Rapport d'avancement de projet

Projet DAHN (avec le soutien du MESRI)

Floriane Chiffoleau

Inria

Le Mans Université

Plan

- Pour commencer
- Stockage de données
- Établir l’encodage
- Création de scripts
- Difficultés rencontrées
- Publication de billets de blog

Pour commencer

Pour commencer

- Inventaire de la correspondance de Paul d'Estournelles de Constant à partir de la première lettre en 1914 jusqu'à la 579ème en 1919 → informations sur le statut (publié, transcrite, etc.), le nombre de pages, la date, le titre, la présence de pièce(s) jointe(s) et dans certains cas, les lettres de réponses de Nicholas Murray Butler
- Version transcrite et mise en forme des lettres de la correspondance qui ont été publiées dans l'ouvrage *En guerre pour la paix* de Stéphane Tison et Nadine Akhund
- Photographies prises aux Archives départementales de la Sarthe de la correspondance entre son début en 1914 et sa fin en 1924, l'année de la mort de Paul d'Estournelles de Constant (à partir de août 1920, la majorité des dossiers sont absents ou incomplets et antérieurement, certaines lettres manquent)

Pour commencer

- « Vers un processus de numérisation harmonisé pour les sources patrimoniales », *Rapport intermédiaire du Dispositif de soutien à l'archivistique et aux Humanités Numériques*
- Le site *Lettres et textes : le Berlin intellectuel des années 1800*
- La documentation sur les règles suivies pour l'encodage des textes présents dans l'édition *Lettres et textes* : **Edition-specific TEI encoding guidelines**
- Le site web du logiciel d'OCR Kraken et le repository GitHub qui contient tout le code

Pour commencer

- Lecture de toutes les lettres transcrites à disposition et de l'inventaire → mise en contact avec les thèmes abordés par d'Estournelles de Constant dans sa correspondance et découverte des principaux protagonistes
- Lecture d'une majorité des éléments présents dans la bibliographie du projet DAHN → familiarisation avec le projet et ses composants, les travaux déjà réalisés et les parties essentielles pour mon travail
- Lecture des règles d'encodage et navigation sur le site *Lettres et textes* → repérage des éléments d'encodage particuliers et de la mise en place des différentes langues et catégories sur le site
- Lecture de toute la documentation à propos de Kraken → compréhension du fonctionnement des éléments d'entrainement, des modules qui s'utilisent, ainsi que du fonctionnement général du logiciel

Stockage de données

Stockage de données

- Deux espaces mis en place pour permettre de stocker les photos, les transcriptions, les scripts et autres fichiers qui constituent notre projet :
  - Un espace Sharedocs Humanum (fichiers statiques)
  - Un repository GitHub (fichiers en constante évolution)

Stockage de données

- Sharedocs Humanum
  - Répertoire = « DAHN »
  - Contient :
    - Photos de la correspondance prises aux AD de la Sarthe, triées par dossiers des AD puis par numéro de lettre avec la date
    - Données d'entrainement pour la transcription avec un dossier par lettre qui comprend les photos de la lettre et les fichiers XML correspondants, ainsi que les fichiers XMLLIST qui permettent de les appeler dans une ligne de commande

Stockage de données

- Repository GitHub
  - DAHNProject
  - Contient :
    - Documentation d’encodage
    - Index
    - Transcription (modèle et exemple)
    - Scripts (modèle, transcription, encodage, etc…)

FloChiff / DAHNProject

Unwatch 2 | Star 0 | Fork 0

Code | Issues 1 | Pull requests 0 | Actions | Projects 0 | Wiki | Security 0 | Insights | Settings

Project DAHN "Digital Edition of historical manuscripts (correspondences)"

Edit

Manage topics

29 commits | 4 branches | 0 packages | 0 releases | 3 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

FloChiff Modifications in the msIdentifier section — Latest commit abf5e05 2 hours ago

Guidelines	Modifications in the msIdentifier section	2 hours ago
Indexes	New entries in the index	4 hours ago
Scripts	Adding the transcription script	21 days ago
Transcription	Modifications made on the manuscript description	3 hours ago
.gitignore	.DS_Store banished!	2 months ago
README.md	README Update	19 days ago

README.md

Établir l’encodage

Établir l’encodage

- Nécessaire de mettre en place des règles pour l’encodage d’égodocuments, qui sont le type de documents sur lequel on va travailler pour la plateforme, avec notamment la structure d’une lettre
- Trois types de documents à réaliser :
  - Guidelines (règles d’encodage)
  - Transcription (exemple avec une lettre)
  - Index (cinq index à créer)

Établir l’encodage

- Guidelines :
  - Reprise des règles d’encodage écrites pour *Lettres et textes*
  - Suppression des éléments qui n’avaient aucun lien avec les égodocuments (drama, front, back, etc.)
  - Ajout d’éléments inhérents à la correspondance de Paul d’Estournelles de Constant
  - Restructuration du plan des guidelines

Établir l’encodage

- Guidelines :
  - Rédaction des guidelines en ODD avec une description et un exemple pour tous les éléments (exemple tiré de *Lettres et textes* ou du projet DAHN)
  - Création d’un <schemaSpec> qui limite le nombre de modules acceptés dans les fichiers de transcription
  - Transformation en deux versions :
    - Un fichier HTML (qui contient les liens vers les TEI Guidelines pour tous les éléments inclus dans le <schemaSpec>
    - Un fichier RNG (qui a été relié à la transcription modèle qui a été créée)

Établir l’encodage

Extrait du contenu du <schemaSpec>

```
<head>List of the modules used on my TEI files</head>
<schemaSpec ident="myTEI" docLang="en" prefix="tei_" xml:lang="en">
    <moduleRef key="core" except="email gloss headItem headLabel l lg meeting postBox said sp speaker stage"/>
    <moduleRef key="tei" except=""/>
    <moduleRef key="header"
        except="appInfo application cRefPattern catDesc catRef category classCode classDecl conversion interpretation listCha
    <moduleRef key="textstructure"
        except="argument back div1 div2 div3 div4 div5 div6 div7 docAuthor docEdition docImprint docTitle floatingText front
    <moduleRef key="namesdates" except=""/>
    <moduleRef key="transcr"
        except="am damage damageSpan ex line listTranspose metamark mod path redo restore retrace secl sourceDoc substJoin su
    <moduleRef key="msdescription"
        except="adminInfo colophon custEvent custodialHist depth dim filiation finalRubric heraldry incipit locus locusGrp ms
    />
</schemaSpec>
```

Guide d'encodage pour l'édition numérique d'égodocuments

Table of contents

Établir l’encodage

- Transcription :
  - Sélection d’une des lettres déjà transcrites qui possèdent à la fois les photos correspondantes et plusieurs éléments récurrents des lettres pour encoder le plus d’éléments possibles (Lettre n°477 du 4 février 1919)
  - Création de l’encodage (header et body) sans annotations
  - Réflexion sur certains éléments dont l’encodage est plus difficile dû à leur absence ou rareté dans les TEI Guidelines

Établir l'encodage

- Éléments particuliers dans la transcription :
  - Éléments récurrents dans le <opener>
    - Quatre éléments difficiles à encoder :
      - Numérotation des lettres (ex: LETTRE N°477)
      - En-tête de lettre
      - Statut de la lettre (ex: Personnelle)
      - Tampon des AD
    - Suivi de certaines des propositions des articles d'*Encoding Correspondence*
    - Choix de balisage exposé dans l'article « Working through minor issues »

Établir l’encodage

Démonstration de l’encodage de l’<opener> avec certains des éléments récurrents

```
<opener>
    <fw type="letterhead" place="align(left)" corresp="#entete-senat"><hi rend="underline">SÉNAT</hi>
    </fw>
    <dateline rend="align(right)">
        <placeName>Paris</placeName>, <date when-iso="1919-02-04">4 Février 1919</date>
    </dateline>
    <title rend="align(center)"><hi rend="underline">LE PRÉSIDENT WILSON<lb/> REÇU A LA CHAMBRE DES
        DÉPUTÉS</hi></title>
    <salute rend="indent">Mon cher Butler,</salute>
</opener>
```

Démonstration de l’encodage du tampon dans le <teiHeader>

```
<additions>
    <p>A stamp has been put on almost every page of the correspondence by the institution in charge of
        the collection. On the stamp is written: <stamp resp="#stamp">Archives de la Sarthe
            <lb/>Propriété publique</stamp></p></additions>
```

Établir l'encodage

- Éléments particuliers dans la transcription :
  - Mise en place d'un système de niveaux dans la transcription
    - Issue #1 et Pull Request #2
    - Utilisation de la documentation du projet WeGA et des TEI Guidelines
    - Ajout d'un attribut *@status* sur le <revisionDesc> afin d'introduire des niveaux de finalisation pour les fichiers et de mettre en ligne au minimum des transcriptions brutes sans annotation
    - Trois niveaux :
      - Proposed = transcription brute
      - Unfinished = annotation en cours
      - Approved = document complètement finie

Établir l’encodage

Exemple d’une utilisation de l’attribut *@status*

```
<revisionDesc status="proposed">
    <change when-iso="2020-04-23" who="#floriane.chiffoleau">Transcription completed</change>
    <change when-iso="2020-04-02" who="#floriane.chiffoleau">Modifications on the encoding</change>
    <change when-iso="2020-04-01" who="#floriane.chiffoleau">Modifications on the encoding</change>
    <change when-iso="2020-03-19" who="#floriane.chiffoleau">Creation of the encoding (Body)</change>
    <change when-iso="2020-03-18" who="#floriane.chiffoleau">Creation of the encoding (Header)</change>
</revisionDesc>
```

Établir l’encodage

- Transcription :
  - Établissement d’un template d’arbre XML pour les lettres
  - Maintien des métadonnées communes à toutes les lettres dans l’arbre et suppression de toutes les informations spécifiques à la lettre utilisée pour créer le modèle d’encodage
  - Structure à utiliser pour y ajouter la version encodée d’une nouvelle lettre et compléter les métadonnées spécifiques

Établir l’encodage

- Index :
  - Mise en place d’index dont les règles d’encodage ont été établies dans les guidelines
  - Une à deux entrées (à titre d’exemple) dans chacun des index créés
  - Cinq index :
    - Bibliographie
    - Contributeurs
    - Lieux
    - Organisations
    - Personnes

Établir l’encodage

Exemple d’une entrée de l’index « Bibliographie »

```
<biblStruct xml:id="w00001">
    <monogr>
        <author>AKHUND Nadine</author>
        <author>TISON Stéphane</author>
        <title level="m">En guerre pour la paix, 1914-1919</title>
        <title level="m" type="sub">Correspondance Paul d'Estournelles de Constant
            et Nicholas Murray-Butler</title>
        <imprint>
            <pubPlace>Paris</pubPlace>
            <publisher>Alma</publisher>
            <date when-iso="2018"/>
        </imprint>
        <biblScope unit="page">546</biblScope>
    </monogr>
</biblStruct>
```

Établir l’encodage

Exemple d’une entrée de l’index « Contributeurs »

```
<person xml:id="floriane.chiffoleau">
    <persName>
        <forename>Floriane</forename>
        <surname>Chiffoleau</surname>
    </persName>
    <affiliation>
        <orgName ref="inria">Institut national de recherche en informatique et automatique (INRIA)</orgName>
        <address>
            <street>2 rue Simone Iff</street>
            <postCode>75012</postCode>
            <settlement>Paris</settlement>
            <country key="FR">France</country>
        </address>
    </affiliation>
    <occupation>Transcription du corpus</occupation>
    <occupation>Encodage du projet</occupation>
</person>
```

Établir l’encodage

Exemple d’une entrée de l’index « Lieux »

```
<place xml:id="l0001" type="city">
    <placeName>Paris</placeName>
    <country>France</country>
    <location>
        <geo>48°51′24″N, 2°21′07″E</geo>
    </location>
</place>
<place xml:id="l0002" type="city">
    <placeName>La Flèche</placeName>
    <country>France</country>
    <location>
        <geo>47°41′59″N, 0°04′34″W</geo>
    </location>
</place>
```

Établir l’encodage

Exemple d’une entrée de l’index « Organisations »

```
<org xml:id="sénat">
    <orgName from-iso="1875-02-24" to-iso="1942-08">Sénat (Troisième
        République)</orgName>
    <desc>Le Sénat sous la Troisième République est l'un des deux organes
        législatifs, le second étant la Chambre des députés, mis en place par les
        lois organiques des 24 et 25 février 1875. Il s'agit d'un bicaméralisme
        strict, les deux chambres ayant les mêmes pouvoirs législatifs. Les lois
        doivent être votées dans les mêmes termes par les deux chambres. Seul signe
        de préséance, la chambre des députés se prononce la première pour les lois
        des finances. En revanche le chef de l'État doit obtenir l'avis conforme du
        Sénat pour procéder à la dissolution de la chambre basse. Politiquement, la
        création de la Chambre Haute est un compromis entre une Assemblée nationale
        où monarchistes (eux-mêmes divisés entre orléanistes et légitimistes) et
        républicains s'opposent. Ces derniers acceptent la présence d'une assemblée
        ayant un caractère conservateur en échange du ralliement des premiers à la
        République.</desc>
    <placeName from-iso="1799">
        <placeName>Palais du Luxembourg</placeName>
        <country>France</country>
        <location>
            <geo>48°50′54″N, 2°20′15″E</geo>
        </location>
    </placeName>
    <idno type="VIAF">148585812</idno>
</org>
```

Établir l’encodage

Exemple d’une entrée de l’index « Personnes »

```
<person xml:id="p0001">
    <persName>
        <roleName type="nobility">Baron</roleName>
        <forename>Paul</forename>
        <nameLink>d'</nameLink>
        <surname>Estournelles de Constant</surname>
    </persName>
    <nationality>French</nationality>
    <birth>
        <date when-iso="1852-11-22"/>
        <placeName>La Flèche (Sarthe)</placeName>
    </birth>
    <death>
        <date when-iso="1924-05-15"/>
        <placeName>Bordeaux (Gironde)</placeName>
    </death>
    <sex value="1"/>
    <occupation>diplomate</occupation>
    <!-- <education>...</education> -->
    <affiliation notBefore-iso="1895" notAfter-iso="1904">Député de <placeName>la Sarthe</placeName></affiliation>
    <affiliation notBefore-iso="1904" notAfter-iso="1924">Sénateur de <placeName>la Sarthe</placeName></affiliation>
    <event when-iso="1909">
        <p>Prix Nobel de la paix</p>
    </event>
    <idno type="VIAF">15798950</idno>
</person>
```

Création de scripts

Création de scripts

- Réalisation d’un modèle de transcription
  - Modèle uniquement à partir des données d’entrainement (scratch)
  - Modèle à partir des données d’entrainement et d’un modèle (finetune)
- Transcription des images en fichiers texte
- Encodage des textes

Création de scripts

- Réalisation d'un modèle uniquement à partir des données d'entrainement (scratch) : **train-letter-newmodel.sh**
  - Création de données d'entrainement à partir de certaines des lettres déjà transcrites et des photos correspondantes
  - Utilisation de l'interface *eScriptorium* (binarisation des images, segmentation manuelle ou avec un modèle, transcription manuelle en effectuant un copier/coller des lettres)
  - Export des données avec XML ALTO qui contient toutes les lignes d'une page dans un seul fichier XML et les indique par des coordonnées géographiques

Création de scripts

- Réalisation d'un modèle à partir des données d'entrainement et d'un modèle (finetune) : **train-letter-finetune.sh**
  - Utilisation des données d'entrainement mentionnées précédemment
  - Utilisation d'un modèle d'alphabet latin mais de langue anglaise qui ne reconnait donc pas les signes diacritiques et un grand nombre de mots
  - La ligne de commande ne diffère pas beaucoup : seul y est rajouté le module pour appeler le modèle et le chemin vers celui-ci

Création de scripts

- Transcription des images en fichiers texte : **script_transcription.py**
  - Fichier d'origine = script d'entrainement de transcription Kraken pour créer des fichiers HTML de plusieurs images en même temps (écrit par Alix Chagué)
  - Réécriture du fichier :
    - La ligne de commande correspond aux arguments à inscrire pour réaliser une transcription
    - Les variables appellent les dossiers d'entrée et de sortie, ainsi que le modèle de transcription

Création de scripts

- Encodage des textes : **script_encodage.py**
  - Accélérer et faciliter le travail d'encodage des textes afin de n'avoir qu'à encoder un nombre restreint de parties du texte et les métadonnées qui ne peuvent être encodées qu'à la main
  - Encodage basique qui se base sur l'utilisation d'expressions régulières et d'un cherche/remplace dans le texte.

Création de scripts

- Exécution
  - Modèle pour la transcription → le résultat n'est pas encore adéquat, les scripts ont besoin d'ajustement
  - Transcription des textes → Ce script sera probablement délaissé pour se concentrer sur une transcription avec l'interface graphique *eScriptorium*
  - Encodage des textes → le script fonctionne sur les versions transcrites utilisées pour la création de données d'entrainement mais il sera nécessaire de voir comment cela fonctionne avec les textes transcrits avec *eScriptorium*

Difficultés rencontrées

Difficultés rencontrées

- Fonctionnement de Kraken dans l'environnement GPU créé sur le cluster RIOC de l'INRIA
- Création d'un modèle de transcription correcte et efficace
- Création de données d'entrainement sur *eScriptorium*

Difficultés rencontrées

- Fonctionnement de Kraken dans l'environnement pour GPU créé sur le cluster RIOC de l'INRIA :
  - Certaines versions des packages installés sur le cluster sont plus anciennes que celles présentes dans Kraken donc certains bugs empêchent la bonne exécution du script d'entrainement
  - Problème récurrent et solutionné récemment donc il n'a pas encore été possible de travailler assez longtemps sur le cluster pour obtenir un résultat convenable

Difficultés rencontrées

- Création d'un modèle de transcription correcte et efficace :
  - Entrainement lancé sur mon ordinateur car impossible (au moment de l'exécution) de travailler sur le GPU
  - Efficacité du script
  - Efficacité des données d'entrainement

Difficultés rencontrées

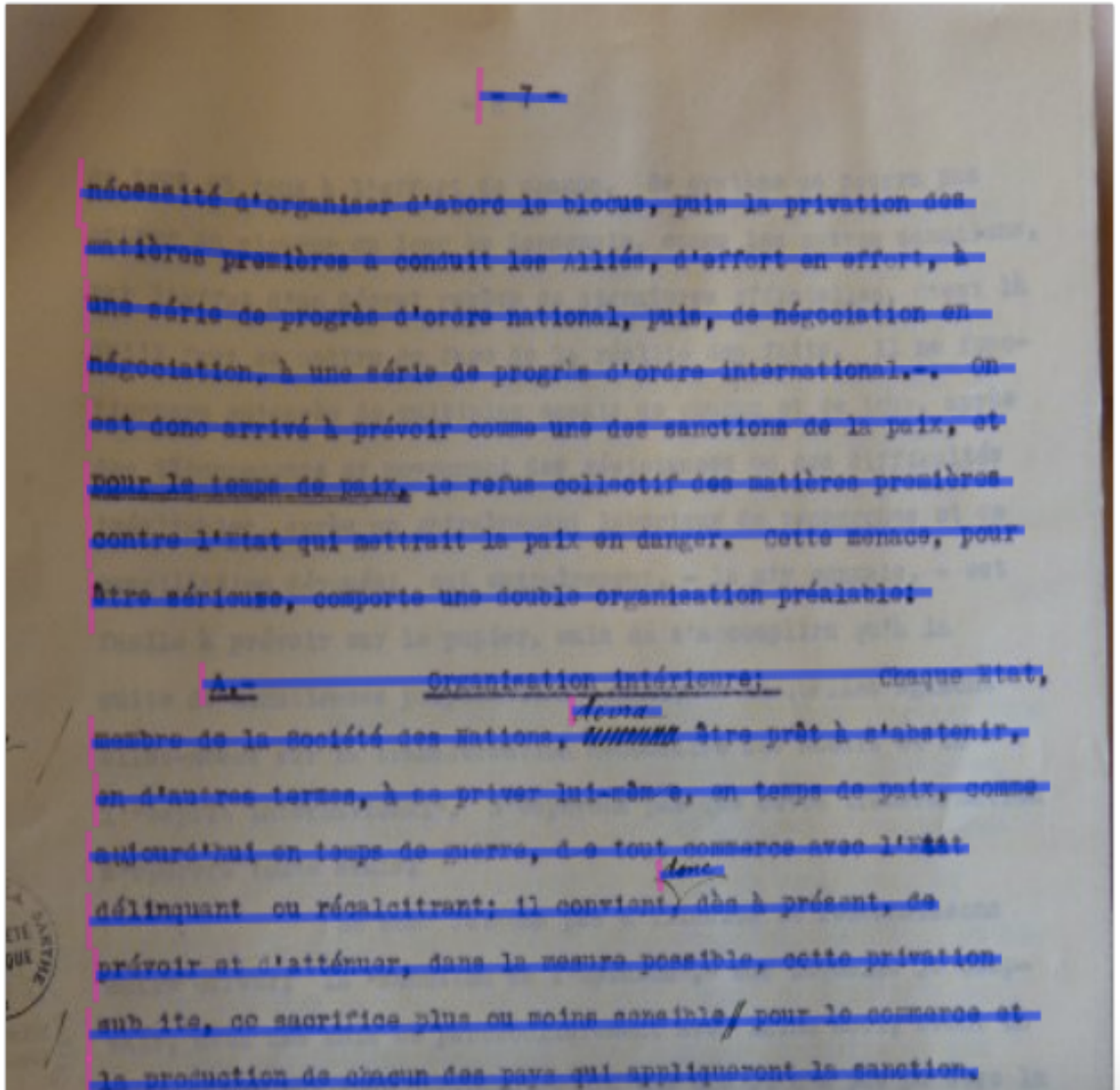
- Création d'un modèle de transcription correcte et efficace :
  - Entrainement lancé sur mon ordinateur car impossible (au moment de l'exécution) de travailler sur le GPU → **exécution du script très long, entrainement prend plus d'1h30**
  - Efficacité du script → **nécessaire de l'ajuster avec de multiples modules pour qu'il fournisse un résultat pertinent**
  - Efficacité des données d'entrainement → **début à 1000 lignes (insuffisant), rajout de 500 (précision améliorée mais toujours insuffisant), nécessité de rajouter encore des lignes d'entrainement**

Difficultés rencontrées

- Création de données d'entrainement sur *eScriptorium* :
  - Trois éléments qui imposent des choix de présentation spéciaux :
    - Rajout de mots au dessus des lignes d'écriture standard
    - Écriture manuscrite de certaines informations (ex : signature)
    - Présence de ratures parfois dans le texte (un ou deux mots mais aussi des lignes entières parfois)

Difficultés rencontrées

- Création de données d'entrainement sur *eScriptorium* :
  - Trois éléments qui imposent des choix de présentation spéciaux :
    - Rajout de mots au dessus des lignes d'écriture standard → **Marqué comme une ligne unique dans la segmentation**
    - Écriture manuscrite de certaines informations (ex : signature) → **Délimité par le signe £, doublé de chaque côté (signe choisi car n'apparait pas dans le texte)**
    - Présence de ratures parfois dans le texte (un ou deux mots mais aussi des lignes entières parfois) → **Délimité par le signe €, un de chaque côté (signe choisi car n'apparait pas dans le texte)**



nécessité d'organiser d'abord le blocus, puis la privation des matières premières a conduit les Alliés, d'effort en effort, à une série de progrès d'ordre national, puis, de négociation en négociation, à une série de progrès d'ordre internationale.-. On est donc arrivé à prévoir comme une des sanctions de la paix, et pour le temps de paix, le refus collectif des matières premières contre l'Etat qui mettrait la paix en danger. Cette menace, pour être sérieuse comporte une double organisation préalable:

A.- Organisation intérieure : Chaque Etat, membre de la Société des Nations, ff devraff €€ être prêt à s'abstenir, en d'autres termes, à se priver lui-mêm e, en temps de paix, comme aujourd'hui en temps de guerre, d e tout commerce avec l'Etat ffdoncff délinquant ou récalcitrant ; il convient dès à présent, de prévoir et d'atténuer, dans la mesure possible, cette privation sub ite, ce sacrifice plus ou moins sensible€.€ pour le commerce et la production de chacun des pays qui appliqueront la sanction.

Publications de billets de blog

Publications de billets de blog

- Blog de publication (créé par Anne Baillot) :
  https://digitalintellectuals.hypotheses.org
- Afin de suivre l'avancée de mon travail sur le projet, de documenter les différentes étapes ou bien de mettre en avant les difficultés rencontrées
- Rédaction en anglais, articles généralement illustrées et sujets assez divers

Publications de billets de blog

- « Encoding an XML Tree model for my corpus », 25 mars 2020 :

  https://digitalintellectuals.hypotheses.org/3360
- « Starting a new project – Discovering its source material », 31 mars 2020 :

  https://digitalintellectuals.hypotheses.org/3398
- « Working through minor issues », 7 avril 2020 :

  https://digitalintellectuals.hypotheses.org/3528
- « How to produce a model for the transcription », 7 mai 2020 :

  https://digitalintellectuals.hypotheses.org/3702

CATEGORY: WORK PROGRESS

DAHN PROJECT / WORK PROGRESS 07/05/2020

0

How to produce a model for the transcription

My work on the DAHN project can be divided into two major parts: transcribing and encoding. I already mentioned the encoding in two previous posts (there and there) and now it is time I talk a little bit more about the other part of my work. In order to encode...

DAHN PROJECT / WORK PROGRESS 07/04/2020

0

Working through minor issues

While I worked on the creation of my XML tree model, I realized that some recurrent elements of the letters will be more difficult to encode, mainly because they are neither mentioned nor explained on the TEI Guidelines. While the guidelines contain many sections and cover a lot of different...

DAHN PROJECT / WORK PROGRESS 31/03/2020

0

Starting a new project – Discovering its source material

At the beginning of this month, I started to work at the French Institute for Research in Computer Science and Automation (INRIA) where I have been hired to work on the DAHN project. The project, that I have already mentioned in my previous post, has two corpora: Letters and texts (already...

DAHN PROJECT / WORK PROGRESS 25/03/2020

0

Encoding an XML Tree model for my corpus

I work on a project called "Digital edition of historical manuscripts" which aims to diffuse on a public platform multiple corpora that all have one thing in common: they are ego documents, which refer to personal writings, memoirs, correspondence and documents alike. My job, personally, is focused on one corpus,...

Conclusion

Conclusion

- Futures tâches à réaliser
  - Développer un modèle adéquat pour la transcription des lettres
  - Effectuer la binarisation, segmentation et transcription des lettres avec *eScriptorium*
  - Exécuter le script d'encodage en suivant pour vérifier son bon fonctionnement avec les données obtenues

Merci de votre attention