

Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives

Floriane Chiffolleau, doctorante à Le Mans Université (3LAM) et Inria (ALMAnaCH)
Anne Baillot, professeure à Le Mans Université (3LAM) et chercheuse à ICAR UMR 5191 à l'ENS de Lyon
1^{er} avril 2022



ALMAnaCH project-team

Inria



3LAM
Langues, Littératures,
Linguistique
Le Mans Université
Université d'Angers



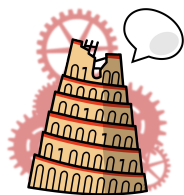
Plan de la présentation

1. Qu'est-ce que le **projet DAHN** ?
2. **Pourquoi** créer une pipeline ?
3. Expérimenter avec la pipeline : travailler sur des **corpus**
4. **Garder trace** du travail effectué sur la pipeline
5. Les **étapes** de la pipeline : de la numérisation à la publication
6. Une **plateforme** pour des éditions scientifiques numériques : **DiSchoIEd**
7. **Ressources** diverses

Qu'est-ce que le projet DAHN ?

Qu'est-ce que le projet DAHN ?

- “Dispositif de soutien à l’Archivistique et aux Humanités Numériques”
- Partenariat entre l’Inria (équipe ALMAnaCh), l’Université du Mans et l’EHESS
- Financé par le Ministère de l’Enseignement Supérieur, de la Recherche et de l’Innovation
- Objectif → Définir une chaîne d’édition scientifique permettant de valoriser des fonds d’archives dans des formats numériques facilitant leur exploitation par la recherche



ALMAnaCH project-team

Inria



**Le Mans
Université**



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION**

*Liberté
Égalité
Fraternité*

Pourquoi créer la pipeline ?

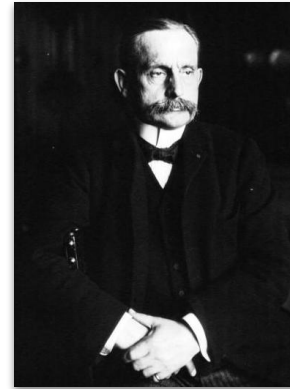
Pourquoi créer la pipeline ?

- Mettre à disposition des outils *open source* librement accessibles
 - Afin de maximiser les possibilités d'utilisation de la pipeline, que ce soit pour des chercheurs ou de simples amateurs, il est nécessaire d'avoir à disposition des outils librement accessibles et réutilisables
- Mettre en lien des logiciels suivant le principe d'interopérabilité
 - Contrer les cas où des outils font une tâche ou une autre mais leur production n'est pas compatibles avec les outils de l'étape suivante
- Aller à l'encontre du système de boîte noire
 - Tous les logiciels utilisés sont présentés, le fonctionnement de chacun est connu et une documentation est toujours disponible

Expérimenter avec la pipeline : travailler sur des corpus

Corpus : Correspondance de Paul d'Estournelles de Constant

- Les correspondants
 - L'expéditeur : Paul d'Estournelles de Constant
 - Le destinataire : Nicholas Murray Butler
- Le contexte
 - Rapports de vie et de faits pendant la Première Guerre mondiale
 - Échange d'avis, d'opinion et de ressenti entre deux pacifistes
- Intérêt du corpus
 - Des égodocuments (journaux personnels, livres de raison, correspondance), à la limite entre les sphères privée et publique
 - Un corpus volumineux : environ 1 500 lettres disponible, 430 pour la période 1914-1918 et une diversité du nombre de pages (certaines lettres font 1 à 2 pages et d'autres font 30 à 40 pages)
 - Les lettres sont tapuscrites : plus grande lisibilité, exploitation plus aisée, utilisation de la pipeline facilitée



Corpus : Lettres et textes des intellectuels berlinois

- Les auteurs :
 - Une dizaine dont August Boeckh, Adelbert von Chamisso et Ludwig Tieck
- Le contexte :
 - Genèse du mouvement littéraire de romantisme allemand
 - Exploration des cercles intellectuels de Berlin au début du XIXème siècle
- Intérêt du corpus:
 - Des égodocuments mais aussi quelques documents plus différents (nouvelles, pièces de théâtre, thèse)
 - Un corpus déjà traité entièrement et publié qui a besoin d'une mise à jour
 - Des lettres et textes dans une toute autre langue et datant d'une toute autre période que le corpus précédent



Garder trace du travail effectué sur la pipeline

Sauvegarder le travail sur la pipeline

- Nombreux fichiers créés avec l'élaboration et le développement de la pipeline
 - Modèles de segmentation/transcription, rapports d'entraînement, *vérités de terrain*
 - Fichiers XML-TEI
 - Scripts et dictionnaires Python
- Outil de sauvegarde : GitHub
 - GitHub → Plateforme d'hébergement de code pour la gestion de versions (de fichiers) et la collaboration (à distance)
 - Repository du projet : <https://github.com/FloChiff/DAHNPProject>
- Un autre outil a également été utilisé, dans une moindre mesure : Sharedocs
 - Sharedocs → Gestionnaire de fichiers pour stocker, échanger, partager, travailler sur des données
 - Rangement des fichiers privées du projet mais partageable avec les autres membres (facsimile des corpus, sauvegarde du site original des BI, etc.)



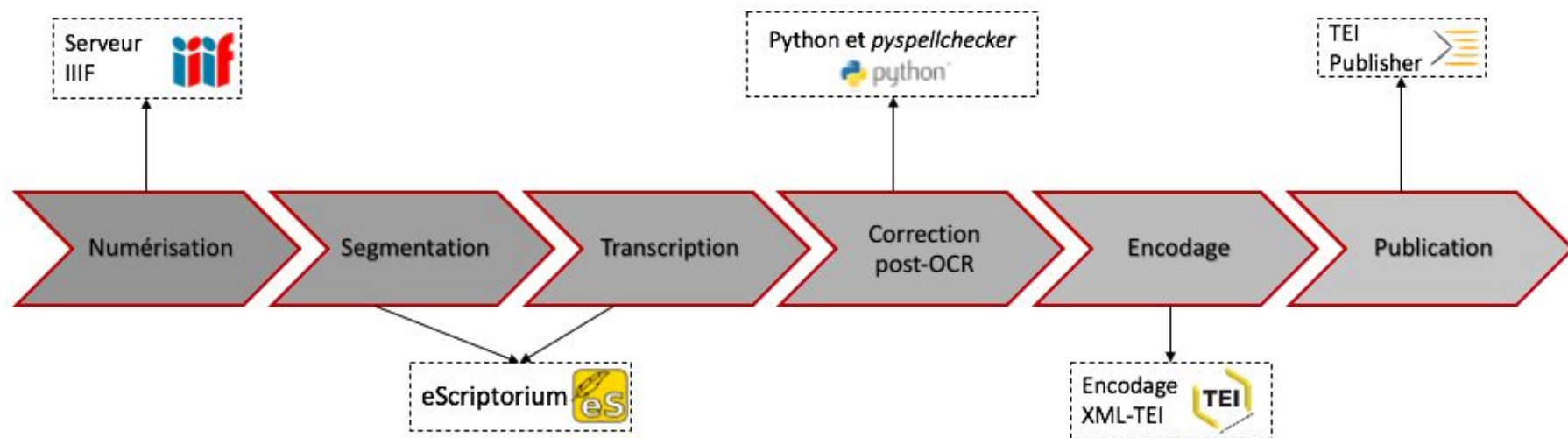
Documenter le travail sur la pipeline

- Posts de blog : Hypothèses
 - Hypothèses → Site internet pour la publication de carnets de recherche scientifique en sciences humaines et sociales
 - Rédaction de posts de blog sur les différentes tâches, avancées ou difficultés que j'ai pu rencontrer durant l'élaboration et le développement de la pipeline
 - Posts mis en ligne sur le blog d'Anne Baillot : <https://digitalintellectuals.hypotheses.org/> dans la catégorie "DAHNPProject"
- Documentation
 - Rédaction de guidelines sur l'encodage d'égodocuments, sur le fonctionnement de la plateforme de publication et sur les méthodes d'utilisation des scripts Python
 - Fichiers mis en ligne sur le repository du projet : <https://github.com/FloChiff/DAHNPProject>



Les étapes de la pipeline : de la numérisation à la publication

Les étapes de la pipeline : Schéma



Les étapes de la pipeline : Numérisation

- Numérisation professionnelle des images par les centres d'archives ou bibliothèques qui les conservent
- Utilisation de serveurs IIF pour stocker les facsimile de nos documents
 - IIF = International Image Interoperability Framework
 - Favoriser l'interopérabilité par la création d'un cadre technique commun pour la diffusion standardisée d'images haute résolution sur le Web afin de les rendre consultables, manipulables et annotables par n'importe quelle application compatible
- Serveur principal choisi pour nos corpus : NAKALA
 - Outil créé par la TGIR Huma-num
 - Service permettant de partager, publier et valoriser tous types de données numériques documentées afin de les publier en accord avec les principes du *FAIR data*



 **Lettre n°1 de Paul d'Estournelles de Constant à Nicholas Murray Butler (15 août 1914)** EN | FR

ID : 10.34847/nkl.2accvml1

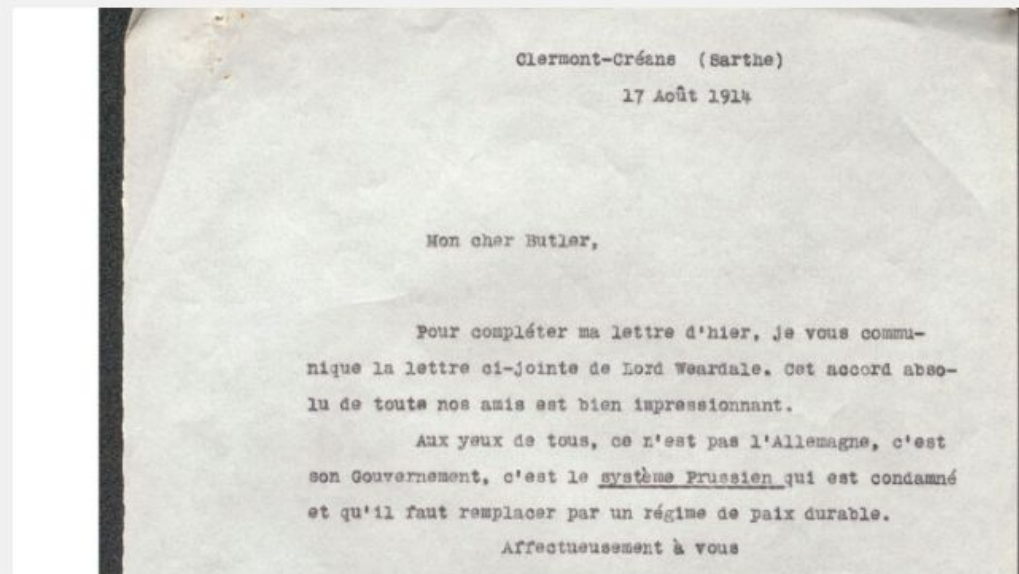
 Publiée

Auteur : Paul d'Estournelles de Constant

Fichiers

-  FRAD072_12j381_le..._003.jpg
-  FRAD072_12j381_le..._004.jpg
-  FRAD072_12j381_le..._003.jpg
-  FRAD072_12j381_le..._008.jpg
-  FRAD072_12j381_le..._009.jpg
-  FRAD072_12j381_le..._001.jpg
-  FRAD072_12j381_le..._002.jpg
-  FRAD072_12j381_le..._006.jpg
-  FRAD072_12j381_le..._001.jpg
-  FRAD072_12j381_le..._005.jpg
-  FRAD072_12j381_le..._002.jpg
-  FRAD072_12j381_le..._007.jpg

Visualisation



ID : 10.34847/nkl.2accvml1/8387d7ba8ddef09a75d0f26bb021f4104e5dec81

Les étapes de la pipeline : Segmentation/Transcription

- Utilisation de Kraken et de son interface web eScriptorium
 - Kraken → logiciel de reconnaissance de texte qui permet également l'entraînement de modèles spécifiques pour des textes imprimés, tapuscrits et manuscrits
 - eScriptorium → interface web pour des projets collaboratifs de transcription automatique
- Création de *vérités de terrain*, c'est-à-dire une segmentation et transcription manuelles d'une partie du corpus, afin de l'utiliser par la suite pour entraîner un modèle qui pourra segmenter/transcrire automatiquement le corpus
- Deux solutions pour la création de modèles:
 - *From scratch* → création de ses *vérités de terrain* et entraînement de son modèle seulement avec ces données
 - *Fine tune* → utilisation de modèles déjà existants, généralement entraîné sur des corpus similaires au sien et sur lequel on ajoute ses propres *vérités de terrain* pour l'adapter à son corpus



Affichage des pages d'un document dans eScriptorium

The screenshot displays the eScriptorium web interface. At the top, the navigation bar includes the eScriptorium logo, 'Home', 'Contact', 'My Projects', 'My Models', and a user profile 'Hello fchiffol'. Below this, a secondary navigation bar shows 'Description', 'Images', 'Edit', and 'Models'. The main content area features a dashed blue border containing a 'Drop images here or click to upload.' instruction. Below this, a toolbar offers actions like 'Select all', 'Unselect all', 'Import', 'Export', 'Train', 'Binarize', 'Segment', and 'Transcribe'. The central part of the interface shows eight document pages, each with a thumbnail, a close button, and a control bar at the bottom. The control bars for pages 4, 5, and 6 are highlighted with red arrows, pointing to a '100%' status indicator.

My Projects My Models Hello fchiffol

Description Images Edit Models

Lettre 1-2

Accès à mes autres projets

Accès à mes modèles téléchargés et/ou créés

Drop images here or click to upload.

Importation des images

Select all Unselect all Selected 0/19 Import Export Train Binarize Segment Transcribe

A grid of eight document page thumbnails, numbered 1 through 8. Each thumbnail includes a close button in the top right corner. Below each thumbnail is a control bar with a play icon, a list icon, and a blue box containing '100%'. Red arrows point to the '100%' boxes for pages 4, 5, and 6.

Statut de la binarisation, segmentation et transcription

Affichage de la segmentation/transcription d'une page dans eScriptorium

The screenshot displays the eScriptorium web interface. At the top, the navigation bar includes the eScriptorium logo, 'Home', 'Contact', 'My Projects', 'My Models', and 'Hello fchiffol'. Below this, a secondary navigation bar shows 'Description', 'Images', 'Edit', and 'Models'. The main content area is titled 'Lettre 1-2' and shows 'Element 1 - P1040234.JPG - (1920x2560) - 0 Bytes'. A toolbar with various icons is visible above the document image. The document image on the left shows a scanned page with several lines of text highlighted in purple. A circular stamp is visible in the bottom left corner of the image. To the right of the image, a dropdown menu is open, listing three transcription options: 'corrected_transcript', 'kraken:modelpec_9290_NFC', and 'manual'. A red arrow points from the text 'Différentes transcriptions disponibles' to this menu. Further right, another dropdown menu is open, showing 'Texte segmenté' and 'Transcription'. A red arrow points from the text 'Texte segmenté' to the first option, and another red arrow points from the text 'Transcription' to the second option. The main content area on the right displays the segmented transcription of the document image, with the text: 'Lettre N° 1 (15 Août 1914)', 'La Démagogie militariste allemande.', 'La mobilisation - La France récolte ce qu'elle a semé.', 'Aux funérailles de Jaurès.', 'Annexe:', and 'Texte de mon discours aux funérailles de Jaurès.'

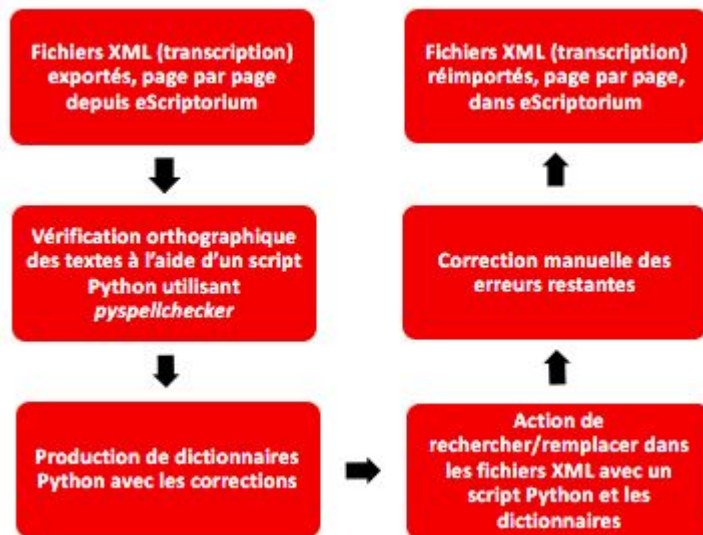
Différentes transcriptions disponibles

Texte segmenté

Transcription

Les étapes de la pipeline : Correction post-OCR

- La transcription automatique ne garantit pas un résultat parfait → nécessité de corriger une partie de la transcription
- Utilisation de scripts Python et d'un module de correction (*pyspellchecker*) pour modifier le texte afin d'avoir la bonne transcription



Les étapes de la pipeline : Encodage



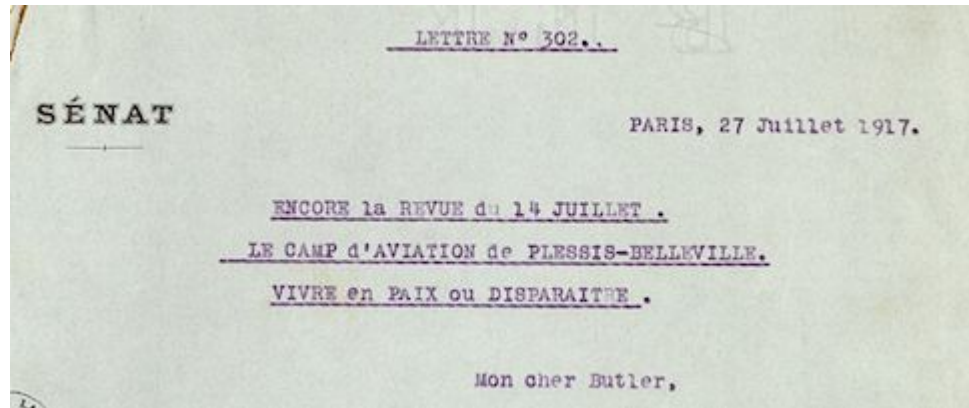
- Encodage XML-TEI
 - Standard pour la représentation de textes sous un format numérique, notamment dans le domaine des sciences humaines et de la linguistique
 - Règles de balisage pour l'encodage des textes selon le type de texte et l'information donnée
- Deux types d'export de la transcription possibles depuis eScriptorium
 - Page XML : transformation du fichier grâce à un script XSL (scénario de transformation) pour qu'ils soient encodés selon les règles de la XML-TEI
 - Texte : exportation de la transcription sous un format .txt pour qu'ils soient encodés par la suite à l'aide d'un script Python

Les étapes de la pipeline : Encodage

- Encodage depuis Page XML (<https://github.com/TEI4HTR/page2tei>)
 - Création du contenu de la balise <sourceDoc> utilisée pour reproduire exactement la transcription selon ce qui se trouve sur le facsimile (régions, lignes, coordonnées)
- Encodage depuis le format texte
 - Création de plusieurs scripts Python pour aider l'encodage
 - Utilisation d'expressions régulières, c'est-à-dire une chaîne de caractères, qui décrit, selon une syntaxe précise, un ensemble de chaînes de caractères possibles.
 - Un script permet de créer le *header* et de compléter les métadonnées et un autre permet d'encoder le *body* selon sa structure (paragraphe, saut de ligne, changement de page, etc.)



Image du début de la lettre n°302 du corpus de Paul d'Estournelles de Constant



Encodage (body) du début de la lettre n°302 du corpus de Paul d'Estournelles de Constant

```
<div type="letter">
  <pb facs=".JPG" n="1"/>
  <head rend="center underline">LETTRE N° 302.</head>
  <opener>
    <fw corresp="#lh-senat" place="margin" type="letterhead">
      <hi rend="underline"><orgName ref="#g0002" type="pec">SÉNAT</orgName></hi>
    </fw>
    <dateline rend="align(right)"><placeName ref="#l0001" type="pec">PARIS</placeName>, <date when-iso="1917-07-27">27 juillet
      1917</date>.</dateline>
    <title rend="align(center) underline">ENCORE la REVUE du 14 JUILLET .<lb/>LE CAMP d'AVIATION de <placeName ref="#l0395"
      type="pec">PLESSIS-BELLEVILLE</placeName>.<lb/>VIVRE en PAIX ou DISPARAITRE.</title>
    <salute rend="indent">Mon cher <persName ref="#p0002" type="pec">Butler</persName>,</salute>
  </opener>
```

Les étapes de la pipeline : Encodage

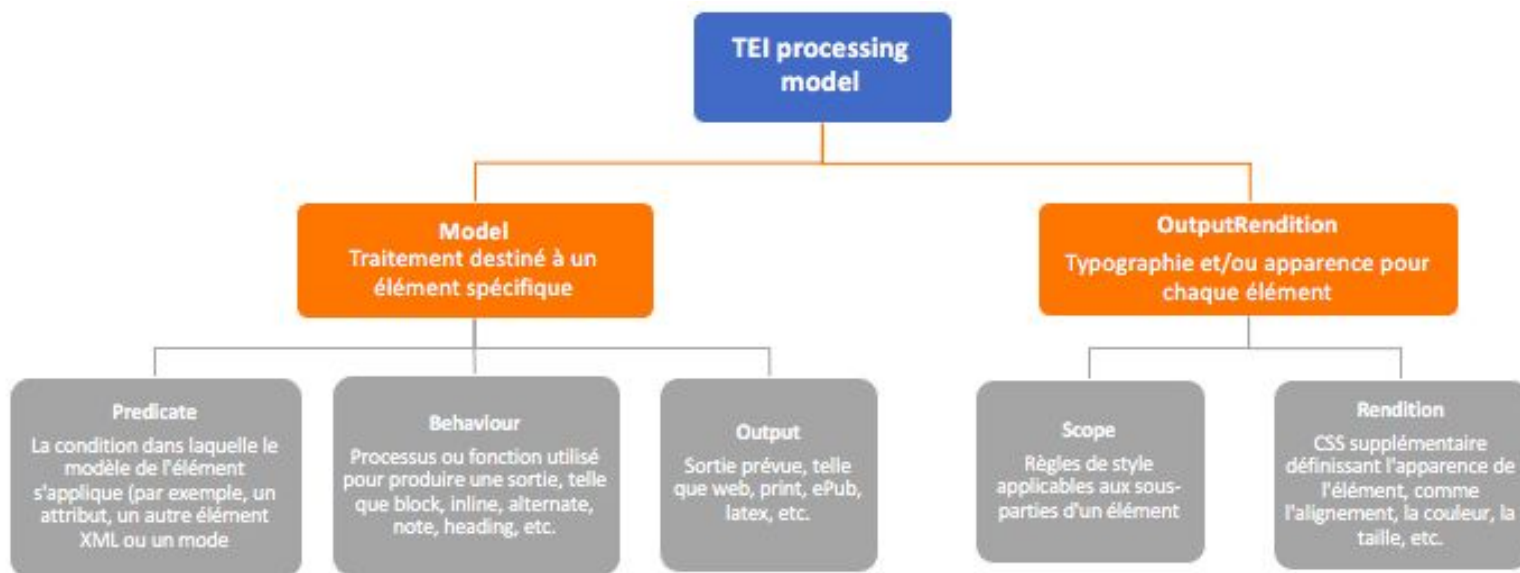
- [Guide d'encodage](#) pour le cas spécifique des égodocuments
 - Mise en place afin de créer une réelle homogénéité entre les éditions qui seront mise en ligne lors de la phase suivante de notre pipeline
 - Mention de toutes les balises qui pourraient être utilisées dans l'encodage d'un égo document et des précautions d'utilisation de telle ou telle balise
- Courte documentation
 - [Marche à suivre](#) pour utiliser les scripts Python d'encodage des fichiers textes
 - [Démonstration](#) du fonctionnement du script principal (balisage du *body*)

Les étapes de la pipeline : Publication

- Publication des fichiers XML TEI avec TEI Publisher
 - Plateforme qui s'appuie sur *exist-db*, un système de gestion de base de données qui s'appuie sur la technologie XML
 - Outil qui donne la possibilité aux chercheurs de publier leurs projets sans avoir à devenir programmeurs et aux développeurs d'avoir à disposition une application qui demande moins de code et qui offre une meilleure maintenance et interopérabilité
- Transformation des fichiers XML TEI
 - Affichage des fichiers au format HTML
 - Utilisation d'une ODD, c'est-à-dire un fichier qui permet de customiser les règles de transformation du document selon la manière dont on veut que chaque élément de l'arbre TEI apparaisse dans l'affichage
 - Export possible des fichiers sous divers formats (LaTeX, ePub, PDF)



Les étapes de la pipeline : Publication



Fonctionnement des règles de transformation d'un élément TEI dans l'ODD

Une plateforme pour des éditions scientifiques numériques : DiScholEd

Une plateforme pour des éditions scientifiques numériques : DiSchoIEd

- Développement complet d'une plateforme pour la publication d'éditions scientifiques numériques
 - **DiSchoIEd** → **D**igital **S**cholarly **E**ditions
 - Hébergée sur un serveur *exist-db* fourni par Huma-num
 - <https://discholed.huma-num.fr/exist/apps/discholed/index.html>
- Contient actuellement six éditions :
 - Les deux corpus précédemment mentionnés
 - Un corpus de papiers appartenant à un des intellectuels berlinois
 - Deux journaux de guerre (Guerre napoléonienne et WWI)
 - Des témoignages de victimes de l'Holocauste
- La plateforme est toujours un "work in progress"



Paul d'Estournelles de Constant



Charles Bruneau



EHRI



Bataille de Leipzig



Adelbert von Chamisso



August Boeckh



DiScholEd - Éditions scientifiques numériques

Trier par
Date

Filtrer selon
Titre

Filtrer

Filters



Correspondance de d'Estournelles de Constant

Ce dossier contient le corpus et les index de la correspondance de Paul d'Estournelles de Constant, ainsi que l'histoire du corpus et des informations à propos du projet.



Lettres et textes: Le Berlin intellectuel des années 1800

Ce dossier contient le corpus et les index des lettres et textes des intellectuels berlinois de 1800 à 1830.

Ressources diverses

Liens vers les ressources du projet

Dépôt GitHub : <https://github.com/FloChiff/DAHNPProject>

Blog Hypothèses : <https://digitalintellectuals.hypotheses.org/category/dahn>

Vérités de terrain DAHN : <https://github.com/HTR-United/dahncorpus>

Collection d'Estournelles : <https://nakala.fr/collection/10.34847/nkl.adeb801d>

Collection Intellectuels Berlinois : <https://nakala.fr/collection/10.34847/nkl.8479g2z4>

Plateforme de publication pour les égodocuments :

<https://discholed.huma-num.fr/exist/apps/discholed/index.html>

Liens vers les ressources évoquées

IIIF : <https://iiif.io/>

Huma-num : <https://www.huma-num.fr/>

Nakala : <https://nakala.fr/>

Kraken : <http://kraken.re/master/index.html>

eScriptorium : <https://escriptorium.paris.inria.fr/>

Python : <https://www.python.org/>

pyspellchecker :

<https://github.com/barrust/pyspellchecker>

XML Alto : <http://www.loc.gov/standards/alto/>

Page XML : <https://doi.org/10.1109/ICPR.2010.72>

Text Encoding Initiative : <https://tei-c.org/>

TEI Publisher : <https://teipublisher.com/index.html>

GitHub : <https://github.com/>

Sharedocs :

<https://documentation.huma-num.fr/sharedocs-stockage/>

Hypothèses : <https://hypotheses.org/>

Merci de votre attention

Des questions ?

Contact : floriane.chiffoleau@inria.fr