ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Kelly Christensen
licenciée ès enseignement musical
diplômée de master musicologie
diplômée de doctorat musicologie

Modélisation des transcriptions ALTO avec la TEI

En complétant le pipeline du projet *Gallic(orpor)a*

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2022

Résumé

Quand des modèles OCR (*Optical Character Recognition*) et HTR (*Handwriting Text Recognition*) extraient les données d'une ressource textuelle numérisée, les informations relatives à la structure physique de l'image risquent de se perdre. Un schéma XML standardisé qui s'appelle ALTO (*Analyzed Layout and Text Object*) a été créé afin de conserver et structurer ces données non-textuelles et géometriques en les tenant en relation avec le contenu textuel. La plupart des modèles OCR et HTR compte sur ce schéma. Cependant ALTO ne convient pas bien à l'édition numérique ni aux traitements automatique du langage. Les éditeurs et les chercheurs en lettres attendent un schéma XML plus courant dans le monde des humanités numériques : la TEI (*Text Encoding Initiative*). Il faut donc un mapping pour transformer un fichier ALTO en fichier TEI sans perdre aucune donnée lors du processus. Cette transformation automatisée permet à conserver les données particulières au schéma ALTO, telles que celles sur la segmentation et sur la structure physique du document numérisé, ainsi qu'à exploiter le contenu textuel de la ressource textuelle. La flexibilité de la TEI et son usage très répandu rendent le schéma idéal pour mieux valoriser les données produites par les modèles OCR et HTR.

Dans le cadre du stage pour obtenir le diplôme de Master 2 « Technologies numériques appliquées à l'histoire », ce mémoire porte sur la modélisation de la transformation de ALTO en TEI. Cette modélisation a été réalisée dans le cadre du projet *Gallic(orpor)a*, financé par la BnF (Bibliothèque nationale de France) lors d'un stage qui a eu lieu au sein du laboratoire ALMAnaCH (Automatic Language Modelling and Analysis & Computational Humanities) entre avril et juillet 2022.

Mots-clés : HTR, OCR, ALTO, TEI, TAL, édition numérique.

Informations bibliographiques : Kelly Christensen, *Modélisation des transcriptions ALTO avec la TEI. En complétant le pipeline du projet* Gallic(orpor)a, mémoire de master « Technologies numériques appliquées à l'histoire », dir. [Noms des directeurs], École nationale des chartes, 2022.

Remerciements

Mes remerciements vont tout d'abord à...

Introduction

L'introduction parlera du traitement en masse des ressources textuelles et le besoin de l'automatiser à l'échelle.

Des institutions patrimoniales génèrent plus en plus d'images numérisées de leurs fonds. Il faut maintenant un pipeline généralisé qui peut transformer une image en une édition numérique structurée pouvant servir à l'étude littéraire, historique, et linguisitique. Le projet *Gallic(orpor)a* vise à construire un tel pipeline.

Peut-être je parlerai dans l'introduction de ce que j'ai fait par rapport au projet ?

Je m'occupais d'une partie de ce pipeline. Ma modélisation en TEI des fichiers ALTO sert à faire la connection entre les transcriptions automatisées des images et les données structurées pour l'édition numérique.

Première partie

Le contexte du projet

Chapitre 1

Traitement automatique des ressources textuelles

1.1 Guildelines *SegmOnto*

Parler des guidelines *SegmOnto* (J.-B. Camps) qui ont été pris pour standardiser la classification des regions et des lignes de texte du corpus.

1.2 Entraînement des modèles HTR

Parler de la génération du corpus d'or pour entraîner les modèles sur les guidelines *SegmOnto*, et comment les modèles HTR fonctionnent.

Chapitre 2

Le pipeline de *Gallic(orpor)a*

2.1 Extraction de texte

Parler de l'application des modèles entraînés pour extraire des données et la sortie produite.

2.2 Transformation d'ALTO en TEI

Parler rapidement de la transformation en TEI des fichiers ALTO. Rapidement parce qu'il est le sujet du mémoire.

2.3 Annotation du texte

Parler rapidement de l'application des outils du traitement automatique du langage au contenu textuel extrait et structuré dans le fichier TEI. Rapidement parce qu'il est l'autre sujet (moins important parce que j'aurai moins de temps pour y travailler) du mémoire.

2.4 Exportation diversifiée

Parler de la transformation du document TEI en divers formats, tel que RDF, IIIF et, à la suite d'une reconversion, ALTO. Ce dernier sert à entraîner de nouveau des modèles.

Chapitre 3

Projet *Gallic(orpor)a*

Est-ce que je devrais commencer avec une présentation du projet et le pipeline (contraire à ce que j'ai fait ci-dessus), ou devrais-je commencer avec une explication du traitement automatique des resources textuelles : les modèles HTR et les normes de classification des regions/lignes de texte ?

3.1 Présentaiton des institutions

3.1.1 BnF DataLab

3.1.2 Inria

3.1.3 École nationale des chartes et Université de Genève

3.2 Présentation du projet

Ici je reproduirais ce que j'ai envisagé dans le chapitre précédent, « Le pipeline de *Gallic(orpor)a* »

3.2.1 Extraction de texte

3.2.2 Transformation d'ALTO en TEI

3.2.3 Annotation du texte

3.2.4 Exportation diversifiée

Deuxième partie

D'une transcription ALTO en édition TEI

Chapitre 4

Structure des données

4.1 ALTO

Parler du schéma ALTO, ses origines, son utilisation, ses avantages, ses desavantages.

4.2 TEI

Parler du schéma TEI, ses origines, son utilisation, ses avantages, ses desavantages.

Chapitre 5

Mapping

5.1 Métadonnées du document numérisé

5.1.1 Le `<teiHeader>`

Parler du `<teiHeader>` et son utilisation.

5.1.2 Extraction des métadonnées

Parler de l'extraction des données de l'IIIF Image API et de l'API SRU (Search/Retrieval via URL) Catalogue général de la BnF.

5.1.3 Exploitation des métadonnées

Parler du mapping des données sortant des API (JSON de l'API IIIF et XML Unimarc de l'API SRU) dans le `<teiHeader>`.

5.2 Transcription de l'image

5.2.1 Le `<sourceDoc>`

Parler du `<sourceDoc>`, son utilisation et pourquoi il marche bien pour gérér les données textuelles et graphiques du fichier ALTO.

5.2.2 Extraction des données du fichier ALTO

Parler de l'extraction des données du fichier ALTO en python.

5.2.3 Exploitation des données du fichier ALTO

Parler du mapping des données des éléments du fichier ALTO vers les éléments TEI, spécifiquement le `<sourceDoc>`.

Chapitre 6

Annotation du texte

Exploitation des données déjà mappées et structurées dans le document préliminaire TEI.

6.1 Une transcription hierarchisée

6.1.1 Le `<body>`

Parler du `<body>` et son utilisation : analyse linguistique ou littéraire. Conserver le texte ainsi qu'il est dans la source originale, pas de correction.

6.1.2 Extraction du texte du `<sourceDoc>`

Parler de l'extraction et manipulation du texte avec le script python, y compris le mapping des éléments du `<sourceDoc>` et la classifcation des guideliens *SegmOnto* vers les éléments du `<body>`. Par exemple, une ligne de texte classifée *HeadingLine* sera balisée dans l'élément TEI `<hi rend="HeadingLine">`.

6.2 Le texte normalisé

6.2.1 Le `<standOff>`

Parler du `<standOff>` et son utilisation.

6.2.2 Le traitement automatique du langage

Parler du TAL et les modèles à utiliser pour lemmatiser, normaliser, et reconnaître des entités nommées le texte segmenté.

Je ne sais pas si je devrais mettre avant cette sous-section une explication de la segmentation que je devrais faire en python afin de passer le texte aux modèles de lemmatisation, etc. Ou s'il n'est pas aussi importante pour mériter une sous-section

Troisième partie

Après le projet

Chapitre 7

Critique

Critiquer ce que j'ai fait / les stratégies du projet *Gallic(orpor)a.*. Parler des autres pistes / projets similaires.

Chapitre 8

Autres utilisations

Parler des autres utilisation de la modélisation en TEI que j'ai faite.

Conclusion

La conclusion : résumer...

Annexe A

Données

Est-ce qu'il faut mettre dans un appendice des données ?

Table des matières