Université de Genève

Certificat de spécialisation en humanités numériques

2021-2022

Numériser et enrichir des procédures criminelles de l’ancien Évêché de Bâle

Travail de diplôme réalisé par Élodie Paupe

sous la direction du Dr Simon Gabay

Date de reddition : 1er août 2022

Date de soutenance : 5 septembre 2022

Élodie Paupe

rue de l’Église 1

2900 Porrentruy

Jura/Suisse

[elodie.paupe@etu.unige.ch](mailto:elodie.paupe@etu.unige.ch)

**Table des matières**

[1. Introduction 4](#_Toc110288883)

[1.1. Présentation du projet et objectifs 4](#_Toc110288884)

[1.2. Présentation du corpus 4](#_Toc110288885)

[1.3. Présentation de la chaîne de traitement 6](#_Toc110288886)

[2. Retour d’expérience sur la chaîne de traitement 6](#_Toc110288887)

[2.1. Étapes 1 et 2 : de l’image au texte : HTériser avec *e-Scriptorium* et Kraken 6](#_Toc110288888)

[2.2. Enrichir le texte : SegmOnto et Pyrrha, vers l’édition TEI 7](#_Toc110288889)

[2.2.1. Vers une édition numérique 7](#_Toc110288890)

[2.2.2. Étapes 3 et 4 : intégrer le vocabulaire contrôlé SegmOnto 7](#_Toc110288891)

[2.2.3. Étape 5 : passage de l’XML alto à l’XML-TEI 10](#_Toc110288892)

[2.2.4. Étape 6 : annotation linguistique à l’aide de Pyrrha 11](#_Toc110288893)

[2.2.5. Étape 7 : annotation éditoriale 12](#_Toc110288894)

[3. Visualiser l’édition numérique : HTML, CSS, JavaScript et images IIIF 13](#_Toc110288895)

[3.1. Étape 8 : choix des modalités d’édition 13](#_Toc110288896)

[3.2. Gestion des images 14](#_Toc110288897)

[4. Conclusion 14](#_Toc110288898)

[5. Bibliographie sélective 16](#_Toc110288899)

**Résumé**

Ce mémoire s’intègre dans le cadre d’un projet de numérisation et d’inventorisation du fonds des Procédures criminelles conservées par les Archives de l’ancien Évêché de Bâle. Il consiste à tester une partie de la chaîne de traitement proposée par Gabay *et al.* (s. d., p. 1) afin d’enrichir les documents numérisés : « On top of the acquisition task (HTR), we describe a procedure to transform the text into minable information, including lemmas, parts of speech, full morphology, named entities and linguistic normalisation. ».

Le corpus défini pour cette recherche est composé d’une vingtaine de dossiers de procès de sorcellerie de la fin du XVIe et du début du XVIIe siècle. Les documents ont été écrits en cursive française par un seul scripteur, le prévôt Henri Farine, actif entre 1580 et 1618. Un modèle HTR sera entraîné avec Kraken à l’aide de l’infrastructure FoNDUE (UniGe 2022), une première expérience a été réalisée en décembre 2021 (Paupe 2022b). Les données seront enrichies par des annotations de diverses natures (description de la mise en page à l’aide du vocabulaire SegmOnto [SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages 2021], lemmatisation, POS-tagging et annotations philologiques).

Les données obtenues seront diffusées dans des formats standards, notamment en XML alto, XML-TEI (Burnard 2015) et IIIF[[1]](#footnote-1), mais permettront une première édition numérique de ces sources juridiques (HTML).

# Introduction

## Présentation du projet et objectifs

En 2022, les Archives de l’ancien Évêché de Bâle ont lancé un vaste projet de numérisation, transcription et médiation de ses sources criminelles qui devrait aboutir en 2025 à la mise en ligne des facsimilés et d’un large nombre de transcriptions dans une nouvelle interface graphique. Les objectifs du projet sont largement orientés vers la médiation : il s’agit de rendre accessible au plus grand nombre des sources difficiles d’accès pour les noms paléographes, mais également de fournir aux chercheur·euses des outils pour lire ces documents (modèles HTR, exercices paléographiques développés avec la plateforme *Ad fontes* de l’Université de Zurich[[2]](#footnote-2), notamment). La recherche de fonds est en cours et ce travail se présente comme une phase pilote qui permet d’explorer le champ de la numérisation des sources patrimoniales.

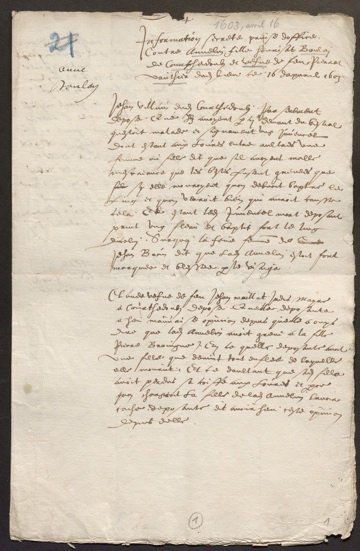


Figure AAEB B 168/15-6.1 f.1

En testant et adaptant la chaîne de traitement proposée par Gabay *et al.* (s. d.), ce travail propose une chaîne largement automatisée de traitement des données qui doit permettre de passer de la numérisation à l’édition en ligne des sources juridiques de l’ancienne principauté épiscopale de Bâle. Étant donné que l’étape du développement de modèle HTR a déjà été réalisée, ce travail se concentre plus particulièrement sur le traitement des données issues de la reconnaissance de texte. Dans le souci pédagogique de rendre cette expérience réplicable, adaptable et transparente, la chaîne d’acquisition et d’enrichissement des données a été découpée en huit étapes principales présentées au point 1.3. Chaque étape se décompose ensuite en un ou plusieurs traitements qui nécessitent le recours à des scripts xlst[[3]](#footnote-3), mais également à des outils externes, comme Pyrrha (Clérice *et al.* 2021a). Un projet Oxygen (aaeb\_procrim.xpr) accompagnant les différentes étapes de la transformation a été créé : à chaque étape du projet, une nouvelle série de documents sont créés dans un nouveau dossier afin de simplifier le développement de la chaîne de travail[[4]](#footnote-4). Les différents scripts écrits pour ce travail sont annotés pour faciliter leur lecture et leur compréhension tandis que le dossier présente les différentes étapes de la chaîne de traitement, justifie certains choix et suggère des pistes d’améliorations.

## Présentation du corpus

Le corpus d’une centaine de pages correspond à environ vingt dossiers d’instruction portant sur des crimes de sorcellerie menés dans l’Évêché de Bâle à la fin du XVIe siècle et au début du XVIIe siècle. Ils ont tous été écrits par le prévôt Henri Farine en cursive française entre 1580 et 1618. Ces documents sont conservés par les Archives de l’ancien Évêché de Bâle à Porrentruy dans la série B 168 (*Criminalia in genere*, soit d’affaires criminelles en général)[[5]](#footnote-5) qui couvre trois siècles (XVe-XVIIe siècles). Comme le précise la notice de la sous-série consacrée aux *Criminalia in sortilegiis, veneficiis et maleficiis*, soit aux crimes de sorcellerie[[6]](#footnote-6), le contenu de chaque dossier d’instruction est variable, mais peut comprendre de la correspondance, des dépositions de témoins, des interrogatoires, des reconfessions (des reformulations d’aveux après interrogatoire), des urphèdres (serment prêté par le condamné libéré ou banni de ne pas chercher à se venger), des notes diverses. Si ces fonds peuvent être trilingues (français, allemand, latin), l’entier de notre corpus est en français (langue d’oïl). Le tableau ci-dessous (figure 2) donne la cote des douze pièces retenues pour ce projet pilote et leur longueur comparée.

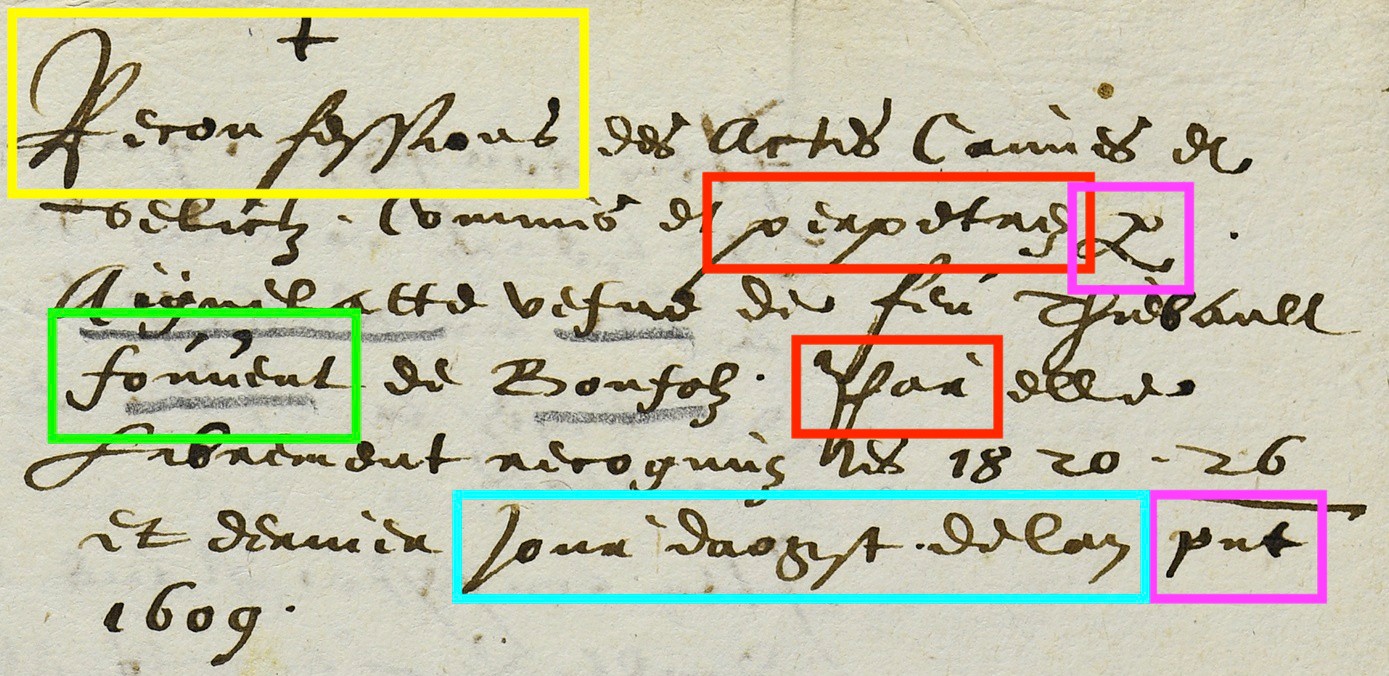
| **Numéro d’inventaire** | **Nombre de pages** |
| --- | --- |
| B 168/15-2.1 | 3 |
| B 168/15-2.2 | 11 |
| B 168/15-2.3 | 4 |
| B 168/15-2.4 | 4 |
| B 168/15-9.2 | 11 |
| B 168/15-9.4 | 4 |
| B 168/15-10.3 | 1 |
| B 168/15-23.1 | 4 |
| B 168/15-23.2 | 4 |
| B 168/15-23.3 | 4 |
| B 168/15-23.4 | 4 |
| B 168/15-23.5 | 3 |
| **Total** | **57** |

Du point de vue graphique, l’écriture de Farine est une cursive française assez régulière et peu originale, même si, dès 1610, le prévôt se met à accentuer les « e » finaux. À noter toutefois quelques traits communs avec la Kurrentschrift germanique, comme le <r> minuscule ou le signe diacritique occasionnel sur les <u/v> (cf. figure 3).

Figure 2 Liste des dossiers formant le corpus.

Les transcriptions réalisées pour ce projet ne sont pas allographétiques (Stutzmann 2011). Étant donné les objectifs de médiation auprès d’un large public poursuivis par l’institution qui conserve ces documents, une telle démarche n’a pas semblé se justifier. La transcription à proprement parler pose peu de problèmes spécifiques (cf. figure 3). Toutefois, les majuscules se divisent en deux groupes. Celles qui ont une morphologie distincte de la lettre minuscule ne posent pas de problème (B, E, P, Q, R, S, T...), contrairement à celles qui ne se distinguent des minuscules que par la taille (A, C, D, G, V...) et sont donc diﬀiciles à différencier. Les <i/j> sont identiques et systématiquement sous forme majuscule au début d’un mot. Par ailleurs, le nombre de jambages peut varier d’un mot à l’autre pour les lettres <m/n>.

Figure 3 En jaune, les différents <s> de « Reconfessions » ; en rouge, <r> minuscule (« perpetrez » ; « Par ») ; en vert, signe diacritique sur <u/v> (nom de famille « fouvent ») ; en jaune, différents <s> ; en violet, abréviations « par » et « present ». AAEB B 168/15-10.3 p. 1.



Farine utilise également quelques abréviations. Les plus fréquentes sont celles du déterminant « ledit » et de la préposition « par » par troncation finale. Dans les transcriptions utilisées pour ce travail, les abréviations sont toutes développées (cf. 2.1 De l’image au texte : HTériser avec e-Scriptorium).

## Présentation de la chaîne de traitement

La première partie de la chaîne de traitement, soit la création d’une vérité de terrain en vue du développement de modèles HTR, a déjà été réalisée en début d’année (Paupe 2022a). Toutefois, ces données ne correspondent pas aux besoins de ce projet. En particulier, elles ne contiennent que la transcription des lignes écrites par le prévôt Farine et ne font pas mention des autres types d’annotations qui figurent sur les documents (notes marginales de Farine lui-même ou d’archivistes). Si la transcription est exploitable, la segmentation du document doit être reprise.

1. Développement d’une vérité de terrain et entraînement de modèles HTR ;
2. Chargement des images JPEG[[7]](#footnote-7) dans *e-Scriptorium* et segmentation à l’aide du modèle par défaut, puis chargement des fichiers alto contenant la vérité de terrain : deux segmentations se superposent, cette situation peu idéale a été préférée à une deuxième option qui aurait nécessité d’ajouter manuellement les zones de texte et les lignes manquant à nos données ;
3. Annotation des zones de texte et des lignes à l’aide du vocabulaire contrôlé SegmOnto et nettoyage des zones de texte vides ;
4. Nettoyage des fichiers alto grâce à HTRVX (Clérice et Pinche 2021) : suppression des lignes vides et correction de l’annotation SegmOnto ;
5. Transformation des fichiers alto en fichier XML-TEI avec conservation des données de spatialisation et ajout de métadonnées ;
6. Annotation linguistique automatique (lemmatisation et Part-Of-Speech) grâce à Pyrrha : cette étape nécessite une manipulation assez lourde des fichiers TEI ;
7. Encodage philologique manuel des fichiers XML-TEI (balisage du texte supprimé, intégration des notes marginales au texte principal) ;
8. Transformation des éditions numériques XML-TEI en pages HTML et intégration au site web créé pour le cours « Numériser le patrimoine II »[[8]](#footnote-8).

Ces différentes étapes seront rappelées dans les sous-chapitres suivants.

# Retour d’expérience sur la chaîne de traitement

## Étapes 1 et 2 : de l’image au texte : HTériser avec *e-Scriptorium* et Kraken

Le corpus présenté ci-dessus fait l’objet de plusieurs expérimentations sur Transkribus (READ-COOP 2021) et *e-Scriptorium* depuis décembre 2021 (Paupe 2022b). Les résultats de l’entraînement de modèles HTR qui seront brièvement résumés ici ont été présentés en juin 2022 dans le cadre du colloque *Documents anciens et reconnaissance automatique des écritures manuscrites* (Paupe 2022a) et dans un billet de blog (Paupe 2022b). Sur Transkribus, le meilleur modèle (AAEB\_v4) a été développé par *fine tuning* en s’appuyant sur le modèle public de Tobias Hodel, « Charter Scripts XIII-XV\_M1 »[[9]](#footnote-9). Avec un CER (*Character Error Rate*) sur set de validation de 5,69 %, ce modèle permet de transcrire les textes de la main du prévôt Farine de façon satisfaisante. En revanche, son efficacité sur des documents écrits par d’autres scripteurs contemporains est très mauvaise.

Le premier modèle entraîné sur *e-Scriptorium* à partir des mêmes données obtient une précision très faible (83,8 %). Deux pistes ont été suivies pour améliorer son efficacité : le recours à la binarisation du corpus avec la méthode Otsu (Gupta *et al*. 2006) et la modification de l’architecture d’apprentissage utilisée par le moteur HTR Kraken sur lequel est construit *e-Scriptorium*. La précision du modèle entraîné sur des données binarisées[[10]](#footnote-10) s’est avérée inférieure (75,1 %). Même si calcul du seuillage qui permet de déterminer quels pixels passent au blanc ou au noir est réalisé pour chaque image, les faibles différences de teintes entre le papier et l’encre d’une partie de notre corpus ont dû fortement affecter l’entraînement. La modification de l’architecture neuronale d’apprentissage de Kraken a, en revanche, permis d’obtenir un modèle plus efficace.

-s '[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do]' -r 0.0001

Figure 4 Architecture recommandée par les développeurs de Kraken pour la reconnaissance d’écritures manuscrites[[11]](#footnote-11).

Avec une précision de 96,1 % après entraînement et de 85,7 % sur le set de test[[12]](#footnote-12), ce modèle est en amélioration, mais ne permet pas encore d’assurer la lisibilité d’une transcription. Les prochaines étapes nécessiteront d’augmenter la vérité de terrain d’une part, mais également de tester le *fine* *tuning* pour autant que l’on trouve un modèle adapté à nos données.

## Enrichir le texte : SegmOnto et Pyrrha, vers l’édition TEI

### Vers une édition numérique

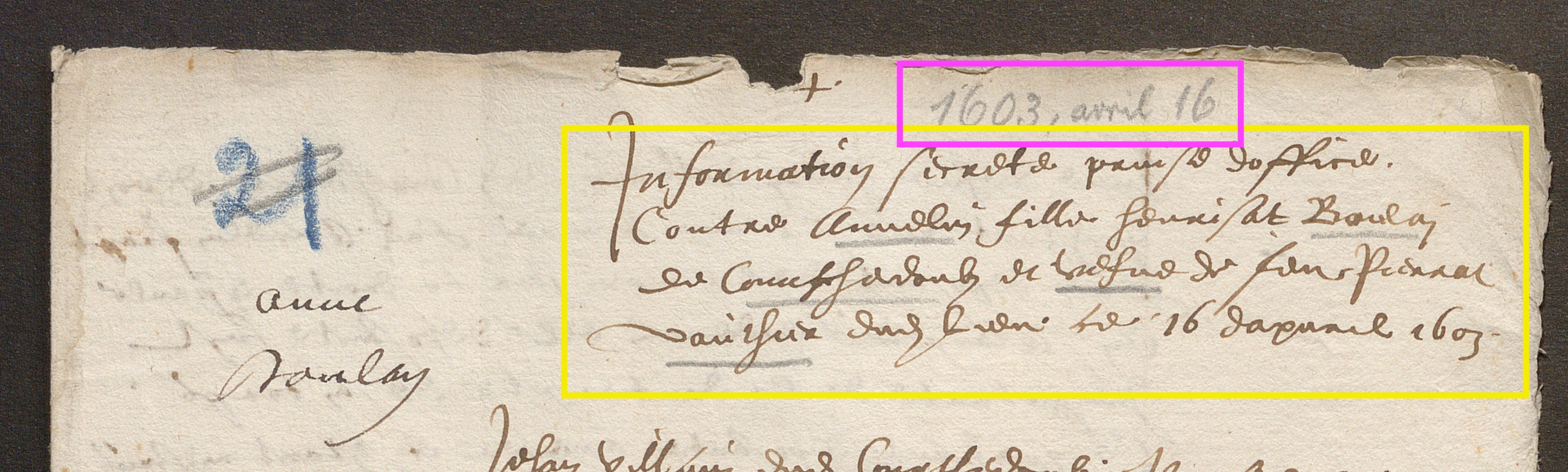
Après la segmentation et la transcription (automatique ou manuelle), les données obtenues sont brutes, dans le sens où elles ne contiennent aucun marquage qui permettrait de préciser le sémantisme de la segmentation : les zones et les lignes sont toutes de même niveau. Le recours au langage de description SegmOnto (*[SegmOnto, A Controlled Vocabulary to Describe](#_bookmark28)* [*the Layout of Pages*](#_bookmark28)[2021](#_bookmark28)) permet de décrire la mise en page à l’aide de valeurs standardisées. Par ailleurs, ces métadonnées permettent ensuite d’assurer une mise en forme minimale des fichiers au moment du passage à l’édition web. Les données de lemmatisation et de Part-Of-Speech (POS) sont ensuite ajoutées au fichier TEI grâce à Pyrrha. Ces métadonnées ne seront pas exploitées dans l’édition web, mais enrichiront les données proposées au téléchargement.

### Étapes 3 et 4 : intégrer le vocabulaire contrôlé SegmOnto

Si la mise en page des documents qui composent notre corpus n’est pas complexe, elle comporte quelques particularités : des inscriptions marginales de différentes natures, plusieurs numérotations de dossier, de page ou de feuillet par différentes mains d’archivistes, l’ajout marginal postérieur à la rédaction du dossier du nom de l’accusé·e et de la date.

Les valeurs standards suivantes ont été sélectionnées pour qualifier les zones de texte : MainZone et MainZone:column pour le texte principal rédigé par le prévôt Farine, MarginTextZone pour les notes marginales du prévôt et MarginTextZone:note pour celles d’autres scripteurs, GraphicZone pour la petite croix présente au début des dossiers, NumberingZone:other pour la numérotation des dossiers ou des feuillets et NumberingZone:page pour la numérotation des pages. Pour les lignes, les valeurs suivantes sont utilisées : DefaultLine, HeadingLine (pour le chapeau introducteur des dossiers) et InterlinearLine. Le vocabulaire contrôlé SegmOnto permet de personnaliser les valeurs de zone (CustomZone) et les lignes (CustomLine). Dans quelques documents figurent les signatures des témoins. Une CustomZone:signature a ainsi été créée pour annoter ces éléments. L’annotation des zones et des lignes se fait directement à partir d’*e-Scriptorium*.

Figure 5 En jaune la zone de l’en-tête de document et en rose la date ajoutée par un archiviste. AAEB B 168/15-2.1 p.1.



Plusieurs types de zones ont fait l’objet de réflexions particulières, comme la mention de la date, utile pour le chercheur, de l’en-tête introductif du dossier qui permet de situer les acteurs de la procédure, le lieu et la date (cf. figure 6) et la présence d’une numérotation du texte principal parfois reléguée loin dans la marge.

La date n’appartient pas au corps de texte principal puisqu’il s’agit d’un ajout récent. La question s’est posée de savoir s’il fallait utiliser la valeur MarginTextZone:note et considérer simplement qu’il s’agissait d’une annotation sans spécifier sa nature ou s’il se justifiait de créer une zone particulière pour cette date pour pouvoir récupérer sa valeur de façon rapide. Afin de demeurer le plus générique possible, la première option a été retenue. Placée très haut dans la page de titre autour de l’en-tête de dossier, cette annotation est la première zone de type MarginTextZone:note à apparaître dans notre corpus : cette position initiale devrait suﬀire à permettre l’extraction de la donnée en cas de besoin.

Si on a finalement considéré que l’en-tête introductif faisait partie du corps de texte principal, son léger retrait à droite, en fait un bloc de texte particulier. Ce statut particulier est marqué par une annotation à la ligne : les lignes de l’en-tête ne sont pas identifiées par une valeur DefaultLine, mais HeadingLine. L’autre possibilité aurait été de personnaliser une zone de texte spécifique, du type CustomZone :header. Toutefois, les résultats peu engageants obtenus par le modèle de segmentation entraîné sur notre corpus qui seront présentés par la suite n’ont pas convaincu de pousser plus loin les expérimentations en la matière.

Le statut de la numérotation marginale présentée dans la figure 6 a également fait l’objet de réflexion. S’agit-il d’une annotation marginale ou d’un redimensionnement de l’espace alloué au corps de texte ? Dès lors, faut-il considérer la présence d’une ou deux zones ?

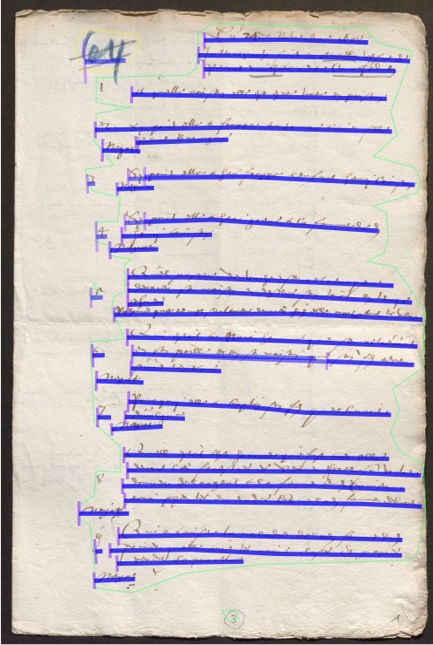
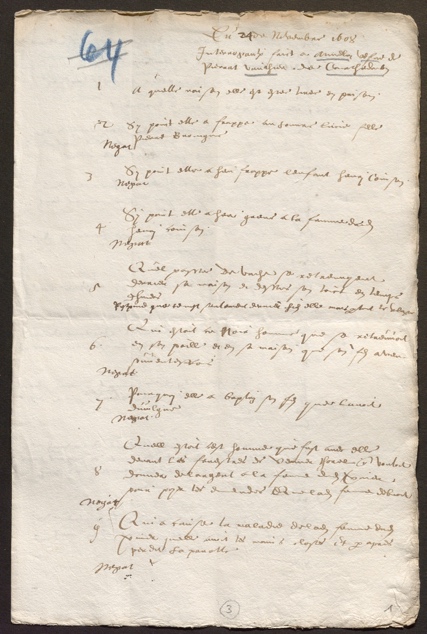


Figure 6 À gauche : les numéros à gauche doivent-ils être considérés comme des annotations ? À droite : le modèle de segmentation par défaut d’e-Scriptorium opère un découpage peu homogène. AAEB B 168/15-2.3 p.1.

Dans le cas présenté ci-dessus (cf. figure 6), on observe que le modèle par défaut de segmentation d’*e-Scriptorium* n’a pas traité de façon homogène les numéros : le premier n’est pas reconnu du tout, le deuxième est relié à la première ligne du paragraphe qui lui correspond, le troisième est identifié comme une ligne nouvelle appartenant à la zone de texte principale et le cinquième est analysé comme une nouvelle zone de texte. Face à ces traitements variés, le choix de segmentation à opérer n’est pas évident : si la solution qui consisterait à relier les nombres à la première ligne de leurs paragraphes respectifs semble évidente du point de vue du sens, il est probable que la position relative de ces derniers ne simplifiera pas véritablement la tâche d’un nouveau modèle de segmentation. En effet, comme on peut le voir pour le cinquième paragraphe par exemple, le nombre est placé face au milieu de ce dernier : sa liaison à la première ligne du paragraphe sera loin d’être horizontale. Un problème similaire apparaît ailleurs avec la mention de « Negat ». Dans ce second cas, la mention a été interprétée comme une note marginale puisqu’il s’agit d’un commentaire sur le texte. La numérotation des paragraphes a toutefois été réglée en reliant le nombre à la première ligne du paragraphe. Une solution *a posteriori* jugée peu pertinente puisqu’il s’agit bien d’une numérotation de paragraphe et pas de lignes.

Une fois les données annotées, ces dernières ont été importées sous forme de fichier alto (/1alto\_import)[[13]](#footnote-13). Afin de supprimer les lignes vides créées par la double segmentation évoquée plus tôt, des scripts xslt ont été élaborés (scripts/1alto\_supprimer\_lignesvides.xsl). Le recours à HTRVX (Clérice et Pinche [2021](#_bookmark22)), un script qui vérifie la validité d’un document XML au regard de certaines normes, permet de contrôler que les données sont bien nettoyées. Dans notre corpus, six zones de texte sont vides : il s’agit des pages qui contiennent un élément de décoration :

(p37) macbook−air : travail de recherche elodiepaupe $ htrvx 2alto\_sanslignesvides/\*.xml −−check−empty

× Detection of empty lines or region in 2

alto\_sanslignesvides/B\_168\_15−10−3\_0001.xml : Empty elements founds( 1 )

===== REPORT

57/57 valid XML files

Le même script est ensuite utilisé pour contrôler que le vocabulaire SegmOnto a bien été appliqué à l’entier du corpus au moment de l’annotation dans *e-Scriptorium*. Afin de gagner un peu de temps, les lignes principales identifiées avec DefaultLine n’ont pas été marquées manuellement dans *e-Scriptorium*, mais *a posteriori* grâce à un script xslt (scripts/2alto\_ajouterDefaultLine.xsl).

Sur la base de ces données, un modèle de segmentation a été entraîné à l’aide d’*e-Scriptorium[[14]](#footnote-14)*. La précision du modèle s’élève à 98,5 %. Toutefois, les tests réalisés sur les trois pages conservées à cet effet montrent un résultat nettement insuﬀisant, comme l’illustrent [la](#_bookmark9) figure 7.

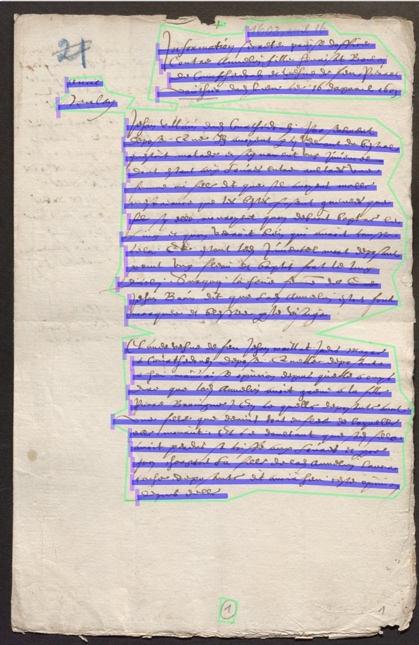
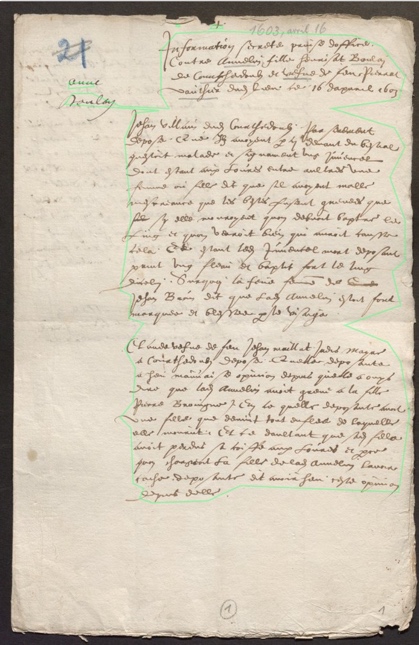
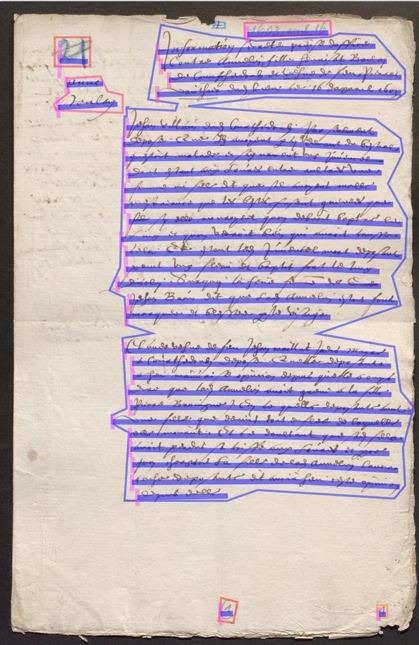


Figure 7 À gauche, la segmentation de référence après correction manuelle. Au centre, le résultat de la segmentation entraînée sur notre corpus. À droite, le résultat de la segmentation avec le modèle par défaut.

Dans le cas où le projet nécessitait de développer un modèle de segmentation, l’architecture d’apprentissage pourrait être modifiée et les données d’entraînement devraient de toute manière être augmentées. Une telle démarche permettra d’entraîner un modèle sans passer par *e-Scriptorium* directement sur le service de calcul hautes performances (HPC) de l’université de Genève[[15]](#footnote-15). Les mauvais résultats de Kraken en matière de segmentation sur de petits échantillons ont déjà été mis en avant par Clérice ([2020](#_bookmark21)).

### Étape 5 : passage de l’XML alto à l’XML-TEI

À ce stade, le projet dispose de fichiers XML alto, mais ce format, s’il a l’avantage de conserver le lien entre image et transcription, ne permet pas d’annoter le document. Suivant la chaîne de traitement proposée par Gabay *et al.* (s. d.), les fichiers alto sont transformés en fichier TEI. Les données spatiales et le texte transcrit sont conservés dans l’élément <sourceDoc>, mais la transcription est répétée dans le <body> du document où elle pourra être annotée.

Un script xslt est nécessaire pour réaliser le passage de l’XML alto à l’XML-TEI. P. Jacsont a réalisé un fichier de ce type pour le projet *Desenrollando el cordel* (Carta et al. [2022](#_bookmark20)). S’il n’a pas pu être réutilisé tel quel, il a servi d’inspiration à celui qui se trouve associé à ce travail de recherche (scripts/3ALTOtoTEI\_EP.xsl). Un effort d’optimisation a été consenti afin de faciliter son réemploi par d’autres chercheur·euses : les métadonnées du <teiHeader> doivent être actualisées, mais le reste du script devrait s’appliquer à n’importe quel fichier alto.

Les métadonnées contenues dans le fichier alto, puis dans le fichier TEI ne sont pas très développées. Un certain nombre d’informations supplémentaires sont ajoutées aux documents à partir de fichiers csv convertis en XML (4TEI/recolement.xml) et d’un deuxième script xslt (scripts/5TEItoTEI\_EP.xsl), notamment la cote des pièces d’archives et les liens des images IIIF. Étant donné que ces documents d’archives sont décrits dans une base de données Oracle hébergée par l’institution qui les conserve, on pourrait imaginer étendre le nombre et la qualité des métadonnées présentées dans le <teiHeader> en suivant la même procédure.

### Étape 6 : annotation linguistique à l’aide de Pyrrha

Afin d’ajouter les informations linguistiques, le corpus doit être lemmatisé. Une étape qui pourrait être réalisée à l’aide d’environnement virtuel Python ou de la plateforme Pyrrha (Clérice *et al.* 2021a). Cette dernière s’avère utile puisqu’elle intègre un modèle pour le moyen français et des outils de correction dont il ne sera pas fait usage dans le cadre de ce projet où les données linguistiques ne seront ni revues ni corrigées.

Si Pyrrha présente des avantages en termes d’ergonomie et de performance, elle ne supporte pas l’importation de fichiers XML. Les fichiers TEI obtenus à l’étape précédente sont transformés en fichier .txt à l’aide d’un script xslt qui ne récupère que le contenu écrit par le scripteur du XVIIe siècle, soit le corps de texte principal et les annotations marginales de sa main (scripts/5TEItoplaintext\_v2.xsl). Toutefois, cette transformation fait disparaître tout le balisage XML-TEI. Afin de faciliter la récupération de cette syntaxe, le nom du fichier, la valeur de l’attribut @xml:id, ainsi que les éléments <lg> et <l> sont codés et exportés à l’aide de symboles qui ne pourront pas être lemmatisés, comme l’illustre la figure 8 :

₳ ₦ fedora\_ug27803420 ₲

$ ¥ eSc\_textblock\_dce4e4e9\_lg\_1

£ Information secrete prinse doffice, ₽

£ Contre Annelin fille henrisat Boulai ₽

Figure 8 Les balises XML sont codées pour ne pas être lemmatisées.

Les 57 fichiers TXT sont assemblés dans un seul document avec une ligne de commande (cat \*.txt > nouveaufichier.txt) et le fichier peut être lemmatisé avec Pyrrha (modèle « Français classique [non modernisé] »). Les données peuvent être téléchargées en XML-TEI. Chaque mot ou signe de ponctuation traité est inséré dans un élément <w> dont les attributs @lemma et @type contiennent les analyses linguistiques relatives respectivement au lemme et à la morphologie ainsi qu’aux Part-Of-Speech (7Pyrrha\_import/). On dispose alors alors d’un seul fichier XML-TEI qui contient le texte lemmatisé de tous les documents du corpus. Le script xslt (scripts/7PyrrhatoXMLDatabase\_EP.xsl) va permettre de décoder le contenu du fichier en lui faisant prendre la forme d’une base de données XML qui pourra ensuite être importée dans les fichiers XML-TEI qui correspondent aux différentes pages du corpus. Une fois que le script a transformé le document, il est nécessaire d’effectuer un chercher/remplacer sur le fichier pour rendre aux caractères échappés (&lt ; | &gt ;) leur valeur (< | >) et ainsi rétablir les attributs cf. figure 9.

|  |  |
| --- | --- |
| &lt;div type="ID"&gt ; <w xml:id="t865" lemma="@latin" type="POS=ADJqua|NOMB.=s|GENRE=m">fedora\_ug27803420</w> &lt;/div&gt ; | <div type="ID"> <w xml:id="t3" lemma="@latin" type="POS=VERppa|MORPH=empty">fedora\_ug27803361</w> </div> |

Figure 9 Décodage des données extraites de Pyrrha.

On obtient alors une base de données XML à trois entrées (cf. figure 10). Le script 8TEI+Database.xsl appliqué sur les fichiers XML-TEI conservés dans le dossier 5TEI+csv/ permet à chaque document XML-TEI composant le corpus de venir récupéré ses données annotées.

<root>

<div type="ID"/>

<div type="MainZone"/>

<div type="MarginTextZone"/>

</root>

Figure 10 Structure de la base de données extraite de Pyrrha.

Suivant les recommandations de la chaîne de traitement (cf. figure 11), les annotations linguistiques devraient ensuite être extraites des éléments <w> et placées de façon structurée par segment (<seg>), par exemple par une unité de phrase, dans un élément <standOff>.

<text>

<body>

<p>

<seg xml:id="s1">

<w xml:id="s1w1">m~ange</w>

</seg>

</p>

</body>

</text>

<standOff>

<listAnnotation type="linguistic">

<annotationBlock corresp="#s1">

<spanGrp>

<span target="#s1w1" ana="#s1ling1"/>

</spanGrp>

<fs xml:id="#s1ling1">

<f name="lemma">

<string>manger</string>

</f>

</fs>

</annotationBlock>

</listAnnotation>

</standOff>

Figure 11 Structure des annotations linguistiques selon Gabay *et al.* (s. d., 6-7, notamment fig. 12).

Dans le cadre de ce travail, les phrases ne sont pas annotées. Pour cette raison, l’élément <annotationBlock> contient directement tous les <span> du document. Par ailleurs, comme l’élément <listAnnotation> ne permettait pas de valider le fichier XML-TEI avec un schéma tei\_all.rng, il a été supprimé. Le fichier xslt 10TEItoStandOff.xsl permet de réaliser cette dernière opération automatique sur les fichiers XML.

### Étape 7 : annotation éditoriale

Une fois les fichiers XML-TEI enrichis par les informations spatiales (cf. ch. [2.2.3](#_bookmark6)) et linguistiques (cf. ch. [2.2.4](#_bookmark10)), un travail éditorial manuel est réalisé sur le des fichiers TEI : il s’agit notamment de relier les notes marginales rejetées en fin de <body> au corps du texte à l’aide de la balise <ref>, mais également d’intégrer toutes les corrections sur le texte qui n’ont pas pu être réalisées jusqu’ici : biffures (<del>), ajouts (<add>) qui seraient passés entre les mails du filet, etc. Les fichiers XML-TEI obtenus par traitement automatique et conservés dans le dossier 11standODD/ sont dupliqués dans un nouveau dossier 12TEI\_annotée/ pour éviter que le lancement de scripts sur les données qui se trouve en amont dans la chaîne de traitement ne viennent effacer le travail d’annotation. Dans le cas où des corrections devaient être apportées automatiquement dans les étapes précédentes, Oxygen dispose d’un comparateur de fichier qui permet de réconcilier deux fichiers.

# Visualiser l’édition numérique : HTML, CSS, JavaScript et images IIIF

## Étape 8 : choix des modalités d’édition

Au terme des manipulations évoquées, le projet dispose d’une édition numérique complexe et interopérable, mais pas encore d’une publication numérique. Pour lire les transcriptions, il convient de développer une interface de visualisation sur le web.

TEI Publisher[[16]](#footnote-16) est l’outil idéal pour ce type de solution puisqu’il permet d’éditer presque automatiquement des fichiers TEI. Même si cette solution est aujourd’hui plébiscitée par différents projets, une solution plus basique qui s’appuie sur un code simple afin d’assurer la maintenance des données à long terme a été privilégiée. La solution de visualisation proposée est ainsi basée sur des langages d’édition du web usuels : HTML, CSS et JavaScript. Ce choix éditorial nécessite toutefois de transformer les données XML-TEI en fichiers HTML (scripts/TEItoHTML.xsl) : on génère ainsi autant de pages web qu’il n’y a de pages dans le corpus[[17]](#footnote-17).

Une première version de l’interface de lecture, autrement dit le site web qui accueille ces données, a été conçue dans le cadre du séminaire « Numériser le patrimoine II ». Entre les deux versions, les deux modifications les plus importantes apportées tiennent à la génération des transcriptions et au format des images mises en ligne. En effet, les transcriptions données à lire étaient générées directement en HTML à partir des fichiers XML alto extraits de *e-Scriptorium*. Les données étaient brutes, voire erronée puisqu’elles ne permettaient pas de savoir qu’un extrait de texte était biffé par exemple. En ce sens, les transcriptions présentées et publiées dans le cadre de ce travail ont subi un traitement éditorial bien plus poussé.

En ce qui concerne les images, ces dernières n’étaient pas publiées selon les standards IIIF, mais simplement reprises d’un dépôt Github et elles ne bénéficiaient pas d’un lecteur Mirador (Mirador 2014-2022), même si un effort avait été consenti pour améliorer l’expérience du lecteur en lui offrant la possibilité de zoomer dans l’image[[18]](#footnote-18).

Les avantages d’une édition web plus basique sont à la fois sa très grande robustesse et le fait que son contenu peut être mis à jour sans grandes difficultés. Cependant, dans le cadre d’un projet patrimonial comme le nôtre, ces avantages sont également des points négatifs : si aucun acteur du projet ne maîtrise les rudiments d’HTML nécessaires à la mise en ligne des nouvelles données, l’absence de CMS et d’automatisation dans la gestion des flux peut rapidement poser problème. Le coût modeste de développement au lancement du projet devient par la force des choses un coût récurrent.

## Gestion des images

Comme cela a été dit précédemment, les images sont mises en ligne au format IIIF sur le serveur dédié de l’Université de Genève. Le format IIIF (International Image Interoperablity Framework) est un ensemble de normes permettant de diffuser des images et des données audiovisuelles sur le web. Les avantages de cette mise en ligne sont de permettre l’édition à la volée des données grâce à une syntaxe particulière intégrée à l’URL. Ainsi, chaque zone d’image associée à une transcription dispose d’une URL spécifique qui permet d’afficher une portion d’image spécifique.

Afin de permettre la mise en ligne et la visualisation des données, les images (JPEG dans notre cas) doivent être livrées avec un Manifest de format JSON qui contient les métadonnées. Les images qui composent nos données ont été organisées en douze collections qui correspondent aux douze pièces d’archives conservées par l’institution qui conserve ces documents. Cette architecture des données a été préférée à une collection unique avec table des matières en raison de la nature des documents : comme toutes les pages ne concernent pas la même procédure, les réunir en une seule collection n’avait pas de sens. Une solution alternative, intermédiaire et sans doute plus pertinente aurait été d’éditer dans une seule collection avec table des matières les documents relatifs à la même procédure, suivant en cela le plan de classement des AAEB. Par exemple, les quatre premiers documents de notre jeu de données (B 168/152.1 ; B 168/15-2.2 ; B 168/15-2.3 ; B 168/15-2.4) se trouveraient réunis dans une seule collection (B 168/15.2) avec quatre entrées dans la table des matières : les collections numériques reflèteraient ainsi mieux l’inventaire. Il est probable qu’en cas de poursuite du projet les manifestes soient corrigés en ce sens.

L’utilisation de Mirador couplé au dépôt des fichiers images sur des serveurs IIIF permet une meilleure lisibilité des facsimilés puisqu’il est possible de zoomer dans l’image de façon aisée et intuitive, contrairement à la solution CSS mise en place antérieurement.

# Conclusion

Au terme de ce travail, une chaîne de traitement a été mise en place qui permet de générer automatiquement des transcriptions et de les transformer en éditions numériques riches à destination d’un public de chercheur·euses avisé·es, mais également du grand public grâce à la publication finale. Les objectifs annoncés au début de ce travail sont donc atteints : des données qualitatives ont été produites et permettent la publication du corpus en ligne et la diffusion d’éditions numériques XML-TEI enrichies et d’éditions XML alto qui pourraient prochainement être déposées comme un jeu de donnée auprès de HTR-United[[19]](#footnote-19).

L’annotation du corpus n’est pas allée jusqu’à la recherche des entités nommées, par exemple, ni du côté de la normalisation de la langue, alors même que le corpus aurait pu s’y prêter. Toutefois, la chaîne de traitement et les modalités de travail sur le corpus explorées permettent de se rendre compte des étapes théoriques qui pourraient permettre de poursuivre l’enrichissement du texte. En déportant les annotations dans des balises extérieurs à l’élément <text> d’un document TEI, il devient aisé d’augmenter les métadonnées sur le texte tout en conservant une part de lisibilité pour l’homme. À partir de là, le champ des possibles s’ouvre et on peut imaginer voir naître des corpus d’une densité exceptionnelle. Si tous les projets ne nécessitent pas le même raffinement des données, le recours à des langages standardisés et ouverts comme la TEI permettent d’envisager une circularité du savoir : les corpus numériques ne sont pas enrichis par un seul chercheur ou une seule institution.

Toutes les étapes de la chaîne de traitement testée dans le cadre de ce travail qui sert également de projet pilote en vue de la mise en ligne des procédures criminelles de la principauté épiscopale de Bâle ne seront pas conservées. En particulier, vu les objectifs poursuivis par le fonds d’archive qui conserve ces pièces, les annotations linguistiques, peu utiles à la mise en ligne des transcriptions seront vraisemblablement abandonnées. Le recours au vocabulaire SegmOnto, en plus de décrire précisément la source en question, permet également de simplifier considérablement l’édition web : le coût d’encodage s’avère ici particulièrement bien rentabilisé.

D’une façon plus générale, la réalisation de ce travail de recherche a permis de revenir sur les bases de la modélisation du projet, de vérifier certaines approches et de prendre conscience de certains changements à apporter au projet prévu initialement.

**Corpus :**

* Les scripts et les données textuelles : <https://github.com/elodiepaupe/elodiepaupe_unige_certificatDH.github.io>
* Le site web tiré du projet : <https://elodiepaupe.github.io/elodiepaupe_unige_certificatDH.github.io/>
* Les images déposées sur le serveur IIIF de l’Université de Genève : <https://iiif.unige.ch/mirador/https://iiif.unige.ch/collection/aaeb/manifest>

**Remerciements**

Un merci tout particulier à Simon Gabay pour son suivi et sa disponibilité au fil du temps et à l’équipe genevoise, Pierre Kunzli, Cédric Viannoz et leurs collègues. Mes remerciements vont également à la Fondation des Archives de l’ancien Évêché de Bâle pour la confiance qui m’y est accordée, en particulier à Damien Bregnard et à Jean-Claude Rebetez. Je remercie également mon directeur de thèse, Jean-Jacques Aubert pour m’avoir offert la possibilité de suivre cette formation.

Ce travail leur doit tout, excepté ses faiblesses qui m’incombent.

# Bibliographie sélective

**Chagué 2022 :** Chagué, Alix, « Conditions de la mutualisation : les principes FAIR et HTR-United », *Humanistica 2022*, mai 2022, Montréal, Canada. [⟨hal-03685731⟩](https://hal.inria.fr/hal-03685731v1)

**Burnard 2015** : Burnard, Lou. *Qu’est-ce que la Text Encoding Initiative ?* 2015. doi : https://doi.org/10.4000/books.oep.1237. url : [http://books.openedition.](http://books.openedition.org/oep/1237) [org/oep/1237](http://books.openedition.org/oep/1237).

**Carta *et al.* 2022** : Carta, Constance *et al.*, *DesenrollandoElCordel,* 2022, url : <https://github.com/DesenrollandoElCordel>, visité pour la dernière fois le 26 juillet 2022.

**Clérice 2020**: Clérice, Thibault, « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin », *Journal of Data Mining & Digital Humanities* 2020 (avr. 2020). doi : [10 . 46298 / jdmdh . 5581](https://doi.org/10.46298/jdmdh.5581). url : <https://jdmdh.episciences.org/6264>.

**Clérice *et al.* 2021a** : Clérice Thibault, Pilla Julien, FrFerry, Camps Jean-Baptiste, ngawangtrinley, architexte, Aditya Jetely, Ariane Pinche, Sai Siddhant, & jhrdt. *hipster-philology/pyrrha: 3.0.0 (3.0.0)*, Zenodo, 2021. doi : /10.5281/zenodo.5144781

**Clérice 2021b :** Clérice, Thibault et Ariane Pinche. *HTRVX, HTR Validation with XSD.* <https://github.com/HTR-United/HTRVX>, version 1.2.0, 2021. doi : [10.5281/zenodo.](https://doi.org/10.5281/zenodo.5359963) [5359963](https://doi.org/10.5281/zenodo.5359963).

**Gabay s.d**. : Gabay, Simon *et al.* « Towards the Fourth Paradigm : From digital facsimiles of historical documents to highly annotated data », 2021.

**Gupta *et al.* 2007** : Gupta, Maya R., Jacobson, Nathaniel P. et Garcia Eric K. « OCR binarization and image pre-processing for searching historical documents », *Pattern Recognition*, 40/2, 2007, 389-397. doi : 10.1016/j.patcog.2006.04.043.

**Mirador 2014-2022** : Mirador, <https://projectmirador.org/>, 2014-2022, visité pour la dernière fois le 28 juillet 2022.

**Paupe 2022a :** Paupe, Élodie « Htériser une cursive du XVIIe s. », *Documents anciens et reconnaissance automatique des écritures manuscrites*, 23-14 juin 2022, Paris.

**Paupe 2022b :** Paupe, Élodie « Lire un manuscrit à l’aide du numérique : Développement d’un modèle HTR ». *LexTech Institute* (fév. 2022). [Blog]. url : https://www.lextechinstitute.ch/lire-un-manuscrit-a-laide-du-numerique-developpement-dun-modele-htr/

**segmonto** : SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages. [https://](https://github.com/SegmOnto) [github.com/SegmOnto](https://github.com/SegmOnto). Version 0.9, Paris/Genève, 2021.

**Stutzmann 2011** : Stutzmann, Dominique, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? » In : *Codicology and Palaeography in the Digital Age 2*, sous la dir. de Franz Fischer, Christiane Fritze et Georg Vogeler, 2011, 247-277. url : halshs-00596970.

1. International Image Interoperablity Framework, [https://iiif.io](https://iiif.io/), consulté pour la dernière fois le 28 mars 2022. [↑](#footnote-ref-1)
2. *Ad fontes*, <https://www.adfontes.uzh.ch/fr/ad-fontes/einleitung>, visité pour la dernière fois le 31 juillet 2022. [↑](#footnote-ref-2)
3. W3Consortium, *XSL Transformation Version 2.0*, https://www.w3.org/TR/xslt/ , visité pour la dernière fois le 1er août 2022. [↑](#footnote-ref-3)
4. Ainsi, on applique le script numéroté 1alto….xsl sur des fichiers contenus dans le dossier numéroté 1alto\_import, le script numéroté 2alto….xsl sur le dossier numéroté 2alto\_sanslignesvides, etc. [↑](#footnote-ref-4)
5. B 168 Criminalia in genere Affaire criminelles en générale, [https://archives-aaeb.jura.](https://archives-aaeb.jura.ch/detail.aspx?ID=77497) [ch/detail.aspx?ID=77497](https://archives-aaeb.jura.ch/detail.aspx?ID=77497), consulté pour la dernière fois le 28 mars 2022. [↑](#footnote-ref-5)
6. B 168/14 à 19 (1546-1670), <https://archives-aaeb.jura.ch/detail.aspx?ID=223580>, visité pour la dernière fois le 28 mars 2022. [↑](#footnote-ref-6)
7. Il n’a pas été possible de charger les images à partir des serveurs IIIF, ces données ont été acquises ultérieurement. [↑](#footnote-ref-7)
8. Le premier état de ce site web est consultable à l’adresse suivante : <https://aaeb-porrentruy.github.io/aaeb-procedurescriminelles.io/>, visité pour la dernière fois le 1er août 2022. [↑](#footnote-ref-8)
9. Le modèle est décrit sur le site de Transkribus, <https://readcoop.eu/model/charter-scripts-german-latin-french/>, visité pour la dernière fois le 31 juillet 2022. Les jeux de données qui ont permis son entraînement sont en partie référencés sur la plateforme HTR-United, <https://htr-united.github.io>, visité pour la dernière fois le 31 juillet 2022. [↑](#footnote-ref-9)
10. Le script employé est joint au dossier (scripts/Otsu Binarisation.ipynb). [↑](#footnote-ref-10)
11. Kraken, « Training », https :kraken.re/master/ketos.html, consulté pour la dernière fois le 1er août 2022. [↑](#footnote-ref-11)
12. Ce set de trois pages a été pris hors corpus et sera prochainement publié. Le modèle est joint au dossier (ressources/best\_model.mlmodel). [↑](#footnote-ref-12)
13. Les dossiers et fichiers mentionnés entre parenthèses renvoient à la structure du projet Oxygen aaeb\_proccrim.xpr joint au dossier. [↑](#footnote-ref-13)
14. Le document AAEB B 168/15-2.1 de trois pages a été conservé pour servir de test set tandis que les 54 autres pages ont servi à l’entraînement. [↑](#footnote-ref-14)
15. *Disposer de ressources de calculs à hautes performances (HPC)*, <https://catalogue-si.unige.ch/hpc>, visité pour la dernière fois le 31 juillet 2022. [↑](#footnote-ref-15)
16. TEI Publisher. The Instant Publishing Toolbox, <https://teipublisher.com/index.html>, consulté pour la dernière fois le 1er août 2022. [↑](#footnote-ref-16)
17. Le dossier zip contient tous les fichiers nécessaires à l’ouverture du site hors ligne. Ce dernier est également en ligne sur https://elodiepaupe.github.io/elodiepaupe\_unige\_certificatDH.github.io/, visité pour la dernière fois le 1er août 2022. [↑](#footnote-ref-17)
18. Le premier site est consultable à cette adresse : <https://github.com/aaeb-porrentruy/aaeb-procedurescriminelles.io>, visité pour la dernière fois le 1er août 2022. [↑](#footnote-ref-18)
19. HTR-United, <https://htr-united.github.io/index.html>, visité pour la dernière fois le 1er août 2022. [↑](#footnote-ref-19)