

Édition à l'ère numérique

HTR et XML TEI

Ariane Pinche¹², Matthias Gille Levenson³²

¹CNRS, ²CIHAM, ³ÉNS Lyon

24 janvier 2023

Table of Contents

- 1 Édition à l'ère numérique : introduction
- 2 Acquisition des corpus : Reconnaissance automatique d'écriture
 - Définition
 - Scores et évaluation des performances d'un modèle HTR
 - Constituer et partager des modèles et des données d'entraînement
 - Recherche et HTR
 - Présentation de Kraken et eScriptorium
- 3 Édition numérique et XML TEI
 - Principes généraux
 - Les éditions scientifiques
 - Qu'est-ce que XML TEI
 - XML
 - TEI

Au commencement était le texte...

Toute édition est le fruit d'une époque et d'une école philologique. Les éditions traditionnelles, héritières de tradition centenaire, sont limitées à leur matérialité et présentent un texte extrêmement lissé et figé.



VIE DE SAINT MARTIN

De saint Martin

1. Mout¹ doit on doucement et volentiers le bien oïr et entendre, car par le bien savoir et retenir [fol.103b] puet l'en sovent a bien venir. Qui bien ne seit ne bien n'entent de bien faire n'a nul talent. Mes del bien nest sovent li biens, del mal li maussi com dist l'Ecriture. Por ce se doit l'en au bien avoirier et le bien feire, si com li saint home furent ça en arriere de cui nos trovons les oevres et les vies [es] Ecritures. Et bien sacent tuit cil qui vivent qe ja n'auront tant de bien fet en totes lor vies qe, qant la mort dont nule rien n'eschape les poindera au cuer, q'il ne cuident petit avoir fait. Dex ! Qe feront dont cil qui riche sont et aise de l'avoir de cest siecle, ne en eus n'ont doucor ne humilité ne misericorde, ainz sont plein d'angoisse et de traisson et de felonie et de si grant

5

10

15

'La Vie de saint Martin s'ouvre sur une majuscule historiée (m) qui représente Saint Martin sur un cheval, l'épée à la main, prêt à couper son manteau pour le partager avec un nécessiteux également présent dans l'illustration.

1 De saint Martin] Ci commence la vie de monseigneur saint Martin C^o, Ci commence la vie saint Martin C^o || ♫ es Ecritures C^C | escriptures C^c

Au commencement était le texte...

L'édition numérique permet de s'affranchir de ces limitations

- Les données peuvent être enrichies
 - ▶ Annotations linguistiques
 - ▶ Variantes graphiques
 - ▶ Versions du texte
- Le travail préliminaire à l'établissement du texte est sauvegardé
- Les données textuelles peuvent être exploitées en dehors du cadre de l'édition
 - ▶ Reprises (correction de l'édition, insertion dans un autre corpus, etc.)
 - ▶ Multiplication des visualisations (imitative, normalisée, etc.)
 - ▶ Analyses statistiques (stemmatologie, stylométrie, dataviz, etc.)

Au commencement était le texte...

- Comment acquérir son texte ?
 - ▶ En transcrivant manuellement
 - ▶ En partant d'un texte nativement numérique
 - ▶ En utilisant l'acquisition automatique de texte (HTR / OCR)
- Comment enrichir son texte ?
 - ▶ Utiliser un système de balisage (XML TEI)
 - ▶ Annoter manuellement
 - ▶ Utiliser des outils d'annotation automatiques (TAL, NER)

Table of Contents

1 Édition à l'ère numérique : introduction

2 Acquisition des corpus : Reconnaissance automatique d'écriture

- Définition
- Scores et évaluation des performances d'un modèle HTR
- Constituer et partager des modèles et des données d'entraînement
- Recherche et HTR
- Présentation de Kraken et eScriptorium

3 Édition numérique et XML TEI

- Principes généraux
- Les éditions scientifiques
- Qu'est-ce que XML TEI
 - XML
 - TEI

Du miracle de l'intelligence artificielle

Les mains peuvent trembler, l'écriture s'effacer,
Mais grâce à l'intelligence artificielle,
La reconnaissance automatique s'éveille,
Les mots restent gravés, le sens préservé.

Les lettres s'enchaînent sur l'écran, comme un ballet,
Et les phrases se forment, dans un mouvement gracieux,
Sans effort, sans douleur, sans risque d'erreur,
La reconnaissance automatique est un miracle.

Elle permet de transcrire les pensées,
De conserver les mémoires, les idées,
Et de les partager avec les autres.

L'intelligence artificielle est à l'œuvre,
Pour améliorer sans cesse la reconnaissance,
Qui devient plus précise, plus fiable, plus sûre.

La reconnaissance automatique d'écriture,
Est un outil précieux pour l'humanité,
Qui nous permet de transcrire notre savoir.

Figure: Poème écrit par ChatGPT

Qu'est-ce que la reconnaissance automatique d'écriture ?



Figure: Prédiction HTR

- Prédiction d'un contenu texte
- à partir d'une image de la source par une
- intelligence artificielle entraînée par un humain
- dans un processus alternant
 - ▶ phases d'interventions humaines
 - ▶ et phases de calcul

Différences entre OCR et HTR

OCR	HTR
Performance : Taux d'erreur sur les caractères inférieur à 2 %, fonctionne uniquement sur les documents imprimés	Performance : Taux d'erreur sur les caractères entre 5 et 10 %, fonctionne sur les documents manuscrits
Outils : Abby (adobe), mais commercial, pas de code ouvert; Tesseract 4 (gratuit, code ouvert)	Outils : Transkribus (commercial) ou Kraken (gratuit, code ouvert)
Fonctionnement : Modèles génériques par langue préexistants et s'appuie sur des fontes de caractères	Fonctionnement : nécessite la constitution d'un corpus d'entraînement pour entraîner un modèle

!! Ce n'est plus aussi tout à fait vrai aujourd'hui...

Un peu d'histoire...

- 1990 : Développement de l'usage de l'OCR
- 2000-2010 : Développement de l'usage de l'HTR (expérimental)
- fin 2010: Augmentation de l'usage de l'HTR dans les projets de recherche avec le développement de deux outils : Trankribus (moteur HTR Pylaia et HTR+) et eScriptorium (moteur HTR Kraken)
- 2022: l'HTR est devenue une étape courante dans les pipelines d'acquisition textuelle
 - ▶ Colloque "Documents anciens et reconnaissance automatique des écritures manuscrites"
 - ▶ DH conference
 - ▶ TEI conference
 - ▶ intéresse aussi bien les projets de recherche que les institutions patrimoniales

Un peu d'histoire...

Nowdays, with model that can reach a CER (character error rate) between 8% and 2% for manuscripts, “from a computer science point of view, the recognition of handwriting seems to be a resolved task. The latest recognition engines allow for the successful recognition of specifically trained hands producing a text as reusable data”

*Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021.
“General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example.” Journal of Open Humanities Data 7(0):13.
doi: 10.5334/johd.46.*

Un peu d'histoire...

Les défis à relever en 2023...

- Faciliter l'accès à l'HTR pour les non-spécialistes
- Mettre en commun les données
- Créer des modèles généraux/génériques pour les grands corpus

"Well-prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently"

Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021.

*"General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example." Journal of Open Humanities Data 7(0):13.
doi: 10.5334/johd.46.*

Les étapes de l'HTR

- Chargement des images
 - ▶ Chargement d'une collection d'image en JPG ou tif en local
 - ▶ Chargement depuis un manifeste iiif (e.g collections issues de Gallica ou de e-Codices)
- Traitement des images (facultatif)
 - ▶ résolution 300dpi
 - ▶ couleur ou niveau de gris
 - ▶ possible binarisation pour réduire le bruit
 - ▶ imagerie multispectrale (dans le cas de documents très abimés)

Les étapes de l'HTR

- Segmentation des zones de l'image



Figure: Bnf, fr. 412, fol.10r

Les étapes de l'HTR

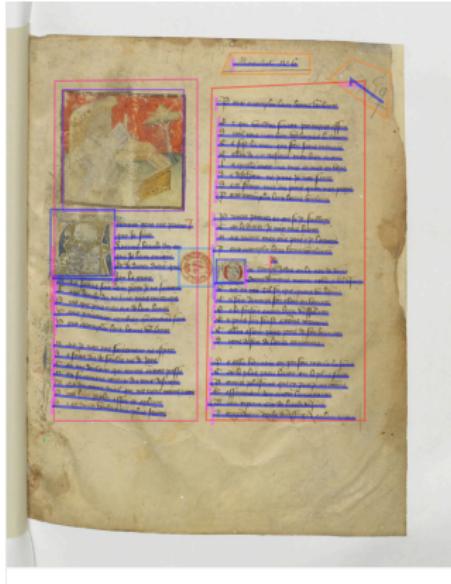
- Segmentation des lignes contenant du texte



Figure: Bnf, fr. 1728, fol.8v

Les étapes de l'HTR

- Prédiction du texte qui se trouve sur l'image



1 uernes gens me prient ¶.
2 que le fac
3 Aucuns beaux diz et
4 que le leur envoye
5 Et de diter dient que
6 A
7 Mais sauau soit leur paix le ne sauroye
8 lay la gracie
9 Faire beaulte diz ne bons, mais toutsuoys
10 Puis que prie men ont de leur bonta
11 Paix y mettray combien quignoult soyse
12 Pour accomplir leur bonne oulante
13 Mais le n ay pas sentement ne espere
14 De faire diz, de soulaus ne de loye
15 Car ma douleur qui toutes autres passe
16 Mon sentement loyeus du tout desuoye
17 Mais du grant dueil qui me tient morne ¶coye
18 Puis bien parler assez et aplaine
19 Si en dray oulementiers plus feroye
20 6259
21 Pour accomplir leur bonne oulante
22 Et qui voudra sauau pourquoi efface
23 Duel tout mon bles oulementier le diroye
24 Ce fist la mort qui ferri sans mercie
25 Celli de qui trestout mon bien ausye
26 Laquelle mort ma mis et met en uoye
27 De dessooper ne puis le nos sante
28 De ce feray mes diz puis quon men proye
29 Pour accomplir leur bonne oulante
30 Princes premes en gre se le falloye
31 Car le dister le nay mie hante
32 Mais maint men ont prie ¶ le lottroye
33 Pour accomplir leur bonne oulante
34 u temps ladis en le cite de Rôme
35 .JI.
36 O
37 ung en yet. Tel fu que quant un hôme
38 Orient Rômaines maint noble ¶ bel usage

Figure: Bnf, fr. 12779, fol.9r

Les étapes de l'HTR

- Export des données (txt, alto, page)

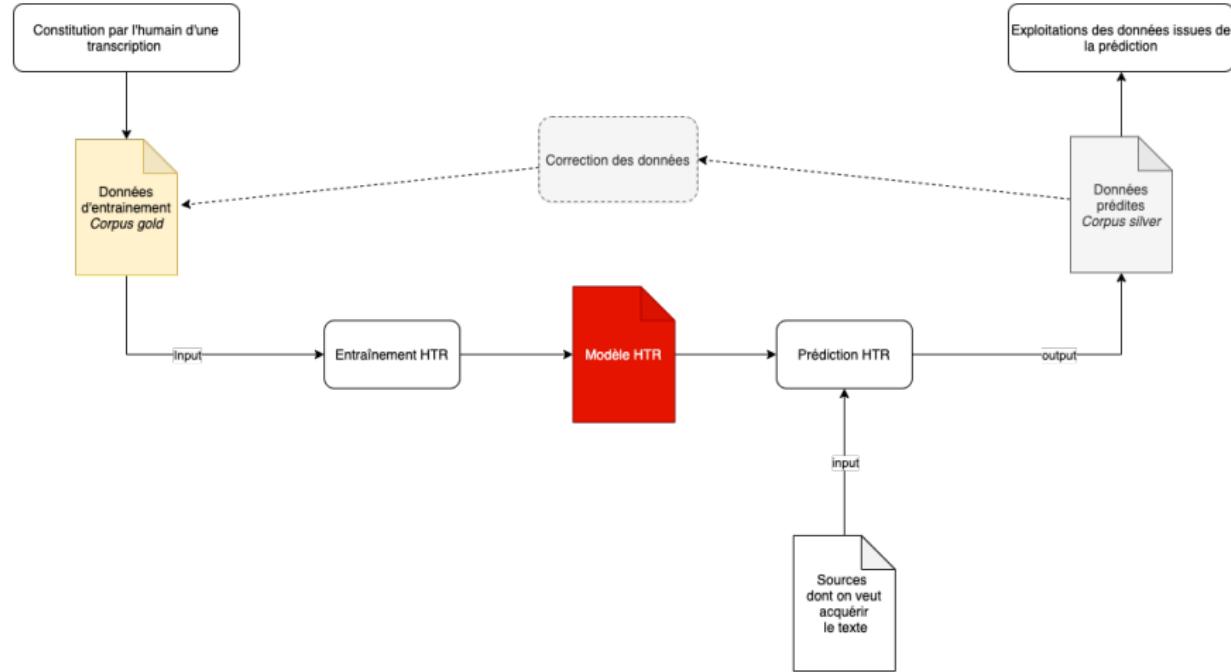
```
<Layout>
    <Page WIDTH="4648" HEIGHT="3407" PHYSICAL_IMG_NR="8" ID="eSc_dummypage_>
        <PrintSpace HPOS="0" VPOS="0" WIDTH="4648" HEIGHT="3407">

            <TextBlock HPOS="693" VPOS="321" WIDTH="1701" HEIGHT="2451"
                ID="eSc_textblock_08b9f915" TAGREFS="BT3852">
                <Shape>
                    <Polygon
                        POINTS="693 413 693 2772 2394 2772 2254 321"/>
                </Shape>

                <TextLine ID="eSc_line_d939596f" TAGREFS="LT1299"
                    BASELINE="746 476 2143 428" HPOS="743" VPOS="352"
                    WIDTH="1400" HEIGHT="156">
                    <Shape>
                        <Polygon
                            POINTS="2078 388 2050 388 2021 386 1993 383 1964 383 1936 380 1908 377 1876 374 1848 374 1820 371 1811
                            />
                    </Shape>
                    <String
                        CONTENT="fors de la ville. Tant fut l'assaut merveilleux et"
                        HPOS="743" VPOS="352" WIDTH="1400" HEIGHT="156"/>
                </TextLine>
            </TextBlock>
        </PrintSpace>
    </Page>
</Layout>
```

Figure: Exemple de fichier Alto

Entrainer un modèle HTR



Pourquoi utiliser ou créer modèle pour l'HTR ?

- Pour accélérer la phase d'acquisition du texte. La prédiction peut servir :
 - ▶ de base à une édition : niveau de précision haut, supérieur à 95 % d'*accuracy*
 - ▶ à de la mise à disposition de texte brut : niveau de précision moyen, entre 90 % et 95 %
 - ▶ de base à des analyses quantitatives : niveau de précision faible, supérieur à 80 % (voir EDER, Maciej, « Mind your corpus: systematic errors in authorship attribution », *Literary and Linguistic Computing*, vol. 28 / 4, décembre 2013, p. 603-614.)
- Où trouver les données pour entraîner un modèle ? Où vérifier si un modèle performant existe déjà ? HTR-united

Principes

Pour évaluer un modèle HTR

- on prépare son corpus
 - ▶ train set (80%) - entraînement
 - ▶ dev set (10%) - évaluation de l'entraînement pendant les cycles d'apprentissage
 - ▶ test set (10%) - données jamais vues pendant l'entraînement
- on compare :
 - ▶ une vérité de terrain (GT) produite par un humain (test set)
 - ▶ à la prédiction des mêmes lignes par le modèle
 - ▶ pour calculer un score qui prend soit la forme :
 - ★ d'un CER (Caracter Error Rate)
 - ★ d'une Accuracy (précision, soit le pourcentage de réussite du modèle)

Types d'erreurs



Calcul du CER

$$CER = \frac{S + D + I}{N}$$

Performances Kraken : Cas d'étude

Modèle *saintMartin* entraîné sur le manuscrit BnF, fr. 412



Performances Kraken : Cas d'étude

Modèle *saintMartin* entraîné sur le manuscrit BnF, fr. 412

- Train : 10 folios - soit 1680 lignes transcrives
- Accuracy : 95.38% sur une même main

Type d'erreur	Nombre total	Tx d'erreurs/ligne
Insertions	1 883	0.76
Délations	725	0.29
Substitutions	1 823	0.73

Performances Kraken : Cas d'étude

Table des erreurs les plus fréquentes

Nb d'erreurs	Vérité de terrain	Prévision
762	[SPACE]	[]
473	[]	[SPACE]
162	[i]	[]
77	[.]	[]
73	[n]	[]

Exemple de prédiction du modèle dans eScriptorium

The screenshot shows the eScriptorium interface. At the top, there are two small circular icons and the text "Line #3". Below this is a text area containing a block of Old French text. A specific line is highlighted with a red box and a red underline under the word "qatant:qeli". The line reads: "qatant:qeli seinz euesques fu enterrez.Nil". At the bottom of the interface, there is a text input field containing the same line of text with the underlined word. Below the input field, there is a small "Toggle history" button and a question mark icon. The footer of the interface includes the text "by apinche (eScriptorium) on Tue Oct 19 2021 10:02:39 GMT+0200" and a series of navigation icons.

Performances Kraken : Cas d'étude

Modèle *Cremma-medieval*

- entraîné sur onze manuscrits différents
- 18385 lignes transcrites
- modèle Bicerin :
 - ▶ 22629 Characters
 - ▶ 1020 erreurs
 - ▶ 95,49% d'accuracy, mais sur le corpus complet qui comporte des mains différentes et des manuscrits compris entre le 13^eet le 14^esiècle

Performances Kraken : Cas d'étude

Type d'erreur	Nombre total
Insertions	317
Délations	229
Substitutions	474

Table des erreurs les plus fréquentes

Nb d'erreurs	Vérité de terrain	Prévision
160	[SPACE]	[]
153	[]	[SPACE]
35	[u]	[v]
34	[i]	[]
14	[u]	[n]
73	[v]	[u]

Entrainement d'un modèle pour les manuscrits médiévaux

Un modèle possède une élasticité limitée, il peut arriver qu'il atteigne ses limites, car le document proposé est trop différent de son corpus d'entraînement.



Figure: Prédiction Manuscrit 15e par le modèle saintMartin



Figure: Prédiction Manuscrit 15e par Bicerin

Personnaliser un modèle

Résolution du problème en personnalisant (finetuning) un modèle existant avec 10 pages, voire 5 pages de la nouvelle source

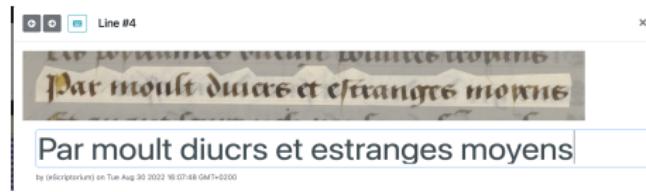


Figure: Prédiction Manuscrit 15e avec un modèle personnalisé à partir de Bicerin

Constituer et partager des modèles et des données d'entraînement

“Well prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently”

Hodel, 2021, p. 7

Constituer des données d'entraînement

- Identifier les besoins :
 - ▶ Produire un corpus lisible sur un document particulier;
 - ▶ Produire une transcription proche de la source avec le moins d'interprétation possible et la plus grande "élasticité" possible pour le modèle;
 - ▶ Gérer la tension entre le besoin de modèle(s) générique(s) et le besoin d'un modèle spécifique pour chaque projet.
- Mettre en place de normes et des ontologies pour transcrire et décrire les sources

Définir des normes de transcription

L'harmonisation des données permettra d'échanger des données et des modèles HTR. Comment faire ?

- Définir des méthodes de transcription adaptées à une problématique de recherche et à l'apprentissage machine.
- Définir le degré de précision recherché dans la transcription
- Utiliser un set de caractères prédéfini et documenter ses choix
- Voir les préconisations de transcription proposées dans la cadre de CREMMA Lab pour les textes médiévaux, et celles de CREMMA pour les transcriptions modernes.

Utiliser des ontologies

- Repérage des différentes zones du document : utiliser un vocabulaire contrôlé, comme SegmOnto.



Figure: Bnf, fr. 412, fol.10r

SegmOnto

<https://github.com/SegmOnto>

Page

- DamageZone
- DropCapitalZone
- FigureZone
- MainZone
- MarginZone
- MusicNotationZone
- NumberingZone
- QuireMarksZone
- RunningTitleZone

Line

- DefaultLine
- DropCapitalLine
- Interlinear
- MusicLine
- HeadingLine

Definitions

<https://github.com/SegmOnto/examples>

Rendre disponible ses données

- Déposer ses données sur un dépôt accessible en ligne
 - ▶ Github
 - ▶ GitLab
- Documenter ses données
 - ▶ Format des données
 - ▶ Nombre de ligne transcrites
 - ▶ Outils de segmentation
 - ▶ Moteur HTR
 - ▶ Langue du corpus
 - ▶ Date
 - ▶ Type de document, d'écriture
 - ▶ Méthode de transcription
- Rendre visibles ses données : intégrer un catalogue, voir HTR-united.

- Développement des outils mis à disposition : Transkribus, eScriptorium, Kraken
- Des projets pionniers : le projet Himanis (2015), le projet ANR Horae (2017) dirigés par Dominique Stutzmann .
- Une technologie qui fait partie des attendus :
 - ▶ Le projet Biblissima+ a consacré un de ses clusters à cette problématique : cluster 3 "*Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites*"



Kraken

- Outil d'analyse de mise en page et d'HTR
- fondé sur de l'apprentissage profond
- développé par Ben Kiessling dans le projet Scripta (PSL);
- Module Python, <https://github.com/mittagessen/kraken>;
- Doc: kraken.re
- Il peut être utilisé directement en ligne de commande ou via l'interface d'eScriptorium

eScriptorium

- logiciel libre qui permet de segmenter un document, de détecter les lignes, de transcrire, d'entraîner un modèle HTR et de l'appliquer à ses sources
- développé dans le cadre du projet Scripta (PSL);
- se branche sur Kraken pour l'analyse de mise en page et HTR;
- nécessite d'être déployé sur un serveur par une institution;
- Code: <https://gitlab.inria.fr/scripta/escriptorium>;
- démos vidéos: <https://scripta.hypotheses.org/escriptorium-video-gallery>.

Utiliser eScriptorium nécessite :

- D'ouvrir un compte sur une instance d'eScriptorium ou d'installer une instance locale d'eScriptorium
- D'avoir accès aux fichiers images de ses sources : des fichiers locaux ou téléchargés directement depuis un site institutionnel en utilisant un manifeste IIIF :
<https://gallica.bnf.fr/iiif/ark:/12148/btv1b84259980/manifest.json>

eScriptorium

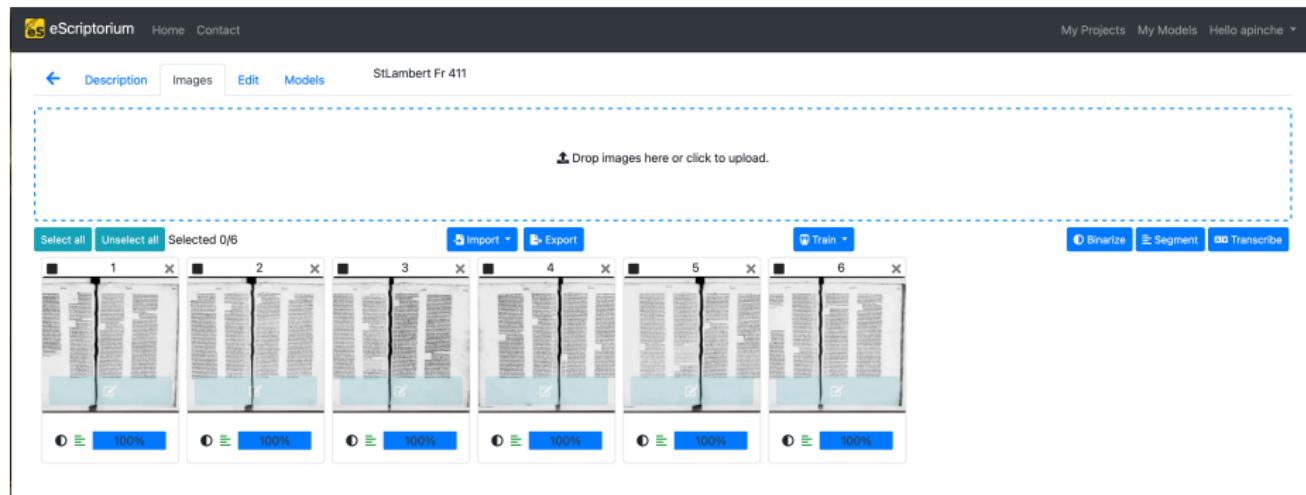


Figure: Interface d'eScriptorium

eScriptorium est une interface web qui permet :

- de segmenter la page d'un document (zones et lignes)
- de transcrire des documents pour créer des données d'entraînement
- d'entraîner un modèle HTR
- d'appliquer un modèle de HTR ou de segmentation à un document



MARTIN

Audit non le mouicin. Eti

Puet len souent abien ueni

Figure: Segmentation et transcription d'un document à l'aide d'eScriptorium

Table of Contents

- 1 Édition à l'ère numérique : introduction
- 2 Acquisition des corpus : Reconnaissance automatique d'écriture
 - Définition
 - Scores et évaluation des performances d'un modèle HTR
 - Constituer et partager des modèles et des données d'entraînement
 - Recherche et HTR
 - Présentation de Kraken et eScriptorium
- 3 Édition numérique et XML TEI
 - Principes généraux
 - Les éditions scientifiques
 - Qu'est-ce que XML TEI
 - XML
 - TEI

L'édition scientifique à l'ère numérique: structuration et annotation des données textuelles en XML-TEI

L'édition numérique permet de consigner de nombreuses informations sur le texte qui seront :

- conservées
- interrogables
- reexploitables

Introduction

L'objet éditorial pourra être démultiplié

- plusieurs visualisations
- corpus avec des annotations linguistiques
- bases de données des variations du texte

Qu'est-ce qu'une édition numérique ?

Il existe différents niveaux d'édition numérique

- Mise à disposition d'un texte structuré
- Mise à disposition d'un texte structuré et enrichi
- Éditions scientifiques et/ou critique avec plusieurs strates d'information

Production de textes structurés

Ce sont les textes les plus représentés :

- Projet Perseus, exemple : Amphitryon de Plaute
- Labex OBVIL, exemple : Mercure Galant

Les éditions scientifiques

Elles ont pour but de rendre accessible et compréhensible (ajout de note de bas de page, d'index, de glossaires) le texte au lecteur, mais aussi de renseigner la manière dont le texte fonctionne (sa transmission, les variantes, les liens vers ses sources)

- Les éditions documentaires ou à visée paléographique, exemple : le Didascalicon d'Hugues de Saint-Victor
- Les éditions génétiques, exemple : Madame Bovary : l'histoire du texte à travers ses brouillons,
- les éditions critiques, exemple : *Li Seint Confessor* de Wauchier de Denain

Et même un peu plus...

Aujourd'hui XML TEI n'est plus utilisé pour de l'édition à proprement parler, mais aussi comme un format de pérennisation de l'information textuelle. Les fondateurs de la TEI réfléchissent à la mise au point d'un schéma d'encodage XML TEI minimal et générique (« comity ») pour favoriser les analyses de corpus et permettre ce qu'on appelle le « distant reading »

BURNARD, Lou, SCHÖCH, Christof et ODEBRECHT, Carolin, « In search of comity: TEI for distant reading », *Journal of the Text Encoding Initiative, Text Encoding Initiative Consortium*, mars 2021.

Et même un peu plus...

Le projet Gallic(orpor) a intégré également la création de fichiers XML TEI pour conserver ses corpus constituer à partir d'HTR.

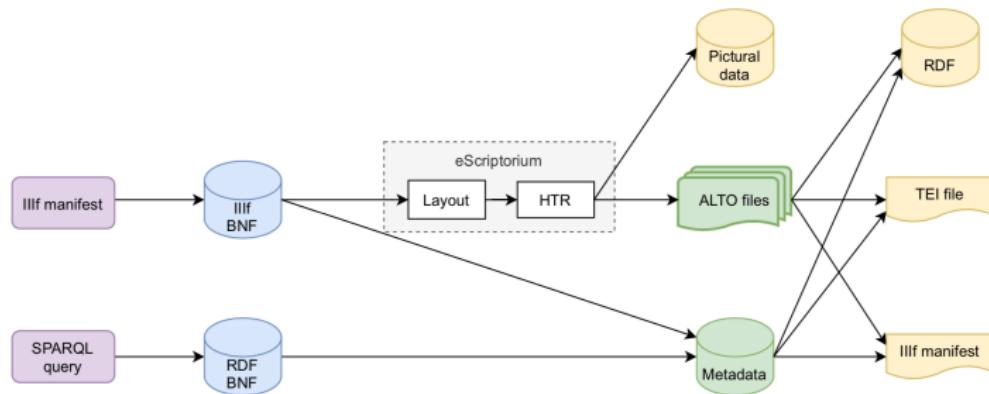


Figure: Protocole de création des données du projet Gallic(orpor)a

XML

XML est un format de données pur, très simple et documenté, conçu pour la **description** des documents textuels. XML ne possède pas de jeu de balises prédéfini.

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

XML : un standard international

Depuis 1998, XML est un langage libre et documenté. XML est également un **langage standard** respectant les recommandations du **W3C** (World Wide Web Consortium), il facilite :

- la lisibilité par les machines ou par l'œil humain;
- l'échange de données;
- la migration vers d'autres plates-formes, d'autres logiciels, d'autres formats.

XML et ses langages associées

XML s'inscrit dans un environnement complet et possède des langages associés qui permettent de parser, de transformer et/ou d'interroger les fichiers XML.

- Xpath est un langage pour parser/naviguer dans les documents XML : TEI/text/body/lg/l
- XSLT est un langage de transformation pour les documents XML
- Xquery est un langage pour interroger les documents XML qui forment alors une sorte de base de données.

XML : Un peu d'Histoire...

XML est né en 1996 sous la tutelle du W3C (World Wide Web Consortium).

- SGML (1970), Standard Generalized Markup Language;
 - ▶ HTML, HyperText Markup Language: affiche des données notamment sur le Web;
 - ▶ XML, eXtensible Markup Language: contient et structure des données textuelles.

XML : principes de base

- Les données sont incluses dans le document XML sous forme de chaînes de caractères délimitées par un balisage les décrivant.
- L'unité de base qui comprend données et balisage est appelée élément.
Exemple : <nomElement>chaineCaracteres</nomElement>
- Les éléments peuvent être vides : <element>texte</element> ou <elementVide/>
- Les éléments XML suivent un principe strict d'arborescence par imbrication.
- les éléments *enfants* héritent des propriétés des éléments *parents*.

XML : principes de base

- Les attributs XML peuvent être multipliés autant que nécessaire
- On ne peut pas ajouter deux fois le même attribut sur un élément.
- Dans un attribut, on peut mettre plusieurs valeurs.

```
<MiseEnValeur rendu='rouge italique' position='centrePage'>  
texte  
</MiseEnValeur>
```

XML : principes de base

Quelques règles importantes :

- À chaque balise de début doit correspondre une fin de balise.
- Les éléments peuvent être imbriqués, mais ils ne doivent pas se recouvrir.
- Il ne doit y avoir qu'un seul élément racine.
- Un élément ne doit pas avoir deux attributs avec le même nom.

Un encodage qui respecte ces grands principes du XML est dit **bien formé**

Bien formé ou pas ?

<paragraphe>du texte</paragraphe>

<paragraphe><article>du</article><nom>texte</nom></paragraphe>

<paragraphe><article>du <nom></article>texte</nom></paragraphe>

<paragraphe type="texte">du texte</paragraphe>

<paragraphe type=texte>du texte</paragraphe>

<paragraphe type="texte">du texte<paragraphe/>

<paragraphe type="texte">du texte<nomPersonnage>nom de personnage</paragraphe>

<paragraphe type="texte">du texte</Paragraphe>

<segment type="texte" type="nombre">du texte</paragraphe>

- Qu'est que TEI
 - ▶ TEI est un set de balises prédéfini pour la description des sources textuelles. Elle comprend plus de 550 éléments et est en constante évolution.
- Quels sont les avantages de TEI ?
 - ▶ XML TEI permet de proposer un vocabulaire commun pour les balises
 - ▶ Le XML TEI s'intéresse au sens du texte plutôt qu'à son apparence;
 - ▶ Le XML TEI est indépendant de tout environnement logiciel particulier;
 - ▶ Le XML TEI a été conçu par la communauté scientifique qui est aussi en charge de son développement continu.
 - ▶ XML TEI est intégralement documentée

Naissance de la TEI

Lou Burnard et Marjorie Burghart, *Qu'est-ce que la Text Encoding Initiative ?, 2015*

"La TEI a été d'abord développée, il y a plus de trente ans, comme un projet de recherche dans le champ alors émergent du « Humanities computing ». L'idée originelle était de proposer un ensemble de recommandations sur la façon dont les chercheurs devraient créer des ressources textuelles « lisibles par ordinateur », qui soient adaptées aux besoins de la recherche – dans la mesure où un consensus existait sur le sujet –, mais qui soient également extensibles, puisque ces besoins changent et évoluent."

Quelques dates

- 1987 : établissement de la *Text Encoding Initiative*;
- 1990 : TEI P1 (proposal 1), dir. Michael Sperberg-McQueen et Lou Burnard;
- 1994 : TEI P3, première version complète;
- 2000 : naissance du TEI Consortium;
- 2001-2004 : TEI P4, introduction du XML;
- 2007-... : TEI P5, abandon de SGML.

La communauté TEI

La communauté TEI est animée par le **TEI consortium**, fondation interdisciplinaire à but non lucratif.

Il se compose des unités suivantes:

- TEI Board of Directors;
- TEI Technical Council;
- Membres institutionnels et individuels;
- TEI Workgroups, par exemple :
 - ▶ TEI Manuscripts Special Interest Group;
 - ▶ Correspondence SIG;

La communauté TEI

La communauté peut échanger et se rencontrer grâce à :

- Une liste de diffusion internationale : TEI-L mailing list;
- Une liste francophone : TEI-FR et un wiki;
- Des members meetings
- Des congrès annuels : TEI Conference;
- Une revue : Journal of the Text Encoding Initiative;
- Des Guidelines ("recommandations") qui documentent notamment chaque élément.

TEI, mode d'emploi

TEI est un set de balises prédéfini et documenté dans les TEI guidelines qui permet de procéder à une description « scientifique » et « sémantique » d'un texte.

Les balises TEI forment un framework, utile à la conception de son propre encodage. **Il est fortement déconseillé d'utiliser l'intégralité de la TEI pour un document.** Il faut concevoir un modèle de données le plus simple possible et adapté à son projet et sa question de recherche.

Pour aller plus loin : *Why do we encode* : E. Pierazzo

TEI, spécialiations

- Il existe des sous-ensembles de la TEI qui permettent d'avoir accès à un set de balises réduits
- Il existe également des personnalisations de la TEI qui s'organisent autour d'une communauté scientifique active: Epidoc