# Réalisation et Evaluation d'un dataset d'entraînement d'un modèle de segmentation

Juliette Janes

20-04-2021

Le but de ce travail est d'améliorer l'OCR du workflow d'encodage des catalogues d'exposition réalisée par Caroline Corbières en 2020 [1]. Pour ce faire, il convient donc, dans un premier temps, de migrer les données travaillées sur Transkribus dans eScriptorium. En effet, Transkribus étant devenu payant en octobre 2020, cela permet de conserver une chaîne de traitement entièrement open-source et accessible pour les potentiels réutilisateurs. Les modèles de segmentation et de reconnaissance réalisés dans Transkribus n'étant pas exportable, il a donc fallu en exporter les fichiers XML de transcription, en format ALTO2 et Page XML (version 2013).

Ainsi, nous aimerions améliorer l'OCR des catalogues d'exposition mais également d'en nommer ses zones de segmentation et de récupérer et conserver l'information typographique des documents. Pour ce faire, il était nécessaire de réaliser un nouveau set de données afin d'entraîner un modèle à partir de celles-ci. Parmi les solutions pour avoir un meilleur modèle, nous nous sommes intéressés à l'injection de pages de catalogues qui ne sont des catalogues d'exposition à ce set de données. En effet, en entraînant un modèle capable de reconnaître et de transcrire une plus grande hétérogénéité de documents, on améliore également la qualité globale du modèle. On aurait ainsi pu récupérer des données différentes et en associer une petite quantité de 10% à notre set de données. Il serait donc tout à fait intéressant d'essayer de réaliser un modèle permettant de segmenter et reconnaître efficacement ces autres types de données, afin de concevoir un outil utilisable par des chercheurs en dehors de notre projet. En suivant cette idée, nous nous sommes donc arrêté sur la création d'un modèle améliorant l'OCR du workflow d'Artl@s tout en étant réutilisable en dehors du projet.

Dans un premier temps, nous avons réalisé un dataset test, contenant trois types de données en proportion égale. Nous nous sommes arrêté sur une quantité de 30 pages, soit 10 pages par types, afin d'évaluer la quantité et le temps de travail sur un échantillon.

## 1 Description des données utilisées

Le premier type de données utilisé pour réaliser ce dataset correspond aux catalogues d'exposition d'Artl@s, préparés par Caroline Corbières [2]. Ceux-ci sont la source de travail principal des historiens de l'art permettant de documenter précisemment une exposition et de retracer la circulation des images mais également d'obtenir des informations sociales,

1. Caroline Corbières, Simon Gabay and Béatrice Joyeux-Prunel, *Worklow to encode exhibition catalogues*, 2020, https ://github.com/carolinecorbieres/ArtlasCatalogues.
2. Caroline Corbières, Simon Gabay and Béatrice Joyeux-Prunel, Ibid

économiques, politiques et géographiques sur le marché de l'art. Il s'agit de documents allant du début du XIXème à la fin du XXème, utilisé partout dans le monde et dont la forme est assez établie internationalement. Les catalogues de ventes de manuscrits sont des données du projet Katabase [3] de Simon Gabay, du milieu du XIXème à aujourd'hui, éditées par des libraires parisiens et rendant compte de leurs ventes. Enfin, le dernier type de données est l'Annuaire de Commerce de 1898, utilisé par le groupe Addresses et Annuaires de Paris Time Machine [4]. Ces données ont été préparées par Gabriela Elgarrista sous la direction de Carmen Brando.

Toutes les données utilisées, catalogues d'exposition, catalogues de vente de manuscrits et annuaires, ont été abbyysés puis insérées dans Transkribus afin d'entraîner les modèles de segmentation et de reconnaissance. Ainsi, il a été possible de récupérer ces données en export de Transkribus en ALTO2 et en PageXML 2013. eScriptorium utilisant du ALTO4 et non ALTO2 et PageXML étant plus riche, il a été décidé de mettre en entrée du logiciel le second format.

Il a fallu ensuite sélectionner les pages du dataset afin d'avoir un échantillon global du corpus. Ainsi, les pages correspondant aux catalogues d'exposition et aux catalogues de vente de manuscrits sont représentatif de la diversité des dates, lieux de production (pour les catalogues d'exposition surtout) et éditeurs, tandis que les pages de l'annuaire de commerce de 1898 proviennent chacune d'une lettre différente de l'alphabet. [5]

# 2 Préparation et Segmentation

Une fois le set de données décidé, celui-ci a été intégré dans eScriptorium. Nous avons alors fait face à un problème d'affichage. Le texte présenté par le logiciel correspondait à une transcription non corrigée. En effet, comme on peut voir sur l'image, la structure du format PageXML fait que la transcription réalisée précédemment par ABBYY et Transkribus est présentée dans le //TextEquiv de la balise //Word, tandis que la transcription corrigée à la main se trouve dans les //TextEquiv des balises //TextRegion et //TextLine. eScriptorium récupérerait donc les transcriptions situées au niveau du mot. Après réflexion et confirmation de Gabriela Elgarrista [6], Il s'agirait d'un problème émanant de la correction manuelle réalisée sur la transcription sur ABBYY dans la première partie de la chaîne. Dans le cas du groupe de travail Addresses et Annuaires, ces modifications ont eu lieu uniquement sur 14 feuilles, qu'il a donc été facile d'évincer du set de données. Cela n'est pas une solution pour les catalogues d'exposition et de vente de manuscrits, ceux-ci étant presque entièrement corrigés à la main. Ainsi, une solution a été trouvée de réaliser une feuille de transformation XSLT permettant de récupérer les transcriptions corrigées au niveau de la ligne et de les intégrer directement au niveau mot [7].

L'an dernier, un travail avait été tenté par Caroline Corbières sur la typographie. L'idée était de récupérer et signaler celle-ci par l'utilisation des balises html correspondantes : <b> pour le gras, <i> pour l'italique. Cette méthode ne fonctionnant pas, il a également

3. https ://github.com/katabase/OCRcat

4. https ://paris-timemachine.huma-num.fr/groupe-adresses-et-annuaires/

5. Pour un détail des pages composant le dataset :`https://github.com/Juliettejns/cataloguesPipeline/blob/main/Dataset_190421.csv`

6. Juliette Janes, *Appropriation du projet, travaux précédents*, 15/04/2021, fichier retour d'expérience du projet overleaf

7. `https://github.com/Heresta/BAO_Stage_DH_ENS_2021/tree/main/CorrectionPageXMLeScriptorium`

```
<TextRegion type="paragraph" id="r_1_1" custom="readingOrder {index:0;}">
    <Coords points="1849,726 4150,726 4150,1037 1849,1037"/>
    <TextLine id="tl_1" primaryLanguage="English" custom="readingOrder {index:0;}">
        <Coords points="1850,727 4149,727 4149,1036 1850,1036"/>
        <Baseline points="1851,951 4149,1036"/>
        <Word id="w_w2aablb1b2b1blab1" language="English" custom="readingOrder {index:0;}">
            <Coords points="1851,727 4149,727 4149,1036 1851,1036"/>
            <TextEquiv>
                <Unicode>(Gâ'TM.iGMEüm</Unicode>
                <!--Mot transcrit par Transkribus et affiché dans eScriptorium-->
            </TextEquiv>
            <TextStyle fontFamily="Times New Roman" fontSize="27.0"/>
        </Word>
        <TextEquiv>
            <!-- texte corrigé à la main de la transcription dans la balise Word, situé dans le TextLine-->
            <Unicode>CATALOGUE</Unicode>
        </TextEquiv>
        <TextStyle fontFamily="Times New Roman" fontSize="27.0"/>
    </TextLine>
    <TextEquiv>
        <!--texte corrigé à la main de la transcription dans la balise Word, situé dans le TextRegion-->
        <Unicode>CATALOGUE</Unicode>
    </TextEquiv>
</TextRegion>
```

FIGURE 1 – Un TextRegion détaillé problématique

fallu réaliser un script python afin de supprimer les balises. [8]. Une autre méthode est en cours de réflexion par un groupe de travail rassemblant Thibault Clérice (École nationale des Chartes), Simon Gabay et Anna Scius-Bertrand (UNIGE) [9].

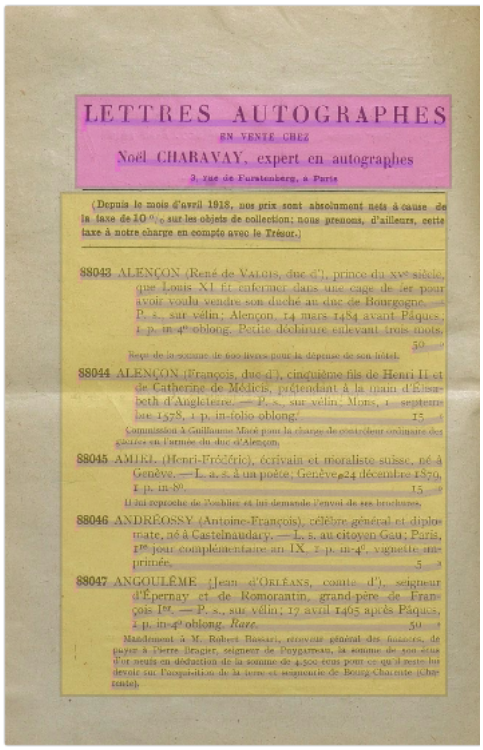



FIGURE 2 – Segmentation Simple

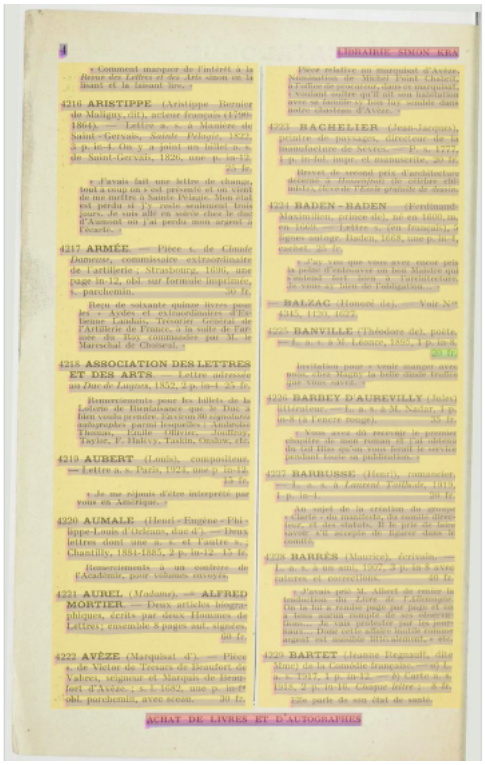


FIGURE 3 – Segmentation Double

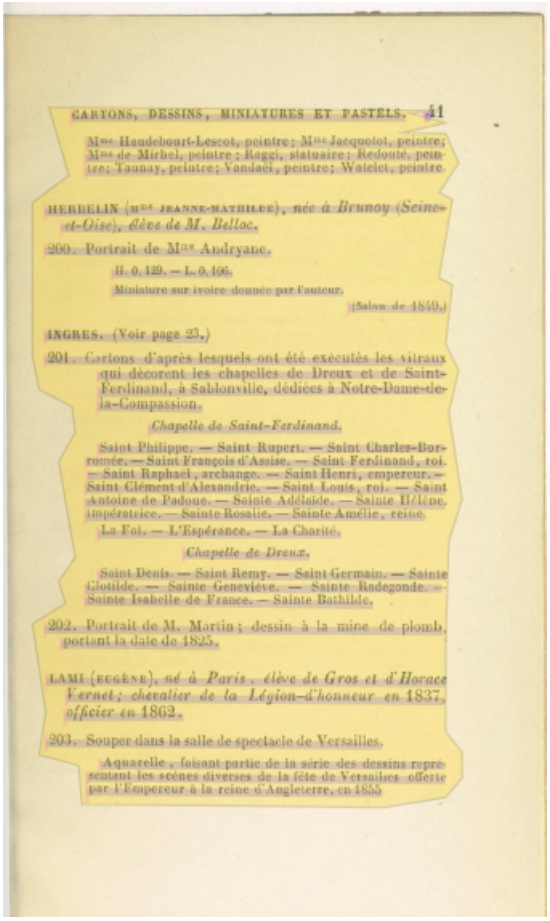
8. https://github.com/Heresta/BAO_Stage_DH_ENS_2021/blob/main/suppressionGrasItalique.py
9. https ://github.com/PonteIneptique/alto-style-classifier

Enfin, afin d'entraîner un modèle de segmentation capable de nommer les zones et les lignes, il a fallu intégrer cela dans le dataset. Pour ce faire, nous nous sommes basés sur le projet SegmOnto [10], réalisé par un groupe de travail tentant de créer une ontologie pour le nommage des zones en HTR basé sur la TEI. Des exemples de ces segmentations simples obtenues sont présentées sur les figures 1 et 2 suivantes. Le but a été de réaliser une segmentation la plus épurée possible afin d'obtenir un segmenteur fonctionnel. Si toutes les lignes de ces documents sont, d'après la documentation SegmOnto, définies par le terme *default*, et ont donc été ajoutées automatiquement, ce n'est pas le cas des zones, qui, plus nombreuses, ont dû être réalisées à la main.

— **Title**, *rose* : permet d'indiquer un titre.
— **Main**, *jaune* : permet d'indiquer une colonne de texte.
— **Numbering**, *violet* : signale la pagination.
— **Running Title**, *rouge* : signale les titres en haut d'une page.
— **Figure**, *bleu* : signale les illustrations et images.

Une fois la création de ces nouvelles zones ainsi que leur nommage réalisé, il a fallu supprimer les anciennes zones de segmentation existantes et non nécessaires et lier les lignes contenues dans les zones à celles-ci.

Après une heure d'entraînement de ce dataset, nous avons obtenu un modèle ayant un taux de précision assez bas de 62%, sans erreurs. Un test réalisé sur 3 pages non préparées a permis d'obtenir une segmentation plus qu'améliorable, comme le montre la figure 4. Une solution pour améliorer ce rendu serait de réaliser un dataset plus large, afin d'avoir un taux d'accuracy plus élévé. On pourrait alors viser un dataset de 78 pages, soit 26 pages par types de données, permettant ainsi d'avoir une page par lettre pour l'annuaire. Auquel cas, il serait peut être intéressant de voir comment automatiser le nommage des zones, celles-ci étant les plus longues à réaliser lors du travail de préparation. Dans ce cas là, le dataset présenté ici ayant été préparé en 2 jours, avec un peu d'entraînement et de prise de rapidité, il serait possible de le réaliser en moins d'une semaine. Avant cela, il serait également possible de tester un dataset avec un seul type de données d'abord pour voir si un modèle serait capable de reconnaître des zones efficacement. Il serait également possible de simplifier encore plus les zones à reconnaître, par exemple en commençant par entraîner un modèle ne prenant en compte que les titres et parties principales, sans s'intéresser aux running titles, par exemple. Une dernière solution serait de commencer à entraîner un modèle sur un dataset de catalogues d'exposition puis d'incorporer au fur et à mesure les autres types de données afin de vérifier la fiabilité du segmenteur produit plus facilement.


CARTONS, DESSINS, MINIATURES ET PASTELS. 41

Mme Haudebourt-Lescot, peintre; Mme Jacquotot, peintre; Mme de Mirbel, peintre; Raggi, statuaire; Redouté, peintre; Taunay, peintre; Vandaël, peintre; Watelet, peintre.

HERBELIN (Mme JEANNE-MATHILDE), *née à Brunoy (Seine-et-Oise), élève de M. Belloc.*

200. Portrait de Mme Andryane.

H. 0,429. — L. 0,106.

Miniature sur ivoire donnée par l'auteur.

(Salon de 1849.)

INGRES. (Voir page 23.)

201. Cartons d'après lesquels ont été exécutés les vitraux qui décorent les chapelles de Dreux et de Saint-Ferdinand, à Sablonville, dédiées à Notre-Dame-de-la-Compassion.

*Chapelle de Saint-Ferdinand.*

Saint Philippe. — Saint Rupert. — Saint Charles-Borromée. — Saint François d'Assise. — Saint Ferdinand, roi. — Saint Raphaël, archange. — Saint Henri, empereur. — Saint Clément d'Alexandrie. — Saint Louis, roi. — Saint Antoine de Padoue. — Sainte Adélaïde. — Sainte Hélène, impératrice. — Sainte Rosalie. — Sainte Amélie, reine.

La Foi. — L'Espérance. — La Charité.

*Chapelle de Dreux.*

Saint Denis. — Saint Remy. — Saint Germain. — Sainte Clotilde. — Sainte Geneviève. — Sainte Radegonde. — Sainte Isabelle de France. — Sainte Bathilde.

202. Portrait de M. Martin; dessin à la mine de plomb, portant la date de 1825.

LAMI (EUGÈNE), *né à Paris, élève de Gros et d'Horace Vernet; chevalier de la Légion-d'honneur en 1837, officier en 1862.*

203. Souper dans la salle de spectacle de Versailles.

Aquarelle, faisant partie de la série des dessins représentant les scènes diverses de la fête de Versailles offerte par l'Empereur à la reine d'Angleterre, en 1855

FIGURE 4 – Segmentation obtenue

10. https ://github.com/SegmOnto