Numériser les archives d'histoire de l'art

Pauline Jacsont, Simon Gabay, Tristan Weddigen

Numériser les archives d'histoire de l'art : la collection de photographies d'Heinrich Wölfflin

Pauline Jacsont, Simon Gabay
Université de Genève
{prenom.nom}@unige.ch

Tristan Weddigen
Université de Zurich
tristan.weddigen@uzh.ch

Résumé

La collection de photographies du célèbre Henrich Wölfflin, léguée et conservée à l'université de Zürich, est une source historique importante pour les spécialistes d'histoire de l'art. Riche de milliers de photographies, de notes manuscrites et imprimées, ou encore d'annotations par les conservateurs, sa souhaitable numérisation reste néanmoins un enjeu de taille. Si la simple mise à disposition de fac-similés en ligne facilite l'accès aux documents, elle ne simplifie aucunement l'accès à l'information qu'ils contiennent. L'expérience que nous avons menée tente de résoudre ce problème, au moyen de nouveaux modèles d'analyse, de mise en page et de reconnaissance de caractères, afin de revaloriser un corpus dont l'importance a longtemps été sous-estimée.

1 Introduction

Comme tous les spécialistes de sciences humaines, les historiens de l'art ont besoin de matériaux de travail, qui ont longtemps pris la forme de collections de photographies, devenues aujourd'hui des bases de données comme celles d'Aby Warburg [1] ou de la Humboldt (Schelbert, 2018). Ces collections revêtent une importance primordiale, longtemps sous-estimée jusqu'à ce que des chercheurs appellent à « dépasser une historiographie à dominante scripturale » et à mobiliser « l'image scientifique » (Griener, 2011, p. 103). Nombre d'entre elles restent malheureusement difficilement consultables, faute de numérisation convenable permettant une exploitation optimale.

Parmi les plus célèbres de ces collections, nous trouvons celle d'Heinrich Wölfflin (1864-†1945), successeur de Jacob Burckhardt à Bâle (1893) puis professeur d'histoire de l'art à l'université de Zürich. À sa mort, il lègue sa *Kunstwissenschaftliche Fotosammlung* à l'institut d'histoire de l'art, où elle est encore aujourd'hui conservée sous la forme de 4 200 photographies de différents formats, 16 600 photographies dans de petites boîtes et 39 000 diapositives de verre (cf. figure 1). La richesse de cette ressource dépasse de loin le simple contenu iconographique, car elle comporte nombre d'annotations manuscrites de la part de Wölfflin (cf. figure 2).

FIGURE 1 – Exemple de photographie de la collection Wölfflin (UZH 0239, 4, VI, 003 R).

Tant du point de vue de son histoire que de son contenu, la numérisation d'un tel fonds s'avère aussi importante que complexe. Étant donnée sa taille, la simple mise à disposition en ligne de fac-similés n'est pas convenable : il importe d'arriver à extraire un maximum d'informations et de les restructurer pour rendre le contenu facilement requêtable, et ainsi valoriser cette source encore trop sous-utilisée par les chercheurs.

FIGURE 2 – Exemple de photographie avec annotation manuscrite en bas à droite (UZH 0009, VII, 002 R).

1. https://warburg.sas.ac.uk/photographic-collection.

2 Description des données

FIGURE 3 – `FirstFolderPage` (UZH 0172, V, 000).

La chaire du *Kunsthistorisches Institut* de Zürich a réalisé différentes campagnes de numérisation et un inventaire de cette collection : chaque document a été regroupé au sein de portfolios (que nous nommerons `FirstFolderPage` ; cf. figure 3) sur lesquels sont inscrits une cote et un titre de dossier, dans lesquels se trouvent des pages avec les photographies d'œuvres (que nous nommerons `IllustrationPage` ; cf. figures 1, 2 et 4). D'un point de vue numérique, ce fonds contient un peu plus de 86 000 images aux formats JPG et TIFF.

FIGURE 4 – *Illustrations Pages* (UZH 0172, V, 003 R).

Ce type de corpus, dont l'objet principal est la reproduction d'œuvres d'art, suscite de nombreuses difficultés pour la reconnaissance de caractères (HTR). Premièrement, la documentation est extrêmement hétérogène : on trouve différentes strates d'annotation, d'époque et de langue distinctes. Deuxièmement, la masse de données textuelles par document est extrêmement limitée : les documents contiennent de nombreux tokens peu fréquents (anthroponymes, cotes…), et le nombre d'exemples pour chaque main reste faible. Troisièmement, il n'existe à notre connaissance aucun projet qui a entrepris une transcription automatique pour ce type de documents : en l'absence de données d'entraînement utilisables, il a donc fallu construire une vérité de terrain (VT) *ex nihilo* afin d'entraîner des modèles de segmentation et d'HTR.

3 Réalisation de la vérité de terrain

L'extraction d'informations nécessite l'établissement d'une VT pour deux tâches liées, mais distinctes : l'analyse de mise en page d'une part (diverses zones textuelles, illustrations, tampons…) et la reconnaissance optique de caractères d'autre part (texte en romain, en gothique, manuscrit…). Afin d'obtenir suffisamment de données pour l'entraînement, nous avons manuellement segmenté et réalisé la transcription graphématique de 559 `FirstFolderPage` et de 548 `IllustrationPage` grâce à l'application *eScriptorium* (Kiessling et al., 2019) [2]. La segmentation de la page (cf. figure 5) a suivi les recommandations du vocabulaire contrôlé *SegmOnto* (Gabay et al., 2021), et a retenu les zones suivantes :

- `TitlePageZone` pour les titres des portfolios (en rouge) ;
- `GraphicZone` pour les illustrations (en turquoise) ;
- `MarginTextZone` pour toutes les zones de texte qui ne se situent pas dans une `GraphicZone` (en violet) ;
- `NumberingZone` pour la cote (en orange) ;
- `DefaultLine:manuscript` pour les lignes standards manuscrites ;
- `DefaultLine:print` pour les lignes standards en caractères d'imprimerie ;
- `HeadingLine:manuscript` pour la titraille manuscrite ;
- `HeadingLine:print` pour la titraille en caractères d'imprimerie.

L'utilisation de *subtypes* (`print` vs `manuscript`) offre la possibilité de distinguer les différents types de ligne, et donc de les séparer lors de la création de modèles HTR, mais aussi au moment de l'application de ces derniers. De la même manière, grâce à l'usage d'une

2. Nous avons utilisé FoNDUE, l'instance de l'université de Genève.

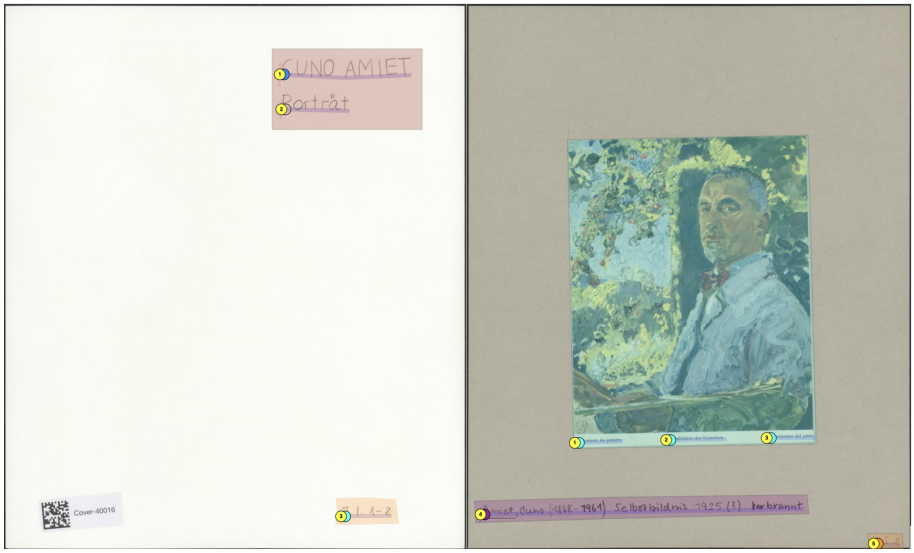


FIGURE 5 – Segmentation avec SegmOnto d'une *First Folderpages* et d'une *Illustrations Pages*.

terminologie précise, nous pouvons conserver un certain nombre d'informations relatives à la mise en page qui seront utiles pour la réalisation d'une base de données à plus long terme, notamment lors de la conversion de XML-ALTO vers d'autres formats (Pinche et al., 2022).

4 Résultats des différents modèles réalisés

La réalisation de modèles performants est une tâche complexe qui dépend de nombreux paramètres : la représentativité du corpus à transcrire, la qualité des transcriptions (Couture et al., 2022), de potentiels prétraitements des images, du recours à la *data augmentation*, etc. Il convient donc de tenter plusieurs expériences pour définir la meilleure méthode.

Étant donné l'extrême hétérogénéité des documents, l'entraînement de modèles de segmentation distincts selon le type de page [3] et de modèle d'HTR selon le type d'écriture [4] a constitué notre hypothèse de départ.

4.1 Les modèles de segmentation

Nous avons réalisé deux modèles de segmentation [5].

— Modèle pour les `FirstFolderPage` : FWIoU 0.76 ;
— Modèle pour les `IllustrationPage` : FWIoU 0.89.

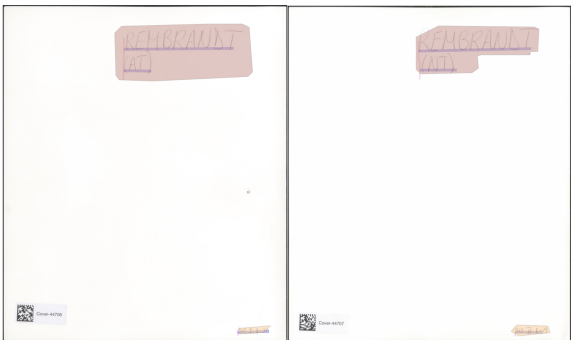
Pour tester notre hypothèse, nous avons ensuite entraîné avec l'ensemble de nos données (`FirstFolderPage` + `IllustrationPage`) un modèle de segmentation qui a obtenu une FWIoU bien supérieure de 0.94.

4.2 Les modèles pour la transcription automatique

Le corpus a été divisé selon chaque type d'écriture (`manuscriptLine` et `printLine`), en trois jeux de données (entraînement, évaluation et test) permettant respectivement d'entraîner le modèle, d'évaluer les résultats à chaque itération lors de l'entraînement, puis de faire un test final sur le modèle avec des données qu'il n'a jamais vues [6].

Ces expériences ont permis d'interroger non seulement la pertinence de réaliser un modèle extrêmement spécialisé selon le type d'écriture et de document traité, mais aussi le bénéfice des prétraitements et de la *data augmentation*. Dans le tableau *infra* (cf. tab. 1), sont consignés les résultats obtenus pour chaque expérience ; nous avons réalisé des modèles selon les différents types de page et d'écriture, mais aussi en combinant nos données avec

3. *FirstFolderPage* et*IllustrationPage*.
4. *manuscriptLine* pour toutes les lignes manuscrites, et *printLine* pour toutes lignes imprimées ou dactylographiées.
5. La métrique d'évaluation utilisée est la FWIoU (*frequency-weighted intersection over union* et non l'IoU (*intersection over union*). En effet, l'intersection sur l'union ne propose pas un score représentatif du résultat car elle ne distingue pas les différentes classes : les moins fréquentes (donc les moins bien reconnues, comme les tampons de bibliothèques) font diminuer le score. La FWIoU, elle, assigne un poids en fonction de la fréquence de chaque classe.
6. Les différents jeux de données sont disponibles en suivant ce lien : https://github.com/FoNDUE-HTR/FoNDUE_Kunsthistorisches-UZH_Archivdatenbank/tree/main/4_Split.



(a) `FirstFolderPage` segmentée à gauche avec le modèle spécialisé et à droite le modèle général.

(b) `IllustrationPage` segmentée à gauche avec le modèle spécialisé et à droite le modèle général.

Evaluation test set → HTR Model ↓	*Manuscript-Lines* Firstfolderpages	*Manuscript-Lines* Illustrationspages	All *manuscriptLines*	*All printLines*	*Manuscript-Lines & printLines* Illustrations pages	*All*
GT1 *Firstfolderpage*	**93.74%**	7.53 %	43.50 %	11.62 %	9.43 %	27.28 %
GT2 *ManuscriptLines* illustrations pages	89.03 %	80.30 %	**88.57%**	18.38 %	35.71 %	52.68 %
GT3 All *manuscriptLines*	**93.93%**	**83.53%**	85.89 %	17.57 %	33.02 %	27.17 %
GT4 *PrintLines*	//	//	//	[0.7] %	//	//
GT5 *Illustrations pages*	12.32 %	85.34%	48.25 %	**71.08%**	74.04%	59.84 %
GT6 *All*	91.13%	73.14%	82.16%	52.84%	64.08%	70.56%
GT7 *ManuscriptLines* & Lectaurep-repertoires	[24.21]%	//	//	//	//	//
GT8 *PrintLines* & Cremma16-17	32.66 %	33.47 %	42.36 %	**67.84%**	54.48 %	50.55 %
GT9 *All* & Lectaurep-repertoires	**59.91%**	41.72 %	54.87 %	33.92 %	35.52 %	41.64 %

TABLEAU 1 – Résultats des différentes expériences réalisées pour mesurer l'efficience de chaque modèle. L'en-tête horizontal correspond au nom du jeu de test utilisé : nous avons voulu ainsi évaluer chaque modèle selon son efficience sur le type de page et le type d'écriture ; l'en-tête vertical correspond aux différents jeux de données utilisés pour l'entraînement du modèle. En gras sont indiqués les meilleurs résultats de chaque modèle, et en couleur les meilleurs résultats par set. Lorsque le résultat d'un test d'entraînement a une précision par caractère inférieur à 50%, les différents tests d'évaluation n'ont pas été réalisés, le résultat du test d'entraînement est alors indiqué entre parenthèses.

d'autres jeux de données [7], ou encore en réalisant quelques simples prétraitements [8].

4.3 Discussion

En ce qui concerne la segmentation, la séparation de notre corpus en deux sous-ensembles, selon le type de document, n'a pas permis une amélioration des résultats ; le modèle de segmentation est ainsi beaucoup plus performant lorsqu'il est entraîné sur un nombre de données plus important, même si celles-ci ont des mises en page très hétérogènes.

L'expérimentation faite avec les modèles HTR, dont les résultats sont exposés dans le tableau 1 révèle que les entraînements avec *data augmentation* [9] ou prétraitement n'apportent pas d'amélioration conséquente. Pour la transcription des `manuscriptLine` c'est le modèle GT3, entraîné avec toutes les `manuscriptLine`, qui obtient le meilleur résultat. Les `printLine` obtiennent dans l'ensemble de moins bons résultats, probablement à cause de la nature extrêmement hétérogène des caractères et du plus faible nombre de lignes par rapport aux `manuscriptLine`. Toutefois, il faut noter que le modèle GT3 se révèle être tout particulièrement performant sur les `manuscriptLine` des portfolios, qui sont toujours réalisées avec la même main, mais l'est beaucoup moins sur les `manuscriptLine` des `IllustrationPage`, qui proviennent de plusieurs mains et sont donc plus irrégulières. En définitive, le modèle GT3, entraîné sur toutes les `manuscriptLine`, et le modèle GT5, réalisé seulement sur un corpus avec des *Illustrations Pages*, obtiennent de bien meilleurs résultats que le modèle GT6 conçu sur l'ensemble des données [10].

Tel que l'a démontré Springmann et al. (2016) construire un modèle extrêmement spécifique, avec seulement un échantillon des données à transcrire, n'est pas bénéfique : cela se constate notamment en comparant les résultats de GT1, très spécialisé car seulement entraîné avec les données manuscrites provenant de la même main, et GT3 entraîné avec toutes les données manuscrites [11]. Les résultats du tableau 1 révèlent aussi que la combinaison de données n'est pas automatiquement bénéfique. Concernant cette dernière affirmation, le cas des `printLine` est tout particulièrement intéressant : le modèle GT5, entraîné avec des `printLine` et `manuscriptLine`, est nettement plus performant que le modèle entraîné avec seulement les `printLine`, mais le modèle GT6, qui contient un plus grand nombre de lignes manuscrites, obtient des résultats plus faibles sur ce même jeu. Ces expériences montrent que la combinaison des données d'entraînement doit être faite de manière réfléchie et modérée.

4.4 Conclusion

Si les premiers résultats de nos expériences peuvent encore être améliorés, ils démontrent que l'extraction d'informations depuis des documents complexes comme ceux du fonds Wölfflin avec des solutions intégralement ouvertes et accessibles à tous est désormais possible, que les outils utilisés par les philologues, les littéraires ou les historiens peuvent aussi servir aux historiens de l'art, et donc qu'il est important de dépasser l'usage des moteurs OCR pour les sources textuelles uniquement.

En ce qui concerne la numérisation du fonds Wölfflin, trois scénarios se dessinent dans l'immédiat pour réaliser la transcription automatique du corpus :

— le premier plus pragmatique consisterait à choisir le modèle GT6 permettant d'obtenir une bonne moyenne de résultats sur l'ensemble des données ; cette solution a l'avantage d'être techniquement la plus simple à réaliser.

— le deuxième serait de transcrire les documents en utilisant les modèles qui ont obtenu les meilleurs résultats selon le type de ligne à transcrire [12], ce scénario pourrait aisément se mettre en place grâce aux *subtypes* (`print` et `manuscript`) utilisés lors de la phase de segmentation ;

7. La *data augmentation* a été faite avec trois VT différentes (mais toujours francophones) : Cremma16-17 qui contient des transcription de livres imprimées du XVI^{e} au XVII^{e} s. ; CremmaMs_20, avec des transcriptions manuscrites du XX^{e} s. ; Lectaurep-repertoires qui est une VT réalisée avec des registres des XIX^{e} et XX^{e} s. de notaires parisiens.

8. Nous avions pour hypothèse que l'utilisation de prétraitement pourrait améliorer la lisibilité de certaines données écrites au crayon à papier. Les scripts concernant la réalisation de ces prétraitements sont disponibles en ligne : `https://github.com/FoNDUE-HTR/FoNDUE_Kunsthistorisches-UZH_Archivdatenbank/blob/main/5_Script_python/ImagesTreatments.ipynb`.

9. Cremma16-17, CremmaMs_20 et Lectaurep-repertoires

10. Précision par caractère de 93.74% vs 91.13% pour les `manuscriptLine`, et 71.08% vs 52.84% pour les `printLine`.

11. Précision par caractère 93.74% vs 93.93%.

12. GT2 pour les `manuscriptLine` (précision par caractère 88.57%) et GT5 pour les `printLine` (précision par caractère 71.08%

— le troisième scénario consisterait à réaliser les transcriptions selon les modèles qui ont obtenu les meilleurs résultats selon le type de page, la distinction entre les différents fichiers se fera à partir de leur nom qui varie suivant le type de document.

Financements

Le présent travail a bénéficié d'un financement COFUND des universités de Zurich et de Genève.

Bibliographie

Thibault Clérice. 2022. You actually look twice at it (YALTAi): using an object detection approach instead of region segmentation within the kraken engine.

Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2021. Archives and AI: An overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage*, 15(1) :4 :1–4 :15.

Béatrice Couture, Verret Farah, Gohier Maxime, and Deslandres Dominique. 2022. The challenges of HTR model training: Feedbacks from the project donner le goût de l'archive à l'ère numérique.

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, and Nicola Carboni. 2021. A controlled vocabulary to describe the layout of pages.

Simon Gabay, Pierre Kuenzli, Jean-Luc Flacone, and Christophe Charpilloz. 2022. FoNDUE: Documentation.

Pascal Griener. 2011. *Pour une nouvelle histoire des images scientifiques. Eugène Müntz, la Renaissance et la nouvelle fonction de la photographie d'art*, Italienische Forschungen des Kunsthistorischen Institutes in Florenz, Max-Planck-Institut. I Mandorli, pages 101–116. Deutscher Kunstverlag, Berlin.

Benjamin Kiessling. 2019. Kraken - a universal text recognizer for the humanities. In *Digital Humanities Conference 2019 - Book of abstracts*, Utrecht, the Netherlands.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, Sydney, Australia.

Ariane Pinche, Kelly Christensen, and Simon Gabay. 2022. Between automatic and manual encoding. In *TEI 2022 conference : Text as data*, Newcastle, United Kingdom.

Georg Schelbert. 2018. Bildgeschichte digital greifbar. die glasdiasammlung des instituts für kunst- und bildgeschichte der humboldt-universität zu berlin. bericht von einem work in progress.

U. Springmann, F. Fink, and K. U. Schulz. 2016. Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. *ArXiv e-prints*.

Phillip Ströbel, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel, and David Schoch. 2022. Evaluation of HTR models without ground truth material. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.