foyers. C'était en grand, la même cérémonie de celle que j'avais organisée, le 14, dans la commune de **Clermont-Créans**. Malgré la pluie, grâce à la marche parfaite de mon auto, je franchis cette nouvelle étape avec une régularité chronométrique, laissant la grande route d'Alençon et c'est elle de la belle forêt de **Perseigne** pour couper ou plus court, par ces jolis villages aux vieux

Labels

LOC ORG PER

Alençon

Cancel OK

Semantic@'s interface (here to view named entity extraction)

Example of TEI output, after NER

Semantic@ performs NER task and inject found named entities into the TEI file

source document: **DAHN Project**

(https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre0569_3octobre1919.xml)

de ne m'a pas empêché<lb/> d'arriver
agricole</orgName> du canton
me, où ma présence était d'autant
lations que le Préfet craignait de se
fusait même de s'y faire représenter ;
aux</persName> - et que,
aux</persName> n'est toujours pas jugé, il n'en est pas moins
on et présumé coupable par le Gouvernement qui l'y a fait<lb/> mettre. Voyez, entre
u-<lb break="no"/>sion de M. <persName>Caillaux</persName> ; le pays qui l'a élu est mis à l'index; ses<lb/>
je ne parle<lb/> pas de la façon dont furent traités pendant la guerre les soldats<lb/> natifs de l'arrondissement de <placeName>Mamers</placeName>
ceux qu'il fallait punir,<lb/> sinon supprimer : <hi rend="underline">les électeurs à <persName>Caillaux</persName> </hi> Une fanfare de cors<lb/>
de chasse accueillit ma descente de voiture ; je visitai rapide<lb break="no"/>ment, sous des parapluies, l'exposition des
bestiaux, peu<lb break="no"/>importante, et je me rendis à la <orgName>mairie</orgName> où
le banquet était pré<lb break="no"/>paré, exactement comme avant la guerre, plus d'une centaine de<lb/> convives, malgré le
temps et les circonstances si défavorables.<lb/> Des allocations cordiales furent échangées et je m'excusai de ne<lb/> pouvoir
prendre place à table. On m'attendait au chef-lieu de<lb/> l'arrondissement, à <placeName>Mamers</placeName>, pour présider une fête semblable,
mais<lb/> plus imposante et doublée d'un vin d'honneur offert par la Muni-<lb break="no"/>cipalité aux soldats démobilisés
rentrés à leurs foyers. C'était<lb/> en grand, la même cérémonie de celle que j'avais organisée, le 14,<lb/> dans la commune
de <placeName>Clermont-Créans</placeName>. Malgré la pluie, grâce à la<lb/> marche parfaite de mon auto, je franchis cette nouvelle étape avec<lb/>
une régularité chronométrique, laissant la grande route d'<placeName>Alençon</placeName><lb/> et<sub>
<del rend="strikethrough">c
<add> </add>
</sub><sub>
<del cert="medium" rend="overwritten">d
<add hand="#annotation" place="across">c</add>
</sub>>elle de la belle forêt de <placeName>Perseigne</placeName>

USING TEI PUBLISHER FOR CREATING A DIGITAL EDITION

- Hosted by *exist-db*, a software for **NoSQL databases** based on XML,
- **Open source**,
- **Automatically generated** and **customizable web application** for digital editions,
- Display collections of corpora in TEI XML or other markup format, with **templates** in HTML format and **ODD** for the TEI processing model,
- **Search engine** based on metadata and document content,
- Files downloadable to multiple formats (PDF, ePUB, XML, etc.).



VISUALIZING A TRANSCRIPTION WITH TEI PUBLISHER

Flat/basic representation of the transcription
(displayed region by region)

Imitative representation of the transcription
(leveraging SVG and coordinates)

<SOURCEDOC>

<BODY>

Diplomatic edition of the source document
(extended markup and additional information,
help discovery)

Facsimile (source image via IIIF protocol)

DIPLOMATIC VERSION

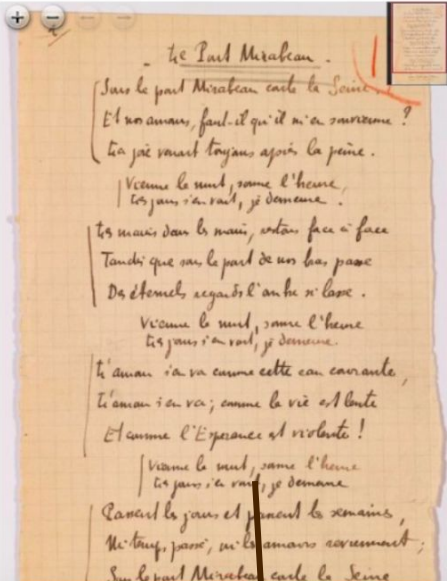
Example

View
Transcription

1

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.
Vienne la nuit, sonne l'heure,
Les jours s'en vont, je demeure.
Les mains dans les mains, restons face à face
Tandis que sous le pont de nos bras passe
Des éternels regards l'ombre si lasse.
Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure.
L'amour s'en va comme cette eau courante
L'amour s'en va, comme la vie est lente
Et comme l'Espérance est violente !
Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure
Passent les jours et passent les semaines
Ni temps passé, ni les amours reviennent ;
Sous le pont Mirabeau coule la Seine
Vienne la nuit, sonne l'heure



FACSIMILE

IMITATIVE VIEW

View
imitative view

1

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.
Vienne la nuit, sonne l'heure,
Les jours s'en vont, je demeure.
Les mains dans les mains, restons face à face
Tandis que sous le pont de nos bras passe
Des éternels regards l'ombre si lasse.
Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure.
L'amour s'en va comme cette eau courante
L'amour s'en va, comme la vie est lente
Et comme l'Espérance est violente !
Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure
Passent les jours et passent les semaines
Ni temps passé, ni les amours reviennent ;

View
Region by region

title

Le Pont Mirabeau.

line_group

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

line_group

L'amour s'en va comme cette eau courante
L'amour s'en va, comme la vie est lente
Et comme l'Espérance est violente !
Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure

FLAT VIEW

Mock-up showing the four different views potentially available in TEI-Publisher
(ex. from "Le Pont Mirabeau" by Guillaume Apollinaire)

PERSPECTIVE AND KEY TAKE-AWAY

- Some missing bricks (or pipes):
 - post-HTR correction
 - text normalization
 - better solutions of complex layouts
- Transparency and flexibility are key features for robust and generalizable pipelines
- TEI provides solution to build pivot files easing the navigation down and up stream

THANK YOU!

