

ÉCOLE NATIONALE DES CHARTES

Juliette Janès

licenciée ès lettres

licenciée ès sciences

Du catalogue papier au numérique

**Une chaîne de traitement ouverte pour
l'extraction d'informations issues de
documents structurés**

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

2021

Résumé

Ce mémoire a été réalisé à la suite d'un stage de quatre mois à Artl@s, projet en histoire de l'art et humanités numériques dirigé par Béatrice Joyeux-Prunel et financé par l'École Normale Supérieure et le centre IMAGO. Ce projet a pour but de rassembler des catalogues d'exposition du XIX^{ème} et XX^{ème} siècle issus du monde entier au sein de la base de données Basart. Un premier travail, réalisé par Caroline Corbières, a permis d'établir une chaîne de traitement permettant d'automatiser la production de ces catalogues, de leur version numérisée à leur versement dans la base. L'objectif de ce stage a été de tester une alternative libre, ouverte et gratuite à ce travail.

Ce mémoire s'attache donc à décrire les différentes briques permettant l'élaboration de cette chaîne de traitement. Il s'intéresse à la problématique de la récupération puis l'annotation d'informations depuis des documents semi-structurés, en ciblant son propos autour des catalogues. Il développe tout au long de ce travail une réflexion autour de la Science Ouverte, l'application de ses principes et son intérêt pour les projets de recherche, en prenant appui sur l'exemple d'Artl@s.

Mots-clés : Artl@s ; Katabase ; catalogues ; histoire de l'art ; documents semi-structurés ; OCR ; Python ; Alto ; XML-TEI ; Science Ouverte.

Informations bibliographiques : Juliette Janès, *Du catalogue papier au numérique : Une chaîne de traitement ouverte pour l'extraction d'informations issues de documents structurés*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Simon Gabay, École nationale des chartes, 2021.

Remerciements

Je voudrais tout d'abord remercier ma tutrice de stage, Béatrice Joyeux-Prunel, pour sa confiance et son écoute. Je tiens également à adresser toute ma gratitude à Simon Gabay, qui par sa patience, sa disponibilité et ses conseils, m'a permis d'atteindre mes objectifs et de grandir pendant ces quatre mois.

Je remercie également Thibault Clérice, mon directeur de mémoire, pour ses nombreuses recommandations et conseils, qui m'ont permis de contourner les problèmes rencontrés au cours de ce stage.

Je me dois de remercier les participants des réunions de projets associés à mon stage, tels que *Visual Contagions*, le projet Gras ou encore SegmOnto, auxquelles j'ai eu l'opportunité de participer. Leur accueil, leurs explications et leurs conseils m'ont permis d'appréhender au mieux mon travail.

Je tiens à remercier plus particulièrement Claire Jahan pour son soutien, son aide et son travail au cours de ce stage. Notre collaboration m'a grandement aidée à mener à bien mes missions et à démarrer de la meilleure façon possible mon travail.

Pour finir, je me dois de remercier ma famille et Mélanie, qui m'ont soutenue et épaulée tout au long de la rédaction de mon mémoire. Merci à Kiki pour sa compagnie et ses miaulements qui ont égayés mon année en distanciel. Enfin, un grand merci à mes relectrices Fanny et Elise pour leur soutien et leur présence : ces années d'études n'auraient pas été pareilles sans vous.

Bibliographie

Histoire des catalogues

- ALBANO (Caterina), *Exhibition*, dir. Oxford University, Oxford et New York, URL : <https://ualresearchonline.arts.ac.uk/id/eprint/8058/>.
- BARBIER (Frédéric), DUBOIS (Thierry), SORDET (Yann) et BROGLIE (Gabriel de), *De l'argile au nuage : une archéologie des catalogues (IIe millénaire av. J.-C. - XXIe siècle)*/ouvrage publié à l'occasion des expositions organisées par la Bibliothèque Mazarine & la Bibliothèque de Genève, Paris 13 mars - 15 mai 2015, Genève 18 septembre - 21 novembre 2015], Editions des Cendres, 2015 (Bibliothèques).
- BERTRAND DORLÉAC (Laurence), *Le Commerce de l'art : de la renaissance à nos jours*, Editions La Manufacture, Besançon, 1992.
- BODIN (Thierry), « Les grandes collections de manuscrits littéraires », dans *Les ventes des livres et leurs catalogues, XVIIe-XXe siècle*, dir. Annie Charon et Élisabeth Parinet, Publications de l'École nationale des chartes, Paris, 2018 (Études et rencontres), p. 169-190, URL : <http://books.openedition.org/enc/1419> (visité le 13/04/2021).
- BODIN (Thierry) et NEEFS (Jacques), « Les autographes. Entretien », *Genesis (Manuscrits-Recherche-Invention)*, 7 (1995), p. 177-184.
- BON (Laurent Le) et HUESCA (Roland), « Propos autour du catalogue d'expo.... Entretien avec Roland Huesca », *Le Portique. Revue de philosophie et de sciences humaines*-30 (juil. 2013), DOI : 10.4000/leportique.2638.
- BURCKHARDT (Leonhard (Basle)), *Katalogos*, en, oct. 2006, URL : <https://referenceworks.brillonline.com/entries/brill-s-new-pauly/katalogos-e610620> (visité le 27/08/2021).
- CASTETS (Sylvie), « Les biennales internationales d'art contemporain et leurs touristes-amateurs, vus sous l'angle de l'utopie », *Études caribéennes*-37-38 (oct. 2017), DOI : 10.4000/etudescaribeennes.11291.
- CHANTE (Alain), « La notion de catalogue : de l'imprimé au numérique », *Culture & Musées*, 21-1 (2013), p. 131-152, DOI : 10.3406/pumus.2013.1735.
- FRANCEARCHIVES, *Collection des catalogues de vente d'autographes et livres anciens imprimés des libraires et des salles de vente*, URL : <https://francearchives.fr/fr/findingaid/2bd6f418cd252890f42e22c8215d45785c96a3fe> (visité le 13/04/2021).
- GALOIN (Alain), *Le Salon de la Rose-Croix*, fr, URL : <https://histoire-image.org/fr/etudes/salon-rose-croix> (visité le 30/08/2021).
- GONCERUT (Véronique), *Les catalogues d'exposition : gros plan sur ces ouvrages d'art devenus incontournables*, mars 2017, URL : <https://blog.mahgeneve.ch/les-catalogues-dexposition/> (visité le 27/08/2021).

- HOUSSAIS (Laurent) et LAGRANGE (Marion), *Marché(s) de l'art en province 1870-1914*, Presses universitaires de Bordeaux, Bordeaux, 2010 (Les Cahiers du Centre François Georges Piset, 8).
- JOYEUX-PRUNEL (Béatrice) et MARCEL (Olivier), « Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History », 4-2 (2015), p. 26.
- LEINMAN (Colette), « Le catalogue d'art contemporain », *Marges. Revue d'art contemporain*-12 (avr. 2011), p. 51-63, DOI : 10.4000/marges.408.
- *Les catalogues d'expositions surréalistes à Paris entre 1924 et 1939*, 2015, URL : <https://brill.com/view/title/31570> (visité le 27/08/2021).
- MIRANDA MENDOZA (Ileana), *L'économie du patrimoine écrit : le marché des autographes*, These de doctorat, Paris 1, 2010, URL : <http://www.theses.fr/2010PA010078> (visité le 27/08/2021).
- PARCOLLET (Remi) et SZACKA (Léa-Catherine), « Écrire l'histoire des expositions : réflexions sur la constitution d'un catalogue raisonné d'expositions », *Culture & Musées*, 22-1 (2013), p. 137-162, DOI : 10.3406/pumus.2013.1755.
- PARINET (Elisabeth), *Les ventes de livres et leurs catalogues, XVIIe-XXe siècle*, Publications de l'Ecole nationale des Chartes, Paris, 2018 (Etudes et rencontres).
- ROCHE (Mélanie), *En attendant "le jour [...] où il n'y aura plus de catalogue à faire" : une histoire matérielle des catalogues de bibliothèques (1789-1993)*, Mémoire d'études de Diplôme de Conservateur des Bibliothèques, Lyon, ENSSIB, 2014, URL : <https://www.enssib.fr/bibliotheque-numerique/documents/64118-en-attendant-le-jour-ou-il-n-y-aura-plus-de-catalogue-a-faire-une-histoire-materielle-des-catalogues-de-bibliotheque-1789-1993.pdf>.
- ROSENBERG (Pierre), « L'apport des expositions et de leurs catalogues à l'histoire de l'art », *Les Cahiers du Musée national d'art moderne*-29 (1989), p. 49-56.
- VAN BALBERGHE (Emile), « Les manuscrits et leur histoire. » *Scriptorium*, 40-1 (1986), p. 123-125, DOI : 10.3406/script.1986.1436.

Artl@s et Katabase

- CORBIÈRES (Caroline), *Du catalogue au fichier TEI : Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition*, dir. Thibault Clérice et Béatrice Joyeux-Prunel, mémoire de master Technologies numériques appliquées à l'histoire, Paris, Ecole nationale des Chartes, 2020.
- DOSSIN (Catherine), KONG (Nicole Ningning) et JOYEUX-PRUNEL (Béatrice), « Applying VGI to collaborative research in the humanities : the case of ARTL@S », *Cartography and Geographic Information Science*, 44-6 (nov. 2017), p. 521-538, DOI : 10.1080/15230406.2016.1216804.
- GABAY (Simon), RONDEAU DU NOYER (Lucie) et KHEMAKHEM (Mohamed), « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *IX Convegno AIUCD*, 2020, URL : <https://hal.archives-ouvertes.fr/hal-02388407> (visité le 13/04/2021).
- GABAY (Simon), PETKOVIC (Ljudmila), BARTZ (Alexandre), GILLE LEVENSON (Matthias) et RONDEAU DU NOYER (Lucie), « Katabase : À la recherche des manuscrits vendus », dans *Humanistica 2021*, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03066108> (visité le 13/04/2021).

- JOYEUX-PRUNEL (Béatrice), « Visual Contagions, the Art Historian, and the Digital Strategies to Work on Them », *Artl@s Bulletin*-8-3 (2019), URL : <https://docs.lib.psu.edu/artlas/vol8/iss3/8>.
- JOYEUX-PRUNEL (Béatrice), DOSSIN (Catherine) et SAINT-RAYMOND (Léa), *Artl@s*, URL : <https://artlas.huma-num.fr/fr/> (visité le 27/08/2021).
- KHEMAKHEM (Mohamed), *Grobid-dictionnaires*, 2020, URL : <https://github.com/MedKhem/grobid-dictionaries> (visité le 27/08/2021).
- RONDEAU DU NOYER (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay*, dir. Thibault Clérice et Simon Gabay, mémoire de master Technologies numériques appliquées à l'histoire, Paris, Ecole nationale des Chartes, 2019.
- TOPALOV (Barbara), GABAY (Simon), JOYEUX-PRUNEL (Béatrice), ROMARY (Laurent) et RONDEAU DU NOYER (Lucie), « Automating Artl@s - extracting data from exhibition catalogues », dans, 2020.

Humanités numériques

- ABAYNARH (Mohammed), FADILI (Hakim El) et ZENKOUAR (Lahbib), « Reconnaissance optique de documents amazighes : approches et évaluation des performances », *Etudes et Documents Berberes*, N° 34-1 (2015), p. 189-198, URL : <https://www.cairn.info/revue-etudes-et-documents-berberes-2015-1-page-189.htm> (visité le 27/08/2021).
- BIBLIOTHÈQUE DE L'INHA, *Bases de données sur le marché de l'art*, URL : <http://bibliotheque.inha.fr/iguana/www.main.cls?p=74469586-3948-11e2-a8f1-ac6f86effe00&v=2ca1bb8c-9a81-11ea-a5ae-5056b21d9100> (visité le 12/06/2021).
- BONHOMME (Marie-Laurence), *Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture*, Rapport exploratoire, Inria, 2018, p. 10.
- BURGY (Florence), GERSON (Steeve) et SCHÜPBACH (Loïc), *Ex imagine ad litteras : Projet d'océsiration de la collection De Bry*, mémoire de recherche réalisé dans le cadre du Master en Sciences de l'Information, Genève, Haute Ecole de Gestion, 2020, URL : https://doc.rero.ch/record/328465/files/BURGY_GERSON_SCHUPBACH_Projet_Recherche_Bodmer_Lab.pdf (visité le 27/08/2021).
- CAMPS (Jean-Baptiste) et PERREAUX (Nicolas), *Reconnaissance optique des caractères et des écritures manuscrites - Projet E-NDP*, févr. 2021, URL : https://outils.lamop.fr/lamop/mp3/E-Ndp/JBC-NP_e-NDP_OCR-et-HTR.pdf.
- CHAGUÉ (Alix) et CHIFFOLEAU (Floriane), *An accessible and transparent pipeline for publishing historical egodocuments*, mars 2021, URL : <https://hal.archives-ouvertes.fr/hal-03180669> (visité le 08/04/2021).
- CUADRA (Ruth) et MICHELS (Suzanne), *Publishing German Sales, A look under the Hood of the Getty Provenance Index*, avr. 2013, URL : <http://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index/> (visité le 27/08/2021).
- JOYEUX-PRUNEL (Béatrice), « Bases de données et gestion de projets en humanités numériques. Les dessous du projet Artl@s », *Biens Symboliques / Symbolic Goods. Revue de sciences sociales sur les arts, la culture et les idées*-2 (févr. 2018), DOI : 10.4000/bssg.242.

PUREN (Marie), *La numérisation du patrimoine. Du projet Gutenberg à Google Arts & Culture*, oct. 2020, URL : <https://hal.archives-ouvertes.fr/hal-03152774> (visité le 27/08/2021).

WEBER (Anne) et APAHAU, ASSOCIATION DES PROFESSEURS D'ARCHÉOLOGIE ET D'HISTOIRE DE L'ART DES UNIVERSITÉS, *Numérisation des catalogues de ventes d'oeuvres d'art de la Bibliothèque de l'INHA*, 2014, URL : <http://blog.apahau.org/numerisation-des-catalogues-de-ventes-doeuvres-dart-de-la-bibliotheque-de-linha/> (visité le 27/08/2021).

OCR

BHATT (Jwalin), HASHMI (Khurram Azeem), AFZAL (Muhammad Zeshan) et STRICKER (Didier), « A Survey of Graphical Page Object Detection with Deep Neural Networks », *Applied Sciences*, 11–12 (janv. 2021), DOI : [10.3390/app11125344](https://doi.org/10.3390/app11125344).

COPPI (Dalia), GRANA (Costantino) et CUCCHIARA (Rita), « Illustrations Segmentation in Digitized Documents Using Local Correlation Features », *Procedia Computer Science*, 38 (2014), p. 76-83, DOI : <https://doi.org/10.1016/j.procs.2014.10.014>.

DIEM (M.), KLEBER (F.), SABLATNIG (R.) et GATOS (B.), « cBAD : ICDAR2019 Competition on Baseline Detection », dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1494-1498, DOI : [10.1109/ICDAR.2019.00240](https://doi.org/10.1109/ICDAR.2019.00240).

GABAY (Simon), *Cours sur l'OCR et GROBID*, en, 2020, URL : https://github.com/gabays/Cours_2020_01_Strasbourg (visité le 27/08/2021).

KARPINSKI (Romain), LOHANI (Devashish) et BELAID (Abdel), « Metrics for Complete Evaluation of OCR Performance » (, juil. 2018), p. 8, URL : <https://hal.inria.fr/hal-01981731>.

KIESSLING (Benjamin), « Kraken - an Universal Text Recognizer for the Humanities », *Digital Humanities* (, 2019), URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html> (visité le 27/08/2021).

MITTAGESSEN, *kraken*, août 2021, URL : <https://github.com/mittagessen/kraken> (visité le 27/08/2021).

OCR4all, URL : <https://github.com/OCR4all> (visité le 15/04/2021).

PENGYUAN (Li), XIANGYING (Jiang) et HAGIT (Shatkay), « Figure and caption extraction from biomedical documents », *Bioinformatics*, 35–21 (2019), p. 4381-4388, DOI : <https://doi.org/10.1093/bioinformatics/btz228>.

REUL (Christian), GÖTTEL (Sebastian), SPRINGMANN (Uwe), WICK (Christoph), WÜRZNER (Kay-Michael) et PUPPE (Frank), « Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification : A Case Study on Daniel Sander's Wörterbuch der Deutschen Sprache », dans *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, Brussels Belgium, 2019, p. 33-38, DOI : [10.1145/3322905.3322910](https://doi.org/10.1145/3322905.3322910).

— « OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings », *Applied Sciences*, 9–22 (janv. 2019), p. 4853, DOI : [10.3390/app9224853](https://doi.org/10.3390/app9224853).

SCRIPTA, *eScriptorium*, URL : <https://gitlab.inria.fr/scripta/escriptorium> (visité le 15/04/2021).

SegmOnto, URL : <https://github.com/SegmOnto> (visité le 08/04/2021).

STOKES (Peter A.), *eScriptorium : un outil pour la transcription automatique des documents*, Billet, URL : <https://ephenum.hypotheses.org/1412> (visité le 14/04/2021). *tesseract*, avr. 2021, URL : <https://github.com/tesseract-ocr/tesseract> (visité le 15/04/2021).

Transkribus : AI powered Platform for Handwritten Text Recognition, URL : <https://readcoop.eu/transkribus/>.

VALENTINE (Greta), *Subject & Course Guides : Optical Character Recognition (OCR) - Getting Started : Other Tools*, URL : <https://guides.lib.ku.edu/ocr/other-tools> (visité le 15/04/2021).

ZRAMDINI (A.) et INGOLD (R.), « Optical font recognition using typographical features », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20–8 (1998), p. 877-882, DOI : 10.1109/34.709616.

Formats et Python

(W3C) (World Wide Web Consortium), *XSL Transformation documentation*, URL : <https://www.w3.org/TR/2017/REC-xslt-30-20170608/> (visité le 27/08/2021).

ALTO : Technical Metadata for Layout and Text Objects (Standards, Library of Congress), URL : <https://www.loc.gov/standards/alto/> (visité le 27/08/2021).

BELAÏD (Abdel), RANGONI (Yves) et FALK (Ingrid), *Représentation des données en XML pour l'analyse d'images de documents*, fr, text, URL : <http://lodel.irevues.inist.fr/cide/index.php?id=147> (visité le 27/08/2021).

BURNARD (Lou), *What is the Text Encoding Initiative ? : How to add intelligent markup to digital resources*, Marseille, 2014 (Encyclopédie numérique), URL : <http://books.openedition.org/oep/426> (visité le 27/08/2021).

NOUSIAINEN (Sami), *Report on File Formats for Hand-written Text Recognition (HTR) Material : CO :OP Community as Opportunity The Creative Archives' and Users' Network*, rapp. tech., National Archives of Finland, 2016, p. 69.

PAGE-XML, URL : <https://github.com/PRImA-Research-Lab/PAGE-XML> (visité le 27/08/2021).

TEI (Consortium), *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*, 2021, URL : <https://zenodo.org/record/5347789#.YTCYsI5KjIU> (visité le 01/09/2021).

Science Ouverte

DEDIEU (Laurence) et MARIE-FRANÇOISE (Fily), *Rendre publics ses jeux de données scientifiques*, rapp. tech., Montpellier, CIRAD, 2015, p. 6, DOI : 10.18167/COOPIST/0059.

JACQUEMIN (Bernard), SCHÖPFEL (Joachim) et FABRE (Renaud), « Libre accès et données de recherche. De l'utopie à l'idéal réaliste », *Études de communication*–52 (2019), DOI : <https://doi.org/10.4000/edc.8468>.

LUPOVICI (Catherine), « Le Digital Object Identifier : Le système du DOI », *Bulletin des Bibliothèques de France*–43-3 (1998), p. 49-54, URL : <https://bbf.enssib.fr/consulter/bbf-1998-03-0049-007>.

- UNESCO, *Vers une recommandation de l'UNESCO pour la Science ouverte*, 2019, URL : https://en.unesco.org/sites/default/files/open_science_brochure_fr.pdf (visité le 26/08/2021).
- URFIST MÉDITERRANÉE, *Les principes FAIR*, URL : <https://doranum.fr/enjeux-benefices/principes-fair/> (visité le 27/08/2021).
- VANHOLSBEECK (Marc), « La notion de Science Ouverte dans l'Espace européen de la recherche », *Revue française des sciences de l'information et de la communication*–11 (2017), DOI : <https://doi.org/10.4000/rfsic.3241>.
- WILKINSON (Mark D.), DUMONTIER (Michel), AALBERSBERG (IJsbrand Jan), APPLETON (Gabrielle), AXTON (Myles), BAAK (Arie), BLOMBERG (Niklas), BOITEN (Jan-Willem), SILVA SANTOS (Luiz Bonino da), BOURNE (Philip E.), *et al.*, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data*, 3–1 (mars 2016), DOI : [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

Acronyms

ALTO Analysed Layout and Text Object.

CSV Comma Separated Values.

DOI Digital Object Identifier.

FAIR Findable Accessible Interoperable Reusable.

GROBID GeneRation Of BIbliographic Data.

HTR Handwritten Character Recognition.

IIF International Image Interoperability Framework.

NER Named Entity Recognition.

OCR Optical Character Recognition.

ODD One Document Does it all.

PAGE Page Analysis and Ground truth Elements.

PDF Portable Document Format.

TEI Text Encoding Initiative.

XML eXtensible Markup Language.

XSLT eXtensible Stylesheet Language Transformations.

Introduction

Au mois de juillet dernier, le ministère de la recherche scientifique et de l'innovation a lancé le deuxième Plan national pour la Science Ouverte pour la période 2021-2024. Celui-ci confirme et accentue des travaux, déjà entamés par un premier plan datant de 2018, qui visent à développer une science plus transparente et accessible. Ces actions se structurent en plusieurs axes : généraliser l'accès ouvert et gratuit aux publications scientifiques, développer le partage des données de la recherche et promouvoir l'ouverture libre aux protocoles produits. Tout cela a pour ambition de mettre les principes de la Science Ouverte au cœur du travail des chercheurs et s'inscrit dans un mouvement plus large de développement de ces pratiques à l'échelle européenne et internationale¹.

La Science Ouverte est un « mouvement qui vise à rendre la science plus ouverte, plus accessible, plus efficace, plus démocratique et plus transparente² ». Plus largement, elle a pour volonté de permettre à tous d'accéder au produit ainsi qu'aux méthodes de la recherche. Souhait relativement ancien de la communauté scientifique, qui a pu se matérialiser par l'apparition des revues académiques au cours du XVIII^{ème} et XIX^{ème} siècle, l'idée de partage de la science s'est renforcé avec l'apparition d'internet. Les années 1990 sont en effet un terreau fertile pour l'établissement de tentatives de mise en accès libre d'archives et de périodiques. À l'origine porté par la communauté scientifique, cette idée est alors récupérée par les décideurs politiques. Tout d'abord concentrée sur le développement d'une édition scientifique libre d'accès et gratuite, l'*Open Access*, dans les années 1990, elle s'élargit peu à peu en prenant en compte les autres éléments qui composent la recherche : les données scientifiques, documents autres que les publications scientifiques qui sont produit au cours de la recherche ou encore les protocoles de production des résultats. La Science Ouverte permet, dans la théorie, plusieurs choses³ :

- accélérer les découvertes scientifiques en partageant l'information
- encourager la collaboration entre équipes
- assurer l'intégrité des données et la reproductibilité des méthodologies
- ouvrir des nouveaux champs d'analyse auxquels n'aurait pas pensé le producteur

Dans la pratique, la Science Ouverte réforme la manière de travailler des chercheurs. En effet, pour parvenir à une certaine transparence de la recherche, il est primordiale de se plier à certains standards, autant au niveau de la structure des données de la recherche que pour les protocoles qui ont conduit à leur production. On réalise alors une transformation complète de faire de la recherche dans le but de rendre accessible de façon libre, gratuite, et donc ouvertes, méthodes, données et publications. Il est donc intéressant de

1. Pour plus d'information sur le Plan national pour la Science Ouverte, voir le site internet correspondant : https://www.ouvrirla.science.fr/category/science_ouverte/

2. Définition donnée par l'UNESCO de la Science Ouverte ou *Open Science* en anglais, https://en.unesco.org/sites/default/files/open_science_brochure_fr.pdf

3. Laurence Dedieu et Fily Marie-Françoise, *Rendre publics ses jeux de données scientifiques*, rapp. tech., Montpellier, CIRAD, 2015, p. 6, DOI : 10.18167/COOPIST/0059.

se demander : en quoi cette restructuration du fonctionnement de la recherche est utile ? Autrement dit, pourquoi faire de la Science Ouverte ?

Artl@s, lancé par Béatrice Joyeux-Prunel, Catherine Dossin et Léa Saint-Raymond, fait partie de ces projets de recherche ayant à cœur le partage libre, gratuit et ouvert des données, protocoles et résultats. Résolument ancré dans la Science Ouverte, il se structure autour de plusieurs grands axes en histoire de l'art dont le principal est Basart. Cette base de données, « libre et accessible à tous⁴ », a pour ambition de rassembler et diffuser des catalogues d'exposition, sources dispersées ou absentes et de les valoriser par le biais des outils et méthodes d'humanités numériques. Mise en ligne courant 2018, elle permet l'interrogation de données, à une échelle globale et sur plus de deux siècles, issues d'une source principale en histoire de l'art. En effet, le catalogue d'exposition, en listant les artistes exposants et les œuvres exposées pour un événement donné, fournit de multiples informations au chercheur. Le projet s'articule autour de deux axes principaux. Dans un premier temps, il s'agit de se détacher des monographies et des grandes figures de l'histoire de l'art, en abordant une approche plus quantitative, à l'instar de l'École des Annales des années 1960, par la visualisation de cette profusion de données. Dans un second temps, le projet adopte une approche dé-coloniale. En traitant des catalogues provenant non seulement de la France, de l'Allemagne et des États-Unis, mais de manière plus large de l'Europe et de l'Amérique, notamment latine, ainsi que de l'Asie, de l'Afrique, du Moyen Orient et de l'Australie, Artl@s contribue à la production d'un récit alternatif de l'histoire des expositions du XIX^e et XX^e siècle. Celui-ci se détache de la rhétorique d'un axe culturel Paris-New-York en histoire de l'art à cette période et met en valeur le développement artistique des régions dites périphériques⁵.

Jusqu'en 2020, les catalogues d'exposition étaient retrançerts à la main par les membres d'Artl@s afin de les intégrer par la suite dans la base de données. Béatrice Joyeux-Prunel a alors souhaité automatiser ce travail, dans l'idée de faire grossir plus rapidement la base de données tout en réduisant ce travail fastidieux et chronophage. En effet, les catalogues d'exposition ont une structure assez figée et normalisée qui mène naturellement à penser à une extraction automatique de ces données. Pour ce faire, la chercheuse s'est entourée d'une équipe pluri-institutionnelle, issue d'autres projets de récupération de données provenant de sources textuelles historiques normalisées et structurées. Ainsi, Caroline Corbières⁶ a élaboré, dans le cadre de son stage de fin d'études, une chaîne de traitement des catalogues d'exposition d'Artl@s allant de l'image à un fichier structuré XML-TEI et un CSV présentant l'information contenue. Ce travail a été réalisé en partenariat avec le groupe de travail ALMAnaCH, de l'INRIA⁷, qui s'intéresse aux minutiers de notaires des Archives Nationales, ainsi que le projet Katabase qui étudie les catalogues de ventes de manuscrits parisiens du XIX^e siècle et est dirigé par Simon Gabay, Maître-assistant en Humanités Numériques à l'Université de Genève. Il s'est basé sur le travail de Lucie

4. Description de Basart sur son site internet : <https://artlas.huma-num.fr/fr/bases-en-acces-libre/>, consulté le 26-08-2021

5. Béatrice Joyeux-Prunel, « Bases de données et gestion de projets en humanités numériques. Les dessous du projet Artl@s », *Biens Symboliques / Symbolic Goods. Revue de sciences sociales sur les arts, la culture et les idées*-2 (févr. 2018), DOI : 10.4000/bssg.242.

6. Caroline Corbières, *Du catalogue au fichier TEI : Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition*, dir. Thibault Clérice et Béatrice Joyeux-Prunel, mémoire de master Technologies numériques appliquées à l'histoire, Paris, Ecole nationale des Chartes, 2020.

7. « ALMAnaCH », INRIA, <https://www.inria.fr/fr/almanach> (visité le 27-06-21)

Rondeau du Noyer, qui a mené une réflexion autour de l'encodage automatique en TEI des catalogues de ventes de manuscrits d'editiones au cours de son stage de fin d'études⁸.

Par conséquent, mon stage de fin d'études au sein d'Artl@s, qui clôture ma deuxième année de master Technologies numériques appliquées à l'Histoire à l'École nationale des Chartes, intervient à un stade avancé du travail d'automatisation des catalogues d'exposition. Ce stage, se déroulant sur quatre mois sous la direction de Béatrice Joyeux-Prunel et la supervision de Simon Gabay, avait pour principal objectif d'améliorer une chaîne de traitement déjà en place et assez fonctionnelle. Mise en place par Caroline Corbières, elle OCRise des catalogues préalablement numérisés avec Transkribus puis récupère et structure les données obtenues en XML-TEI avec GROBID.

L'OCRisation des catalogues, c'est-à-dire la récupération des données textuelles issues des images numérisées, est réalisée via l'outil Transkribus⁹. Ce moteur de reconnaissance de textes manuscrits¹⁰, créé en 2015 par l'Université d'Innsbruck en Autriche est devenu payant en 2020. Or, Béatrice Joyeux-Prunel souhaite une chaîne de traitement qui ne nécessite pas de logiciels payants, afin que celle-ci puisse être reproductible. La structuration de ces données est par la suite réalisée par GROBID¹¹. Il s'agit d'un logiciel de *machine learning* développé par l'INRIA permettant d'encoder des données textuelles structurées, de type dictionnaires. Cependant, la gestion de cet outil semble compromise. En effet, le développeur de GROBID, Mohamed Khemakhem, a terminé sa thèse et il est assez difficile de trouver des personnes ayant ses compétences qui acceptent de travailler pour la recherche française.

Ainsi, au cours de mon stage, je me suis intéressée dans un premier temps à l'amélioration de l'OCRisation des catalogues, c'est à dire à la récupération de l'information textuelle contenue dans les catalogues. Pour ce faire, j'ai organisé la migration des jeux de données d'entraînement des modèles d'OCRisation depuis Transkribus vers un autre outil de transcription, eScriptorium. Par la suite, j'ai travaillé à l'élaboration d'un système de scripts Python permettant de récupérer les données textuelles des catalogues issues de l'OCR et de les structurer afin de les rendre exploitable.

Ce mémoire présente donc plus largement une réflexion sur l'intérêt de la Science Ouverte dans le cadre de la mise en place de cette nouvelle chaîne de traitement de catalogues. Quels sont les apports réels de la restructuration complète d'une chaîne de traitement, pourtant fonctionnelle, dans le but de la rendre intégralement ouverte, libre et gratuite ? Il décrit les différentes briques qui construisent la chaîne et leur élaboration et développe certains points techniques et le travail de circonspection qui l'entoure.

Dans un premier temps, nous nous attacherons à définir nos jeux de données ainsi qu'à présenter un état de l'art de leur traitement dans le cadre de la recherche en Humanités Numériques, notamment en nous attardant sur le travail déjà réalisé au sein d'Artl@s. Puis, nous aborderons le travail effectué au cours de ce stage autour de l'amélioration de l'OCR. Enfin, nous allons nous intéresser à l'élaboration du système d'extraction et de structuration des données textuelles.

8. Lucie Rondeau Du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay*, dir. Thibault Clérice et Simon Gabay, mémoire de master Technologies numériques appliquées à l'histoire, Paris, Ecole nationale des Chartes, 2019.

9. Accessible ici : <https://readcoop.eu/transkribus/>

10. Handwritten Text Recognition ou HTR

11. Disponible ici : <https://grobid.readthedocs.io/>

Première partie

La numérisation des catalogues : problèmes et enjeux

Chapitre 1

Le catalogue : une source primaire pour les historiens

Il convient dans un premier temps de présenter les catalogues en tant qu'objets, afin de saisir tout leur intérêt et le pourquoi de leur utilisation. Plus largement, ceux-ci appartiennent au genre des documents semi-structurées, organisés sous forme de listes. Cet type d'agencement est employé pour la plupart des données du savoir humain : dictionnaire, inventaires, catalogues... Ainsi, si notre réflexion se concentre sur les catalogues en particulier, et notamment ceux d'exposition et de ventes de manuscrits, elle s'inscrit dans un champ de réflexion plus large concernant un ensemble de documents varié et d'un intérêt scientifique important.

1.1 Définition globale

1.1.1 Histoire du catalogue papier

Un catalogue est une liste d'items inventoriés et organisés¹. Ainsi, il s'agit d'un document multifonctions, dont la forme s'adapte pour des usages et utilisateurs différents : bibliothèques, musées, commerces.... Cette adaptabilité du catalogue nous constraint donc dans un premier temps à dresser un portrait global de son utilité au cours du temps, avant d'avoir un aperçu de sa versatilité.

Les catalogues, en tant que listes descriptives et ordonnées, apparaissent dès l'invention de l'écriture, puisque sa vocation originelle est d'inventorier des éléments. Dans le milieu de la culture, on peut noter les catalogues thématiques de Ninive, qui sont les plus anciens portés à notre connaissance. Ces tablettes d'argiles s'attachent à décrire et organiser la collection de la Bibliothèque royale d'Assurbanipal, dernier roi d'Assyrie, de 668 à 627 avant Jésus-Christ. Autre catalogue antique, les *Pinakes* de la bibliothèque d'Alexandrie, rédigés par Callimaque de Cyrène (305-240 avant Jésus-Christ). Composés de cent-vingt rouleaux, chaque livre de la bibliothèque royale d'Alexandrie y est répertorié et classé par sujet, puis par auteur. Ils sont considérés par certains spécialistes comme une forme embryonnaire du catalogue moderne².

1. SORDET, « Pour une histoire des catalogues de livres : matérialités, formes, usages », In : Frédéric Barbier, Thierry Dubois, Yann Sordet et Gabriel de Broglie, *De l'argile au nuage : une archéologie des catalogues (IIe millénaire av. J.-C. - XXIe siècle)*[ouvrage publié à l'occasion des expositions organisées par la Bibliothèque Mazarine & la Bibliothèque de Genève, Paris 13 mars - 15 mai 2015, Genève 18 septembre - 21 novembre 2015], Editions des Cendres, 2015 (Bibliothèques), p. 15-47

2. Mélanie Roche, *En attendant " le jour [...] où il n'y aura plus de catalogue à faire " : une histoire*

Au Moyen-Âge, le catalogue est un manuscrit répertoriant les éléments d'une collection, essentiellement les livres d'une bibliothèque, en suivant des codes précis de classement. Au XV^{ème} siècle, la production des catalogues s'adapte rapidement à l'apparition de l'imprimerie. En effet, il est possible de retrouver des catalogues imprimés, répertoriant notamment les livres imprimés par les libraires-imprimeurs dès les débuts de l'imprimerie. Cette nouveauté induit le développement d'une certaine uniformité dans la structure des documents ainsi que l'affirmation de codes précis de catalogage. Les catalogues sont alors des « catalogues inventaires », c'est-à-dire que chaque élément de la collection décrite correspond à une notice. Concernant l'époque moderne, l'historiographie considère le catalogue de la bibliothèque Bodléienne d'Oxford, publié en 1605, comme le premier catalogue dictionnaire. En effet, le document ne se contente pas de décrire chaque livre par une notice, il permet également de naviguer dans la collection par plusieurs points d'accès, tels que des thèmes ou des auteurs...³

Le XIX^{ème} siècle marque la consécration du catalogue dictionnaire comme instrument de recherches dans les bibliothèques et par là même le déclin des catalogues inventaires. En effet, les utilisateurs trouvent plus facilement l'objet de leur recherche lorsque celui-ci est décrit par plusieurs éléments et non pas par une unique notice. C'est à cette période que fleurissent de nombreuses initiatives de catalogage des collections complètes des musées, bibliothèques, etc, à l'instar du catalogue de la Bibliothèque nationale de France, impulsé par Prosper Mérimée en 1858, mis en œuvre par Léopold Delisle à partir de 1875 et publié de 1897 à 1981.

Ainsi, si la définition de ce terme est déjà élaborée dans le monde des bibliothèques, son usage se démocratise dans une multitude de secteurs. Il est alors utilisé en science, pour désigner des livres contenant des listes tabulées répertoriant des informations précises sur des objets astronomiques, ou encore dans le commerce. En effet, les premiers catalogues commerciaux apparaissent au XIX^{ème} siècle dans le cadre du développement de la vente par correspondance dans les grands magasins. Dans le monde de l'art, le mot est également utilisé avec l'épithète raisonné, pour décrire un inventaire mentionnant toutes les œuvres d'un artiste ou en tant que « catalogue d'exposition », répertoriant toutes les œuvres exposées dans une exposition. C'est donc le contexte d'utilisation du mot catalogue qui déterminera la fonction précise de ce document multi-facette⁴.

1.1.2 Étymologie

Le mot « catalogue » est emprunté au bas latin *catalogus*, énumération, liste, lui-même issus du grec *κατάλογος*. Le mot *κατά* signifie de haut en bas, tandis que le mot *λογος* provient de *legein*, rassembler, apparenté à *legere*, lire. Il s'agit donc d'une « liste établie dans un ordre donné, de noms de personnes ou de choses formant une collection »⁵. Le terme est utilisé, au IV^{ème} avant Jésus-Christ, pour désigner le registre militaire répertoriant les hommes disponibles. On le retrouve également un peu plus tard, chez Platon,

matérielle des catalogues de bibliothèques (1789-1993), Mémoire d'études de Diplôme de Conservateur des Bibliothèques, Lyon, ENSSIB, 2014, URL : <https://www.enssib.fr/bibliothèque-numérique/documents/64118-en-attendant-le-jour-ou-il-n-y-aura-plus-de-catalogue-a-faire-une-histoire-materielle-des-catalogues-de-bibliothèque-1789-1993.pdf>.

3. Alain Chante, « La notion de catalogue : de l'imprimé au numérique », *Culture & Musées*, 21-1 (2013), p. 131-152, DOI : 10.3406/pumus.2013.1735.

4. Colette Leinman, « Le catalogue d'art contemporain », *Marges. Revue d'art contemporain*-12 (avr. 2011), p. 51-63, DOI : 10.4000/marges.408.

5. Centre national de Ressources textuelles et lexicales

pour désigner une liste de citoyens reconnus aptes à exercer la magistrature⁶.

Il est, par la suite, employé ponctuellement au cours du Moyen Âge pour décrire des listes de livres. À cette période, les mots « inventaire », « registre » ou encore, tout simplement, « liste », lui sont préférés pour nommer ces objets décrivant le contenu, item par item, d'une collection.

Avec la normalisation de la structure de ces documents suite à l'utilisation de l'imprimerie, on assiste une réduction des termes employés pour les désigner, qui se resserre autour d' « index », de « bibliothèque », de « bibliographie » et de « catalogue ». La définition respective de ces termes au sens où on l'entend aujourd'hui intervient au XIX^{ème} siècle. Dès lors, un catalogue désigne, pour les bibliothèques, la liste descriptive des documents qu'elle contient ainsi que le livre qui contient cette liste. Ce nom est par la suite réemployé en dehors de ce domaine pour décrire tout livre présentant une liste organisée énumérant des items.

1.1.3 De l'utilisation des catalogues

Ainsi, décrire le catalogue en temps qu'objet, c'est saisir la multitude de ses usages et comprendre la similitude de ses objectifs. On peut classifier ceux-ci en plusieurs points, communs à tous les types de catalogues⁷.

Tout d'abord, le catalogue exprime une volonté de regroupement, puisqu'il assemble des éléments ensemble. Il évoque également l'idée d'une série repérable, puisqu'il s'agit de la description d'objets précis d'une collection dans le but de les retrouver. Ensuite, ce document transmet une idée de dénombrement, c'est-à-dire qu'il énumère ces mêmes éléments, exprimant une volonté d'en faire le compte. Ceci est directement en lien avec la description du catalogue en temps qu'inventaire exhaustif d'un groupe d'objets. Il permet également de catégoriser ces objets. Il démontre donc un besoin de nommer et classifier des données, en leur collant des étiquettes précises.

À ces différents points, inhérents à la structure de type liste de ces documents, s'ajoutent des idées plus abstraites. En effet, une certaine neutralité se dégage des catalogues, de part leur aspect d'inventaire, ce qui induit un document objectif est s'attachant principalement aux faits. Sans que ce danger soit totalement écarté, la structure du catalogue préserve grandement son contenu de toute subjectivité, et en fait un outil de travail de choix pour les chercheurs. Enfin, le catalogue traduit une idée de monstration. Ce terme signifie l'action, le fait de montrer quelque chose, soit ici l'acte d'exposer sur le papier les éléments composants un événement ou un lieu. En décrivant tous les éléments d'une collection, que ce soit une bibliothèque, un musée, une exposition ou encore une vente, le catalogue est un témoin, voire le seul objet écrit permettant de reconstruire l'événement. À partir du milieu du XIX^{ème} siècle, il est de plus en plus considéré en tant qu'outil au service de l'image de l'événement. Cette idée est particulièrement visible dans le cadre du développement des catalogues de ventes de commerce, qui émergent pendant cette période et ont un rôle double de description des contenus à vendre et « présentation de soi » et de l'image de l'entreprise⁸.

6. Leonhard (Basle) Burckhardt, *Katalogos*, en, oct. 2006, URL : <https://referenceworks.brillonline.com/entries/brill-s-new-pauly/katalogos-e610620> (visité le 27/08/2021).

7. Ces idées sont catégorisées et décrites de façon plus exhaustive dans l'article d'Alain CHANTE, dont j'ai repris les principales idées.

8. A. Chante, « La notion de catalogue... ».

1.2 Le catalogue d'exposition

1.2.1 Un objet au centre d'un évènement : l'exposition

Traditionnellement, l'historiographie considère la première exposition, de son nom « L'Exposition », comme ayant eu lieu en 1667. Cet évènement fait suite à une décision de 1663 d'organiser une présentation des œuvres des membres de l'Académie royale de peinture et de sculpture⁹, tous les trois ans, en l'honneur de Louis XIV¹⁰. Cette volonté de présenter les productions royales s'inscrit dans un mouvement plus large d'ouverture au public des collections historiques et artistiques privées au cours du XVII^{ème} siècle en Europe qui est matérialisé par la création de musées, comme le Louvre en France ou encore la Gallerie Uffizi en Italie.¹¹ L'ancêtre du catalogue d'exposition apparaît donc à la suite de cette première exposition et est édité pour la première fois en 1673 par l'Académie royale de peinture et de sculpture, sous-division de l'Académie royale des Beaux-Arts. Ce premier catalogue est appelé livret, un nom qui perdure encore aujourd'hui - bien que vieilli - pour définir les catalogues d'exposition¹².

Ces expositions continuent d'avoir lieu régulièrement au Palais-Royal, en semi-public, puis au Louvre à partir de 1692. Le nom de ces manifestations, « Salon », provient du Salon Carré du Louvre, où elles ont lieu de 1725 à 1848. Au milieu du XVIII^{ème} siècle, le pouvoir politique réalise les possibilités économiques qu'offre le Salon et l'exposition de ces œuvres, face à un marché de l'art dans un état déplorable. Tournehem, surintendant des Bâtiments de 1747 à 1751, réorganise ces manifestations, qui deviennent bisannuelles et accessibles à un public plus large. L'idée est alors de susciter un goût de l'art chez la population qui permettrait de développer des échanges économiques ayant pour vitrine principale l'exposition. En plus de permettre aux peintres d'assurer leurs ventes, Tournehem réduit également le nombre d'œuvres présentées, donnant ainsi une illusion de rareté, permettant d'augmenter les prix. Ces initiatives permettent le développement du marché de l'art parisien dès le milieu des années 1750¹³. Celui-ci est rapidement concurrencé par le marché de l'art de Londres, qui se développe dans les années 1760, à l'instar de l'exposition de la *Royal Academy* de Londres, créée en 1769, ou de celle de la *British Institution* en 1805. Autour de ces deux pôles se cristallise une vision dite « académique » de l'art, du nom de l' « Académie des Beaux-Arts » mêlant néo-classicisme¹⁴ et romantisme¹⁵.

Face au succès grandissant du Salon, place centrale de la vie artistique du pays, de nombreuses expositions fleurissent au cours du XIX^{ème} siècle à l'initiative de Musées,

9. Celle-ci, fondé en 1648, sous le règne de Louis XIV et avec l'impulsion de Mazarin, ambitionne, sur le modèle italien, de former et rassembler les artistes du royaume.

10. « Qu'est-ce qu'un catalogue d'exposition ? » In : C. Leinman, *Les catalogues d'expositions surréalistes à Paris entre 1924 et 1939*, 2015, URL : <https://brill.com/view/title/31570> (visité le 27/08/2021), p. 25-60

11. Caterina Albano, *Exhibition*, dir. Oxford University, Oxford et New York, URL : <https://ualresearchonline.arts.ac.uk/id/eprint/8058/>.

12. C. Leinman, « Le catalogue d'art contemporain »...

13. JOLLET, « Il gagne de l'argent : L'artiste et l'argent au XVIII^{ème} siècle », In : Laurence Bertrand Dorléac, *Le Commerce de l'art : de la renaissance à nos jours*, Editions La Manufacture, Besançon, 1992, p. 167-178

14. Courant pictural de la fin du XVIII^{ème} et du début du XIX^{ème} siècle qui se caractérise par des tons sombres, des sujets issus de l'histoire antique et de la mythologie grecque et romaine et d'une technique qui ne laisse pas apparaître les coups de brosse.

15. Courant pictural évoluant de 1770 à 1870 dont les sujets jouent avec les impressions et sentiments (pour aller plus loin : http://ww2.ac-poitiers.fr/dsden17-pedagogie/sites/dsden17-pedagogie/IMG/pdf/35_-_Le_romantisme-2.pdf).

Académies ou Sociétés des Amis des Beaux-Arts locaux. C'est le cas de la *Société des Amis des Arts de Strasbourg*, fondée par des amateurs d'art de la ville en 1832. Celle-ci prospère tout le long du XIX^{ème} siècle et organise de nombreuses expositions, issues des collections privées de sociétaires, d'artistes locaux ou de participants récurrents non régionaux. Stimulant un marché artificiel qu'elle abreuve de peintures que ses propres membres achètent, elle possède le monopole du marché de l'art de la région jusqu'en 1909. Le monopole artistique du Salon Parisien est définitivement brisé en 1880, alors que l'État se retire de son organisation. Le « Salon des artistes français », géré par la Société des Artistes français, succède donc au Salon de l'Académie des Beaux-Arts. Parallèlement, on assiste à une éclosion de nouveaux salons, sous l'égide d'autres sociétés d'artistes, ainsi qu'à l'implantation des Beaux-Arts dans les Expositions Universelles¹⁶.

Ces grandes manifestations, qui permettent aux États de présenter leurs avancées techniques et scientifiques aux autres pays, jouent un rôle important dans le développement d'un marché de l'art global. En effet, si la peinture n'était pas présente dans la toute première *Great Exhibition* de Londres de 1851, les œuvres d'art trouvent rapidement leur place au sein de ces expositions, permettant de donner une façade internationale aux productions exposées par les différents pays présents. Cette idée se retrouve dans l'émergence des peintures françaises au sein du marché de l'art américain. Pendant toute la première partie du XIX^{ème} siècle, les expositions américaines s'intéressent principalement aux peintres nationaux, à l'exception de quelques manifestations réalisées notamment par Joseph Bonaparte, réfugié à New-York avec sa collection de peintures du XVI^{ème} et XVII^{ème} siècles, dans les années 1820. Les collectionneurs et mécènes américains interviennent donc assez tardivement sur le marché de l'art européen, avec la première participation officielle de l'art américain à l'Exposition Universelle de Paris de 1867. Les peintures présentées, prêtées par des collectionneurs américains, sont le prétexte d'un voyage à Paris, qui leur permet de découvrir le Salon Parisien. Se développe alors au cours des années 1870 toute une série d'échanges artistiques entre l'Europe, principalement Paris et Londres, et New York, qui aboutit, au début du XX^{ème} siècle, à l'organisation d'un marché de l'art en trois capitales¹⁷.

Autre impact de la déliquescence du monopole du Salon Parisien, les galeries privées d'art se développent à partir de 1890-1900. Elles agissent, de la même façon que le Salon auparavant, comme intermédiaires entre l'artiste et le public. Dans ce but, elles sont à l'origine de nombreuses expositions, notamment au cours de l'entre-deux-guerres où elles jouent un rôle prépondérant dans le marché de l'art global. Ce rôle est par la suite repris par les Biennales et les Foires artistiques, à partir de 1945.

La première Biennale - manifestation artistique ayant lieu tout les deux ans - à Venise, est créée en 1895, à l'instigation du conseil municipal de la ville. Considérée comme l'archétype des Biennales, elle se compose de pavillons nationaux, dans lesquels chaque pays exhibe ses représentants artistiques. Construite comme une exposition universelle de l'art contemporain, elle est également la première d'une longue série, qui se poursuit avec la Biennale de São Paulo, créée en 1951, ou encore celle de La Havane, en 1984. Ces deux événements se posent à contre-courant de l'axe artistique Europe-États-Unis. Si l'ambition première de la Biennale de São Paulo était de faire découvrir les artistes occidentaux au Brésil, celle de La Havane a pour volonté de présenter les peintres dits du « Sud » et

16. Laurent Houssais et Marion Lagrange, *Marché(s) de l'art en province 1870-1914*, Presses universitaires de Bordeaux, Bordeaux, 2010 (Les Cahiers du Centre François Georges Pariet, 8).

17. FIDELL-BEAUFORT, « Le marché américain », In : L. Bertrand Dorléac, *Le Commerce de l'art : de la renaissance à nos jours...*, p. 155-173

de véritablement se dresser comme manifestation majeure pour les pays délaissés par le discours artistique traditionnel. Cette idée permet également la création, à partir des années 1990, de près d'une centaine de Biennales partout dans le monde et l'émergence d'un modèle d'exposition moins européenocentré et tourné vers les artistes des pays dits émergents. Le développement des zones marginales de la production artistique entraîne ainsi la progressive disparition de la structure pavillonnaire des expositions, dans le but de mettre en avant ces nouveaux artistes¹⁸.

Au tournant du XX^{ème} siècle, face à cette restructuration du marché de l'art autour de l'exposition temporaire, les musées, jusqu'alors présentant uniquement des collections permanentes, se joignent au mouvement. Ces nouvelles expositions sont favorisées par le développement des prêts d'œuvres entre musées, qui deviennent fréquents à partir de 1918. C'est l'invention des expositions temporaires, qui peuvent prendre plusieurs formes¹⁹ :

- Des expositions monographiques, à propos d'un artiste en particulier. On considère la première exposition monographique comme étant celle à propos de Courbet, suite à sa mort en 1882.
- Des expositions synthèse, ou thématiques, réunion d'œuvres autour d'un thème précis. La première d'entre elles, *Peintres français* a lieu en 1904.
- Des expositions dossiers : petites expositions à but didactique décrivant une œuvre précise
- Des expositions scientifiques : présentent techniques et restaurations

En conclusion, les expositions, manifestations éphémères présentant une réunion d'œuvres d'art à un public, sont un objet mouvant, à la frontière entre les musées et le marché de l'art. L'histoire des expositions reste marquée par une historiographie balbutiante, cet évènement n'ayant commencé à être étudié réellement qu'à partir de 2000, malgré quelques ouvrages de littératures allemandes et anglo-saxonnes, dans les années 1960 et 1990, qui ont pu avoir pour ambition de développer ce champ de recherche. Si des programmes, à l'instar du « Catalogue raisonné des expositions du Centre Pompidou²⁰ » ou encore des études précises sur des expositions en particulier, ont été lancés dans les années 2010, les résultats restent encore lacunaires, face à l'immensité de la tâche. Ainsi, il est assez difficile, pour le moment, de sortir du schéma traditionnel des Salons Parisiens du XIX^{ème} siècle et d'une production artistique réservée à l'axe New-York - Londres - Paris au XX^{ème} siècle²¹. Le travail réalisé par Artl@s sur les catalogues d'exposition de part le monde sur cette période n'en ressort que plus important, avec pour objectif de construire une nouvelle vision globale du marché de l'art moderne du XIX^{ème} et XX^{ème} siècle.

18. Sylvie Castets, « Les biennales internationales d'art contemporain et leurs touristes-amateurs, vus sous l'angle de l'utopie », *Études caribéennes*–37-38 (oct. 2017), DOI : 10.4000/etudescaribeenenes.11291.

19. Pierre Rosenberg, « L'apport des expositions et de leurs catalogues à l'histoire de l'art », *Les Cahiers du Musée national d'art moderne*–29 (1989), p. 49-56.

20. Le carnet Hypothèses du projet, répertoriant les avancées et travaux, est disponible ici : <https://histoiredesexpos.hypotheses.org/>

21. Remi Parcollet et Léa-Catherine Szacka, « Écrire l'histoire des expositions : réflexions sur la constitution d'un catalogue raisonné d'expositions », *Culture & Musées*, 22–1 (2013), p. 137-162, DOI : 10.3406/pumus.2013.1755.

1.2.2 Le catalogue d'exposition : une structure normalisée dans le temps et l'espace

On peut décrire le catalogue d'exposition comme le complément documentaire et la seule archive visible d'une présentation éphémère d'œuvres. La structure de l'exposition est rarement rapportée sur le papier. Le catalogue s'attache plutôt à reproduire une liste fidèle des œuvres exposées au cours de l'évènement, chacune étant accompagnée de notices précises et développées²². Ce document, obéissant à un schéma précis de structuration d'informations factuelles et objectives, tient rapidement lieu d'outil de repérage et de support. Il permet au visiteur, et potentiel acheteur, de naviguer au sein de l'exposition tout en prenant des notes sur les œuvres. En effet, à l'instar des Salons Parisiens du XVIII^{ème} siècle, la plupart des expositions présentent des œuvres à vendre, et le catalogue joue donc un rôle important dans cet aspect commercial de l'évènement²³.

Le terme de livret met en exergue le catalogue d'exposition en tant qu'objet. En effet, il s'agit d'un livre, le plus souvent imprimé, décrivant un évènement précis, une exposition. Y sont donc indiqué le titre de l'exposition, la date et le lieu de l'évènement, mais également une liste des exposants et des œuvres exposées. Il est aussi possible d'y retrouver, pour les exposants, leur nationalité, leur lieu de naissance, leur adresse ou même leur maître, et, pour les œuvres, leur titre, leur médium, leurs dimensions, la description de leur sujet, leur prix, leurs acheteurs...²⁴ La structure de ces informations est susceptible de varier en fonction de l'exposition qu'elle décrit. En effet, différents types d'expositions existent, et la variation de leur forme influe sur l'aspect du catalogue. Ainsi, dans le cas d'une rétrospective, sur un artiste en particulier, le catalogue d'exposition contient une liste numérotée des œuvres présentées. Lorsque plusieurs artistes sont présents, à l'instar des Salons, le catalogue peut être structuré sous la forme d'une liste alphabétique des auteurs, pour lequel chacune des œuvres exposées sera mentionnée.

L'apparition systématique des catalogues décrivant les expositions va donc de pair avec un développement du marché de l'art, qui dépasse rapidement les frontières. Comme les éléments présentés au sein de ceux-ci l'indiquent, il s'agit de faire le lien entre artistes et collectionneurs, d'où la mention de l'adresse personnelle de ces derniers dans les catalogues d'exposition les plus anciens. Le catalogue est à l'époque un objet constitué de trois parties : le texte de présentation de l'exposition (aussi appelé préface), la liste des œuvres (c'est à dire le corps du catalogue),

22. *Ibid.*

23. Véronique Goncerut, *Les catalogues d'exposition : gros plan sur ces ouvrages d'art devenus incontournables*, mars 2017, URL : <https://blog.mahgeneve.ch/les-catalogues-dexposition/> (visité le 27/08/2021).

24. B. Joyeux-Prunel et Olivier Marcel, « Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History », 4-2 (2015), p. 26.

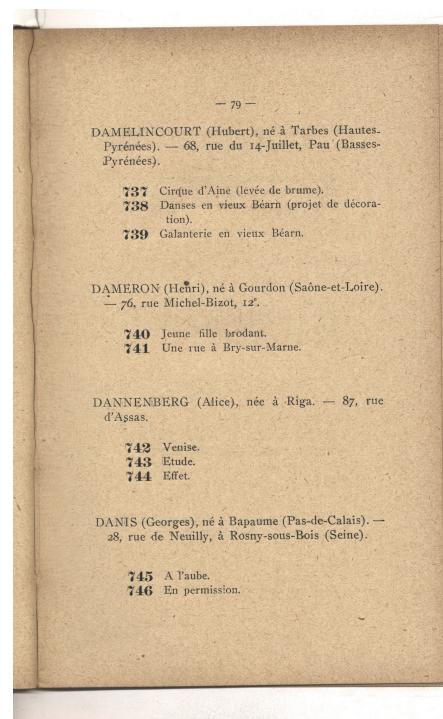


FIGURE 1.1 – Un exemple typique de page de catalogue d'exposition (*Catalogue [...] exposés, Salon des Indépendants*, 1913, p. 79)

puis les reproductions (fac-similés et images des éléments présentés)²⁵. De fait, les illustrations apparaissent dans les catalogues d'exposition dès 1819, deux ans seulement après l'arrivée de la presse lithographique à Paris²⁶. Le catalogue dit « traditionnel » évolue peu à peu au cours du XIXème et du début du XXème siècle, au fur et à mesure de sa propagation, sur un modèle semblable dans le monde, pour atteindre une forme intermédiaire avec les mouvements dada et surréalistes. Cette période marque une rupture dans l'histoire des catalogues, opérant une transformation de son aspect matériel (mise en page et typographie) et de son contenu, faisant disparaître ou réduire à son minimum la description de l'exposition afin de donner la part belle au texte²⁷. Cette transformation coïncide avec la démocratisation de la culture au sortir de la guerre qui se matérialise par l'ouverture au grand public des musées, des galeries et du marché de l'art. Les catalogues d'exposition se veulent alors créatifs et sont vu comme un vecteur de présentation et de communication de l'exposition. À partir de 1960, les catalogues deviennent des livres d'art en soi et font l'objet d'expérimentations de la part des commissaires, des musées et d'autres organisateurs d'exposition. Leur conception est de plus en plus réfléchie dans l'idée de marquer l'esprit et susciter l'envie des visiteurs²⁸. Ainsi, chaque œuvre exposée est accompagnée dans le catalogue d'une notice précise et développée.

1.2.3 Une source primaire pour la recherche en histoire de l'art

Le catalogue d'exposition est une source privilégiée pour l'historien de l'art car il permet de suivre l'évolution de l'activité artistique dans le temps et l'espace. Il est énormément utilisé dans le cadre d'études monographiques, puisqu'il permet de rechercher ou confirmer une information précise, comme le nom d'une œuvre, la date d'une exposition, etc. Plus largement, il donne une documentation précise, fournie et détaillée d'un événement passé, un aperçu de l'évolution des idées artistiques et des relations entre les personnes gravitant autour d'une exposition : artistes, collectionneurs, mais aussi journalistes, critiques, marchands d'art, écoles et curateurs.

Le catalogue du XX^{ème} siècle peut être aussi vu comme une véritable somme scientifique. D'après Pierre Rosenberg²⁹, le premier catalogue de ce type, *Peintre de la Réalité*, date de 1934. Issu d'une exposition ayant eu lieu au Musée de l'Orangerie, il a été dirigé par Charles Sterling. Le travail de rédaction des notices sur chacune des œuvres de l'exposition est tel que le catalogue est considéré, en 1989, comme un manuel pour la peinture française du XVII^{ème} siècle. Chaque catalogue est, à partir de cette période, l'occasion de récupérer le plus d'informations possibles sur l'histoire de l'œuvre, sa naissance, ses différents propriétaires, les prix d'achat au cours du temps, etc. Un travail minutieux pour récupérer ces données est donc entrepris, notamment en utilisant les catalogues d'exposition du XIX^{ème} siècle, où toutes ces informations sont contenues. Les catalogues raisonnés constituent un autre exemple d'utilisation des catalogues d'exposition commerciaux par les chercheurs en histoire de l'art. Ces ouvrages se veulent des listes exhaustives de toutes les œuvres connues d'un artiste. Ils sont, par exemple, utilisés afin de déterminer si un tableau est faux. Ces outils commerciaux, élaborés par des historiens de l'art, mobilisent

25. C. Leinman, « Le catalogue d'art contemporain »...

26. V. Goncerut, *Les catalogues d'exposition : gros plan sur ces ouvrages d'art devenus incontournables...*

27. C. Leinman, « Le catalogue d'art contemporain »...

28. V. Goncerut, *Les catalogues d'exposition : gros plan sur ces ouvrages d'art devenus incontournables...*

29. P. Rosenberg, « L'apport des expositions et de leurs catalogues à l'histoire de l'art »...

également les informations qui peuvent être contenues dans les catalogues d'exposition.

Cet objet fournit donc des données sociales, économiques, géographiques et même politiques sur le marché de l'art pour une période allant du XVIII^{ème} à nos jours³⁰. Cette vision est cependant à nuancer, puisque le catalogue, bien qu'étant vu comme une donnée d'autorité, peut contenir des erreurs. Il revient donc à l'historien de l'art de croiser ses sources. Cependant, les catalogues restent un objet important dans la recherche en histoire de l'art, puisqu'ils fournissent des informations concrètes et précises sur le marché de l'art. Rassemblées, les données qu'ils contiennent offrent la possibilité de reconstruire l'évolution des idées artistiques, retrouver la provenance d'une œuvre et, surtout, de reconstruire les réseaux locaux et internationaux de circulation du marché.

1.3 Le catalogue de ventes de manuscrits

1.3.1 Au cœur d'un marché du manuscrit encore peu étudié

On définit le « manuscrit autographe » comme un écrit manuscrit, de la main de son auteur, à opposer à une copie. Son commerce connaît son plein essor au XIX^{ème} siècle. À noter qu'il s'inscrit dans le commerce du livre, puisque les manuscrits sont essentiellement traités par des libraires.³¹

Avant le XIX^{ème} siècle, le marché est limité aux manuscrits médiévaux. En effet, l'intérêt de ces documents aux yeux des collectionneurs est dû à leurs peintures et leurs enluminures. Ils sont alors vu comme des objets d'art. Les manuscrits autographes, composés matériellement de papier et d'encre et présentant uniquement des écritures, sont donc considérés comme ayant peu d'intérêt dans ce premier marché du manuscrit. Le marché des manuscrits modernes et contemporains fleurit dès les années 1830 et se développe considérablement dans les décennies suivantes. C'est à cette période qu'ont lieu les principales ventes de lettres et autographes de Mme de Sévigné, jusque là gardés dans sa famille, et de grandes collections d'autographes.

Si les autographes séduisent toujours à la fin du XIX^{ème} siècle, ils sont cependant également accompagnés d'un mouvement d'intérêt pour les manuscrits littéraires. La plupart de ces documents présents sur le marché sont issus du XIX^{ème} siècle. En effet, les manuscrits des siècles précédents sont rares à cause du peu d'intérêt qu'il leur était porté auparavant. Ce changement est probablement dû, d'après Bodin³², au legs par Victor Hugo de ses manuscrits à la Bibliothèque nationale de Paris en 1881³³ ainsi qu'à la vente aux enchères des manuscrits d'Edmond Goncourt et de son frère en 1897³⁴. Le marché des manuscrits autographes et littéraires reste assez peu prisé pendant le XX^{ème} siècle³⁵,

30. B. Joyeux-Prunel et O. Marcel, « Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History »...

31. FranceArchives, *Collection des catalogues de vente d'autographes et livres anciens imprimés des libraires et des salles de vente*, URL : <https://francearchives.fr/fr/findingaid/2bd6f418cd252890f42e22c8215d45785c96a3fe> (visité le 13/04/2021).

32. Thierry Bodin, « Les grandes collections de manuscrits littéraires », dans *Les ventes des livres et leurs catalogues, XVIIe-XXe siècle*, dir. Annie Charon et Élisabeth Parinet, Publications de l'École nationale des chartes, Paris, 2018 (Études et rencontres), p. 169-190, URL : <http://books.openedition.org/enc/1419> (visité le 13/04/2021).

33. Le legs est effectif en 1886, date où Léopold Delisle, en sa qualité d'administrateur général de la Bibliothèque nationale, récupère les manuscrits

34. *Ibid.*

35. T. Bodin et Jacques Neefs, « Les autographes. Entretien », *Genesis (Manuscrits-Recherche-Invention)*, 7 (1995), p. 177-184.

cependant un regain d'intérêt est remarqué au début du XXI^{ème} siècle. En effet, la valorisation des archives privées a entraîné le développement d'un marché spéculatif du manuscrit d'autographe, à tel point que les archives publiques se retrouvent confrontées au problème du démembrement des fonds privés.

Très peu de recherches ont été réalisées sur le marché des manuscrits autographes et littéraires jusqu'à présent. Outre quelques études anglo-saxonnes et le projet Kata-base, les principales publications sont issues du milieu des vendeurs d'autographes. Ainsi Thierry Bodin, marchand et expert en manuscrits autographes, est considéré comme un des principaux spécialistes du domaine et est fréquemment invité dans les colloques et séminaires. Les chercheurs ont commencé à s'intéresser progressivement au sujet suite au développement, depuis deux décennies, des spéculations autour des manuscrits littéraires et autographes. Ils ont alors commencé un travail de recensement des manuscrits, en dépouillant les catalogues de ventes³⁶.

1.3.2 Le catalogue de vente de manuscrit : inventaire et commerce

Ces ventes de grandes collections d'autographes, mentionnées précédemment sont, dès les années 1830, accompagnées de gros catalogues de vente. Ils s'inspirent des catalogues de ventes publiques de livres, les deux types de documents étant réalisés, pour la plupart, par des libraires. Ceux-ci s'imposent dès la seconde moitié du XVII^{ème} siècle en France, prenant exemple sur l'usage qu'il en est fait depuis plus d'un siècle aux Pays Bas. Les catalogues contiennent des notices descriptives des documents vendus (ou en vente), parfois des reproductions de ceux-ci, des annotations, les prix et même les acheteurs³⁷. Véritables « Ouvrages de consommation commode »³⁸, les catalogues permettent aux acheteurs potentiels de se faire une idée sur la vente. Ils peuvent être publiés par des archives ou des libraires, soit régulièrement, comme une présentation des fonds de ceux-ci, soit épisodiquement, dans le cadre des ventes aux enchères.

Parmi eux, un intérêt particulier est porté dans notre corpus pour la dynastie Charavay. Son fondateur, Gabriel Charavay (1818-1879), est un ouvrier bonnetier qui, à la suite d'achats de manuscrits et brochures autographes, en devient expert. Il rachète en

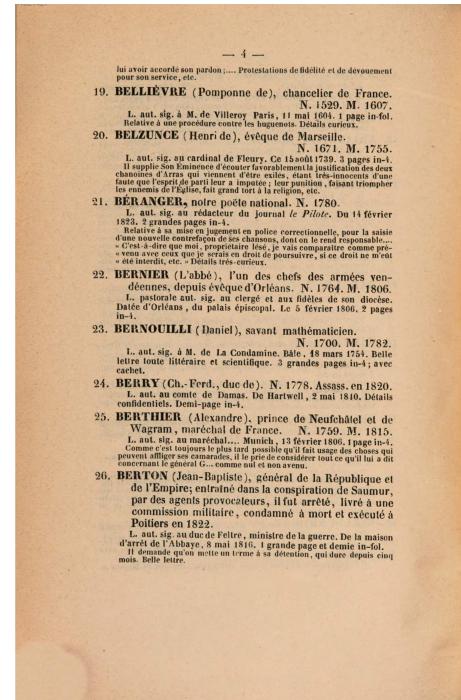


FIGURE 1.2 – Un exemple typique de page de catalogue de ventes de manuscrits (*Catalogue [...] manuscrits*, Charavay, 1846, p. 4

36. Ileana Miranda Mendoza, *L'économie du patrimoine écrit : le marché des autographes*, These de doctorat, Paris 1, 2010, URL : <http://www.theses.fr/2010PA010078> (visité le 27/08/2021).

37. CHARON, « Avant propos », In : Elisabeth Parinet, *Les ventes de livres et leurs catalogues, XVIIe-XXe siècle*, Publications de l'Ecole nationale des Chartes, Paris, 2018 (Etudes et rencontres), p. 5-9

38. MASSON, « Typologie des catalogues de ventes », In : *Ibid.*, p. 119-127

1865 le cabinet d'autographes d'Auguste Laverdet (1805-1865), libraire parisien. Celui-ci est par la suite géré sur plus d'un demi-siècle par la famille et devient un des centres du commerce du manuscrit autographe et littéraire. C'est dans ce contexte que *La Revue des Autographes*, ou *Revue des autographes, des curiosités de l'histoire et de la biographie* est fondée en 1866 par Gabriel Charavay. Il s'agit d'un catalogue hybride, donnant des informations sur le marché ainsi qu'une liste détaillée des manuscrits à vendre, avec leur prix fixe. Sa structure évolue vers un catalogue présentant principalement les manuscrits à vendre aux acheteurs potentiels. Cette tendance est visible notamment dans la construction des pages du catalogue, passant d'une structure en simple colonne à deux colonnes, entre 1873 et 1893, ce qui permet aux lecteurs de naviguer plus rapidement dans les items à vendre et repérer ce qui les intéresse.³⁹

1.3.3 Une source primaire pour le chercheur

Le catalogue de ventes de manuscrits est une source de données non négligeable pour le chercheur. Il peut être utilisé dans le but d'identifier les manuscrits et établir leur authenticité, en retracant leur chemin dans le cadre de différentes ventes. Il donne également un aperçu du succès des auteurs sur une période, en mesurant la quantité de ses manuscrits vendus et leur prix. Grâce aux prix mentionnés, il est également possible de reconstruire, à un niveau plus large, l'histoire de l'évolution des prix des manuscrits. Enfin, les fac-similés contenus dans les catalogues permettent de reconstruire des manuscrits disparus, ou vendus dans une collection privée non accessible⁴⁰. Si les catalogues de ventes de manuscrits en eux-même sont donc particulièrement utiles pour l'historien et le paléographe, ils restent encore peu utilisés par les chercheurs dans les années 1980⁴¹, notamment de part la masse gigantesque de documents et l'absence de catalogue de catalogues. À l'heure du passage des catalogues, véritable réservoir de métadonnées, en numérique, les catalogues de ventes de manuscrits restent peu employés. Les catalogues numériques de manuscrits sont par ailleurs plutôt réalisés par les bibliothèques patrimoniales, directement à partir de leur catalogue⁴² et il n'existe pas d'initiatives pour numériser et rendre interrogable les catalogues de ventes de manuscrits.

Pour ce qui est de l'apport des manuscrits eux-même dans la recherche, ces outils de travail, qu'il s'agisse de correspondances ou de travaux préparatoires, sont une source reconnue pour l'établissement d'éditions critiques, permettant de retracer l'évolution de la construction d'une œuvre littéraire.

39. Simon Gabay, L. Rondeau Du Noyer et Mohamed Khemakhem, « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *IX Convegno AIUCD*, 2020, URL : <https://hal.archives-ouvertes.fr/hal-02388407> (visité le 13/04/2021).

40. *Ibid.*

41. Emile Van Balberghe, « Les manuscrits et leur histoire. » *Scriptorium*, 40-1 (1986), p. 123-125, DOI : 10.3406/script.1986.1436.

42. Un exemple de ce type de catalogues de manuscrits électronique : <https://bu.univ-amu.libguides.com/c.php?g=511748&p=4045678>

Chapitre 2

Humanités numériques et catalogues

Les catalogues sont donc semi-structurés sous la forme d'entrées à l'instar de nombre de documents dits « de type encyclopédiques », tels que les bibliographies, dictionnaires, annuaires... Cette structure standardisée d'informations factuelles permet de penser à une récupération automatique des données, par le biais notamment des outils des humanités numériques. Ainsi, les premières tentatives de transformations de ces réservoirs d'informations en numérique, dans le but de les interroger plus rapidement intervennent dès le début du développement de ce genre de méthodes.

Le présent chapitre met donc en lumière les liens entre catalogues, et plus largement les documents semi-structurés sous la forme d'entrées, et humanités numériques ainsi que les problèmes sous-jacents. Dans un premier temps, nous nous intéressons aux différents projets impliquant ce genre de documents et les enjeux et problèmes qui en résultent, puis nous dresserons un historique de la question plus précise de la numérisation des catalogues. Un dernier point concerne plus particulièrement l'alliance pluri-institutionnelle Artl@s-Katabase-INRIA qui tente de résoudre ces problèmes et les solutions trouvées pour les contourner.

2.1 Un enjeu ancien au sein des humanités numériques

2.1.1 Enjeux et problèmes du traitement numérique des documents semi-structurés

Les documents semi-structurés sous forme d'entrées sont parmi les plus denses en information. Leur structure permet, en effet, à l'instar d'un tableau, de rassembler sur un petit espace un grand nombre d'informations tout en restant compréhensible pour l'œil humain. Les images 2.1, 2.2 et 2.3 correspondent respectivement à des entrées de catalogues d'exposition, de vente de manuscrits et de bibliographie et illustrent bien cette caractéristique des documents structurés. Les informations y sont présentées sous la forme d'une liste. Chaque type de donnée est signalé d'une façon différente, afin de bien distinguer chaque élément. Ainsi, chaque entrée est divisée par un espace, chaque information présente dans l'entrée est signalé par un saut de ligne, un changement typographique (gras, italique, exposants, etc)... Les données font donc sens grâce à une information graphique riche et variée.

BERNARD (Mlle DELPHINE), de Nancy, élève de M. Maréchal, de Metz.	
1. Portrait de Mlle***.	(Pastel).
2. <i>Idem</i> de Mlle W...	<i>Id.</i>
3. Étude d'après nature.	<i>Id.</i>

FIGURE 2.1 – Catalogue [...] Nancy, 1843, p.3

426. Lettre de Cassini (Jacques) à M. de Boulogne,
10 mars 1740. *Autogr. signé.*
427. 13 lignes de la main de Galilée Galilei et signées,
24 août 1601.
428. Lettre de Christian Huygens, 27 décembre 1694.
Autogr. signé.

FIGURE 2.2 – Catalogue de feu M. de Bruyère, Chalabre, 1833, p.102

60. SUPPLÉMENT au traité de l'inamovibilité des pasteurs du second ordre. In-8° de 5 feuillets 1/2. Impr. d'Egron, à Paris. — A Paris, chez Brajeux, chez Baudouin frères. Prix 1—25 Voyez numéro 2776 de 1821.
61. LEÇONS IDÉOLOGIQUES pour apprendre à la jeunesse à contracter des habitudes sociales et des habitudes morales. Par M. Brun. In-12 de 8 feuillets 1/8. Imp. de Bailleul, à Paris.—A Paris, chez Renard. Prix, 2—0 Le faux-titre porte : <i>Habitudes sociales, Habitudes morales.</i>

FIGURE 2.3 – Bibliographie de France, 1822, p.6

On se retrouve ici face à un problème. Cette organisation, tout à fait compréhensible par l'œil humain, est beaucoup plus compliquée à saisir pour un ordinateur. Comment traiter alors ces données afin d'obtenir ces mêmes informations ? Comment signaler qu'un saut de ligne, un espace, un changement typographique correspond à une nouvelle information ? Tout l'enjeu de la numérisation et l'exploitation informatique de ces documents semi-structurés est donc de réussir à passer cette information graphique non textuelle en information sémantique.

Il existe un certain nombre de projets de recherche centrés autour du traitement numérique de documents historiques semi-structurés, de part la quantité de données qu'ils contiennent. Ceux-ci peuvent être source d'inspiration méthodologique afin de gérer au mieux ce problème d'information non textuelle primordiale à la compréhension des documents.

Le groupe « Annuaires et Adresses » du consortium Paris Time Machine travaille sur l'*Annuaire des propriétaires et des propriétés du département de la Seine et de Paris*, publié entre 1894 et 1937 et qui recense toutes les adresses parisiennes et leurs propriétaires. Il s'agit ici de documents imprimés, structurés et composés d'entrées comme des catalogues. Le groupe a travaillé sur les annuaires des années 1898, 1903, 1913 et 1923. Leur chaîne de traitement est également composée de deux grands points : Dans un premier temps, les pages numérisées sont transcris automatiquement. Le résultat obtenu est ensuite analysé par des scripts python d'extraire chaque entrée, puis chaque adresse, nom de personnes possédant la maison, etc. Ici on utilise donc une méthode se basant sur la variation des caractères - majuscule, minuscule, numérotation - pour reconnaître à quelle type d'information chaque élément correspond.

Le projet BasNum¹, « Numérisation et analyse du Dictionnaire universel de Basnage de Beauval : lexicographie et réseaux scientifiques », démarré depuis 2018, s'intéresse également aux mêmes problématiques. Partenariat entre le laboratoire Litt&Arts² de l'Université de Grenoble, le laboratoire LATTICE³ de Sorbonne-Nouvelle et l'INRIA. Il s'intéresse à un dictionnaire du XVI^{ème} siècle, imprimé structuré sous la forme d'entrées,

1. Une description du projet est disponible sur le site de l'Agence Nationale de la Recherche : <https://anr.fr/Projet-ANR-18-CE38-0003>

2. Arts et pratiques du texte, de l'image, de l'écran et de la scène
3. Langues, textes, traitements informatiques, cognition

comme les catalogues et les annuaires. Sa chaîne de traitement présente deux étapes : la transcription automatique des pages puis la relecture du résultat obtenu à l'aide d'un outil développé par l'INRIA, GROBID. Ici la technique employée se base sur la distribution des différents éléments dans la page pour reconnaître chaque entrée, mot, étymologie, définition...

2.1.2 Bref historique de la gestion numérique des catalogues

La numérisation - c'est-à-dire la copie électronique d'un fichier papier - des catalogues s'inscrit dans le mouvement plus large de numérisation des documents, qui débute dans les années 1970. Le développement des nouvelles techniques de numérisation est alors considéré aux États-Unis comme un moyen de donner accès au plus grand nombre aux collections patrimoniales. L'émergence d'internet, qui permet la circulation de ces collections numériques, entraîne ainsi la naissance de nombreuses campagnes de numérisation. C'est par exemple le cas de l'*American Memory*⁴, projet de la *Library of Congress*, ouvert en 1994 qui a pour but de donner accès à une multitude d'archives écrites, sonores et vidéos du domaine public. De même, Gallica est lancé par la Bibliothèque nationale de France en 1997 afin de rendre accessible numériquement ses documents libres de droit⁵. Ainsi, nombre de bibliothèques patrimoniales vont réaliser des campagnes de numérisation de leurs collections, notamment dans les années 2000. C'est la période des grands chantiers de mise en libre accès des données sur internet. Parmi les éléments accessibles se trouvent notamment les catalogues, ces documents aux usages multiples, qui peuvent se trouver dans les fonds des bibliothèques patrimoniales nationales. Cependant, si les catalogues font partie des objets numérisés, il n'y a pas véritablement de politique spécifique pour ces données pendant cette période.

Dans le cadre de ce mouvement d'accessibilisation des collections, la numérisation spécifique des catalogues contenus dans les fonds des musées, galeries et bibliothèques spécialisées dans l'art intervient plus tardivement. La première de ces initiatives centrées sur les catalogues d'art est le *Getty Provenance*, lancé au début des années 1980 par le *Getty Research Institute*. D'autres ont suivi, tels que la *Royal Academy of London*⁶, qui donne un accès numérique à deux cents cinquante catalogues, la bibliothèque du Palais du Belvédère⁷ ou encore l'Institut national d'Histoire de l'Art. Ce dernier a entamé la numérisation de plus de vingt mille catalogues de ses fonds en 2011, dans l'idée de les rendre plus accessibles au public tout en permettant la conservation de leur contenu⁸.

Cette accélération du processus de numérisation des collections, et plus précisément des catalogues, peut être liée à l'émergence de l'*Open Data*, ou données ouvertes. Ce mouvement, qui promeut l'idée de libre accès aux données publiques, se développe au milieu des années 2000. Il reprend les idées de plusieurs mouvements, tel que l'*Open Source* des informaticiens américains des années 1990, visant un partage des données dans le but de produire plus, ou encore l'*Open Science* - ou Science Ouverte - des chercheurs européens

4. Accessible ici : <https://memory.loc.gov/ammem/index.html>

5. <https://hal.archives-ouvertes.fr/hal-03152774/document>

6. Accessible ici : <https://www.royalacademy.org.uk/art-artists/search/exhibition-catalogues>

7. Accessible ici : <https://digitale-bibliothek.belvedere.at/viewer/>

8. Anne Weber et APAHAU, Association des professeurs d'archéologie et d'histoire de l'art des universités, *Numérisation des catalogues de ventes d'œuvres d'art de la Bibliothèque de l'INHA*, 2014, URL : <http://blog.apahau.org/numerisation-des-catalogues-de-ventes-doeuvres-dart-de-la-bibliotheque-de-linha/> (visité le 27/08/2021).

de la même période qui souhaitent partager les données de la recherche au grand public et à la communauté scientifique. L'*Open Data*, qui reprend donc ces différents principes, est porté par Obama dans l'idée que le partage de ces données permettrait une meilleure transparence des actions de l'État. Le but est de rendre la population moins méfiantes vis-à-vis de celui-ci et de développer tout un argumentaire pour la mise en place de l'*Obamacare*. L'idée est reprise en Europe par les collectivités territoriales, qui donnent accès, notamment pour la France via le portail data.gouv, à de nombreuses données, dont certaines issues du monde de la culture et de la recherche.

Si les années 2000 et 2010 ont donc permis le développement de véritables campagnes de numérisation de catalogues d'art, très peu de ces données sont interrogeables. Or, dans le cadre de la Science Ouverte, il ne s'agit pas seulement de rendre accessibles ces données, sous la forme d'images de pages de catalogues sur internet, mais également de les rendre compréhensibles pour le grand public. Une réflexion doit donc être menée sur la façon dont les lecteurs peuvent accéder aux informations. Les catalogues contiennent des données spécifiques, structurées sur des objets précis, que ce soit un artiste, une œuvre, un collectionneur... Il est alors tout à fait possible de réfléchir au développement de bases de données à partir de ces catalogues, qui permettraient au lecteur de rentrer dans le document en interrogeant directement les données.

Il existe de nombreuses bases de données reposant sur des informations issues de catalogues. La base « Salons »⁹, créée en 2006, en fait partie. Partenariat entre l'INHA et le musée d'Orsay, elle a pour but de présenter le contenu des livrets des Salons artistiques français, et principalement parisiens, entre 1673 et 1914. Elle se concentre donc sur les Salons organisés par la Société des artistes français, mais également par les Beaux-Arts et sur quelques salons régionaux. Composée actuellement de 321 livrets, elle s'appuie sur le travail de prestataires privés spécialisés tant pour la numérisation des catalogues que leur exploitation. Le détail de la chaîne de traitement n'est pas présenté. D'autres bases de données d'art payantes existent, tel que SCIPIO¹⁰, *Sales Catalog Index Project Input Online*, un projet de *WorldCat* qui répertorie des catalogues de ventes d'art nord-américains et européens de la fin du XVI^{ème} siècle à nos jours.

Le *Getty Provenance Index* figure de nouveau parmi les premières tentatives d'exploitation de ces documents. Sa chaîne de traitement est présentée en 2013¹¹. dans le cadre du projet German Sales Catalogs, qui vise à rendre interrogeable 8 700 catalogues allemands, suisses et autrichiens sur une période allant de 1900 à 1945. Afin d'intégrer les informations issus de ces documents dans la base de données, il est nécessaire dans un premier temps de numériser les catalogues, puis de les OCRiser. La transcription automatique récupérée est traitée par des scripts ou algorithmes d'intelligence artificielle. Ceux-ci analysent l'information textuelle et la structurent, en signalant si par exemple il s'agit du nom de l'auteur, du nom de l'œuvre, etc. Enfin, les informations structurées sont injectées dans la base de données. La méthodologie de cette chaîne de traitement est reprise par de nombreux projets de recherche, à l'instar du projet Artl@S.

9. Accessible ici : <http://salons.musee-orsay.fr/>

10. Accessible ici : https://help-fr.oclc.org/Discovery_and_Reference/WorldCat-org/Databases_in_detail/SCIPIO?sl=fr

11. Ruth Cuadra et Suzanne Michels, *Publishing German Sales, A look under the Hood of the Getty Provenance Index*, avr. 2013, URL : <http://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index/> (visité le 27/08/2021).

2.2 Un travail ancré dans un projet de plusieurs années...

2.2.1 ...Au sein d'une organisation pluri-institutionnelle...

Actuellement, deux équipes mènent des travaux sur l'extraction d'informations issues de catalogues : celle de Katabase pour les catalogues de ventes de manuscrits et celle d'Artl@s pour les catalogues d'exposition. Si mon stage s'est déroulé administrativement au sein de la seconde équipe, j'ai été amenée à travailler conjointement au sein de ces deux projets. En effet, elles se sont associées dans l'objectif commun de travailler à la transcription et à l'encodage automatique de leurs données complémentaires.

Artl@s

Artl@s¹² regroupe des projets de recherches autour de la mondialisation artistique et culturelle, dans le but de décentraliser les sources en histoire de l'art. Créé en 2009 par Béatrice Joyeux-Prunel, alors maîtresse de conférence à l'École normale supérieure de Paris, Catherine Dossin, *associate professor* à l'Université de Purdue et Léa Saint-Raymond, post-doctorante à l'ENS, il est issu du constat d'un manque de bases de données sur des catalogues d'exposition mondiaux.

Basart¹³ est donc issu de cette observation. Projet phare d'Artl@s, il s'agit d'une base de données géoréférencées regroupant des catalogues d'exposition issus du monde entier, allant du XIX^{ème} siècle à nos jours. Financée par l'Agence Nationale de la Recherche, l'Université Paris-Sciences-Lettres et le labex TransferS¹⁴, elle a été créée en 2011, mise en ligne en 2016 et rendu accessible publiquement en 2018. Elle est structurée sous la forme d'une base de données PostGIS, soit une base PostgreSQL associée à un SIG, Système d'Information Géographique, permettant le traitement des données géographiques présentes et donc de réaliser des cartes. On obtient ainsi une base interrogeable qui permet d'accéder aux données issues des documents par plusieurs biais : expositions, exposants et œuvres. Il est possible de cibler le résultat obtenu pour une période donnée. Celui-ci est présenté sous la forme d'une liste, exportable en csv, d'une carte et d'un graphique. Jusqu'en 2020, les données présentes dans Artl@s, qui sont alimentées par des contributeurs bénévoles, étaient saisies à la main par les membres du projet. À la suite du stage de Caroline Corbières, qui a proposé une chaîne de traitement automatique des catalogues, les données ont été intégrées plus rapidement dans la base. Ainsi, il y a actuellement plus de 3 000 catalogues ainsi que leur contenu injectés dans Artl@s, correspondant à 222 villes différentes, près de 50 000 entrées et 110 000 œuvres répertoriées.

Artl@s met également à disposition des chercheurs et des institutions des outils afin que ces derniers puissent créer eux-mêmes leurs propres bases de données. Des ateliers de formation sont également proposés, ainsi que des séminaires, des colloques et une revue. Parmi les autres projets numériques se trouvent GEOMAP¹⁵, une géographie du marché de l'art parisien du XIX^{ème} et XX^{ème} siècles sous la forme d'une carte géoréférencée

12. B. Joyeux-Prunel, « Bases de données et gestion de projets en humanités numériques. Les dessous du projet Artl@s »...

13. Accessible à cette adresse <https://artlas.huma-num.fr/fr/bases-en-acces-libre/>

14. Le labex TransferS est un laboratoire d'excellence en sciences humaines et sociales regroupant les composantes de l'Ecole Normale Supérieure de Paris et du Collège de France

15. Accessible à cette adresse : <https://paris-art-market.huma-num.fr/>

et le répertoire des pensionnaires de l'Académie de France à Rome¹⁶, qui fournit une biographie individuelle et collective des lauréats du Prix de Rome de 1666 à 1968. Enfin, le centre IMAGO¹⁷, label européen d'excellence Jean Monnet, fédère un projet de recherche et d'enseignement autour de la circulation des images en Europe et leur rôle dans la construction d'une identité commune. Le projet *Visual Contagions*¹⁸, lancé en 2020, par Béatrice Joyeux-Prunel, alors titulaire de la chaire d'humanités numériques de l'université de Genève, a pour but l'étude de la mondialisation visuelle et de la manière dont les images circulent.

Katabase

Katabase s'inscrit à la suite d'un projet financé par le FNS, le Fonds national suisse de la recherche scientifique, « L'écriture privée au XVII^{ème} : étude philologique des manuscrits de Mme de Sévigné »¹⁹. Il s'échelonne de 2015 à 2018, sous la direction conjointe de Marc Escola, de l'Université de Lausanne et d'Alain Corbellani, de l'Université de Neuchâtel et vise à proposer une nouvelle édition des lettres de la marquise, ainsi qu'une analyse linguistique et éditoriale du corpus. Au cours de ce travail, il est constaté qu'il n'existe pas d'étude spécifique concernant ses manuscrits et que l'absence d'un tel catalogue freine l'étude comparative des éditions successives de l'œuvre de Mme de Sévigné. Ainsi, un premier travail de recherche, à partir de catalogues de ventes de manuscrits, est réalisé et rassemblé sous la forme d'un catalogue électronique répertoriant les différentes sources utilisées²⁰.

Katabase est donc un projet de recherche créé à la suite de ce travail par Simon Gabay, alors collaborateur scientifique de l'Université de Neuchâtel en charge des humanités numériques. Depuis 2018, Son principal objectif est de poursuivre ce recensement des manuscrits du XVII^{ème}, dont l'utilité a fait ses preuves au cours du projet « L'écriture privée au XVII^{ème} ». Le travail se divise en deux parties : une étude des manuscrits eux-mêmes et une exploitation des catalogues de ventes de manuscrits. Ces deux éléments sont peu à peu devenus indépendants. L'étude des manuscrits du XVII^{ème} siècle fait partie du projet e-editiones tandis que les catalogues de ventes de manuscrits se concentrent de Katabase.

En 2019, afin d'interroger rapidement ses sources, Simon Gabay a travaillé en partenariat avec Mohamed Khemakhem et Laurent Romary de l'équipe ALMANACH de l'INRIA, et Lucie Rondeau du Noyer, stagiaire TNAH, a l'automatisation de sa chaîne de traitement. Celle-ci permet de passer d'un catalogue numérisé à un fichier XML-TEI interrogable en plusieurs étapes : le catalogue est OCRisé afin d'obtenir les données textuelles, qui sont encodées par GROBID. Les balises obtenues sont réorganisées pour correspondre à l'encodage voulu par le projet. Cette chaîne de traitement est reprise et améliorée en 2020 par Caroline Corbières au cours de son stage.

16. Accessible à cette adresse : <https://acad-artlas.huma-num.fr/>

17. Accessible à cette adresse <https://www.imago.ens.fr/>

18. Accessible à cette adresse <https://www.unige.ch/visualcontagions/>

19. Un descriptif du projet est disponible à cette adresse : <http://p3.snf.ch/projects-157169>

20. Ce catalogue est trouvable à l'adresse suivante : https://f.hypotheses.org/wp-content/blogs.dir/5238/files/2018/11/Sources_sevigneennes.pdf

2.2.2 ...Menant à la création d'une première chaîne de traitement automatique de leurs catalogues

La chaîne de traitement²¹ réalisée par Caroline Corbières dans le cadre de son stage permet donc de passer d'une image issue d'un catalogue papier à des données qu'il est possible d'injecter dans Basart. Pensée de façon à ce qu'elle soit reproductible par les membres d'Artl@s, elle se divise en plusieurs étapes décrites par le schéma ci-contre :

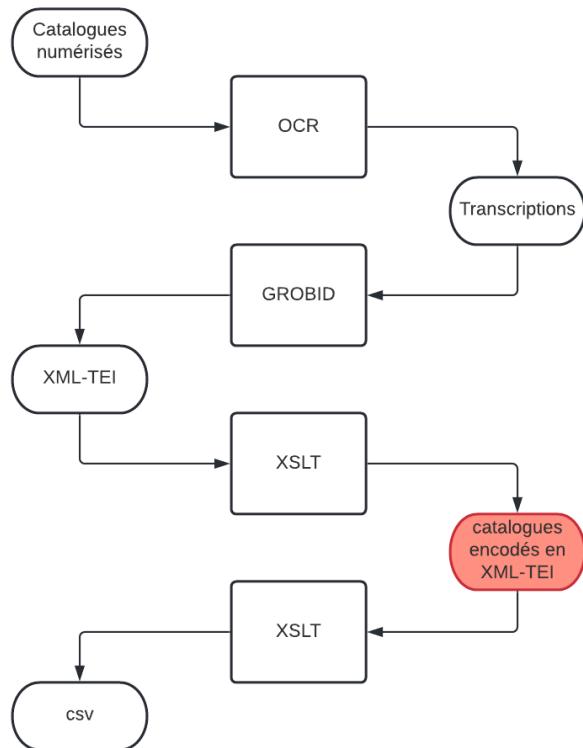


FIGURE 2.4 – Chaîne de traitement actuelle d'Artl@s

Transkribus

Comme les chaînes de traitement citées précédemment, les catalogues numérisés, ici sous la forme d'images PDF ou JPG, sont dans un premier temps transcrits automatiquement. L'outil utilisé pour se faire est Transkribus²², un logiciel OCR spécialisé dans les documents historiques développé depuis 2013 par un consortium de recherche européen s'articulant autour de l'Université d'Innsbruck, Autriche et du projet READ, « Recognition and Enrichment of Archival Document ». Financé par la Commission Européenne, il permet de charger des images et d'entraîner des modèles de reconnaissance des caractères. Le format d'entrée privilégié est le format PDF tandis que les formats de sortie sont les formats Word, texte brut et des variations de textes encodés en XML. Ici, les données sont récupérées sous la forme d'un fichier XML, qui est ensuite retransformé en PDF afin de pouvoir utiliser GROBID.

21. Cette chaîne de traitement et la documentation qui lui est associée sont disponibles dans un dépôt : <https://github.com/carolinecorbieres/ArtlasCatalogues>

22. Disponible ici : <https://readcoop.eu/transkribus/>

GROBID

GROBID²³ est une librairie de machine learning, développée par Patrice Lopez et Luca Foppiano à partir de 2008 et mise en accès libre et gratuit en 2011. Utilisé au départ pour extraire et analyser des métadonnées bibliographiques d'articles scientifiques, son nom vient de GeneRation Of BIbliographic Data. Elle évolue progressivement vers de nouvelles utilisations, à l'instar du module *GROBID-Dictionnaries*²⁴, développé à partir de 2018 par Mohamed Khemakhem, doctorant à Paris VI et associé de l'INRIA, au sein de l'équipe ALMANACH, sous l'impulsion de Laurent Romary. Cette dernière version de GROBID se concentre sur l'extraction et l'encodage de sources lexicographiques et autres documents structurés sous la forme d'entrées, comme les catalogues.

Codé en java, GROBID se base sur une technologie CRF, *Conditional Random Fields*, qui mobilise des modèles en cascade. Il a pour volonté d'extraire et encoder l'information en reconnaissant le général puis en allant dans le détail.

Un fichier pivot : XML-TEI

La sortie de GROBID est un fichier XML-TEI. Le XML, *eXtensible Markup Language*, est un langage d'encodage formel dont la publication officielle au World Wide Web Consortium date de 1998. Il permet de représenter des données textuelles en les structurant par des balises. La TEI, *Text Encoding Initiative*, est à la fois un format de balisage, se basant sur du XML, et la communauté académique qui structure celui-ci. Elle vise à établir des recommandations précises pour l'encodage de documents textuels numériques dans le cadre de la recherche en sciences humaines et sociales depuis 1987. Ainsi, utiliser un fichier XML-TEI au sein de la chaîne de traitement d'Artl@s permet de lui assurer une certaine pérennisation et une interopérabilité des données. Les documents sont conformes à un standard, ce qui permet leur réutilisation et leur compréhension par des personnes en dehors du projet mais également la garantie d'une certaine durabilité.

Si *GROBID-Dictionnaires*, un module construit pour l'encodage des dictionnaires, est aussi capable de traiter des catalogues, de part leur structure en entrées commune, la sortie en XML-TEI utilise cependant des balises XML-TEI correspondant aux dictionnaires. Ainsi, il est nécessaire, pour avoir un fichier XML conforme à la fois aux bonnes pratiques préconisées par la TEI, mais aussi aux balises créées spécialement pour le traitement des catalogues par Laurent Romary et Caroline Corbières, de transformer le résultat obtenu par GROBID. C'est ce à quoi servent les feuilles de transformation XSLT, *eXtensible Stylesheet Language Transformation*, qui est un langage de transformation de fichiers XML. Une feuille XSL permet de passer de transformer un fichier XML en un autre fichier du même format utilisant d'autres balises ou encore de passer à un fichier d'un tout autre format, tel que HTML, LaTeX, etc.

Dans la chaîne de traitement, XML-TEI est employé comme un format pivot. Ce terme désigne un langage spécifique choisi pour le partage ainsi que le stockage des données. Ainsi, le centre de la chaîne de traitement est le fichier en XML-TEI : il permet d'associer toutes les données sous une forme structurée, à leurs métadonnées dans un même unique fichier mais également de d'extraire et exploiter ces informations par la suite à partir de ce document. Si d'autres formats pivot existent, à l'instar du csv, mentionné précédemment ou de bases de données relationnelles par exemple, XML-TEI a

23. Accessible ici : <https://grobid.readthedocs.io/en/latest/>

24. Le code de GROBID Dictionnaires est disponible dans le dépôt github ci-joint : <https://github.com/MedKhem/grobid-dictionaries>

l'avantage de permettre une structuration précise et profonde de l'information textuelle associée à la possibilité de son réemploi. De plus, il s'agit d'un langage standard, grâce au vocabulaire et préconisations de la TEI, qui permet une compréhension et une réutilisation facile des données par d'autres chercheurs. Enfin, point non négligeable, XML est un format flexible et lisible qui ne nécessite pas forcément de logiciel spécifique pour accéder à la donnée, au contraire des bases de données. Une question se pose cependant : si le fichier TEI est ensuite converti en csv afin de corriger les informations plus facilement, pourquoi ne pas faire du csv le format pivot de la chaîne de traitement et ne pas passer par un encodage en XML ? Ici, l'intérêt principal du XML est donc la possibilité d'associer données structurées et métadonnées, dans un cadre où celles-ci sont particulièrement nécessaires afin de saisir le catalogue, puisqu'elles décrivent notamment l'exposition associé à celui-ci.

Ces dernières briques permettent d'obtenir un fichier XML-TEI pour lequel sont encodées toutes les informations textuelles d'un catalogue. À celui-ci sont associées des métadonnées, au sein même du fichier, grâce au **TeiHeader**, qui permet d'indiquer les éléments correspondant à l'élaboration du catalogue papier et de son pendant numérique : titre, auteur, éditeur, date de l'évènement, auteur du document électronique... Une dernière feuille XSL permet également de transformer ce fichier pivot en fichier au format csv²⁵, permettant aux membres d'Artl@s de corriger plus facilement les données obtenues. Le résultat final est alors inséré dans Basart.

25. Le format csv, pour *comma separated values*, est un format texte libre de droit représentant des données tabulaires sous la forme de valeurs séparées par des virgules, comme son nom l'indique.

Deuxième partie

Extraction de l'information issue de l'image

Chapitre 3

Pourquoi, comment et quoi OCRiser ?

Si les prémisses pour une transcription automatique datent de 1870-1930, la première véritable machine capable de lire des caractères alphanumériques est élaborée à la fin des années 1920 par Gustav Tauschek. Par la suite, cette technique se développe jusque dans les années 1990. Elle est alors employée pour de nombreux domaines, à l'instar de la cryptoanalyse et des passeports¹. Ces grandes améliorations au cours du XX^{ème} siècle permettent de généraliser l'emploi de cette méthode et son application dans le domaine culturel, comme dans le cas d'Artl@s.

La transcription automatique des catalogues a donc été jusqu'à présent réalisée au sein de la chaîne de traitement par le biais du logiciel Transkribus. Face à la privatisation de cet outil, ainsi qu'à son manque de transparence, il était donc important de réaliser ce travail sur une autre plateforme. Or, l'étape de transcription fonctionne grâce à des modèles entraînés sur des jeux de données spécifiques et Transkribus n'offre pas la possibilité d'extraire ces éléments afin de les injecter dans un nouveau logiciel plus ouvert, gratuit et libre. Ainsi, cette migration hors de Transkribus nécessite de procéder à une campagne de réentraînement complète des modèles de transcription. Pour ce faire, il est nécessaire de détailler le fonctionnement de la transcription automatique ainsi que les formats et différents logiciels ouverts et libres employables.

3.1 Qu'est-ce que la transcription automatique ?

3.1.1 Vocabulaire et étapes d'une transcription automatique

L'OCR, *Optical Character Recognition*, aussi appelé Reconnaissance Optique de Caractères (ROC), est un processus qui consiste à convertir un ensemble de signes graphiques, caractères alphanumériques comme ponctuation et espaces, encodés sous la forme d'une image en texte². L'HTR, *Handwritten Text Recognition*, ou reconnaissance de texte manuscrit, ne lui est pas très différente, si ce n'est qu'elle est plus récente et s'intéresse, comme son nom l'indique, à l'écriture manuscrite. En effet, ces données-ci sont plus difficiles à étudier pour un OCR, plutôt spécialisé en reconnaissance de l'écriture imprimée.

1. Florence Burgy, Steeve Gerson et Loïc Schüpbach, *Ex imagine ad litteras : Projet d'océsiration de la collection De Bry*, mémoire de recherche réalisé dans le cadre du Master en Sciences de l'Information, Genève, Haute Ecole de Gestion, 2020, URL : https://doc.rero.ch/record/328465/files/BURGY_GERSON_SCHUPBACH_Projet_Recherche_Bodmer_Lab.pdf (visité le 27/08/2021).

2. Jean-Baptiste Camps et Nicolas Perreux, *Reconnaissance optique des caractères et des écritures manuscrites - Projet E-NDP*, févr. 2021, URL : https://outils.lamop.fr/lamop/mp3/E-Ndp/JBC-NP_e-NDP_OCR-et-HTR.pdf.

Après des essais industriels peu fructueux dans les années 1980, l’HTR ne voit donc le jour qu’à partir de 2010, grâce aux progrès de l’Intelligence Artificielle. Il est alors employé par de nombreux programmes scientifiques en humanités numériques. Face à des données plus complexes, notamment lors de l’extraction d’informations textuelles de documents historiques, les logiciels de transcription automatique de type HTR abordent des stratégies différentes des OCR. Le premier tente de reconnaître les caractères de manière individuelle, puis de valider par une reconnaissance lexicale les mots obtenus tandis que le second travaille à la reconnaissance d’un mot entier.³. Ainsi, dans le cas de documents historiques, à l’instar des catalogues d’exposition, où les caractères imprimés peuvent être en mauvais état, il est judicieux d’utiliser de l’HTR.



FIGURE 3.1 – Fonctionnement d’un OCR

La figure 3.1 présente les différentes étapes réalisées par un OCR afin d’obtenir la transcription d’une image. La première étape, de Prétraitement, prépare l’image à son traitement par l’OCR. Il s’agit de la rendre la plus lisible possible et de supprimer au maximum le bruit. Pour ce faire, plusieurs outils sont à disposition⁴ :

- la rotation de l’image : la plupart des images ayant été numérisées, il n’est pas rare d’obtenir une image qui ne sera pas droite. Cela peut influencer le résultat obtenu en sortie, bien que les OCR actuels se défassent peu à peu de ce problème.
- niveau de gris : ces pages numérisées le sont très souvent en couleurs, ce qui peut également impacter le résultat sur des versions plus anciennes d’OCR. Il convient alors de les transformer en nuances de gris afin de réduire ce risque.
- la binarisation de l’image : une image est la plupart du temps de type bitmap, c'est-à-dire qu’elle est composée d’une matrice, soit un tableau de point à plusieurs dimensions. Dans le cas où il s’agit de deux dimensions, à l’instar des formats jpeg, gif, png ou tif⁵, le point est appelé un pixel. Lorsque l’image est en couleur (RGB), le pixel aura une valeur spécifique de niveau de rouge, bleu et vert, qui formera sa couleur. Si l’image est en noir et blanc, un pixel qui portera de l’information aura une valeur 1 et un pixel sans information en aura une de 0. Un OCR va détecter les variations de valeur des pixels pour reconnaître l’information textuelle. Ainsi, il est plus facile pour lui de travailler sur une image codée de façon binaire et où chaque pixel ne peut prendre que les valeurs 1 et 0. Comme les éléments précédents, la binarisation n’est plus obligatoire grâce aux progrès des OCR, mais elle reste cependant conseillée.

3. Mohammed Abaynarh, Hakim El Fadili et Lahbib Zenkouar, « Reconnaissance optique de documents amazighes : approches et évaluation des performances », *Etudes et Documents Berberes*, N° 34-1 (2015), p. 189-198, URL : <https://www.cairn.info/revue-etudes-et-documents-berberes-2015-1-page-189.htm> (visité le 27/08/2021).

4. S. Gabay, *Cours sur l’OCR et GROBID*, en, 2020, URL : https://github.com/gabays/Cours_2020_01_Strasbourg (visité le 27/08/2021).

5. La différence entre ces différents formats réside principalement dans leur taille et leur type de compression.

La deuxième étape est appelée le *layout analysis*, traduisible par **analyse de mise en page**⁶. L'OCR repère la structuration de la page et les différents éléments qui la composent. Il reconnaît également les lignes, qu'il matérialise de deux façons différentes : une ligne sera à la fois une *baseline*, c'est-à-dire une ligne suivant le bas de chaque lettre sur toute une ligne, mais aussi un cache, soit un masque qui recouvre tous les éléments composants la ligne. Une fois cette étape terminée, l'OCR va reconnaître chaque caractère de la ligne. La transcription obtenue est vérifiée, notamment par des modèles de langues ou par la reconnaissance lexicale des mots pendant l'étape de **post-traitement**.

Ainsi, un OCR a besoin de passer par plusieurs étapes pour fonctionner. Si certaines d'entre elles sont manuelles, à l'instar du **prétraitement** et du **post-traitement**, d'autres ne le sont pas. Jusqu'à récemment, **segmentation** et **reconnaissance de caractères** étaient associées au sein d'un seul et même modèle, qui gérait dans un premier temps la mise en page (*Layout Analysis*), puis la transcription proprement dite. Cependant, afin d'obtenir un meilleur résultat, il est maintenant possible de diviser ce modèle en deux. En effet, entraîner spécifiquement un modèle de segmentation permet de le rendre plus performant. Or, avoir une meilleure segmentation, c'est-à-dire réussir à analyser la structure de la page et l'emplacement des caractères à reconnaître, permet de grandement améliorer le résultat de la transcription à laquelle il aboutit. Il est donc nécessaire d'entraîner dans un premier temps un modèle de segmentation puis un modèle de reconnaissance de caractères.

3.1.2 L'entraînement d'un modèle

Si il est maintenant courant d'utiliser des systèmes capables de s'adapter aux données sans avoir à créer des modèles spécifiques à ceux-ci, ce n'est pas le cas pour les documents historiques, un peu plus complexes. Ainsi, il est nécessaire dans ce contexte de lancer des campagnes d'entraînement de modèles spécifiques aux données. Les OCR se basent sur le *machine learning*, ou entraînement automatique. Branche de l'intelligence artificielle, ce domaine englobe toutes les méthodes permettant de créer des modèles ou algorithmes à partir de données. L'idée est de fournir suffisamment d'informations à l'ordinateur pour que celui-ci soit capable d'appréhender les données et de les analyser. Concrètement, dans le contexte de la reconnaissance de caractères, cela signifie que l'on fournit à la machine un certain nombre de pages numérisées ainsi que leur transcription afin qu'il puisse apprendre à réaliser lui-même cette opération d'extraction du texte issu d'image⁷.

L'ordinateur s'entraîne alors à reconnaître les données, en comparant la transcription fournie avec l'image qui lui est associée. Cette étape est répétée une multitude de fois et porte le nom d'*epoch* en anglais et itération en français. Pour chaque étape, un modèle est créé et le meilleur d'entre eux est retenu. En effet, la comparaison entre la transcription réalisée par le modèle et celle fournie permet d'obtenir un score d'*accuracy*, qui détermine le niveau de reconnaissance des différents caractères. L'idée est alors de récupérer le meilleur modèle possible. Il faut cependant faire attention, car un modèle trop entraîné peut n'être utilisable que sur les données spécifiques qui lui ont été présentées et ne pas être adaptables. On appelle cela le surapprentissage - *Overfitting* - qui est la conséquence d'un surentrainement - *Overtraining* du modèle sur des données particulières. Si

6. Au sein de ce mémoire, le terme **segmentation** est employé pour désigner cette étude de la mise en page plutôt qu'une segmentation des caractères.

7. F. Burgy, S. Gerson et L. Schüpbach, *Ex imagine ad litteras : Projet d'océsiration de la collection De Bry...*

ce processus est utile à certains moments, comme pour un modèle entraîné exclusivement pour une tâche unique, précise et limitée, ce n'est pas le cas si l'on souhaite obtenir un résultat généralisable et adaptable sur divers documents.

Ainsi, il est extrêmement important de bien choisir les données que l'ordinateur traite pour créer un modèle. Un grand nombre de données permettant d'obtenir de meilleures performances. Cependant, il ne faut pas oublier de prendre en considération l'adaptabilité du modèle et sa capacité à pouvoir traiter à la fois les données spécifiques présentées, mais aussi des données qui peuvent être un peu plus différentes, tout en restant similaires. C'est notamment très important pour les documents historiques tels que les catalogues d'exposition dont l'aspect des pages peut varier. L'idée étant de réaliser un modèle de segmentation et un modèle de reconnaissance de caractères pour deux types de catalogues différents, il serait alors intéressant de réaliser un modèle commun pour les catalogues. Concrètement, cela veut dire associer aux catalogues d'expositions et de ventes de manuscrits d'autres documents à entrées du XIX^{ème}. L'ordinateur sera capable de traiter un grand nombre de catalogues différents, tout en améliorant les résultats pour les catalogues d'exposition et de ventes de manuscrits, puisqu'il aura étudié plus de données. En réalisant un modèle qui n'est pas spécifique aux catalogues d'exposition, on obtiendrait donc un meilleur résultat en sortie d'OCR pour la chaîne de traitement d'Artl@s tout en créant un outil réutilisable par les chercheurs dans le cadre d'autres projets sur des catalogues ou documents structurés à entrées.

3.2 Les données utilisées

3.2.1 Quelles données de travail ?

Pour réaliser ce modèle commun pour les catalogues, il est donc dans un premier temps nécessaire de dresser un état des lieux des données utilisables, que l'on fournira à l'ordinateur. Les premières données à disposition sont les cinq cent pages de catalogues d'expositions et de ventes de manuscrits traités par les équipes d'Artl@s et d'e-ditiones⁸ et qui ont été utilisées pour réaliser les modèles dans Transkribus. Ces cent cinquante pages de catalogues d'exposition et trois cent cinquante pages de catalogues de ventes de manuscrits s'échelonnent du XIX^{ème} au XXI^{ème}, comme il est possible de le voir dans le tableau 3.1⁹. La proportion importante de données dans la seconde moitié du XIX^{ème} s'explique par une attention toute particulière accordée à cette période lors de la première campagne d'agglomération des documents. Les catalogues d'exposition présents concernent essentiellement des expositions de groupe, à l'instar du « Salon des Indépendants » ou des différentes « Biennales », à l'exception de l'exposition monographique de Courbet de 1882. Pour ce qui est des catalogues de ventes de manuscrits, ceux-ci sont principalement produits par la famille Charavay. À ces deux grands jeux de données s'ajoutent quelques pages de romans et pièces de théâtre du XIX^{ème} siècle. On appelle ces documents qui diffèrent du jeu de données principal, l'*In-domain*, de l'*Out-of-domain*. Il s'agit donc de données assez similaires les unes aux autres, de part leur période de production, qui ont cependant une structure distincte. En les intégrant dans le jeu d'entraînement, le modèle produit apprend à les traiter et sera donc capable de s'adapter plus facilement aux variations des données qu'il devra analyser.

8. Celles-ci sont disponibles dans un dépôt github : <https://github.com/katabase/OCRcat>

9. Pour une version détaillée du jeu de données, voir le tableau le tableau des données disponibles en annexe A.2.

type siècle \	Manuscrits	Expositions	Annuaire	Autres	Totaux	Proportions
19ème	276	83	150	15	524	76%
20ème	41	64	0	0	105	15%
21ème	56	0	0	0	56	8%
Totaux	373	147	150	15	685	100%
Proportions	55%	21%	22%	2%	100%	100%

TABLE 3.1 – Description numérique du corpus

Afin de rendre le modèle le plus adaptable possible pour tout type de catalogue du XIX^{ème} siècle, il est par la suite intéressant d'y associer des données similaires issues d'autres projets de recherches. Ainsi, Carmen Brando et Gabriela Elgarrista, du groupe Annuaire du consortium Paris Time Machine, décris précédemment¹⁰, ont accepté de nous fournir leurs jeux de données dans ce but. En effet, les annuaires sont des documents structurés et composés d'entrées, ils se prêtent donc particulièrement à une utilisation dans le cadre de l'élaboration d'un modèle commun de catalogues. Il s'agit de près de cent cinquante pages issues de l'« Annuaire des propriétaires et propriétés de Paris et du département de la Seine¹¹ ». Chaque lettre est représentée dans le corpus par une dizaine de pages.

Toutes ces données correspondent à la vérité terrain, ou *groundtruth*. Ce terme désigne, en apprentissage automatique, les données vérifiées fournies à l'ordinateur. En l'occurrence, pour l'OCR, chaque donnée correspond à un couple, composé d'une image numérisée, ici une page de catalogue, et de sa transcription, corrigée manuellement. Intervenant sur des données ayant déjà été utilisées pour la construction d'un modèle, je n'ai donc pas eu besoin de réaliser et corriger ces transcriptions et ai pu les récupérer directement pour réaliser mes modèles.

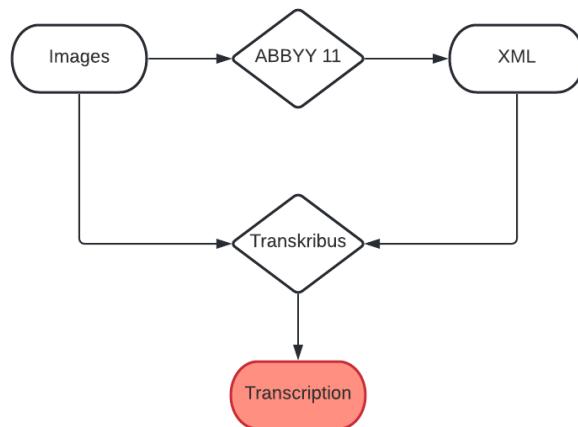


FIGURE 3.2 – Chaîne de production de la vérité terrain

Ces données ont été construites en suivant des chaînes de traitement identiques, ce qui permet d'autant plus de les associer au sein d'un même jeu de données d'entraînement.

10. Pour plus d'informations, voir le 2.1.1

11. Disponible ici : <https://catalogue.bnf.fr/ark:/12148/cb32697229h>

Cette chaîne de production est schématisée par la figure ci-contre. Les pages de catalogues numérisées, au format pdf, jpeg, jpg ou png, sont tout d'abord traitées par ABBYY 11. Logiciel commercial développé depuis 1993 par la société ABBYY¹², c'est un outil fonctionnel, principalement centré sur l'OCRisation de documents modernes d'entreprise. Si son logiciel est intéressant pour la segmentation et la reconnaissance de caractères de documents récents, les méthodes d'entraînement de modèles pour les imprimés historiques ne sont pas très pratiques. Cependant, il est parmi les outils d'OCR les plus faciles quand à la correction manuelle¹³. Il est ainsi conseillé par la documentation Transkribus de préparer les données à transcrire manuellement en les passant tout d'abord dans ABBYY. La transcription est ensuite récupérée sous la forme d'un XML puis le couple image-transcription est ajouté à Transkribus, où un modèle a été entraîné. Les transcriptions récupérées dans le cadre de mon travail sont donc des fichiers qui sont passés par XML ainsi que par Transkribus. Le choix de Transkribus dans tout ces projets s'explique notamment par son utilisation massive dans le cadre de la recherche en humanités numériques, ce qui permet un partage facile des données, comme c'est le cas ici.

3.2.2 Quel format pour ces données ?

Les transcriptions du couple transcription-image peuvent prendre plusieurs formats. Le plus simple correspond au format plein texte, soit « .txt ». Il tend à être remplacé par des formats XML spécialement structurés pour les données d'OCR, à l'instar du PageXML et de l'ALTO.

PageXML (Page Analysis and Ground-truth Elements)

Page¹⁴ est un format XML créé spécialement pour stocker des données OCR, développé par le projet européen PRImA, *Pattern Recognition and Image Analysis*, s'appuyant sur un laboratoire de recherche de la School of Computing Science and Engineering de l'Université de Salford au Royaume-Uni¹⁵. Présenté pour la première fois en 2010, il est utilisé dans de nombreux projets d'OCRisation de documents historiques, tel que Europeana Newspapers¹⁶, projet européen visant à rendre accessible plus de 20 000 pages de journaux, ou encore READ¹⁷, qui s'occupe de Transkribus¹⁸.

Ce format permet non seulement de stocker les transcriptions effectuées, mais également tout le pré-traitement réalisé (binarisation, etc.), ainsi que l'aspect général de la page et son évaluation.

La structure d'un fichier PageXML intègre donc tout ces éléments. Une balise **Metadata** renseigne l'auteur, la date, les étapes de *preprocessing* et toutes autres métadonnées. La transcription en tant que telle est stockée dans une balise **Page** qui donne le nom de l'image associée, ainsi que ses caractéristiques. L'ordre de lecture de la transcription est

12. Accessible ici : <https://www.abbyy.com/>

13. Sami Nousiainen, *Report on File Formats for Hand-written Text Recognition (HTR) Material : CO :OP Community as Opportunity The Creative Archives' and Users' Network*, rapp. tech., National Archives of Finland, 2016, p. 69.

14. Dont la documentation est disponible ici : <https://github.com/PRImA-Research-Lab/PAGE-XML>

15. *Ibid.*

16. <http://www.europeana-newspapers.eu/>

17. Accessible ici : <https://readcoop.eu/>

18. *Ibid.*

également enregistré à l'aide d'une balise **ReadingOrder**. Pour chaque élément transcrit, le fichier stocke les informations textuelles sur plusieurs niveaux :

- **TextRegion** : ce niveau correspond à un groupe de lignes cohérent.
- **TextLine** : ce niveau correspond à une ligne.
- **Word** : ce niveau correspond à un mot.
- **Glyph** : ce niveau correspond à un caractère.

Ces différents niveaux s'emboîtent les uns dans les autres et peuvent être facultatifs, à l'instar de **Glyph** et **Word**. Ils ont tous un identifiant et un attribut faisant référence à leur ordre de lecture. Pour chaque niveau, la transcription est stockée dans une balise **Unicode**, associée à d'autres informations, telles que les coordonnées du niveau, qui se trouvent dans **Coords**. Dans le cas des **TextLine**, une balise **Baseline** est également ajoutée afin d'indiquer les coordonnées du trait qui suit le dessous de chaque caractère composant la ligne. Enfin, sur le niveau le plus bas, une balise **TextStyle** récupère les informations sur le style : la police, la taille des caractères, etc¹⁹.

ALTO (Analyzed Layout and Text Object)

ALTO²⁰ est un format XML développé par le projet européen METAE, *Meta Data engine*, entre 2000 et 2003²¹. Y sont associés quatorze partenaires européens et un partenaire américain, ABBYY. Sa première version, ALTO 1 est présentée en 2004. Il est déployé depuis 2010 par la Bibliothèque du Congrès (United States Library of Congress). Ce format est utilisé par de nombreux projets dont la Bibliothèque du Congrès pour son projet Chronicling America²² et European Newspapers.

ALTO 2 sort en 2014, puis ALTO3 en 2016 et ALTO4 en 2020. Ces versions permettent de donner plus d'informations sur les étapes de *preprocessing* et le *layout* des pages transcrives, mais restent plus pauvres que PageXML.²³

ALTO est structuré en trois grandes parties : la partie **Description** permet de donner des informations sur le fichier transcrit et les étapes du processus réalisé, **Style** décrit les polices et structure des paragraphes (dans le cas de la dernière version d'ALTO) et enfin **Layout** présente le contenu de la page. Cette dernière partie est organisée sous la forme de plusieurs balises **TextBlock**, correspondant aux différentes parties structurant la page. Au sein de ces balises s'emboîtent des **TextLine** qui correspondent à une ligne. Dans ces dernières se trouvent des **String**, où se situent les transcriptions pour chaque ligne, stockées en temps que valeur de l'attribut **CONTENT**. Chacun de ces trois éléments possède également pour attribut un identifiant, **ID**, ses coordonnées et, dans le cas de la **TextLine**, les coordonnées de la baseline²⁴.

Il existe d'autres formats créés pour les données issues des OCR, comme hOCR²⁵ ou ABBYY XML²⁶. Cependant, ALTO et PageXML restent les plus répandus et les plus

19. Un document PageXML annoté, issus des données travaillées et réalisé avec l'aide de Claire Jahan, stagiaire Artl@s est disponible en annexe A.3.

20. La documentation d'ALTO est disponible ici : <https://www.loc.gov/standards/alto/>

21. *Ibid.*

22. Accessible ici <https://chroniclingamerica.loc.gov/>

23. *Ibid.*

24. Un document ALTO annoté, issus des données travaillées et réalisé avec l'aide de Claire Jahan, stagiaire Artl@s est disponible en annexe A.3.

25. hOCR est un format mélant des balises XML et HTML qui fait partie des sorties de plusieurs OCR. Il réutilise des balises HTML dans le but de représenter les informations typographiques.

26. ABBYY XML est un format XML associé à un PDF permettant de stocker les données OCR. Il

utilisés comme format de sortie et d'entrée au sein des logiciels d'OCR²⁷. En effet, l'avantage de ces deux formats est qu'il s'agit de XML, ce qui leur confère nombre de qualités - universalité, facilité de création, d'édition et d'archivage, compacité... - au contraire des autres formats existants. De plus, ils ont été pensés et construits dans le but spécifique de contenir les données issues de l'OCR, ce qui permet de stocker plus d'informations que du PDF ou du texte, tout en se pliant à des standards communs compréhensibles et réutilisables par la communauté scientifique. Ainsi, les données décrites précédemment ont des transcriptions qui ont été sorties de Transkribus sous ces deux formats : PageXML 2013 et ALTO2. Une version ALTO4 est également disponible.

3.3 Les logiciels utilisés

3.3.1 Le choix d'un logiciel OpenSource

	Kraken (associé à eScriptorium)	OCRopus	Tesseract	Calamari	OCR4all
Formats structurés	PageXML-ALTO	hOCR-texte	hOCR-PDF	PageXML	PageXML
Pré-traitement	✓	✓	✓	✓	✓
Segmentation nommée	✓	✗	✗	✗	✓
Documents historiques	✓	✗	✗	✗	✓
Correction Manuelle	✓	✗	✗	✗	✓
Post-traitement	✗	✓	✓	✓	✗

TABLE 3.2 – Comparaison de différents OCR *Open Source*

Le tableau 3.2 présente les différents logiciels *OpenSource* existants pouvant potentiellement remplacer Transkribus et les compare. Tesseract²⁸ est connu comme étant un des OCR les plus performant et fonctionne en lignes de commande dans le terminal, à l'instar d'OCRopus²⁹ et Calamari³⁰. Comme eScriptorium, interface graphique de Kraken, OCR4all³¹ a l'avantage d'être développé spécialement pour le traitement des

est développé par ABBYY, sorti en 2002 et qui fait partie des outputs de son logiciel d'OCRisation.

27. Abdel Belaïd, Yves Rangoni et Ingrid Falk, *Représentation des données en XML pour l'analyse d'images de documents*, fr, text, URL : <http://lodel.irevues.inist.fr/cide/index.php?id=147> (visité le 27/08/2021).

28. Développé par les laboratoires Hewlett-Packard de 1985 à 1994, Tesseract devient OpenSource en 2005 avant d'être racheté par Google en 2006. (<https://github.com/tesseract-ocr/>)

29. Maintenant appelé OCRopy, moteur OCR lancé en 2007 par Thomas Breuel, du *Deutsches Forschungszentrum für Künstliche Intelligenz*, avec le soutien de Google.

30. Moteur OCR récent, lancé en 2018, et basé sur OCRopus et Kraken. Son code est accessible ici : <https://github.com/Calamari-OCR>.

31. Projet de l'université de Wurzburg en Allemagne lancé en 2019. Il s'agit d'un moteur d'OCR spécialisé en documents historiques et associé à une interface graphique permettant d'utiliser l'OCR sans avoir des compétences en informatique. Il intègre les logiciels Calamari et Kraken et a pour ambition d'y ajouter Tesseract.

documents historiques et de proposer la possibilité d'accéder aux données via des GUI, *Graphical User Interface*³². Grâce à ces interfaces graphiques, il est possible de corriger directement et manuellement les erreurs de segmentation. Un des avantages principaux de Kraken, Calamari et OCR4all, trois logiciels issus de la même base, est de supporter des formats XML propres à des données issues de l'OCR, tel que PAGEXML et ALTO. Comme expliqué précédemment, cela permet de stocker plus d'informations sur les pages transcrrites tout en ayant des documents standardisés et partageables. Si chaque logiciel offre la possibilité de réaliser le pré-traitement au sein de celui-ci, ce n'est pas le cas du post-traitement. Ainsi, Kraken et OCR4all ne possèdent pas de modèles de langue, qui permettent de vérifier la qualité des mots reconnus par l'OCR et de les corriger.

Nombre d'expériences de comparaison de la qualité des résultats de chacun de ces logiciels signalent la prévalence de Tesseract³³. Cependant, l'impossibilité de récupérer des données structurées standardisées pour l'OCR et de corriger manuellement les résultats empêche d'employer directement cet outil. Face à ce résultat, il faut alors trouver un compromis entre un meilleur résultat d'OCR et l'application des principes de partage et de structuration de la Science Ouverte. Autre point non négligeable, Kraken permet, au contraire des autres logiciels, la récupération des modèles d'OCR. Cela permet à la fois de conserver les entraînements réalisés et ne pas avoir à les effectuer de nouveau, au contraire de ce qu'il se passe avec la migration depuis Transkribus, et de permettre leur distribution à la communauté scientifique. Si la performance est moindre, elle s'adapte donc aux enjeux de la Science Ouverte. De plus, le développement de eScriptorium présente aussi l'intérêt d'être utilisé par une communauté grandissante ce qui permet d'assurer sa pérennisation et son développement dans le temps. Chaque entraînement de modèle et ajout de données au sein du logiciel permet d'améliorer les résultats généraux de celui-ci. Ainsi cette plateforme jeune est amenée à grandir et ses performances à se perfectionner. D'où le choix de migrer les données de Transkribus vers eScriptorium³⁴ et Kraken, un outil entièrement *Open Source*, bénéficiant à la fois d'une interface graphique et de lignes de commandes plus malléables et permettant de récupérer l'intégralité des données et les modèles dans des formats interopérables et standards.

Kraken

Kraken est un moteur OCR et analyseur de disposition *OpenSource* développé en python³⁵ sous la licence *Apache License 2.0*³⁶. Cette idée d'*OpenSource* permet à la fois de la transparence, de par le partage du code à tous, mais également de s'assurer une collaboration dans le développement du logiciel. Il se base sur le projet OCropus³⁷, logiciel

32. Il s'agit d'un logiciel qui permet l'affichage sous une forme graphique des actions proposées par l'outil. Ici concrètement, pas besoin de lignes de commande pour réaliser des entraînements.

33. Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner et Frank Puppe, « OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings », *Applied Sciences*, 9–22 (janv. 2019), p. 4853, DOI : 10.3390/app9224853 et F. Burgy, S. Gerson et L. Schüpbach, *Ex imagine ad litteras : Projet d'océsiration de la collection De Bry...*

34. Pour ce qui est d'eScriptorium, le travail a été réalisé sur la version web de l'outil, en utilisant les serveurs de l'INRIA : <http://traces6.paris.inria.fr/>. Cependant il est également possible de télécharger et installer le logiciel directement sur l'ordinateur.

35. Son code est entièrement accessible sur le github de son créateur : <https://github.com/mittagessen/kraken>

36. Accessible ici : <https://www.apache.org/licenses/LICENSE-2.0.html>

37. Accessible ici : <https://github.com/ocropus/ocropy>

libre de reconnaissance de caractères développé depuis 2007 par le Centre de Recherche Allemand pour l'Intelligence Artificielle (DFKI), avec l'aide de Google, dans le but de permettre la numérisation à grande échelle de livres³⁸. À cela, Kraken entend aller plus loin puisqu'il donne accès à la fois au programme de reconnaissance de caractères mais également aux modèles. Le logiciel est développé par Benjamin Kiessling, ingénieur de recherches à l'EPHE, l' École Pratique des Hautes Études. Si son cadre de production est tourné vers des documents historiques en hébreu et en arabe, il est cependant utilisable sur des textes en écriture latine.

Il se base sur un réseau neuronal qui reconnaît les lettres d'une image numérisée dans une page déjà segmentée. Ainsi, cet OCR fait partie de ceux nécessitant un modèle de segmentation et un modèle de reconnaissance de caractères. La segmentation et la transcription de documents est réalisable en utilisant les modèles par défaut du logiciel. Il est aussi possible de créer et entraîner ses propres modèles, permettant ainsi à l'utilisateur d'obtenir le résultat le plus performant possible³⁹. Un des intérêts de Kraken est de pouvoir gérer les différents paramètres d'entraînement, notamment en manipulant le réseau de neurones, mais également de pouvoir récupérer ces modèles en format *.mlmodel*. Ce type de modèle est partageable, réutilisable et particulièrement employé dans le cadre du *machine learning*.

L'utilisation de Kraken se base exclusivement sur des lignes de commande, réalisées en bash⁴⁰ dans le terminal. Ainsi, cet outil nécessite un système d'exploitation linux et quelques compétences en bash afin de le faire fonctionner. Kraken peut prendre en entrée tous types de formats d'images ainsi qu'un certain nombre de formats XML pour la transcription (ALTO, PageXML, hOCR...), qui sont également récupérables en sortie.

eScriptorium

eScriptorium⁴¹ est développé dans le cadre du programme « Scripta-PSL. Histoire et pratique de l'écrit. » de l'Université Paris-Sciences-Lettres (PSL)⁴². Il est dirigé par Daniel Stökl Ben Ezra⁴³ et Peter Stockes⁴⁴. eScriptorium a pour but principal de rendre possible la transcription de tous types de documents par le biais de l'apprentissage automatique. Dans la pratique, cela se traduit notamment par la réalisation d'une application, interface graphique *OpenSource* à l'outil Kraken.

eScriptorium permet également d'ajouter en entrée des images au format IIIF⁴⁵. Signifiant *International Image Interoperability Framework*, cela désigne à la fois une communauté et les spécifications techniques qu'elle établit dont l'ambition est de définir un

38. Cela concerne Google Books, Internet Archives et des projets de bibliothèques.

39. Benjamin Kiessling, « Kraken - an Universal Text Recognizer for the Humanities », *Digital Humanities* (, 2019), URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html> (visité le 27/08/2021).

40. Bash, pour *Bourne-Again Shell*, est un interpréteur de lignes de commandes (*shell*), permettant à l'utilisateur d'accéder, par le biais d'une interface texte dans le terminal, aux différents services que propose l'ordinateur.(<https://doc.ubuntu-fr.org/shell>)

41. Son code source est disponible dans le dépôt <https://gitlab.inria.fr/scripta/escriptorium>

42. Pour plus d'informations, le site du projet : <https://gitlab.inria.fr/scripta/escriptorium> et le carnet de recherches d'eScriptorium : <https://escripta.hypotheses.org/1>

43. Directeur d'études à l'EPHE en sciences historiques et linguistiques s'intéressant à la philologie et la linguistique de l'hébreu et de l'araméen ancien.

44. Directeur d'études en humanités numériques et computationnelles appliquées à l'écrit ancien dans la même école.

45. <https://iiif.io/>

cadre d'interopérabilité pour la diffusion et l'échange d'images sur internet. Ces recommandations sont suivies par de nombreuses bibliothèques et institutions patrimoniales, à l'instar de la BNF, ce qui permet de récupérer un grand nombre d'images uniquement grâce à un url de manière aisée.

eScriptorium a plusieurs avantages. En tant qu'application graphique de Kraken, il donne la possibilité de réaliser les mêmes actions mais de façon plus pédagogique et intuitive, sans nécessiter de compétences informatiques particulières. Ainsi, il prend en entrée des images de tout type de formats, associées à leur transcription en PageXML ou Alto, donne accès à une interface visuelle permettant la correction manuelle de la segmentation et de la reconnaissance de caractères, entraîne et lance les différents modèles réalisés et extrait les données obtenues en texte, ALTO ou PAGE et les modèles. Cela rend plus aisément le travail de préparation manuelle et de correction de la vérité terrain. Il permet aussi de partager les documents plus facilement à l'aide de collections, favorisant la collaboration.

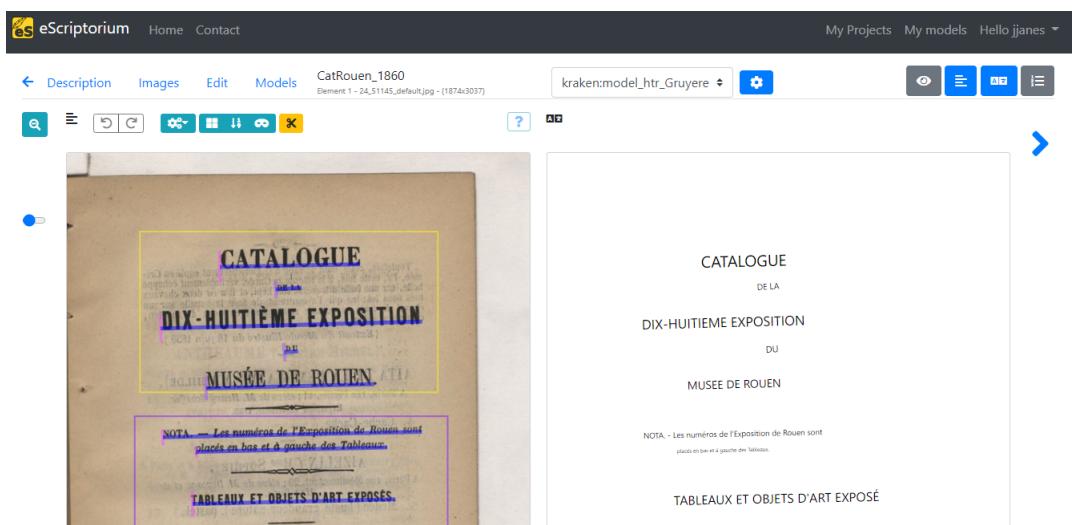


FIGURE 3.3 – Capture d'écran de l'interface eScriptorium

L'image 3.3 est une capture d'écran d'eScriptorium, qui permet de comprendre les différentes possibilités qu'offre une interface graphique d'OCR. Ici, on peut voir l'image OCRisée, à gauche, la segmentation réalisée, sous la forme de lignes violettes et de polygones, et la transcription correspondante à droite. Il est également possible de voir des informations précises sur le document traité dans l'onglet **Description**, l'intégralité des images traitées dans **Images** ou encore les modèles de segmentation et transcription produits et utilisés dans **Modèles**. L'onglet **Description** rend possible l'ajout d'un vocabulaire précis de description des zones et lignes, que l'on peut ensuite appliquer sur la segmentation. **Images** donne aussi l'opportunité de réaliser des étapes de post-traitement (binarisation par exemple), d'appliquer les différents modèles sur toutes les images ou une partie des images du corpus et de télécharger les transcriptions obtenues dans le format de son choix (alto, texte ou page).

3.3.2 De Transkribus à eScriptorium : Organisation de la migration des données

Ainsi il convient donc maintenant de présenter la migration des données réalisée de Transkribus vers eScriptorium. Celle-ci est décrite par le schéma⁴⁶ ci-joint, qui présente toute la chaîne de production par laquelle sont passées les données.

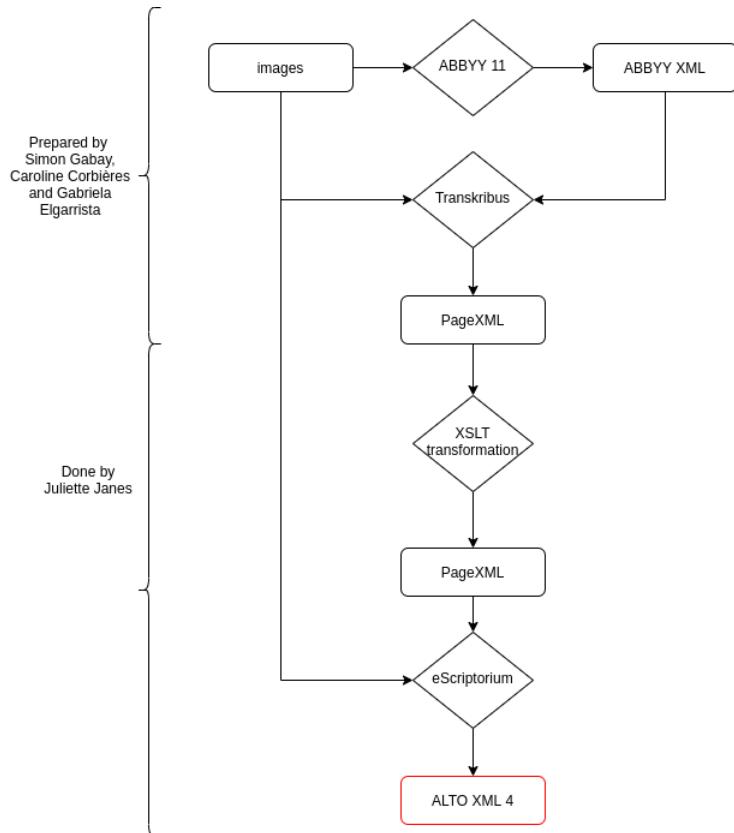


FIGURE 3.4 – Chaîne de production complète des données de la vérité terrain

Dans un premier temps, il a fallu décider quel format utiliser pour la migration Transkribus-eScriptorium. Transkribus permet d'obtenir en sortie de l'ALTO2, du PageXML ou du PDF tandis qu'eScriptorium prend en entrée ALTO4 et PageXML. Ainsi, afin de passer un fichier Alto issu de Transkribus dans eScriptorium, il faut recourir à une chaîne de traitement spécifique. Le programme pdfalto⁴⁷ de Patrice Lopez permet de transformer un document PDF en ALTO3 puis le programme python d'Alix Chagué Aspyre⁴⁸ permet de transformer ce fichier ALTO3 en ALTO4 et de l'insérer dans eScriptorium. Si le fichier est restructuré afin de convenir aux besoins d'eScriptorium, il n'en reste pas moins qu'il est basé uniquement sur des données issues d'Alto2. Ainsi, les informations contenues à l'issue de ce travail restent assez pauvres, au contraire d'un fichier PageXML. Or, jusqu'à il y a peu, eScriptorium n'avait pas la possibilité de prendre en entrée ce format. Le développement d'une entrée pour les transcriptions PageXML dans eScriptorium est relativement récent et intervient donc dans le but de faciliter le travail de migration

46. Ce schéma est issu du dépôt présentant le travail de production de données : <https://github.com/Juliettejns/cataloguesSegmentationOCR>

47. <https://github.com/kermitt2/pdfalto>

48. <https://github.com/alix-tz/aspyre-gt>

depuis Transkribus. Cela permet d'utiliser un format contenant plus d'informations et également d'éviter de perdre trop de données.

Cependant, une fois les images et transcriptions au format PageXML ajoutées à eScriptorium, il y a un problème. Les transcriptions ne sont en effet pas corrigées. Ainsi, on se retrouve avec des fautes dans la transcription enregistrée par eScriptorium. De ce fait, comme expliqué précédemment, différentes transcriptions sont stockées dans le fichier Page, sur plusieurs niveaux (régions, lignes, mots...). Les corrections faites sur les transcriptions dans ABBYY ont été enregistrées dans les niveaux régions et lignes, tandis que la transcription récupérée par eScriptorium est située dans le niveau mot. Ainsi, il s'agit d'une transcription erronée sans correction manuelle. Pour corriger cela, nous avons réalisé une feuille de transformation XSLT⁴⁹, qui permet de remplacer les transcriptions du niveau mot par celles au niveau de la ligne. On obtient alors une structure de ce type⁵⁰ :

```
<TextRegion type="paragraph" id="r_2_1" custom="readingOrder {index:1;}">
  <Coords points="437,471 1012,471 1012,522 437,522" />
  <TextLine id="tl_2" primaryLanguage="English" custom="structure {type:default;} readingOrder {index:0;}">
    <Coords points="438,472 1011,472 1011,521 438,521" />
    <Baseline points="438,509 445,509 488,510 558,499 779,508 862,508 918,507 1011,506" />
    <Word id="w_w2aab1b3b2b1blab1">
      <Coords points="438,472 1011,472 1011,521 438,521" />
      <TextEquiv>
        <Unicode>19. – Baigneuse vue de dos.</Unicode>
      </TextEquiv>
      <TextStyle fontFamily="Times New Roman" fontSize="11.0" />
    </Word>
  </TextLine>
</TextRegion>
```

FIGURE 3.5 – *Catalogue[..] Courbet*, 1882, p.2

Une **TextLine** ne possède alors plus qu'une seule balise **Word**, contenant l'intégralité de la transcription. On perd donc les données sur des mots précis, que ce soit les transcriptions comme les coordonnées des mots dans la page. Il aurait pu être intéressant, pour perdre moins d'informations, de recorriger les données directement dans eScriptorium. Cependant, cela aurait nécessité beaucoup de temps et d'énergie. Le choix a donc été fait de transformer les données avant leur intégration dans eScriptorium, au risque de perdre de l'information. À noter que, contrairement à Kraken, qui traite les transcriptions jusqu'au niveau **Glyph**, eScriptorium s'arrête au niveau **TextLine**. Ainsi, les données perdues en entrée ne seraient de toute façon pas récupérées en sortie d'eScriptorium.

Ce dernier point permet donc de conclure que les données produites en sortie d'eScriptorium auront autant d'informations stockées en Alto qu'en PageXML. Ainsi il a été décidé d'extraire ces données en ALTO4, de par leur standard et leur facilité de lecture à l'œil nu.

49. Cette feuille de transcription a été réalisée en collaboration avec Claire Jahan et est disponible dans notre dépôt commun, avec de plus amples informations techniques : https://github.com/Heresta/BAO_Score_DH_ENS_2021/tree/main/RetraitTranscriptionPAGEXML

50. Cette image est issue du fichier PageXML transformé récupérable ici : https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/1_Data/Cat_exhibition/first_data/page_transforme/Cat_Courbet_1882-1_typo_0002.xml

Chapitre 4

L’entraînement d’un modèle commun de reconnaissance de texte pour les catalogues

Une fois cette description de la migration des données à disposition de Transkribus à eScriptorium faite, il convient donc de s’intéresser à la production de jeux d’entraînement. Comment les construire de façon à représenter au mieux la variété d’aspect des catalogues et autres documents semi-structurés, et réaliser des modèles aptes à traiter la diversité des formes que prennent ces données ? En utilisant Kraken, à la différence de Transkribus, il s’agit non pas de réaliser un unique modèle d’OCR mais d’entraîner un premier modèle dit « de segmentation », qui analyse la mise en page du document, puis un modèle de reconnaissance de caractères. Ainsi, nous allons réfléchir ici aux nouvelles possibilités qu’entraîne cette division tout en détaillant le processus de création de jeux de données, l’entraînement et le choix des modèles ainsi que leur mise en accès libre, ouvert et gratuit suivant les principes de la Science Ouverte.

4.1 Réalisation de modèles de segmentation.

4.1.1 L’initiative SegmOnto : un vocabulaire commun pour le nommage des zones

Comme mentionné précédemment, la segmentation, soit l’analyse de la structure d’une page, *Layout Analysis*, fonctionne avec un système de zones et lignes. Ainsi qu’il est possible de l’observer sur l’image 4.1, les lignes correspondent au dessous des caractères, tandis que les zones, dont le contour est ici représenté en rouge, encadrent aléatoirement des groupes de textes de type paragraphes. Comme visible ici, les zones n’ont pas d’utilité propre en dehors du fait qu’elles permettent la reconnaissance des parties de la page où il y a du texte. Or, eScriptorium ainsi que les formats Alto et PageXML offrent la possibilité de nommer ces différentes zones et lignes. Il peut donc être intéressant d’entraîner un modèle capable non seulement de détecter les zones de texte mais également de reconnaître les différentes parties de la page. Stockée en sortie par ALTO et PAGEXML, cette information permettrait un traitement plus poussé du résultat.

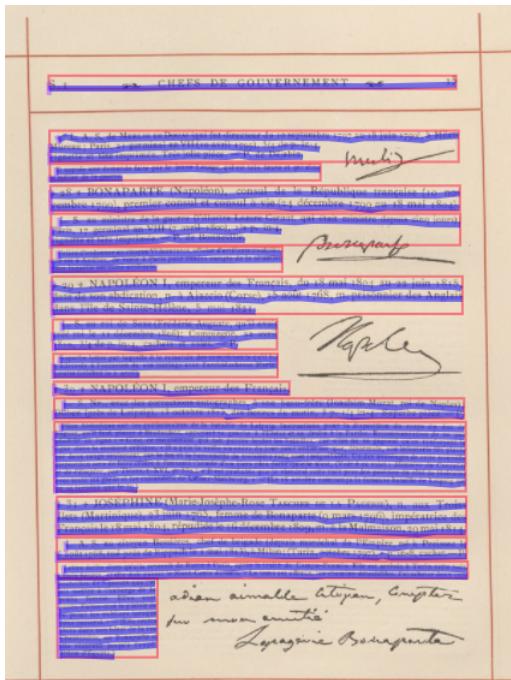


FIGURE 4.1 – Catalogue Bovet, 1887, p.79

Cette idée est à la base de l’initiative SegmOnto. Groupe de travail associant plusieurs projets d’humanités numériques d’OCRisation de documents historiques¹, elle vise, depuis 2020, à mener une réflexion commune autour de l’élaboration de normes et d’un vocabulaire pour la segmentation. En est issue une première version², permettant de typer et structurer une page en fonction des éléments qui la compose. Associant des chercheurs issus d’une grande variété de domaines, l’objectif du vocabulaire est de permettre la description de documents diverses, allant de manuscrits médiévaux aux imprimés du XIX^{ème} siècle. Cette recherche d’une certaine universalité oblige donc les termes employés à être les plus génériques possibles. L’idée finale est de permettre, par le biais de ce vocabulaire commun, le partage des données produites afin de réaliser des grands jeux de données sans problèmes. Cela permettrait d’entraîner des modèles de segmentation, capables d’analyser la structure d’une page, beaucoup plus performant.

Ce travail ambitionne de réaliser une ontologie, d’où son nom. Ce terme définit « un vocabulaire commun pour les chercheurs qui ont besoin de partager l’information dans un domaine³ ». Par rapport à un simple vocabulaire commun standardisé, une ontologie cherche à décrire de manière formelle tout un domaine de connaissance, ici le traitement des documents historiques en OCR, en identifiant les types d’objets qui le composent, leurs propriétés et relations. Ces données sont le plus souvent structurées en RDF,⁴ un langage extensible de représentation de connaissances appartenant à la famille des langages du

1. En font partie des chercheurs de l’École nationale des Chartes, d’Almanach à l’INRIA, de l’Observatoire de Paris ou encore e-editions.

2. Le dépôt du projet, contenant documentation et exemples, est disponible ici :<https://github.com/SegmOnto>

3. BnF-Professionnels : Normes, formats, modélisation - Elaboration d’ontologies [en ligne]. Disponible sur : <http://www.bnf.fr/PAGES/infopro/normes/no-acOntologies.htm> (consulté le 19/01/2010)

4. *Ressource Description Framework*. Pour plus d’informations : <https://rbdd.cnrs.fr/local/cache-zotspip/GZ5TD2GB/TP%20Introduction%20aux%20ontologies%20RDFS%20WL%20inferences.pdf>

web sémantique et fournissant des éléments de base pour la réalisation d'ontologies ou vocabulaires.

Zones		Lignes
Damage	Numbering	Default
Decoration	Running Title	DropCapitalLine
DropCapital	Seal	Interlinear
Figure	Signature	MusicLine
Main	Stamp	Rubric
Margin	Table	
MusicNotation	Title	

TABLE 4.1 – Terminologie SegmOnto

Les zones

Comme visible dans le tableau 4.1, SegmOnto a répertorié quatorze types de zones, à partir de ses propres exemples. Si certaines d'entre elles sont encore à définir d'autres sont déjà bien décrites. C'est le cas de *Main*, qui signale les corps de texte. Il s'agit de la zone la plus fréquente dans les documents, puisqu'elle se prête assez bien à tout type de données. Il est possible qu'elle soit en plusieurs parties, dans le cas de corps de texte sur plusieurs colonnes, ou qu'elle contienne d'autres zones. Celles-ci, telles que *Numbering*, pour la pagination, *Margin*, pour les éléments hors du corps du texte, *Damage*, pour les parties abîmées, et *Running Title*, pour les titres courants, sont également particulièrement fréquentes et génériques et ne posent pas vraiment de problèmes quand à leur définition.

Au contraire, certains éléments demandent encore des définitions. C'est le cas des zones *Decoration* et *Figure*, qui tendent à se confondre. Cependant, d'après leur définition, il est possible de distinguer des différences entre elles. *Decoration*⁵ serait alors employée pour signaler les ornementations, tandis que *Figure*⁶ correspondrait plutôt à l'illustration scientifique d'un texte. Ainsi, la première zone comprend les données de type décoratives, que l'on peut retrouver dans un roman ou une pièce de théâtre par exemple, tandis que la seconde s'attache à décrire une donnée précise présentée sous la forme d'une image. Ce problème se retrouve également dans la description de la zone *Title*. D'après la définition donnée⁷, on pourrait y associer le titre du document ainsi que les titres des différentes parties du document (par exemple les titres de parties, sous-parties, chapitres, etc.). Cependant, cela entraînerait une confusion dans l'apprentissage automatique réalisé par l'ordinateur, de par la diversité de données différentes pour une seule et même zone. Ainsi, celle-ci peut plutôt être comprise comme signalant le titre du document général et donc permettre à l'ordinateur de détecter la page de titre et bien la différencier des autres pages.

Enfin, les dernières zones sont spécifiques à des documents et donc un peu moins génériques que les précédentes, et plus porteuses de sens. C'est le cas de *DropCapital*, pour

5. Sa définition est disponible ici :<https://github.com/SegmOnto/examples/blob/main/zones/Decoration/Decoration.md>

6. Sa définition est disponible ici :<https://github.com/SegmOnto/examples/blob/main/zones/Figure/Figure.md>

7. Accessible ici :<https://github.com/SegmOnto/examples/blob/main/zones/Title/Title.md>

les initiales ornées, *MusicNotation*, pour les portées, *Seal*, pour les sceaux, *Signature*, pour les signatures, *Stamp*, pour les cachets ou encore *Table* pour les tableaux.

Les lignes

Les lignes sont également typées à la manière des zones, comme on peut le voir dans le tableau 4.1. La majorité d'entre elles correspondent à des lignes *Default*. Il s'agit de lignes sans particularité quelconque, qui ne sont pas situées dans des zones induisant un typage spécifique des lignes.

Certaines lignes donc spécifiques à un type de zone. C'est le cas de *DropCapitalLine*, qui permet de typer une ligne située dans une *DropCapital*, soit une initiale. La même situation intervient pour une ligne typée *MusicLine*, qui correspond à une portée, et est donc située dans une zone *MusicNotation*.

Enfin, certains typages particuliers de lignes peuvent intervenir au sein de zones non spécifiques, par exemple des *Main*. C'est le cas des lignes *Rubic*, qui correspondent aux titres au sein du corps du texte, soit des titres de chapitres, de scènes, d'actes. *Interlinear* caractérise, comme son nom l'indique, des lignes qui sont en dehors des lignes *Default*.

La décision d'utiliser ce vocabulaire spécifique pour analyser puis nommer la structure des pages dans eScriptorium permet donc d'obtenir un *dataset* qui se plie aux exigences communes d'un groupe de travail. Cela donne la possibilité de partager des données plus facilement, notamment pour entraîner des modèles plus performants, mais aussi de développer un jeu de données dont le nommage des zones a été documenté et va probablement se pérenniser sous la forme d'un vocabulaire contrôlé qui pourrait être un standard dans la description de documents historiques en HTR. Enfin, SegmOnto étant encore récent, mon travail intervient à un stade où ce vocabulaire est dans un état théorique. La réalisation d'un jeu de données se basant sur celui-ci permet donc de l'appliquer pour la première fois et tester ses limites et ses avantages, ainsi que ses fonctionnalités.

4.1.2 Préparation des données : application et adaptation de ce vocabulaire aux catalogues

Ainsi, l'application de SegmOnto sur nos données correspond à une des premières applications de ce vocabulaire. Il convenait donc dans un premier temps de l'adapter et le tester. Le tableau 4.2 répertorie les différentes zones utilisées sur les données. Ainsi, la plupart des lignes du corpus correspondent à des lignes *Default*, puisque s'agissant d'imprimés, elles sont très structurées et ne présentent pas de particularités. Si certaines ont rarement été utilisées, à l'instar de *Title*, *Stamp* ou *Figure*, d'autres correspondent à la plupart des données rencontrées au cours du traitement du corpus. C'est le cas de *Numbering*, *RunningTitle* et bien entendu *Main*. À ces différents termes s'ajoutent deux nouveaux éléments : *Entry* et *EntryEnd*.

Ces deux nouveaux types, spécifiques à notre travail, ont été créés afin de coller au mieux aux données étudiées. En effet, les catalogues et autres documents qui composent notre corpus ont tous une structure sous la forme de répertoires d'entrées. Aucun terme décrit dans SegmOnto ne correspond à ce type d'éléments. Cependant, l'ajouter à notre travail permet d'améliorer significativement la qualité des données obtenues en sortie d'HTR. En effet, elle donne la possibilité de récupérer les coordonnées précises de chacune des entrées, rendant le travail d'extraction d'informations beaucoup plus facile par la suite.

Zones	
<i>Main</i>	Corps du texte
<i>Title</i>	Titre du document
<i>Numbering</i>	Pagination
<i>Running Title</i>	Titre courant
<i>Figure</i>	Image
<i>Margin</i>	Texte en dehors du corps du texte
<i>Stamp</i>	cachet ou tampon
<i>Entry</i>	Entrée de catalogues (ajoutée)
<i>EntryEnd</i>	Fin d'entrée de catalogues (ajoutée)
Lignes	
<i>Default</i>	Ligne par défaut

TABLE 4.2 – Terminologie utilisée sur les données

Ainsi, *Entry* a été créée et associée aux termes issus de SegmOnto dans le but de définir spécifiquement une entrée de catalogue ou d'annuaires. Certaines entrées commençant sur une page et terminant sur la suivante, il a été décidé d'ajouter un second élément, *EntryEnd*, afin de décrire ces fins d'entrées⁸.

Cet ajout d'une nouvelle classe au sein de SegmOnto est à l'encontre du principe même de « contrôlé ». Cependant, on réalise là des tests permettant de déterminer la viabilité et la reproductibilité du vocabulaire. Ainsi, l'appliquer sur des jeux de données précis permet de réaliser ses limites et ses possibilités. Ici, on teste donc l'éventualité d'un nouveau type de zone, qui permettrait de signaler un objet particulier spécifique au document traité. En l'occurrence, il s'agit là des entrées, zone au centre de notre travail sur les données semi-structurées, qu'il est donc primordial de représenter lors de l'analyse de la mise en page. En effet, plutôt que d'ajouter **Entry** au vocabulaire, ce qui n'aurait de sens que pour les documents de type liste, pourquoi ne pas ajouter une zone « privée » que chaque projet pourrait utiliser à sa guise selon les données traitées. Dans le cas de pièces de théâtre, il pourrait s'agir d'une réplique, pour des poèmes, d'une strophe, etc. Cela permettrait non seulement d'ajouter une sous-division, plus précise, donnant une profondeur supplémentaire à l'analyse de la mise en page mais également de permettre une véritable application du vocabulaire pour chaque type de documents. Afin de conserver un vocabulaire commun et permettre le partage des données plus facilement, il conviendrait donc de trouver un terme précis permettant de représenter ce type de zones et éviter une profusion de noms. **Subdivision**, par exemple, permettrait de mettre l'accent sur la notion de profondeur et pourrait s'appliquer à de nombreux documents différents. Ainsi, au sein d'un projet précis, on signalerait que **Subdivision** désigne les strophes, entrées, répliques. Cela permet de garder un cadre standard tout en donnant une certaine plasticité au vocabulaire, l'adaptant à tout type de données à un niveau plus poussé.

8. Des exemples sont disponibles en annexes, B.1.1. Ils permettent de se faire une idée de l'aspect des entrées pour chaque type de donnée du corpus.

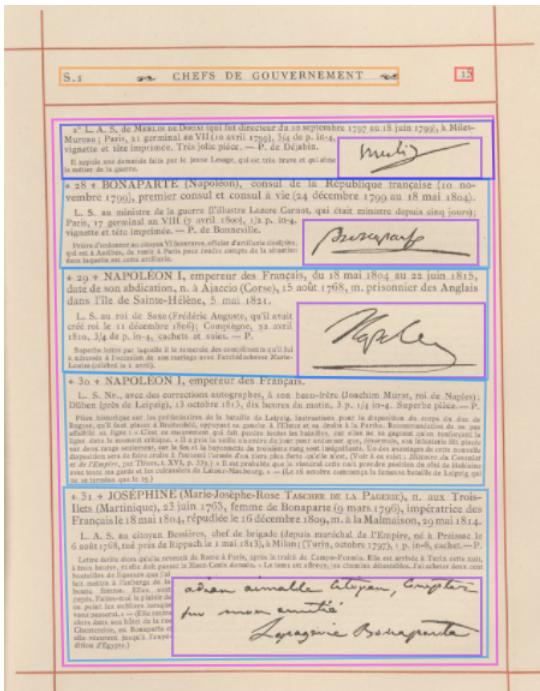


FIGURE 4.2 – *Catalogue Bovet*, 1887, p.79, sortie eScriptorium

La figure 4.2 présente le résultat obtenu en appliquant les recommandations de SegmOnto⁹ sur la page de l'image 4.1. Comme il est possible de le remarquer sur celle-ci, les pages en entrée d'eScriptorium utilisent la segmentation¹⁰ uniquement pour déterminer où se trouvent les données textuelles. L'image présente ici la nouvelle structuration de la segmentation à la suite de l'application des principes de SegmOnto¹¹. La figure 4.3¹² est une image issue du jeu de données de pièces de théâtre du XVII^{ème} réalisé par Claire Jahan dans le cadre de son stage à Artl@s¹³. Elle se base également sur SegmOnto¹⁴. Les deux genres de documents emploient des zones similaires, telles que *Main*, *Numbering* ou encore *Running Title*, non présent dans l'exemple ci-contre. Dans le cas des pièces de théâtre, d'autres éléments sont utilisés, tels que *Decoration*, *Signature* ou encore *Drop Capital*, ainsi que les lignes *Drop Capital* et *Rubric*. Au contraire, les catalogues, quant à eux, emploient les éléments *Entry* et *EntryEnd* afin de décrire au mieux les données. Ainsi, le vocabulaire de SegmOnto est capable de finement décrire deux types de données extrêmement différentes, bien que nécessitant l'ajout d'un nouvel élément *Entry*. Cela permet de voir l'utilité de ce vocabulaire en ce qu'il permet de véritablement analyser l'organisation de la page et nommer les différents éléments qui la composent.

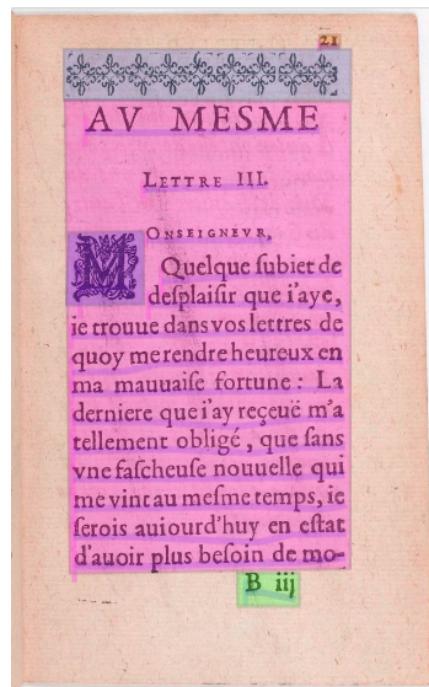


FIGURE 4.3 – *Lettre du Sieur de Balzac*, 1624 p.21

-
9. D'autres exemples sur les différents types de données du corpus sont disponibles en annexes, B.1.2.
10. Les zones sont entourées par un contour rouge.
 11. En rose : *Main*, en orange : *Running Title*, en rouge : *Numbering*, en violet : *Figure*, en bleu ciel : *Entry*, en bleu foncé : *EntryEnd*.
 12. *Lettres du Sieur de Balzac*, 1624, p.21, <https://gallica.bnf.fr/ark:/12148/btv1b86262420/f41.item>
 13. Son dépôt de travail est disponible ici : <https://github.com/e-ditiones/OCR17plus>
 14. En rose : *Main*, en orange : *Numbering*, en bleu : *Decoration*, en vert : *Signatures*, en violet : *DropCapital*.

4.1.3 Une preuve de concept pour le nommage des zones

Trois modèles de segmentation¹⁵ ont été réalisés, au fur et à mesure de la production de données issues d'eScriptorium. Le premier d'entre eux a été entraîné directement sur eScriptorium afin d'avoir une idée de la quantité de travail à fournir. Les modèles suivants ont été entraînés directement avec Kraken¹⁶.

Chaque jeu d'entraînement est traditionnellement divisé en plusieurs sous-jeux. Un premier, appelé *train*, contient le plus grand nombre des données et correspond à ce qui est fourni à l'ordinateur pour apprendre à reconnaître les différentes segmentations. Un second, beaucoup plus petit, comprenant le plus souvent environ 10% des données, est appelé *val*. Il permet à l'ordinateur de s'entraîner dessus et, en comparant avec le résultat fourni, de mesurer la qualité du modèle produit. Enfin, un dernier jeu de données, facultatif et de taille équivalente au *val*, *test* correspond à des pages qui permettent à l'utilisateur de tester directement le modèle. Le plus intéressant est de garder un *test dataset* fixe pour tous les entraînements de modèles réalisés, afin de pouvoir comparer les meilleurs modèles de chaque entraînement entre eux, et pouvoir déterminer le meilleur de tous.

Le premier jeu de données utilisé pour réaliser le premier modèle produit, **Abondance**, est composé de 30 pages. Il est constitué de 10 pages de catalogues d'exposition, 10 pages de catalogues de ventes de manuscrits et 10 pages de l'annuaire de 1898. L'idée était de produire un jeu représentatif du corpus. Ainsi, chaque période et chaque type de catalogue est représenté au sein de ce mini-corpus. Un set d'entraînement a été réalisé à partir de ce jeu de données, composé de 10% de *data test*, soit 3 pages, 10% de *data val*, soit 3 pages et 80% de *data train* soit 24 pages. Cela a permis d'entraîner une série de modèle dont le meilleur a pour taux d'*accuracy*, c'est-à-dire de précision, 62%. Ce résultat est très bas et l'application de ce modèle sur le set de données *test* dans eScriptorium l'indique clairement, puisqu'il est impossible pour lui de reconnaître les différentes zones.

Le deuxième modèle, **Beaufort** a été réalisé à partir d'un jeu de données de 150 pages. Il est constitué de 50 pages de catalogues d'exposition, 50 pages de catalogues de ventes de manuscrits et 50 pages issues de l'annuaire. Ici encore, le but est d'avoir un panel de données représentatives des données de travail disponible. Un set d'entraînement a été réalisé en conservant les mêmes proportions que pour **Abondance**. Le *data test* est composé des trois images *test* d'**Abondance**, complété de 12 autres données afin d'obtenir les 10%. Le modèle obtenu possède une accuracy de 69%. Appliqué dans eScriptorium sur les données *test*, il permet de visualiser une nette amélioration des résultats : si la plupart des zones ne sont pas reconnues, il n'en reste pas moins que *Main*, pour le corps du texte, ainsi qu'*Entry* et *EntryEnd*, pour les entrées, sont particulièrement bien ciblées par le modèle. Ainsi, **Beaufort** fournit un premier aperçu du résultat et est en quelque sorte une preuve de concept de l'initiative SegmOnto. Un modèle, entraîné sur des données segmentées et nommées, est capable de reconnaître et analyser une page et sa structure et de nommer les différents éléments qui la compose.

15. Ceux-ci sont disponibles ici : https://github.com/Juliettejns/cataloguesSegmentationOCR/tree/main/4_Models/Segment. Pour un aperçu de leur application sur des pages de catalogues, Voir en annexe (B.1.3).

16. Ce travail est réalisé par un GPU, unité de traitement graphique, plus puissante qu'un CPU, unité centrale de traitement, que contiennent les ordinateurs. Dans l'attente de la construction de l'infrastructure genèvoise, les commandes ont été lancé sur le GPU de l'École nationale des Chartes par Jean-Baptiste Camps, en sa qualité de membre du groupe de travail SegmOnto et Thibault Clérice, directeur de ce mémoire.

Chaource, troisième modèle produit, est issu d'un jeu de données de 277 pages¹⁷. Il est composé de 50 pages d'annuaire, 97 pages de catalogues de ventes de manuscrits et 130 pages de catalogues d'exposition. Le choix de cette répartition des données a été décidé suite au résultat obtenu lors de l'application du modèle **Beaufort**. En effet, celui-ci a permis de récupérer un résultat assez positif quand à la reconnaissance des entrées et fin d'entrées pour les annuaires et encouragé l'idée d'un résultat similaire pour les catalogues ayant des entrées particulièrement espacées. Ainsi, les données ajoutées correspondent à des catalogues ayant de larges espaces entre chaque entrées, soit essentiellement des catalogues d'exposition. Cette décision était également induite par la suite de mon stage, qui nécessitait d'avoir un modèle de segmentation suffisamment performant sur les catalogues d'exposition pour réaliser de l'extraction de données sur les résultats de l'HTR. Le jeu d'entraînement réalisé suit les mêmes proportions que les précédents, en gardant le *data test* de **Beaufort** tout en lui associant 12 nouvelles données. Le modèle obtenu, **Chaource**, a été appliqué sur le jeu de données *test* et les résultats visualisés dans eScriptorium sont grandement satisfaisants quant au traitement des entrées des catalogues d'exposition. Malgré tout, les zones plus marginales dans le jeu de données, à l'instar de *Running Title* ou *Numbering*¹⁸, restent non-reconnues et nécessiteraient un plus grand nombre de données¹⁹ mais les données les plus importantes, permettant de réaliser une extraction de données par la suite sont disponibles sur un certain nombre de catalogues.

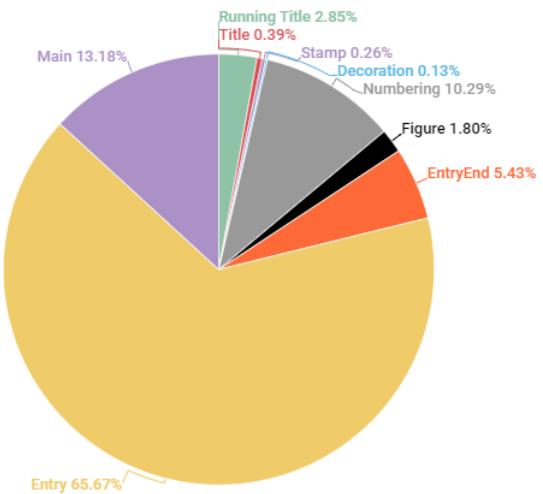


FIGURE 4.4 – Répartition des zones dans le *dataset Chaource*

4.2 Réalisation de modèles de reconnaissance de caractères

4.2.1 La création de jeux de données représentatifs du corpus de travail

En parallèle, un travail a été réalisé pour l'élaboration d'un modèle de transcription fonctionnel. Pour ce faire, les données issues de eScriptorium, segmentées selon les principes de SegmOnto ont été utilisées. Un certain nombre de jeux de données, de plus en plus grands, a donc été réalisé de la même façon que pour les modèles de segmentation. Cependant, à la différence de ces derniers, l'entraînement des modèles de transcription ne nécessite pas obligatoirement d'ordinateur extrêmement puissant et a donc été réalisé directement sur ma machine.

17. Celui-ci a été réalisé avec l'aide de Claire Jahan et Simon Gabay. Un manuel de segmentation réalisé dans ce contexte est par ailleurs disponible en annexes (C.2).

18. La version numérique de la visualisation ci-contre est disponible en annexes (B.1)

19. Ces données étant reconnues par le modèle de segmentation de Claire Jahan sur les pièces de théâtre du XVII^{ème}, l'idée serait d'associer les deux sets de données, ce qui formerait un corpus de 800 pages, afin d'entraîner un modèle commun capable d'analyser et nommer ces zones.

Comme mentionné précédemment, les données utilisées dans le cadre de ce travail sont déjà transcris, puisqu'ayant servi à l'élaboration de modèles de transcription sur Transkribus. Ainsi, il a suffit de récupérer les fichiers Alto, avec leur nouvelle segmentation, en sortie d'eScriptorium et de sélectionner les données afin de créer des jeux d'entraînement les plus représentatifs des documents étudiés. Ceux-ci fonctionnent en effet de la même façon que pour l'entraînement d'un modèle de segmentation et nécessitent une division des données entre deux ou trois jeux : *train*, *val* et *test*. Or, la diversité importante de notre corpus de travail induit une grande multiplicité de typographies, formes, polices, tailles pour les caractères étudiés. Ainsi, il est particulièrement important de faire très attention à disposer d'un jeu de données qui puisse entraîner un modèle capable de traiter tout ces cas de figures le mieux possible.

Afin de gérer au mieux ce problème, un *split*, c'est-à-dire un programme python qui divise les données de façon à ce que les différents jeux d'entraînement soit représentatifs, a été mis au point. Dans un premier temps, les données utilisées lors de l'entraînement sont décrites dans un fichier csv. Une des colonnes permet de définir, pour chaque page du *set*, un sous-corpus :²⁰. Par exemple, une page de catalogue d'exposition du XIX^{ème} est décrite comme faisant partie du sous-corpus **19-exp**, tandis qu'une page de catalogue de ventes de manuscrits du XXI^{ème} appartient au sous-corpus **21-man**. Au sein du programme python, il est alors possible de décider à quelles proportions doivent correspondre chacun de ces sous-corpus dans chaque jeu d'entraînement. Ainsi, on peut décider que les catalogues d'exposition doivent être à 80% d'entre eux dans le *train*, 10% dans le *test* et 10% dans le *val*, ou encore que 50% des pages de pièces de théâtre d'*Out-of-domain* doivent se trouver dans le *test*, etc. Cela permet également de réaliser un *dataset test* fixe pour tous les entraînements, afin d'appliquer les modèles créés sur les mêmes données et donc de réellement pouvoir les comparer²¹. Il est alors possible de réaliser un jeu d'entraînement réellement représentatif des données de travail, en jouant sur les proportions, en fonction du corpus à notre disposition. Une fois le jeu d'entraînement réalisé, il suffit alors de lancer l'entraînement d'un modèle par ligne de commande en fournissant à Kraken les données sous la forme de couples image-transcription au format ALTO4.

Le premier modèle de transcription²² réalisé, **Abondance**, a été entraîné sur 30 pages. Il correspond au même *dataset* que celui utilisé pour l'entraînement du modèle de segmentation **Abondance** et a donc été divisé de la même façon entre *test*, *train* et *val*, puisqu'il s'agit d'un tout petit jeu de données. **Beaufort**, le second modèle de transcription, a été réalisé à partir du même jeu de données. Ces deux modèles ont été entraînés en même temps afin de déterminer l'impact d'une repolygonisation sur la qualité de la transcription automatique, d'où le jeu de données réduit. Par le terme « repolygonisation », on entend

20. Les différents sous-corpus existant sur les données de travail étant **19-man**, **20-man**, **21-man**, **19-exp**, **20-exp**, **19-an** ou **19-autres**, pour l'*Out-of-domain*

21. C'est ce qui a été fait avec la création de plusieurs jeux de données *test* distincts. Les jeux *test_15* (https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/3_Scripts_training_construction/test_15.txt) et *test_30* (https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/3_Scripts_training_construction/test_30.txt) permettent d'adapter son jeu de données *test* fixe en fonction de la taille du corpus d'entraînement. Il s'agit de jeux où les proportions de pages de catalogues d'exposition, de ventes de manuscrits et d'annuaires sont identiques. Le jeu *test_20expo* (https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/3_Scripts_training_construction/test_20_catExpo.txt) est exclusivement composé de pages de catalogues d'exposition. Cela permet d'obtenir une visualisation la plus précise possible de la qualité du modèle pour la suite des missions du stage et l'extraction de données à partir de ces résultats.

22. Tout les modèles de reconnaissance de caractères réalisés au cours du stage sont disponibles ici : https://github.com/Juliettejns/cataloguesSegmentationOCR/tree/main/4_Models/HTR

la restructuration par le logiciel de transcription automatique de la zone des *TextLine*, soit le masque recouvrant les caractères que lit le modèle de reconnaissance. En effet, les données utilisées sont issues d'un long processus de production et sont passées par un certain nombre de logiciels différents. Ainsi, la façon dont sont traitées les informations par les logiciels et par les formats de stockage (Alto, PageXML, etc) peut induire une perte de la qualité des données. Repolygoniser, c'est donc essayer de récupérer les données perdues afin de tenter d'obtenir un modèle de transcription automatique plus performant. Ainsi, **Abondance** a été entraîné sans repolygonisation au contraire de **Beaufort**. Sur les 3975 caractères qui composent le jeu de données *test*, le modèle sans repolygonisation a 190 caractères faux, tandis que le modèle avec repolygonisation a 1304 fautes. **Beaufort** aurait dû obtenir un meilleur score que **Abondance**, d'après la définition de la repolygonisation. Or, ce n'est pas le cas. Ce résultat est d'autant plus étonnant en ce qu'il est particulièrement bas pour le modèle avec repolygonisation mais également que celui-ci a un score plus bas que le modèle sans repolygonisation, ce qui devrait être le contraire. Il a donc été considéré que, face à l'évolution des scripts de repolygonisation pour l'OCR, domaine assez récent, il valait mieux laisser de côté ce travail, et se concentrer, sur la création d'un modèle d'OCR valide.

Un troisième modèle a donc été réalisé, sans repolygonisation, à partir de 99 pages. **Chaource** a été entraîné sur 33 pages de catalogues d'exposition, 33 pages de catalogues de ventes de manuscrits et 33 pages d'annuaires. Le jeu d'entraînement a été divisé en 80%, 10%, 10% de façon à avoir les données *test* fixées.

4.2.2 Comment évaluer ces modèles : outils et méthodes

Une fois ces modèles entraînés, il est nécessaire de les évaluer, afin de déterminer leur qualité et leur précision. Pour ce faire, plusieurs outils sont à notre disposition. Dans un premier temps, Kraken calcule lui même un *Accuracy Rate* que l'on pourrait traduire par taux de précision. Ce pourcentage correspond au nombre de caractères corrects prédicts par le modèle par rapport au nombre de caractères total sur le *dataset val*. Cette donnée permet de se faire une idée de la performance du modèle. En général, un bon modèle a un *Accuracy Rate* d'environ 98 à 98%. Il est également possible d'obtenir le nombre de mots corrects prédicts par le modèle par rapport au nombre de mots total sur le jeu de données. Il s'agit du WAC, *Word Accuracy*.

Il existe cependant d'autres méthodes permettant d'obtenir un score beaucoup plus parlant de la qualité du modèle de reconnaissance de caractères.

On parle de CER, *Caracter Error Rate*²³ dans le cas des HTR. En effet, ceux-ci reconnaissent les caractères un par un, ce qui entraîne donc un calcul du taux d'erreur au niveau du caractère. Il permet de déterminer la différence entre la transcription produite par le modèle et celle fournie à la machine. On peut typer ces erreurs en fonction de leurs caractéristiques :

Le calcul du CER est un petit peu plus compliqué qu'une simple mesure du nombre d'erreurs de caractères comme présenté ici. En effet, pour l'obtenir, il est nécessaire de mesurer la distance de Levenshtein. Cette métrique permet de déterminer la différence entre deux chaînes de caractères. Son résultat correspond au nombre minimum de caractères nécessaire pour changer un mot en un autre. Comme il est possible de le voir dans la

23. Romain Karpinski, Devashish Lohani et Abdel Belaid, « Metrics for Complete Evaluation of OCR Performance », (juil. 2018), p. 8, URL : <https://hal.inria.fr/hal-01981731>.



FIGURE 4.5 – Les types d’erreurs

figure 4.6, on passe du mot « Peinture » au mot « Jointure » lorsque 2 caractères ont été modifiés. Ici, la distance de Levenshtein est donc de 2. Plus les deux mots sont différents, plus le nombre de changements nécessaires pour passer de l’un à l’autre sera élevé et plus la distance de Levenshtein sera grande.

- 1. PEINTURE**
- 2. JEINTURE**
- 3. JOINTURE**

FIGURE 4.6 – Exemple du calcul d’une distance de Levenshtein

Le CER repose donc sur cette idée : le résultat de la transcription faite par le modèle est comparé à la vérité terrain en calculant la distance de Levenshtein de chaque caractère. Ainsi, on calcule le nombre de caractères à transformer pour passer du texte de la vérité terrain au texte en sortie de l’OCR. On obtient alors un pourcentage qui indique le taux d’erreur du modèle. Plus le pourcentage est près de 0, plus l’OCR est bon²⁴. Le WER, *Word Error Rate*, ressemble au CER à la différence qu’il s’agit de calculer le nombre d’erreurs au niveau du mot. Cette valeur est beaucoup plus importante que le CER dans le cas de la transcription automatique de documents historiques, puisqu’ici on s’intéresse tout particulièrement aux mots. Le calcul est donc réalisé de la même façon que le CER mais au niveau du mot. Afin de se faire une idée du WER, il faut multiplier par 4 le CER puisqu’un mot sera faux à partir du moment où un seul caractère sera faux. Ainsi le taux d’erreur au niveau du mot est beaucoup plus élevé que le taux d’erreur au niveau du caractère. Un bon modèle aura pour WER 10%.

Ces différentes métriques ont été calculées pour les trois modèles présentés précédemment. Ainsi, **Abondance** a une *accuracy* de 95,22% et un CER de 4,7%. **Beaufort** a une *accuracy* de 67% et un CER de 33%. Ainsi, un mot sur trois de la transcription d'**Abondance** sera faux, presque tout les mots du résultat de **Beaufort** le seront²⁵. Enfin, **Chaource** a une *Accuracy* de 97,19% et un CER de 2,8%. Cela signifie qu’un mot sur dix du résultat obtenu avec ce modèle sera faux²⁶. Comme signalé précédemment, il

24. Un modèle est considéré bon lorsque le CER est de 1-2%

25. En effet, en réalisant un rapide calcul du WER, comme expliqué plus haut, celui-ci devrait être d’environ 33% pour **Abondance** et de presque 90% pour **Beaufort**.

26. Ici, le WER calculé manuellement serait d’environ 10%.

s'agit d'un bon modèle, sans pour autant être excellent. Il est donc possible de l'utiliser dans la chaîne de traitement d'Artl@s comme modèle de reconnaissance de caractères. Cependant, il est également préférable de produire plus de données afin d'entraîner un nouveau modèle plus performant.

4.3 Réflexions autour des modèles et jeux de données produits

4.3.1 Utiliser les modèles entraînés : mise en application

Ce travail permet ainsi d'obtenir un couple de modèles pour l'HTR assez fonctionnels :

- le modèle de segmentation **Chaource**, entraîné sur 274 pages
- le modèle de reconnaissance de caractères **Chaource**, réalisé à partir de 100 pages

Ils ont été appliqués l'un après l'autre sur le jeu de données *test* dans eScriptorium. Cela permet de visualiser l'aspect du résultat d'un HTR utilisant ces deux modèles. La plupart des transcriptions sont assez bonnes mais nécessitent un certain nombre de corrections, en moyenne une par ligne. Afin de mesurer de façon plus objective les résultats, le CER a été calculé en comparant la transcription issue des deux modèles et celle réalisée manuellement. On obtient alors un *Character Error Rate* de 3,45% pour le couple et une *accuracy* de 96,55%. Ce résultat confirme que l'utilisation de ces deux modèles entraîne des transcriptions qui ne sont pas mauvaises, mais devraient être meilleures en comparaison de la quantité de données fournies.

La question est donc de savoir pourquoi les résultats ne sont pas aussi bons qu'escompté. La réponse à cette interrogation est importante puisqu'elle permet de déterminer comment améliorer le modèle : est-il nécessaire d'ajouter plus de données ou le problème réside-t-il ailleurs ? Une hypothèse émise concerne la hauteur des **baselines**. Ces objets géométriques sont des traits qui suivent le dessous des caractères d'une ligne et font parti de la segmentation. Le modèle de reconnaissance de caractères les utilise comme guide lors de la transcription. Les données qui ont été utilisées pour construire les modèles sont passées par de nombreux logiciels auparavant disposant de leur propre hauteur de **baseline** celle-ci pouvant varier. Par exemple, la **baseline** de Transkribus sera en moyenne 5 pixels plus haut que celle d'eScriptorium. Ainsi, les modèles obtenus ici ont été entraînés à partir de données issues de Transkribus, leurs **baselines** sont plus hautes que celles normalement traitées par Kraken. La question serait donc : est-ce que cette différence impacte la qualité de la sortie de l'HTR ?

Une expérience a été réalisée pour répondre à cette question. L'idée est de comparer un modèle de reconnaissance de caractères entraîné sur des données ayant des **baselines** issues de Transkribus à un autre réalisé à partir de données natives d'eScriptorium. Pour ce faire, il est nécessaire de produire de nouveau un jeu de données afin d'entraîner ce nouveau modèle. Il a donc été décidé dans un premier temps de réaliser un mini-modèle sur les 30 premières pages des données et de le comparer à **Beaufort**, lui-même entraîné sur ces mêmes pages²⁷. Afin de réaliser ce nouveau modèle, **Epoisse**, on a tout d'abord

27. Un troisième modèle, **Danablu**, a également été entraîné à partir de ces 30 mêmes pages (avec les lignes non corrigées issues de Transkribus) afin de vérifier que le problème ne résidait pas dans l'entraînement de **Beaufort**. En effet, comme mentionné précédemment, il s'agit d'un logiciel développé

récupéré les zones en appliquant sur les pages le modèle de segmentation **Chaource**. Par la suite, on a utilisé le modèle par défaut de Kraken, **blla**, afin d'obtenir des lignes natives d'eScriptorium, et donc à la bonne hauteur pour le logiciel. Enfin, on a ajouté la transcription et entraîné le modèle à partir du jeu de données obtenu. Plusieurs tests ont été réalisés afin de départager les résultats des deux modèles de reconnaissance de caractères. Pour chacun d'entre eux, on les a appliqués tout d'abord sur des données segmentées uniquement par le modèle de segmentation **Chaource**, puis sur des données segmentées en zones par **Chaource** et en lignes par **blla**²⁸. Pour chaque cas, on calcule le CER en sortie.

Étonnamment, le résultat ne correspond pas à celui prévu. En effet, les CER sont tous à peu près équivalents, aux alentours de 4%²⁹ dans le cas où la segmentation correspond au modèle utilisé³⁰. Pourtant, le résultat a été grandement amélioré en utilisant **blla** pour la reconnaissance des lignes dans le cas des données de Claire Jahan. Cependant, contrairement à ces données, qui sont des pages issues de pièces de théâtre imprimées du XVII^{ème} siècle, l'application de **blla** sur les catalogues entraînent une quantité importante de bruit. En effet, le modèle de segmentation par défaut reconnaît en tant que lignes des caractères situés à l'envers de la page et visibles en transparence. Ainsi, il est tout à fait probable que cela entraîne une réduction de la qualité des données issues de ce couple. Il pourrait être alors intéressant d'entraîner un nouveau modèle avec des données issues de **blla** corrigées afin d'obtenir une segmentation de lignes natives d'eScriptorium mais sans trop de problèmes de reconnaissance. Cependant, il n'a pas été possible de faire cela car il aurait fallu reprendre chaque ligne des 274 pages traitées, ce qui s'avère énormément de travail.

En conclusion, face à l'ampleur du travail demandé pour obtenir des données natives d'eScriptorium, il a été décidé de conserver le couple de modèles **Chaource** pour l'HTR. L'application des modèles a cependant été revue dans la volonté d'optimiser au mieux les données transcris en sortie. Ainsi, j'avais l'ambition initiale de réaliser directement la transcription automatique des images en ligne de commande sur Kraken, c'est-à-dire de lancer les modèles de segmentation puis de reconnaissance de caractère ensembles en récupérant directement les fichiers ALTO produits. Au vu de la situation, il a été décidé de réaliser manuellement la transcription dans eScriptorium. En premier lieu, on ajoute les images dans le logiciel, puis on applique le modèle de segmentation **Chaource** sur celles-ci. Une fois cela fait, on réalise une étape de correction manuelle des lignes et zones obtenues, dans l'idée d'améliorer le résultat. Enfin, le modèle de reconnaissance de caractères **Chaource** est appliqué. Ainsi, il n'est pas possible pour le moment de réaliser une transcription automatique directe des images avec les modèles obtenus. Il

dans le cadre de programmes de recherche sur des langues non-latines, qui nécessitent donc une **topline** au lieu d'une **baseline**. Concrètement, cela veut dire que la ligne de lecture ne suit pas le bas des lettres mais le haut des lettres, puisqu'il s'agit d'hébreu et d'arabe. Ainsi, on a testé, en ré-entraînant un modèle sur les mêmes données et signalant bien qu'il s'agit là de caractères latins, si l'entraînement de **Beaufort** n'avait pas eu lieu sur des **topline** au lieu de **baseline**. Le résultat étant strictement identique pour les deux modèles, ce n'est donc pas le cas. Par conséquent, le problème réside ailleurs.

28. Pour une description détaillée de ces tests, voir les tableaux en annexes B.2. Le premier d'entre eux répertorie la qualité des modèles produits à partir du jeu de données de 30 pages. le second réalise le même travail pour les modèles ayant été entraîné avec plus de données. Le dernier répertorie les couples Segmentation-Reconnaissance et les comparent.

29. Les résultats pour chaque test sont mentionnés dans le tableau cité plus haut.

30. C'est-à-dire pour le test modèle **Chaource** en segmentation et en reconnaissance de caractères et pour le test modèle **Chaource** en zones, modèle **blla** en lignes et modèle Fourme en reconnaissance de caractères.

est nécessaire de passer par une correction manuelle entre les étapes segmentation et reconnaissance de caractères. En conséquence, il s'agit d'optimiser les modèles obtenus en ajoutant plus de données, jusqu'au moment où il sera possible de lancer l'HTR directement en ligne de commandes et d'obtenir une transcription acceptable³¹ en sortie.

4.3.2 Vers un accès libre et ouvert de ces modèles et jeux de données

En parallèle de ce travail d'amélioration des modèles créés, toute une réflexion sur l'accessibilité des données produites a été construite. Ces données, c'est-à-dire les images et transcriptions associées, peuvent être considérées comme des données de la recherche. On les définit comme étant des « enregistrements factuels utilisés comme sources principales pour la recherche scientifique³² ». Associées, elles forment un jeu de données ou *dataset* en anglais qui est décrit comme une « agrégation de ces mêmes données organisées en un ensemble cohérent formaté pour être communicables, interprétables et adaptées à un traitement informatisé³³ ». Chacune de ces données a besoin de métadonnées, qui sont les décrivent : le titre, l'auteur, manière dont elles ont été produites, etc. Comme mentionné précédemment³⁴, le partage des données et protocoles de production a de nombreux avantages théoriques. Nous allons ici mener une réflexion sur la mise en place de cette diffusion libre, gratuite et universelle des données produites puis nous nous intéresserons aux apports que celle-ci a permis.

Partager les données d'une institution ou d'un projet, comme c'est le cas ici, est une véritable stratégie de recherche. En effet, l'intégralité de la production de données se tourne vers cette ambition et est grandement impactée par cela. Le partage de ces données s'appuie sur de nombreux critères juridiques, scientifiques, humains et techniques. Des outils ont été développés par le milieu de la recherche afin de mener à bien ce travail. Parmi eux se trouvent les principes FAIR³⁵. Développés dans un article du journal *Scientific Data* en 2016³⁶, ce sont de recommandations permettant à la communauté scientifique de rendre accessible simplement des données fonctionnelles et réutilisables³⁷. Son acronyme signifie :

- *Findability* : Faciliter la découverte des données (en leur donnant un identifiant pérenne, en les décrivant avec des métadonnées et en les déposant dans un entrepôt de données)
- *Accessibility* : Permettre l'accès aux données (en ayant un protocole de communication standard, libre et ouvert)

31. Cela correspondrait à un modèle de reconnaissance de caractères ayant un CER de 1% qui permettrait de ne plus avoir à corriger manuellement la segmentation et la transcription, les fautes étant mineures.

32. OCDE, 2007, <https://www.oecd.org/fr/science/inno/38500823.pdf>

33. L. Dedieu et F. Marie-Françoise, *Rendre publics ses jeux de données scientifiques...*

34. Voir Introduction

35. La documentation de l'initiative est accessible ci-joint : <https://www.go-fair.org/fair-principles/>

36. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al., « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data*, 3-1 (mars 2016), doi : 10.1038/sdata.2016.18.

37. URFIST Méditerranée, *Les principes FAIR*, URL : <https://doranum.fr/enjeux-benefices/principes-fair/> (visité le 27/08/2021).

- *Interoperability* : Permettre l'exploitation des données quelque soit le système informatique utilisé
- *Reusable* : Permettre la réutilisation des données par d'autres projets (grâce à une licence de réutilisation et un suivi des standards scientifiques).

Notre travail de mise en accès libre des données s'appuie donc en partie sur ces principes et recommandations.

Dans un premier temps, il a fallu choisir où déposer notre jeu de données. Pour ce faire, nous avons utilisé github³⁸. Il s'agit d'un site web permettant d'héberger et de gérer des projets de développement sous la forme de dépôts, qui permettent le stockage centralisé et organisé de données. Cet outil est particulièrement employé dans les projets d'humanités numériques que ce soit pour le stockage de jeux de données en cours d'élaboration ou pour le développement de protocoles de traitement. Il permet également de télécharger facilement les données, est accessible librement et donne la possibilité d'ajouter de la documentation sur la production. Ainsi, il est particulièrement judicieux de l'utiliser pour donner accès à nos données. Celles-ci étant produites au fur et à mesure, l'utilisation de github offre la possibilité de les ajouter en versionnant le projet. Cela permet à chaque ajout de données de créer une version spécifique du dépôt à cet instant et donc d'avoir une trace du développement du jeu de données dans le temps. Une fois le jeu de données terminé, il a été également passé sur Zenodo³⁹, un répertoire de travaux de recherches créé par le CERN en 2013. Ce site fonctionne par dépôt de projet pour stocker les jeux de données et logiciels produits par la recherche, de la même façon que github. Cependant, il ne permet pas de modifier et ajouter des éléments. Cela permet donc d'avoir une version fixe des données mais également d'obtenir un DOI, *Digital Object Identifier*⁴⁰. Ici sous la forme **10.5281/zenodo.5458350**, il s'agit d'un identifiant numérique qui permet de répertorier une ressource numérique, ici le dépôt, et de lui associer des métadonnées. Cela permet d'obtenir une version et une identification plus pérenne du dépôt.

Comme indiqué dans les critères FAIR, les données mises en lignes doivent être interopérables. Ainsi, il faut qu'elles soient standardisées et que leur protocole de production soit décrit et reproductible. Il a donc fallu mettre en place tout une campagne de vérification de la qualité des données avant de les ajouter au dépôt. En effet, certains fichiers alto peuvent contenir des erreurs. Certaines zones ou lignes peuvent par exemple ne pas avoir été nommées correctement, des lignes ne sont pas liées à leur zone ou encore la transcription est manquante. Pour corriger ces problèmes et empêcher l'ajout de données problématiques au dépôt, j'ai ajouté un github action⁴¹. Celui-ci permet de vérifier que les fichiers alto ajoutés sont bien formés et renvoie une erreur lorsque ce n'est pas le cas. Le *dataset* et sa production sont documentés par de multiples *README.md* ainsi que par un fichier csv répertoriant les métadonnées de chaque page : leur provenance, leur date de production, etc. Si les données sont ajoutées avec leurs zones *Entry* et *EntryEnd*, un programme python permet de s'en débarasser afin d'obtenir des fichiers alto complètement conformes à la terminologie SegmOnto et donc réutilisables dans n'importe quel

38. Le dépôt contenant le jeux de données produit est disponible ici : <https://github.com/Juliettejns/cataloguesSegmentationOCR>

39. La version Zenodo (<https://zenodo.org/>) du dépôt contenant jeux de données et modèles est disponible ici : https://zenodo.org/record/5458350#.YTW_GY5KjIU

40. Catherine Lupoivici, « Le Digital Object Identifier : Le système du DOI », *Bulletin des Bibliothèques de France*–43-3 (1998), p. 49-54, URL : <https://bbf.essib.fr/consulter/bbf-1998-03-0049-007>.

41. Il s'agit d'une feuille XSLT et un programme python emprunté au projet SegmOnto et réalisé par Thibault Clérice et Ariane Pinche.

corpus réunissant divers jeux de données qui appliquent ses principes.

Le dernier grand point important de la mise à disposition des ressources sur internet est la question de la licence. Une licence de diffusion permet de fixer les conditions de l'utilisation du jeu de données : droit d'utilisation, de modification, de partage... Il est obligatoire d'apposer une licence à notre dépôt avant sa publication. La licence choisie, *Creative Commons Attribution* (ou CC-BY)⁴², a été créée en 2002 spécifiquement pour ce genre de situation et permet la diffusion libre et la réutilisation des données dans la mesure où l'auteur est cité. Chaque *dataset* utilisé pour former notre corpus de travail possède cette licence, d'où les crédits indiquant les diverses personnes ayant travaillé à l'élaboration de ces données⁴³. La question de la licence se recoupe également avec d'autres problèmes légaux spécifiques à nos données. En effet, le jeu de données est composé d'images, issues d'autres sites internet, à l'instar de Gallica, et de transcriptions qui ont été entièrement réalisées dans le cadre de notre projet. Un premier problème est donc lié aux droits de diffusion et de réutilisation des sources primaires. Ainsi, les images issues de catalogues de plus de cent ans sont passées dans le domaine public et sont donc également CC-BY. Au contraire, les images plus récentes restent la propriété intellectuelle de leur auteur, ce qui doit être indiqué. Il reste tout de même possible de les partager car il ne s'agit là que d'extraits et non de catalogues entiers. Enfin, le cas des pages d'annuaire est assez particulier, puisque se rapportant à la sensibilité des données, notamment privées. En effet, les personnes concernées par les informations contenues dans les documents ont droit de regard sur leur diffusion et peuvent demander le retrait de ces pages d'internet. Pour gérer ce problème, un mail de contact⁴⁴ est disponible dans le dépôt.

Comme mentionné précédemment, cette mise en ligne des données et de leur protocole permet leur réutilisation et la reproduction de leur production. Dans la pratique, ici, cela permet de donner accès de nos jeux de données à la communauté scientifique. Ainsi, cela rend possible, dans un premier temps, l'entraînement des modèles issus de ces données. En effet, pour pouvoir obtenir ces modèles, il a fallu utiliser le GPU de l'École nationale des Chartes, et le transfert de données s'est donc fait via le dépôt github. Donner accès aux données permet également de rendre le projet visible et offre donc à d'autres personnes la possibilité de tester nos *dataset*. En effet, n'importe qui peut télécharger les données et les réutiliser dans le cadre d'un entraînement. Ainsi, de nouvelles tentatives d'entraînement, notamment par des ingénieurs informaticiens ou encore sur d'autres logiciels d'OCR sont possibles et peuvent être un moyen de faire face aux problèmes de qualité des modèles obtenus mentionnés précédemment. Dans la pratique, l'*OpenData*, a l'échelle de notre projet permet effectivement d'accélérer le projet et d'encourager la collaboration, d'assurer la qualité des données et la reproductibilité des méthodes de production.

42. <https://creativecommons.org/licenses/by/4.0/deed.fr>

43. Voir <https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/README.md#credits>

44. Ici : <https://github.com/Juliettejns/cataloguesSegmentationOCR/blob/main/README.md#contacts>

Chapitre 5

Aller plus loin dans la récupération d'informations

Est-il possible de récupérer plus d'informations depuis une image ? Jusqu'à présent, nous avons décrit des méthodes permettant d'extraire l'information textuelle. Nous sommes allés plus loin avec la récupération de la mise en page et des différents éléments qui la composent. Il reste cependant encore un nombre important de données présentes dans les images qui n'ont pas été sorties. C'est le cas, par exemple, des images et des tableaux mais également de la typographie. Tous ces éléments sont importants pour comprendre l'information sémantique, présentée dans les documents. Ce chapitre permet de réfléchir sur l'intérêt de reconnaître ces informations non textuelles, fait le point sur les méthodes d'extraction de ces données et développe une réflexion sur leur stockage au sein des fichiers ALTO ou PAGEXML en sortie de l'OCR et leurs potentielles utilisations. Dans le cas des catalogues, on s'intéresse aux illustrations et à l'emphase typographique.

5.1 Les illustrations

5.1.1 Panorama des images dans les catalogues

Bien que notre étude s'intéresse principalement au contenu textuel de notre corpus, de nombreuses illustrations se trouvent au sein de ces documents. Dans le cas des catalogues, ce sont principalement des fac-similés d'œuvres, pour les expositions, et de signatures, pour les manuscrits. Quelques publicités imagées existent également dans les pages d'annuaires.

L'image 5.1 est un exemple typique de l'aspect que peuvent prendre ces objets au sein d'un catalogue d'exposition. La page est composée d'une seule image, présentant le tableau, accompagnée d'une légende en dessous, donnant le nom de son auteur et le titre de l'œuvre. Le catalogue d'où est issu cette page contient tout une dernière partie, à la suite de la liste des œuvres exposées, qui présente des illustrations de ce type. En effet, comme mentionné précédemment¹, le catalogue, à partir du XX^{ème}, s'étoffe de ce type de nouvelle partie contenant des images des tableaux exposés. Cela va notamment de pair avec le développement des catalogues en tant qu'outil permettant au visiteur de se repérer dans l'exposition et, en tant que potentiel acheteur, de faire sa sélection : les images donnent un aperçu des peintures que l'on peut voir lors de l'évènement. L'image 5.2 correspond à une illustration courante dans les catalogues de ventes de manuscrits.

1. Voir 1.2.3

Ce genre de figure est située à côté de l'entrée qui lui correspond, afin de véritablement illustrer le document. Il peut s'agir de signatures, comme c'est le cas ici, ou encore de morceaux de manuscrits, soit des éléments de texte rédigé manuellement. L'idée ici est également de présenter aux potentiels acheteurs les produits disponibles à la vente en leur permettant d'accéder au mieux aux manuscrits, par le biais de ces images. Cela est d'autant plus visible que ce genre d'illustrations se trouve sur la même page que la description du manuscrit en question, et très souvent au niveau de l'entrée correspondante.

Ainsi, même si la base de notre travail concerne le texte, récupérer l'information illustrée peut être intéressant. En effet, comme mentionné précédemment, ces fac-similés sont parfois les uniques témoins de l'existence et de l'aspect d'une peinture ou d'un manuscrit. Les récupérer automatiquement permet alors d'avoir une vision globale des différentes images disponibles au sein de nos documents. Cela donne aussi la possibilité d'ajouter un niveau de plus à Basart, en complétant les données disponibles dans la base d'illustrations, auquel le chercheur pourra avoir accès. Pouvoir récupérer ces données permettrait aussi de se faire une idée globale de la circulation des thèmes dans les peintures, ce qui pourrait se lier au projet *Visual Contagions*².

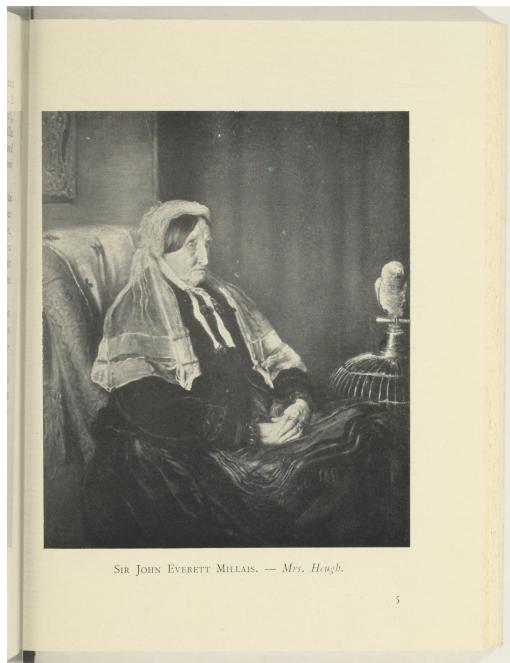


FIGURE 5.1 – Catalogue des œuvres exposées, Palais du Luxembourg, 1915, p.3

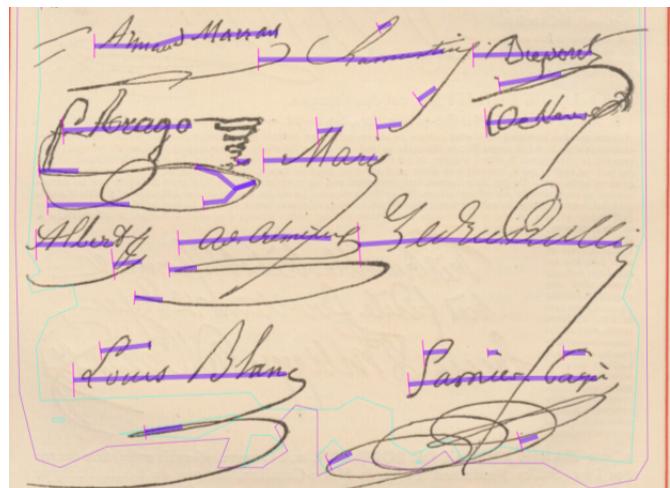


FIGURE 5.2 – Catalogue de manuscrits, Bovet, 1887, p.18

5.1.2 Comment récupérer les images ?

Une des solutions pour récupérer ces images serait de s'appuyer sur le *Layout Analysis*. En effet, grâce à la terminologie SegmOnto, il est possible au cours de la segmentation d'obtenir des régions qui sont nommées *Figure*. Cela nécessiterait cependant d'entraîner un modèle avec un jeu de données contenant beaucoup plus d'images³ **Chaource**, le

2. Voir 2.2.1

3. Un bon exemple serait le projet *Chronicling America* de la Bibliothèque du Congrès américaine, qui vise à OCRiser des journaux du XIX^{ème}. Un article sur l'extraction spécifique des images présente la construction d'un gros corpus de données principalement pour l'entraînement d'un modèle capable de

modèle actuel, possède effectivement très peu d'images dans son *dataset* et a encore de grandes difficultés à les reconnaître et à les cibler. Une fois le modèle réalisé, il suffirait alors de récupérer les coordonnées des régions contenant des images pour les découper. Cette partie peut se faire soit à l'aide d'un rapide programme python, soit en ajoutant les coordonnées dans l'url du IIIF⁴, lorsque le catalogue traité en est issu.

Chaque type de catalogue a cependant son lot de problèmes quant à la gestion et à la récupération de ses illustrations. Comme il est possible de le voir sur l'image 5.2, les figures des catalogues de ventes de manuscrits sont des signatures et du texte écrit à la main. Techniquement, il s'agit là à la fois d'une image et d'un texte. L'OCR peut donc légitimement reconnaître ces éléments comme du texte. C'est justement ce qui se passe ici. Il est en effet possible de distinguer les *baselines*, lignes violettes, sous les différentes signatures. Pour contourner ce problème, l'idée serait d'ajouter au jeu de données servant à entraîner le modèle un maximum d'images de ce type. En effet, l'intégralité de notre corpus de travail se compose d'imprimés et les images que l'on souhaite reconnaître sont manuscrites. On peut donc espérer réussir à entraîner le modèle à catégoriser les éléments manuscrits comme des images. Cela demanderait cependant un corpus assez important correspondant à ces critères et reste une hypothèse.

Un autre problème concerne le lien entre image et légende : comment associer une image à un titre, un manuscrit, un peintre, lorsque ceux-ci sont mentionnés à côté ? Il s'agit là aussi d'une question actuelle, sur laquelle travaille un certain nombre de groupes de chercheurs⁵. Ainsi, en étudiant les différents travaux déjà publiés, plusieurs solutions apparaissent pour résoudre ce problème. Chacune mobilise la segmentation et le *Layout Analysis*. Dans un premier temps, l'idée est simplement d'associer l'image à une zone qui l'enserre, elle et sa légende, afin de les associer directement. C'est par exemple possible avec les catalogues de ventes de manuscrits. Ici, l'illustration, de par sa position à côté de sa description, est située au sein de la zone *Entry*. Il suffit alors d'associer les deux éléments en sortie d'OCR. Cette solution serait également envisageable avec l'exemple présenté (figure 5.1) pour les catalogues d'exposition. La page contient une unique image et une unique légende. Il suffirait alors d'encadrer ces deux éléments dans une zone *main*, puisqu'il ne s'agit pas ici de zone *Entry*. Enfin, en dernier recours, on pourrait penser à créer un nouveau type de zone dans notre terminologie dans le cas où plusieurs images et légendes se trouvent sur une page sans entrées. Cette zone pourrait enserrer à la fois l'image et la légende et ainsi permettre leur association sur le même modèle que les solutions citées plus tôt. Tout cela reste hypothétique et demande certainement une grande quantité de données préparées pour réaliser un modèle viable.

Si cette stratégie peut s'avérer fonctionnelle⁶ elle est définie comme « traditionnelle⁷ » vis-à-vis des avancées actuelles du domaine. Ainsi, un certain nombre d'articles

reconnaître et récupérer les images des documents (illustrations, cartes, publicités, etc.). Sa réalisation a cependant nécessité l'aide de bénévoles qui ont taggé un grand nombre de données. (<https://dl.acm.org/doi/pdf/10.1145/3340531.3412767>)

4. Voir 3.3.1

5. Cette question a par exemple été traitée par le projet *Chronicling America*, toujours dans le même article. Ici le *dataset* a été préparé non seulement en segmentant les images mais aussi en segmentant les légendes correspondantes, afin d'obtenir un modèle capable non seulement de reconnaître une image mais aussi de repérer la légende qui lui est associée.

6. Elle a par ailleurs été appliquée dans la chaîne de traitement de Claire Jahan, autre stagiaire Artl@s. Celle-ci permet la récupération de décossements des pièces de théâtre du XVII^e siècle et la création d'une base de données contenant ces illustrations.

7. Jwalin Bhatt, Khurram Azeem Hashmi, Muhammad Zeshan Afzal et Didier Stricker, « A Survey of Graphical Page Object Detection with Deep Neural Networks », *Applied Sciences*, 11–12 (janv. 2021),

publiés récemment proposent des réflexions autour de l'amélioration de la reconnaissance des images⁸. Sont proposées de nouvelles méthodes, mobilisant notamment du *Deep Learning*. Il s'agit d'un sous-domaine de l'intelligence artificielle, dérivé de l'apprentissage automatique, dont font parti les OCR, qui s'appuie sur un réseau de neurones artificiels. Autrement dit, l'idée est d'extraire automatiquement les images en laissant la machine analyser l'aspect de la page à partir de coefficients de corrélation. Lorsqu'une région de la page n'est pas structurée en lignes horizontales, il s'agit alors potentiellement d'un élément graphique. Les résultats obtenus sont beaucoup plus satisfaisants, comparés à ceux de Transkribus⁹. Bien que ces différentes techniques se développent et que leurs résultats soient prometteurs, il est cependant nécessaire pour les employer de se reposer sur des ingénieurs informaticiens. Ainsi, dans ce contexte, la meilleure solution actuellement est de réaliser un *dataset* contenant un grand nombre d'images¹⁰.

5.2 L'emphase typographique

5.2.1 L'emphase typographique dans les catalogues

La typographie est à la fois une technique d'impression basée sur l'utilisation de caractères en relief et l'art d'utiliser différents types de caractères dans un but esthétique et pratique. Ici on s'intéresse à la seconde définition. Dans ce cadre, une police d'écriture ou de caractères désigne une famille, un assortiment de caractères typographiques ayant un dessin particulier commun¹¹. Une police possède plusieurs caractéristiques qui peuvent être amenées à varier :

- la graisse (l'épaisseur du trait) ;
- la forme ou le style (italique, gras, italique et gras, penché...) ;
- la taille (le corps) ;
- la casse (majuscule ou minuscule).

La fonte est un ensemble de caractères appartenant à la même famille, la même graisse, la même taille et la même forme. Dans ce cadre, l'emphase typographique correspond à l'exagération d'un élément (mot ou phrase) par l'utilisation d'une fonte différente de celle utilisée pour le reste du texte. Cela permet de marquer l'élément en question et mettre l'accent dessus.

L'emphase typographique est primordiale pour les catalogues. En effet, celle-ci peut être utilisée pour mettre en valeur certaines parties des entrées. Ainsi, dans un grand nombre de nos documents, la forme (italique, gras...) des polices d'écriture varie et permet de déterminer le type d'information qui est transmis.

DOI : 10.3390/app11125344.

8. Ce travail s'inscrit dans un mouvement plus large qui améliorer la numérisation des éléments dits « graphiques » soit les illustrations mais aussi les tableaux, graphiques, formules mathématiques et chimiques et schémas scientifiques. (exemple : Li Pengyuan, Jiang Xiangying et Shatkay Hagit, « Figure and caption extraction from biomedical documents », *Bioinformatics*, 35–21 (2019), p. 4381-4388, DOI : <https://doi.org/10.1093/bioinformatics/btz228>)

9. Dalia Coppi, Costantino Grana et Rita Cucchiara, « Illustrations Segmentation in Digitized Documents Using Local Correlation Features », *Procedia Computer Science*, 38 (2014), p. 76-83, DOI : <https://doi.org/10.1016/j.procs.2014.10.014>.

10. Dans le cas du projet *Chronicling America*, on parle d'un corpus de 50 000 images.

11. AUGER, *La typographie*, Coll. Que sais-je ?, Presses Universitaires de France, 1980

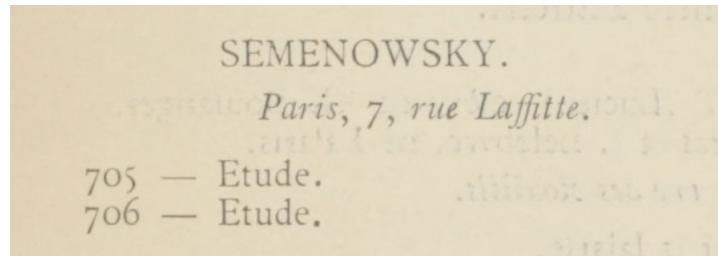


FIGURE 5.3 – *Catalogue des œuvres exposées*, Beaux-Arts de Nancy, 1892, p.004

Les images 5.3 et 5.4 donnent un aperçu de l'utilité de ces emphases dans les catalogues d'exposition et de ventes de manuscrits. Les entrées de ces documents possèdent une densité importante d'informations en un espace réduit. Ainsi, l'emphase typographique permet d'organiser les données et de les structurer. L'italique est utilisé pour les titres d'œuvres et les noms dans les catalogues de ventes de manuscrits et pour certaines informations complémentaires dans les catalogues d'exposition : en fonction du document, il peut servir à présenter l'adresse, le nom du maître du peintre, le titre de l'œuvre, etc. L'utilisation du gras est plus ciblée. Elle sert essentiellement à délimiter le début d'une entrée, en mettant en valeur le nom de l'auteur du manuscrit (image 5.4), le nom du peintre (image 5.3) ou la numérotation des œuvres (image 5.5).

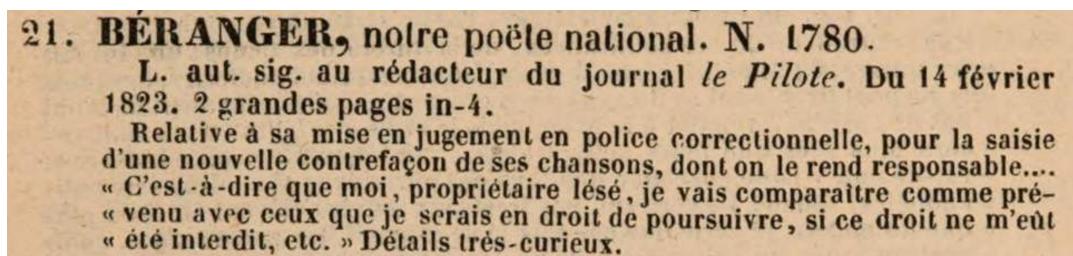


FIGURE 5.4 – *Catalogue de manuscrits*, Charavay, 1845, p.15

Ainsi, détecter ces variations dans la typographie permet d'obtenir des informations sur ce qui est décrit en fonction du type de catalogue. Cela permet également de signaler le début d'une nouvelle entrée : lorsqu'il y a plusieurs lettres en majuscules et en gras, on peut potentiellement penser qu'il s'agit du nom de l'auteur et donc du début d'une nouvelle entrée. Dans certains documents, il est impossible de distinguer le nom d'un auteur de ses informations biographiques, en italique, sans reconnaître l'emphase typographique.

5.2.2 État de l'art : récupérer l'information typographique en HTR

Très peu d'études ont été spécifiquement réalisées sur la reconnaissance de l'emphase typographique en OCR. Cependant, celle-ci rentre dans des tentatives plus larges de récupération de l'information typographique : polices, taille, etc. Ainsi, des groupes de travail¹² tentent, dès les années 1990, de reconnaître les différentes polices d'écritures au sein d'OCR, à l'aide de modèles créés à partir d'une centaine de fontes et de calculs

12. A. Zramdini et R. Ingold, « Optical font recognition using typographical features », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20–8 (1998), p. 877-882, DOI : 10.1109/34.709616.

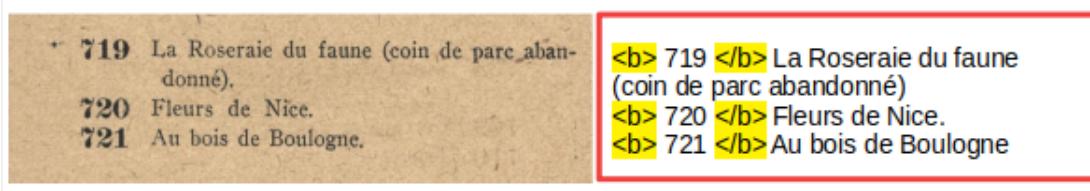


FIGURE 5.5 – Exemple de transcription d'un catalogue

statistiques¹³. Ces travaux intègrent le traitement de l'emphase typographique par l'étude des textures des différentes polices dans le but de les reconnaître. Les études les plus récentes sur les polices d'écriture se basent sur des réseaux de neurones¹⁴.

Le groupe de travail associant Artl@s et e-ditiones a déjà tenté de récupérer l'emphase typographique. La dernière solution qui ait été trouvée a été appliquée au cours du stage de Caroline Corbières en 2020. L'idée était d'entraîner un modèle capable de reconnaître les mots en gras et en italique dans Transkribus et de transmettre l'information dans les fichiers alto¹⁵. Afin de réaliser ceci, il a fallu préparer les jeux de données permettant l'entraînement du modèle en associant les balises **** aux mots en gras et **<i>** aux mots en italique.

Comme le montre l'image 5.5, l'idée est donc, lors de la préparation des jeux de données pour l'entraînement du modèle, d'encadrer le texte en italique ou en gras des balises ouvrantes et fermantes correspondantes, à la manière du langage HTML. Le modèle obtenu a été entraîné à partir de 300 pages de catalogues d'exposition et de ventes de manuscrits datant du XIX^{ème} siècle. Cela permet d'obtenir un résultat ayant un taux d'erreur de 0,89%. Cependant, si ce modèle est particulièrement fonctionnel et reconnaît l'information typographique, il est nécessaire de repasser sur la transcription obtenu à l'aide de programme python afin de bien corriger les balises ouvrantes et fermantes. Tout cela nécessite donc une relecture attentive en sortie d'OCR¹⁶.

5.2.3 La création d'un modèle de reconnaissance de l'emphase typographique

C'est donc dans ce cadre que mon stage s'est intégré au sein d'un groupe de travail pour la reconnaissance de l'emphase typographique ayant pour ambition de réaliser un modèle capable de récupérer ces informations depuis des documents historiques¹⁷.

13. Essentiellement de l'inférence Bayésienne, méthode d'inférence statistique par laquelle on calcule les probabilités de diverses causes hypothétiques à partir de l'observation d'événements connus.

14. C. Reul, Sebastian Göttel, U. Springmann, C. Wick, Kay-Michael Würzner et F. Puppe, « Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification : A Case Study on Daniel Sander's Wörterbuch der Deutschen Sprache », dans *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, Brussels Belgium, 2019, p. 33-38, DOI : 10.1145/3322905.3322910.

15. C. Corbières, *Du catalogue au fichier TEI : Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition...*, p. 37

16. *Ibid.*, p. 45

17. Ce groupe est composé d'Anna Scius-Bertrand, collaboratrice scientifique à l'HSE de Fribourg et doctorante en vision artificielle à l'EPHE, Simon Gabay et Thibault Clérice, tous deux cités précédemment. Il est à l'origine d'un article accepté à ICDAR 2021 et disponible en annexe (C.1).

type siècle \	Manuscrits	Expositions	Autres	Totaux
19ème	102	85	33	220
20ème	9	58	0	67
Totaux	111	143	33	287

TABLE 5.1 – Description numérique des pages du corpus de travail

La réalisation du jeu de données BIR

Le jeu de données BIR (*Bold-Italic-Regular*)¹⁸ est composé de 285 pages issues de catalogues d'exposition d'Artl@s et de ventes de manuscrits de Katabase. À cela s'ajoute quelques pages d'un dictionnaire français-latin du XIX^{ème} siècle. Si le *dataset* est composé essentiellement de documents du XIX^{ème} siècle en langue française, il possède également quelques pages datant du XX^{ème} siècle et 29 autres issues de la Biennale de Sao Paulo (en portugais) et de celle de Venise (en italien). Ces données sont issues des mêmes dépôts que les corpus de travail pour la création de modèle d'HTR et ont donc été préparées suivant la chaîne de traitement décrite précédemment¹⁹. Elles ont été sélectionnées parmi une quantité de pages afin d'avoir un *dataset* représentatif des différents catalogues possédant de l'emphase typographique.

Le *dataset* utilisé est composé de couples de *tokens* - c'est à dire des mots - associés à leur emphase typographique (italique, gras ou *regular*, soit sans emphase). L'idée était tout d'abord d'utiliser les fichiers PageXML qui contiennent la transcription des pages et les coordonnées de chaque mot pour former ces *tokens*. Cependant, le nombre de mots au niveau **Line**, qui correspondent donc la transcription corrigée, ne coïncide plus avec le nombre de mots au niveau **Word**, qui contient les coordonnées. Cela est dû aux nombreuses corrections manuelles faites sur ABBYY²⁰. Un programme python a donc été réalisé par Anna Scius-Bertrand pour corriger les mauvaises transcriptions au niveau mot par leur correction correspondante du niveau ligne. Je me suis occupée manuellement des lignes n'ayant pas pu être corrigées automatiquement par ce programme en coordonnant une campagne d'annotation²¹ à l'aide de l'outil VGG²². Celui-ci permet de charger une image et, pour chaque ligne problématique, de *tokeniser* les mots tel que présenté dans l'image 5.6 en leur associant un attribut (**g** pour gras, **i** pour italique, **vide** pour sans emphase)²³.



FIGURE 5.6 – Catalogue de vente de manuscrits, Charavay, 1843, p.18

Documents	35
Pages	285
Mots	88 019
- gras	2 106
- italique	5 745
- sans emphase	80 168

TABLE 5.2 – Le *dataset* BIR
(Issu de l'article)

18. Le jeu de données est disponible dans le dépôt github suivant : <https://github.com/asciusb/BIR-database>

19. Voir 3.2.1

20. À l'instar des problèmes lors de la migration Transkribus-eScriptorium décrits dans le 3.3.2

21. J'ai été aidée pour ce faire par Simon Gabay et Ljudmila Petkovic. Un manuel d'annotation de l'information typographique a été réalisé dans ce contexte et est disponible en annexes (C.2).

22. Disponible ici : <https://www.robots.ox.ac.uk/~vgg/software/via/via.html>

23. Pour une explication plus détaillée, voir le manuel d'annotation en annexe (C.2).

Le modèle obtenu

Ce jeu de données a été utilisé par Anna Scius-Bertrand pour entraîner deux réseaux de neurones, MobileNetV2 et Xception. Deux *splits* distincts ont été réalisés sur les données afin de constituer des jeux différents de *train*, *val* et *test*. Dans un premier cas, 50% du *dataset* **BIR** est utilisé pour entraîner les réseaux de neurones, 25% sont utilisés pour valider les résultats obtenus et 25% servent à tester le modèle en sortie. Dans le second cas, les 35 documents composant le jeu ont été divisés en 5 sous-corpus²⁴ ce qui a permis de réaliser les trois *datasets* d'entraînement de façon plus représentative. Pour finir, cela permet d'obtenir le réseau de neurone avec le meilleur score (ici MobileNetV2, d'une précision de 95%), qui est capable pour une image contenant un mot, de donner une prédiction indiquant si le token est potentiellement en italique, en gras ou sans emphase.

Ajouter le modèle obtenu dans la chaîne de traitement ?

Ainsi, nous avons à présent à disposition un réseau de neurones qui reconnaît l'emphase typographique. Nous avions donc pour ambition d'intégrer celui-ci au sein de notre chaîne de traitement. L'idée était d'appliquer le modèle MobileNetV2 sur le fichier alto contenant la transcription en sortie de l'OCR. Pour ce faire, il faut appliquer le modèle sur chaque token que contient le fichier XML. Il s'agit donc, pour chaque balise **Word** que contient le fichier alto, de récupérer ses coordonnées puis de découper l'image associée afin d'obtenir une figure contenant uniquement le mot correspondant. Une fois cela fait, on applique le modèle sur l'image obtenue ce qui permet de récupérer une prédiction de l'emphase typographique du mot : g pour gras, i pour italique, rien pour sans emphase. L'idée serait soit d'ajouter ce résultat directement dans la balise *Style* du fichier alto, soit, si ce travail est réalisé au sein d'une chaîne de traitement d'extraction de données, d'utiliser cette donnée pour déterminer plus efficacement ce qu'est le mot étudié (un nom, une adresse, etc.).

Cependant, plusieurs problèmes se sont posés. Premièrement, face aux résultats médiocres des modèles d'OCR, il est nécessaire, comme expliqué précédemment, de passer par eScriptorium. Or ce logiciel permet d'obtenir en sortie un fichier pour lequel le niveau le plus bas est **Line**, au contraire du fichier en sortie de Kraken qui va jusqu'au niveau **Glyph**. Ainsi, il est impossible d'obtenir les coordonnées de chaque mot pour récupérer son token imaginé. De plus, je me suis trouvée face à des difficultés temporelles. J'ai ainsi récupéré assez tardivement le réseau de neurones. Or, il a fallu, à la fin de mon stage, faire des choix sur les éléments que je pouvais traiter et n'ai donc pas pu me pencher assez longtemps sur l'apprentissage de la gestion d'un réseau de neurones avec python, chose que je n'avais jamais faite auparavant. Ce qui a été décrit précédemment reste donc un prototype de chaîne de traitement, qui n'a pas pu être testé mais semble faisable avec plus de temps à disposition.

24. Chaque sous-corpus contenant 7 documents.

Troisième partie

Annotation de l'information récupérée

Chapitre 6

De l’ALTO à la TEI : encodage automatique de données sous la forme d’une application

Une fois l’information extraite des images, il s’agit de les annoter et structurer, de façon à ce que celle-ci soit exploitable. Pour ce faire, l’idée est de réaliser un programme capable de récupérer l’information textuelle contenue dans les fichiers ALTO et de la structurer, à l’aide de l’analyse de la mise en page du document, sous forme XML-TEI. On réalise alors un prototype, c’est-à-dire un exemplaire incomplet et non définitif d’un potentiel produit final. Cela permet de tester la faisabilité du projet, en d’autres termes, voir s’il est possible d’encoder automatiquement des catalogues par le biais d’un programme unique.

On travaille ici uniquement sur les catalogues d’exposition. Cette décision s’explique par le besoin de produire des données pour Artl@s, associé aux résultats de l’OCR. En effet, comme décrit plus tôt, ces documents sont beaucoup plus réceptifs que les catalogues de ventes de manuscrits aux modèles de segmentation produits. Face à la quantité réduite de données d’entraînement, ceux-ci sont pour l’instant capables de reconnaître uniquement les entrées étant séparées par de larges espaces, ce qui est le cas d’un grand nombre de catalogues d’exposition. Ainsi, on se concentre sur ces documents spécifiques afin de se faire une idée des possibilités qu’offre un programme entièrement construit de rien pour l’extraction et la structuration d’entrées issues de documents structurés.

6.1 La construction d’un prototype

6.1.1 Le choix d’une application python

Python est un langage de programme interprété¹. Développé à partir des années 1980 par Guido van Rossum, un ingénieur néerlandais, il sort véritablement en 1991 sous une licence libre proche de la licence BSD². Il fonctionne sur à peu près tous les systèmes informatiques (mac, linux, windows, etc) à partir du moment où il est téléchargé.

1. Cela signifie que le code est lu ligne par ligne par un interpréteur qui le traduit dans un langage compréhensible pour la machine.

2. BSD pour *Berkeley Software Distribution Licence*, une licence pour les logiciels qui permet leur réutilisation sans restriction.

L'objectif de python est d'offrir à la fois un langage de haut-niveau³ et une syntaxe simple et intelligible à l'œil nu.

Python est donc un langage qui n'est pas propriétaire⁴, fiable, intuitif, facile à comprendre et à maîtriser. Autre grand avantage, ce langage est adaptable à de multiples contextes et situations. En effet, il se base sur l'utilisation de bibliothèques spécialisées, ou librairies. Cela correspond à des morceaux de codes réalisant des tâches particulières qu'il est possible de charger et utiliser simplement. Python a également l'intérêt de posséder une grande communauté scientifique d'utilisateurs. C'est par exemple le langage le plus employé en humanités numériques, de par sa facilité d'apprentissage et sa flexibilité. Il s'agit également du langage de programmation enseigné au sein du master TNAH, ce qui me permet d'en posséder les bases. Python est donc le langage de programmation le plus accessible, autant pour moi que pour les chercheurs en humanités numériques. Ainsi, de par ses diverses fonctionnalités, on peut conclure que Python est un langage parfait pour le prototypage⁵ : simple, flexible, multiplateforme, compréhensible par le plus grand nombre, réutilisable...

Une autre idée aurait été de réaliser une chaîne de traitement en XSLT afin de passer des fichiers ALTO contenant la transcription à un fichier XML-TEI. Comme mentionné précédemment⁶, XSLT est un langage de transformation pour les fichiers XML. Il permet donc de passer d'un fichier XML à un document d'un autre format (HTML, etc) ou de passer d'un type de fichier XML à un autre. Ici, il s'agit de récupérer les informations en ALTO et de les structurer en TEI. Ce langage a besoin d'un processeur, ou moteur, pour fonctionner. Dans notre cas, il s'agit du processeur *Saxon*, qui est privé (et donc payant) et est intégré au logiciel *Oxygen*⁷. Cela ne permettrait donc pas d'obtenir une chaîne de traitement *OpenSource*, puisque nécessitant un logiciel payant. Une solution pour contourner ce problème serait d'appliquer les feuilles de transformation XSLT sur les fichiers XML en utilisant Python. En effet, certaines librairies, dont *lxml*⁸. Autant donc utiliser Python comme langage principal de programmation⁹.

Ainsi, le choix de la langue de programmation utilisée pour le prototype a donc été grandement influencé par mes compétences personnelles. En effet, il s'agit d'un langage que je maîtrise¹⁰. Malgré tout, Python est un outil particulièrement utilisé dans le monde de la recherche et plus particulièrement en humanités numériques. L'employer dans le cadre de ce projet ne pose donc pas vraiment un problème. La décision s'est donc rapidement portée sur l'utilisation de ce langage comme langue de programmation de base du

3. Un langage de haut niveau est un langage de programmation tourné autour de la résolution d'un problème qui permet d'écrire des programmes utilisant des mots usuels de langues naturelles. Dans le cas de python, comme pour beaucoup d'autres langages de programmation, il s'agit de l'anglais.

4. Se dit d'un langage privé

5. Le prototypage est la démarche visant à réaliser un prototype

6. Voir 2.2.2

7. *Oxygen* est un éditeur XML. Il permet de lire, visualiser et travailler sur des documents XML et des feuilles de transformation XSLT et possède donc un moteur intégré *Saxon*. Il s'agit d'un logiciel payant et privé. (<https://www.oxygenxml.com/>)

8. *lxml* est une librairie python qui donne accès à des fonctions permettant de traiter les documents XML directement dans Python. Parmi celles-ci se trouvent notamment la possibilité d'appliquer des feuilles de transformation XSLT sur des fichiers XML. (<https://lxml.de/>)

9. Quelques feuilles de transformation XSLT sont par ailleurs appliquées dans le prototype final. (https://github.com/Juliettejns/extractionCatalogs/blob/main/fonctions/Restructuration_alto.xsl)

10. J'aurais pû apprendre un autre langage pour réaliser l'application, cependant, face à l'utilisation presque commune de Python dans la recherche, cela semble un peu superflu et contraignant pour l'élaboration d'un prototype en deux mois.

prototype.

6.1.2 La structuration du programme

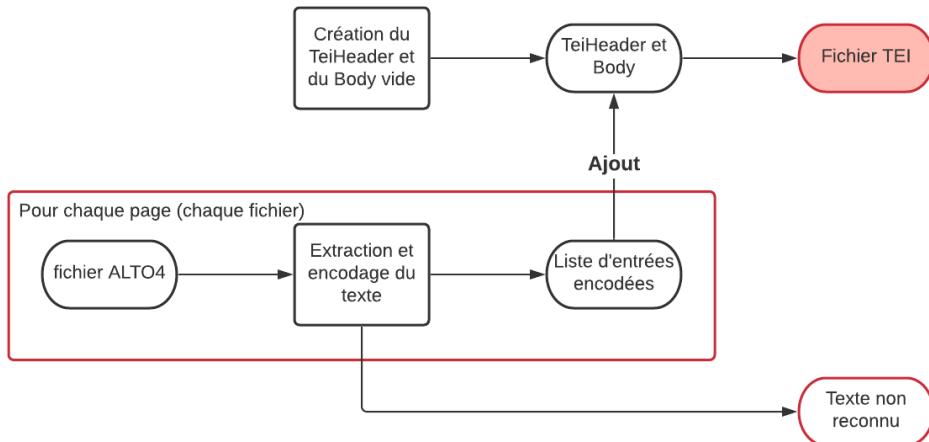


FIGURE 6.1 – Fonctionnement schématique du prototype d'encodage

Pour comprendre le fonctionnement de ce programme, il convient d'expliquer la structure du produit sorti, le fichier XML-TEI. Ce document reprend le schéma créé par Caroline Corbières lors de son stage, sur la base du travail d'encodage des catalogues de ventes de manuscrits réalisé par Lucie Rondeau du Noyer. En tant que fichier TEI, le résultat se divise en deux grandes parties :

- Un **TeiHeader** : Cette première grande partie du document correspond aux métadonnées. Elle contient donc toutes les informations sur le fichier lui-même, sa source première et sa réalisation. Ainsi, on y retrouve le nom de la personne qui a réalisé la chaîne de traitement, diverses données sur le catalogue en temps que source papier - titre, auteur, date et lieu de l'évènement dont il est issu, dans le cas des catalogues d'exposition - et sur la construction du fichier TEI - logiciels ayant permis la formation, nom du correcteur, etc.
- Un **body**, soit, en français « corps » : de par son nom, on peut aisément saisir que cette partie contient le corps du texte, soit l'information textuelle extraite des fichiers alto. S'y trouvent toutes les entrées¹¹, encodées, d'un catalogue.

Ainsi, le prototype réalisé peut être schématisé en deux parties assez distinctes, comme visible sur l'image 6.1. Dans un premier temps, le programme créé un objet contenant un squelette de fichier XML-TEI. Celui-ci est formé des balises du **TeiHeader**, tel qu'il doit l'être selon le schéma d'*Artl@s*, suivies d'une balise **body**. Il s'agit de balises vides, qui n'encadrent rien. On parle alors de balises auto-fermantes. Une fois cela fait, il est alors possible de s'occuper de l'information textuelle à récupérer et encoder. Pour ce faire, chaque fichier ALTO4 est traité l'un après l'autre de la même façon. Pour chaque page, on récupère le contenu textuel présent sur les lignes situées dans des régions *Entry*. Le texte est encodé entrée par entrée afin d'obtenir une liste d'entrées encodées. Il suffit alors d'intégrer cette liste dans le squelette XML-TEI à l'intérieur du **body**. Les régions

11. Ces entrées sont encodées par un élément **entry**, qui ne fait pas partie des balises TEI. Celui-ci a été ajouté spécialement pour traiter les catalogues en XML-TEI suite à une mûre réflexion sur leur encodage réalisée par Caroline Corbières et Laurent Romary.

EntryEnd sont un peu plus compliquées à gérer. Lorsqu'une page en contient une, elle est traitée à part et associée directement au **body**.

Ainsi chaque fichier alto4 permet d'obtenir une liste d'entrées encodées, qui sont ajoutées les unes après les autres dans le **body**, jusqu'à obtenir un élément XML contenant toutes les entrées de toutes les pages d'un catalogue. Il est alors possible d'imprimer cet objet en sortie sous la forme d'un document XML-TEI. Un autre fichier texte, obtenu en sortie, contient l'information textuelle des entrées n'ayant pas été reconnues par le programme.

Comme il est possible de le voir dans le dépôt contenant le code de ce prototype¹², le programme est structuré sous la forme de dossiers. En effet, l'intérêt de python réside également en ce qu'il permet de réaliser des applications avec des modules, c'est-à-dire plusieurs fichiers contenant des morceaux de codes et structurés sous la forme de différents dossiers. Ainsi, un fichier *run.py* permet de lancer le programme, qui lui-même utilise des fonctions contenues dans les dossiers *test*, dans le cas des vérifications de la qualité des fichiers d'entrée et de sortie, et *fonctions*, pour extraire et encoder les données textuelles.

6.2 L'extraction de données depuis les fichiers alto

Afin de présenter au mieux comment les données sont extraites depuis les fichiers alto puis encodées, il s'agit dans un premier temps de faire un panorama des différentes entrées existantes que l'on va traiter.

6.2.1 Recensement des différentes typologies de catalogues

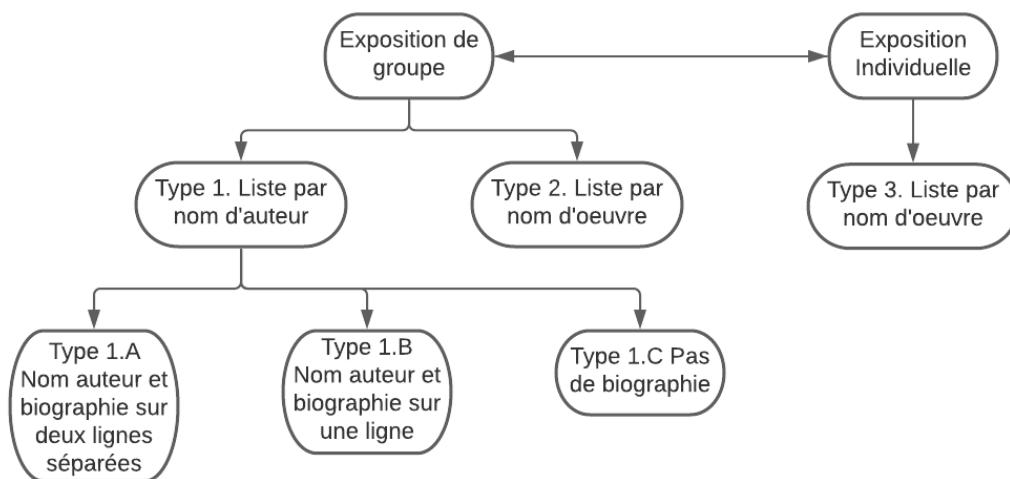


FIGURE 6.2 – Schéma récapitulatif des types de catalogues

Les catalogues, bien qu'étant des documents structurés, ne sont pas toujours identiques, car de provenance et de périodes diverses. Ainsi, pour pouvoir les traiter le mieux

12. <https://github.com/Juliettejns/extractionCatalogs>

possible, il est nécessaire, dans un premier temps, de réaliser un panorama¹³ des formes que les entrées peuvent prendre au sein des données de *Basart*.

Le schéma 6.2 récapitule les différents types de catalogues que l'on peut rencontrer et les entrées qui leur sont associées. Comme expliquer précédemment¹⁴, il existe deux grands types d'expositions, qui influent sur la forme des catalogues. Les premières sont des expositions de groupes, c'est-à-dire que plusieurs peintres sont exposés, tandis que les secondes sont monographiques et ne concernent qu'un seul peintre. Les catalogues des expositions individuelles sont donc des listes d'œuvres (**Type 3**). Les expositions de groupe peuvent avoir des catalogues similaires (**Type 2**), essentiellement en Angleterre et aux Etats-Unis, mais leur version la plus commune correspond à des entrées par auteurs (**Type 1**). Étant donné que la plupart des catalogues à encoder pour *Basart* sont de **Type 1** et qu'il s'agit du type que le modèle de segmentation réussit le mieux à diviser en entrées, il a été décidé de se concentrer sur l'extraction et l'encodage de ceux-ci. Ainsi nous présenterons ici une description détaillée de l'aspect des entrées de ces catalogues¹⁵.

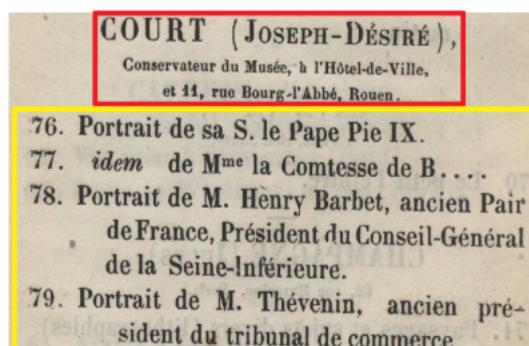


FIGURE 6.3 – Catalogue de l'exposition annuelle du musée de Rouen, 1853, p.12

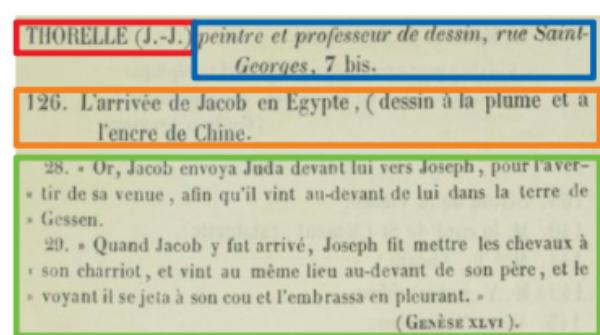


FIGURE 6.4 – Catalogue d'exposition des Beaux Arts de Nancy, 1849, p.11

L'image 6.3 est un exemple de l'aspect général d'une entrée de **Type 1**. Il est possible de distinguer ici deux grandes parties. Une première, en rouge, concerne l'auteur et une seconde, en jaune, présente la liste numérotée des œuvres qu'il a exposées lors de l'évènement. La structure décrite peut être subdivisée en plusieurs parties, tel que présentées dans la figure 6.4. Si chaque entrée contient au moins le nom de l'auteur, en rouge, et le nom de l'œuvre, en orange, il peut aussi être indiqué des informations biographiques, en bleu, et de multiples informations sur l'œuvre en question, en vert. Chacuns de ces éléments ont également un aspect qui peut être variable et que l'on va décrire.

13. Ce panorama se base sur un travail de recensement réalisé par Barbara Topalov, doctorante au sein de *Visual Contagions* ainsi qu'un article en cours de publication : « Automating Artl@s, extracting datas from exhibition catalogs » .

14. Voir 1.2.2.

15. Des exemples des catalogues de **Type 2** et **Type 3** sont cependant disponibles en annexe A.1.

FEURE (E. de).

- 90 — Fleur du mal.**
- 91 ... Feux follets.**
- 92 — Vision.**
- 93 — Décoration.**

FIGURE 6.5

L'image 6.3 est un exemple du **type 1.A**. La première partie donne le nom de l'auteur, la seconde des informations sur celui-ci, telles que son adresse de domiciliation, son école ou sa nationalité. La dernière partie est composée des œuvres de l'auteur présentées lors de l'exposition, sous la forme d'une liste numérotée. L'image 6.4 correspond au **type 1.B**. Elle est composée de deux parties distinctes, la première présentant le nom de l'auteur, puis des informations complémentaires sans sauter de lignes, et la seconde comportant la liste numérotée des œuvres exposées par l'auteur¹⁶. L'image 6.5 est un exemple de **type 1.C**. Ici, l'entrée se résume au nom de l'auteur et à la liste des œuvres qu'il a exposées.

Ces types sont donc variés et peuvent cohabiter au sein d'un même catalogue. L'idée est de sélectionner l'aspect prédominant en entrée de programme afin d'avoir le plus d'information textuelle encodée correctement. Pour ce faire, quatre termes ont été créés :

- **Nulle** : pour les catalogues ayant uniquement des entrées similaires au **type 1.C**.
- **Simple** : pour les catalogues ayant essentiellement des entrées de **type 1.B**.
- **Double** : pour les catalogues ayant essentiellement des entrées de **type 1.C**.
- **Triple** : pour les catalogues ayant des entrées similaires au **type 1.B** pour lesquel la ligne numérotée contient à la fois le nom de l'œuvre et des informations complémentaires, qu'il est possible de diviser¹⁷.

La volonté de la partie « extraction et encodage des informations » du prototype est donc de reconnaître ces différentes parties, de les séparer et de les encoder.

6.2.2 Expressions régulières et structure des entrées de catalogues

Il s'agit donc de repérer, pour chaque entrée, les parties qui la composent : nom d'auteur, informations biographiques, titre des œuvres exposées... En tant que document structuré, chacun de ces éléments au sein d'un même catalogue aura un aspect similaire. On peut alors parler de motif. C'est par exemple le cas pour les noms de peintres, qui sont fréquemment présentés par une suite de lettres majuscules ou encore les œuvres exposées, dont la ligne peut commencer par un numéro¹⁸. Ainsi, il est intéressant de penser la récupération de ces informations en jouant sur les motifs : plusieurs lettres majuscules au début de la première ligne de l'entrée peuvent faire penser qu'il s'agit là du nom de l'auteur tandis que des numéros en début de ligne correspondent potentiellement à une œuvre, etc.

Cette reconnaissance de chaînes de caractères par le biais de motifs peut être réalisée, en informatique, grâce à des « expressions régulières¹⁹ ». Également appelées regex²⁰, il s'agit d'une suite de caractères - aussi appelée motif ou *pattern* en anglais - qui décrit un fragment de texte. Elle peut être composée à la fois de caractères normaux (alphanumériques ou non) et de « métacaractères », c'est-à-dire des caractères qui ont une signification particulière. C'est par exemple le cas de l'élément « ^ » qui signifie que le motif recherché

16. On trouve également ici une information complémentaire sur l'œuvre mais cela est facultatif dans ce type de catalogue.

17. Ce dernier type est encore en phase de test complet et a été créé spécialement pour gérer la production d'un catalogue des Indépendants. On peut donc ne pas le compter comme un type à part entière mais plutôt comme un test intermédiaire.

18. Ces types de motifs sont visibles dans les images 6.3, 6.4 et 6.5.

19. Pour plus d'informations sur le sujet : https://python.sdv.univ-paris-diderot.fr/16_expressions_regulieres/.

20. Une contraction de leur nom en anglais : *regular expressions*.

se trouve en début de ligne et ne correspond pas à « accent circonflexe ». Les expressions régulières sont utilisables en python, par le biais d'une librairie spécialisée²¹. Ainsi elles sont tout à fait utilisables au sein de notre chaîne de traitement. Il s'agit même d'un outil de choix pour l'extraction de données, notamment en humanités numériques. Elles ont par exemple été utilisées pour le traitement des annuaires par le groupe Annuaire de Paris Time Machine, afin de repérer les différents éléments qui composent une entrée de ces documents.

Afin d'illustrer l'utilité ainsi que la construction d'une expression régulière, nous allons prendre en exemple les noms de peintres. L'image 6.6 correspond à la structure la plus basique et la plus commune que peuvent prendre les noms d'artistes dans les catalogues d'exposition. Le nom y est en majuscules, suivi du prénom, en minuscules et entre parenthèses, puis d'une virgule. Cette structure peut varier quelque peu en fonction des catalogues, notamment au niveau de la ponctuation utilisée.

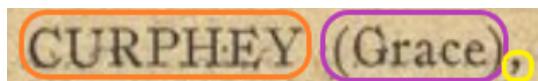


FIGURE 6.6 – Catalogue [...] Indépendants, 1913, p.78



FIGURE 6.7 – Exemple de regex

L'image 6.7 correspond à une expression régulière capable de reconnaître et extraire l'information textuelle que l'on voit sur la figure 6.6. L'élément en rouge signale que l'on recherche une chaîne de caractères située en début de ligne. Ce qui est en orange signifie que l'on souhaite récupérer au moins deux lettres en majuscules qui se suivent²², soit le nom de famille. Le carré violet signale qu'à la suite de ces lettres en majuscules doivent se trouver un élément du type « (lettre en majuscule puis plusieurs lettres en minuscule) »²³, soit le prénom. Enfin, en jaune, on retrouve la virgule, qui sépare dans ce cas-ci le nom du peintre de sa biographie.

Si l'expression régulière présentée ici permet effectivement de récupérer les noms de peintre de forme « NOM (Prénom) », il ne s'agit pas là d'une regex réellement utilisée dans le programme. En effet, les résultats issus de l'OCRisation des catalogues contiennent un certain nombre d'erreurs, ce qui ne permet pas d'utiliser des expressions régulières aussi précises. Par exemple, il peut fréquemment y avoir des caractères de type parenthèses ou autres au début de ces lignes, ce qui influe donc sur la structure des regex construites. Le but est vraiment de pallier à un maximum d'éventualités. Ainsi, on signalera au début de ce type d'expressions régulières qu'avant le nom en majuscule, il y a un risque d'avoir une parenthèse.

Des expressions régulières sont ainsi construites afin de cibler les différents éléments d'une entrée. Si elles sont assez simples pour le nom du peintre et les titres d'œuvres, qui sont particulièrement reconnaissables, de par les majuscules et nombres, ce n'est pas le cas pour le reste des informations. Ainsi, la biographie est plutôt indiquée comme étant

21. La documentation de la librairie `re` est disponible ici : <https://docs.python.org/3/library/re.html>

22. `[A-Z]` signifie toutes les lettres majuscules entre A et Z et `*` indique que l'on souhaite la répétition d'un motif (ici les lettres en majuscule) plus de deux fois.

23. Les parenthèses correspondent à des métacaractères en regex. Le slash avant celles-ci permet donc de signaler que l'on souhaite rechercher les parenthèses en tant que caractères, soit leur seul littéral, et non en tant que métacaractères.

l'élément juste après le nom de l'auteur, en jouant sur la séparation entre les deux. Dans le cas traité précédemment, il s'agit donc de décrire que tout ce qui est entre le nom et la première œuvre de l'entrée correspond à la biographie. Le problème est plus compliqué avec les informations complémentaires sur des œuvres particulières. La solution trouvée à ce jour est traité en récupérant les lignes qui ne commencent pas par des numéros dans la liste des œuvres. Elle reste cependant perfectible car ne prend pas en compte grand nombre d'informations complémentaires. C'est par exemple le cas des informations concernant les matériaux de construction des œuvres, qui se situent la plupart du temps à la fin de la ligne numérotée, sous la forme « (matériaux) ». Si plusieurs moyens pour récupérer ces données sont testés²⁴, le temps a joué en notre défaveur et n'a pas permis d'obtenir un produit final pour cette partie. En effet, le problème ici est de distinguer les éléments entre parenthèses qui concernent des matériaux et sont donc des informations complémentaires de ceux qui donnent l'emplacement du lieu peint²⁵ et font parti du titre de l'œuvre. Pour se faire, l'idée serait de réaliser un petit lexique des différents matériaux existants et de simplement faire vérifier par le code que l'élément entre parenthèses ressemble à l'un d'eux, ce qui permettrait de déterminer s'il s'agit d'informations complémentaires ou du titre.

Un dernier problème sur le point des expressions régulières concerne la variété des motifs et séparateurs existants au sein des catalogues. En effet, si la plupart des catalogues suivent un schéma précis, rentrant parmi les différentes catégories décrites précédemment, chacun d'entre eux garde des spécificités propres. Dans le cas des noms de peintres, ceux-ci pourront prendre plusieurs formes. Ainsi, la structure « NOM (Prénom) », qui nous avons vu n'est pas unique. Le prénom peut ne pas être entre parenthèses, le séparateur peut varier entre des points, des virgules, des tirets cadratins, le nom peut n'avoir que sa première lettre en majuscule, etc²⁶. Ces variations existent également pour le reste des éléments qui composent une entrée. Pour pallier à ce problème, différentes expressions régulières ont été réalisées en fonction des catalogues qui ont pu être traités au cours de l'élaboration du programme. Ceux-ci ont été rassemblés dans un même fichier²⁷ et il suffit de les activer ou les désactiver afin d'utiliser les expressions régulières qui correspondent au catalogue que l'on veut traiter.

6.3 En sortie : résultats et perfectionnement du prototype

6.3.1 Présentation du fichier TEI obtenu

Comme indiqué dans l'image 6.1, la sortie finale du programme est donc un fichier XML-TEI. Celui-ci se conforme au schéma déterminé par Caroline Corbières sur la base du travail de Lucie Rondeau du Noyer. Se conformer à cette structure était obligatoire

24. C'est ce qui a été testé avec le type de catalogue **Triple**.

25. On retrouve beaucoup plus fréquemment des éléments entre parenthèses indiquant des noms de lieux ou le département où la peinture a été réalisée que des matériaux. Ainsi, dans un souci de relecture plus rapide du catalogue encodé obtenu, il a été choisi de ne pas intégrer cette idée au code pour le moment.

26. Pour un détail des variations existantes, voir le manuel de typologie des catalogues en annexe (C.2).

27. Le fichier en question est disponible ici : https://github.com/Juliettejns/extractionCatalogs/blob/main/fonctions/instanciation_regex.py.

car notre travail s'inscrit dans un cadre plus large de production de données. Nos catalogues encodés doivent donc correspondre à la même forme que les autres catalogues déjà produits, dans le but d'associer ces nouvelles données à Basart. Certaines petites améliorations ont pu cependant être réalisées dans un souci de praticité, mais celles-ci n'endommagent pas la structure globale du fichier et des différentes balises.

```

<entry n="1" xml:id="CatSalonRoseCroix_1893_e1"
source="https://gallica.bnf.fr/iiif/ark:/12148/bpt6k54703109/f12/full/full/0/default.jpg">
<desc><name>AMAN-JEAN.</name><trait>
<p>– 15, quai Bourbon</p>
</trait></desc>
<item n="1" xml:id="CatSalonRoseCroix_1893_e1_i1">
<><num>1</num><title>– Affiche de la seconde geste
esthétique.</title></item>
<item n="2" xml:id="CatSalonRoseCroix_1893_e1_i2">
<><num>2</num><title>– Rêverie.</title></item>
</entry>
```

FIGURE 6.8 – Exemple d'entrée encodée

L'image 6.8 est un exemple²⁸ de ce à quoi ressemble une entrée encodée en sortie du prototype. Ainsi, comme on peut le remarquer ici, une entrée est effectivement encadrée par une balise **entry**. Le préambule de l'entrée, contenant le nom du peintre - dans **name** - et ses potentielles informations biographiques - dans **trait** - est encadré d'une balise **desc**. Celle-ci est suivie par autant de balises **item** qu'il y a d'œuvres listées. Pour chaque œuvre, le numéro mentionné est encadré d'une balise **num** et le titre l'est par **title**. Lorsqu'il y a des informations complémentaires sur une œuvre, ce qui n'est pas le cas ici, elles sont encodées par un nouvel élément **desc** situé à l'intérieur de la balise **item**. Chaque **entry** et **item** est affublé d'un attribut **@n** qui numérote chaque élément et d'un **@id** qui permet d'obtenir un identifiant unique par entité. Un dernier attribut, **source**, a été ajouté dans le programme pour récupérer la page numérisée où se trouve l'entrée encodée. Ici, il s'agit d'un catalogue produit à partir des informations IIIF, on renvoie donc directement à la page numérisée dans gallica.

Le résultat obtenu nécessite des corrections manuelles et n'est pas tout de suite conforme à son schéma. Par exemple, il est nécessaire de compléter à la main les métadonnées dans le **TeiHeader**. Ce choix de laisser cette partie du fichier XML-TEI sous la forme d'un squelette vide s'explique par la nécessité de toujours modifier différemment ces informations. Ainsi, il était plus simple de permettre à la personne ayant traité le fichier de directement ajouter ces données elle-même. En dehors de cela, il est également nécessaire de relire et corriger les entrées encodées. Faire cela permet non seulement d'obtenir des fichiers corrects et de meilleure qualité mais également de réaliser un panorama des erreurs obtenues afin de potentiellement réarranger le code pour les éviter par la suite. Cette correction est réalisée grâce aux erreurs indiquées dans le logiciel Oxygen par l'ODD²⁹. Celui-ci peut être défini comme un langage de définition et de maintenance pour la TEI

28. Cet exemple est issu du fichier XML-TEI disponible dans le dépôt https://github.com/Juliettejns/TEIcatalogs/blob/main/catalogs/exhibCat_SalonRoseCroix_1893/TEI_CatSalonRoseCroix_1893.xml. Il s'agit de l'encodage de la première entrée du catalogue de l'exposition Rose Croix de 1893.

29. Les initiales de *One Document Does it all*.

et par extension le fichier XML contenant ce langage qui permet de spécifier l'aspect et les différentes balises que doit prendre le fichier XML-TEI qui lui est associé. À ceci se joint une documentation poussée des choix d'encodage faits au sein du projet. Ici, nous reprenons donc l'ODD réalisé par Caroline Corbières au cours de son stage³⁰.

En l'associant au fichier XML-TEI obtenu en sortie du prototype, chaque erreur est repérée et mentionnée dans le logiciel et il suffit alors de la corriger. Si une grande partie des fautes concerne des problèmes d'OCR et sont seulement corrigables manuellement, d'autres sont plus compliquées. La plupart d'entre elles sont des problèmes résolvables en modifiant le programme³¹, ce qui n'a pas été fait par manque de temps. D'autres correspondent à des situations spécifiques assez rares qui ne semblent pas nécessiter un développement particulier. Ceci prendrait en effet plus de temps de réaliser un programme capable de corriger ces problèmes que de les remanier manuellement. C'est notamment le cas des lignes concernant plusieurs œuvres en même temps et qui sont donc numérotées de cette façon « numéro à numéro ». Enfin, certains éléments peuvent être signalés comme des erreurs mais n'en sont pas. Cela est dû à la structure particulière des catalogues. L'ODD de vérification du XML-TEI est réglé de façon à vérifier que les numéros des œuvres se suivent bien. Or, certaines œuvres peuvent avoir mal été numérotées³² ou encore certains numéros sont présents plusieurs fois³³. Ainsi, il ne s'agit pas d'erreurs mais simplement de la transposition même du catalogue et de ses problèmes. L'ODD n'étant pas aussi malléable, il reconnaîtra ces situations comme des erreurs et il est donc nécessaire de procéder à une vérification humaine.

6.3.2 Améliorations et limites des résultats

Le programme obtenu fonctionne ainsi pour un type spécifique d'entrées. En tant que prototype, il s'agit essentiellement de montrer les possibilités que permet un programme python de récupération et de structuration de l'information textuelle contenue dans les fichiers en sortie d'OCR. Cependant, cela n'empêche pas qu'il est impossible de traiter, extraire et structurer les données de certains catalogues uniquement avec des expressions régulières, ce qui est fait ici. Ainsi, dans certains cas, les œuvres ne sont pas numérotées. La seule façon de les distinguer est donc de reconnaître l'emphase typographique : à chaque élément en gras en début de ligne, on peut supposer qu'il s'agit d'une œuvre. Dans d'autres cas, il n'y a aucun changement visible autre que des sauts de lignes entre les différents éléments d'une entrée. C'est assez fréquent pour les catalogues surréalistes, qui aiment à expérimenter. Il est alors beaucoup plus difficile de trouver un moyen de gérer ce problème, qui ne semble pas résolvable par le biais d'un simple programme python : il est nécessaire de revenir à l'image, comme GROBID peut le faire, afin de déterminer l'emplacement de chaque élément dans l'entrée.

D'autres initiatives ont été testées pour améliorer les résultats du prototype, en plus

30. Celui-ci est disponible ici, https://github.com/carolinecorbieres/ArtlasCatalogues/tree/master/5_ImproveGROBIDoutput/ODD. Plusieurs versions d'ODD cohabitent afin de cibler les différents types d'erreurs dans un premier temps, puis d'obtenir un résultat conforme.

31. C'est par exemple le cas des entrées n'ayant pas le prénom du peintre, dans le cas des catalogues de type **Double**. Ceux-ci sont mal reconnus et nécessiteraient simplement l'ajout d'une nouvelle condition permettant de les repérer et les traiter avec de nouvelles expressions régulières spécifiques à ce genre d'entrées, ce qui est assez courant.

32. L'œuvre « 110 » suit l'œuvre « 105 » par exemple.

33. Cela peut être une erreur dans l'impression du catalogue ou alors, plus fréquent, des œuvres qui sont numérotées en « numéro », « numéro bis », « numéro ter », etc.

de la typologie des erreurs. Parmi elles, la principale vise à instaurer tout un panel de vérifications de la qualité des fichiers d'entrées, donc alto. Chaque fichier est ainsi étudié afin de vérifier sa validité et si chaque **TextBlock** est conforme. Dès que le fichier n'est pas valide - des balises ne sont pas fermées par exemple -, qu'une région n'est pas taggée³⁴ ou encore que des **TextLines** ne se trouvent pas au bon emplacement³⁵, cela est signalé à l'utilisateur dans le terminal. Il a alors la possibilité de vérifier l'erreur dans le fichier alto et de déterminer si celle-ci est corrigable. Auquel cas, elle est manuellement supprimée et le programme est relancé. Cette procédure permet d'obtenir des fichiers alto de bonne qualité, qui permettront de récupérer un meilleur encodage XML-TEI, celui-ci se basant sur la structure des **TextBlocks**.

En plus d'être potentiellement mal formé, un fichier alto peut avoir ses régions et lignes dans le désordre. Par exemple, la première entrée ne se trouve pas forcément au début du fichier. En effet, en passant par eScriptorium pour nettoyer les données, la structure du document est transformée : chaque élément corrigé est traité comme un nouvel objet et se retrouve donc à la fin du fichier alto. Ainsi, l'ordre des différents éléments ne correspond plus à l'ordre de lecture des entrées et des lignes mais à l'ordre dans lequel celles-ci ont pu être corrigées. Or, l'ordre de ces objets a une importance extrême pour le prototype, puisqu'il se base sur ceux-ci pour extraire les différentes informations. Il est donc nécessaire d'obtenir un fichier alto bien ordonné avant de traiter les informations textuelles qu'il contient. Pour se faire, une feuille de transformation XSLT³⁶ a été créée permettant de restructurer dans le bon ordre³⁷ chaque région puis chaque ligne. Elle est appliquée sur chaque page traitée au sein du programme python avant le travail d'extraction d'informations.

Il peut également intéresser de réfléchir à aller plus loin dans l'encodage des données et donc à l'extraction des diverses informations. En effet, au contraire de GRO-BID, qui analyse les différentes parties d'une page par le biais de leur position dans l'image, ce prototype travaille directement sur l'information textuelle. Cela permet donc une extraction potentiellement plus profonde des données, à un niveau plus bas. Par exemple, dans le cas de l'image 6.10³⁸, il serait possible d'encoder le nom et prénom de l'auteur séparément mais également de décrire de quel type de données il s'agit pour chaque élément situé dans la biographie : lieu de naissance, maître, médailles et adresse. Cela permettrait notamment de réaliser le travail de récupération des différentes données pour leur ajout dans Basart beaucoup plus facilement. Cet exemple permet aussi d'illustrer l'utilité d'un modèle permettant de reconnaître l'emphase typographique dans le cadre de ces ajouts : pour réaliser ce nouveau niveau plus fin d'encodage, il serait nécessaire de la reconnaître. Ici, par exemple, les éléments en italique correspondent au maître du peintre. Si cette idée peut s'avérer particulièrement intéressante, elle oblige le programme à se spécialiser encore plus pour un type spécifique de catalogues. En effet, comme présenté précédemment, chaque catalogue a ses particularités au niveau même de la structure des différents éléments qui composent une entrée. En effet, les noms et prénoms ne sont pas toujours

34. Cela signifie qu'elle n'est pas nommée. On ne sait pas s'il s'agit d'un *Main*, *Entry* ou autre, il manque donc des informations au sein du document.

35. On vérifie qu'il n'y pas de lignes en dehors des régions et que la plupart des lignes sont dans des régions taggées *entry*.

36. Ce fichier est disponible ici : https://github.com/Juliettejns/extractionCatalogs/blob/main/fonctions/Restructuration_alto.xsl

37. Les éléments sont triés dans le bon ordre en s'aidant de leur coordonnées verticales.

38. Il s'agit d'un exemple typique de la variété de données que peut contenir la biographie d'un peintre dans une entrée.

indiqués de la même façon, tout comme l'information biographique ou complémentaire aux œuvres. Ainsi, de la décision d'un encodage plus fin, bien que permettant une importation plus facile des données dans Basart, résulte une suppression de la tentative de traiter tout type de catalogues. Il a donc été préféré de s'en tenir à un encodage moins précis.

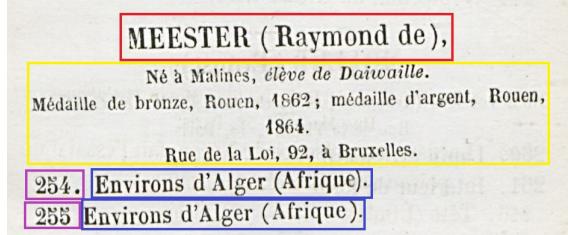


FIGURE 6.9 – Elements récupérés,
Catalogue [...] exposées, Beaux
Arts de Rouen, 1869, p.4

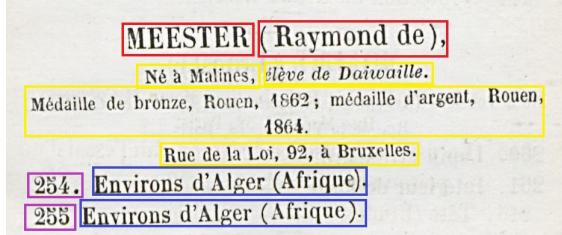


FIGURE 6.10 – Elements récu-
pétables, *Catalogue [...] exposées*,
Beaux Arts de Rouen, 1869, p.4

Chapitre 7

De l'image à la TEI : Un prototype de chaîne de traitement complète

L’assemblage des différentes briques réalisées, modèles d’OCR, réflexions autour des formats de stockage et prototype d’extraction de données, permet de réaliser une chaîne complète de traitement des catalogues. Il est alors possible de passer de l’image numérisée au fichier XML-TEI structurant les données présentes. Ce dernier chapitre décrit donc cette chaîne de traitement et les choix qui ont été faits pour la former et présente des exemples de production de données réalisés grâce à cette chaîne. Il est l’occasion de mener une réflexion globale autour des décisions prises pour réaliser cette chaîne, de la fonctionnalité et l’utilité de celle-ci dans le cadre d’un projet tel qu’Artl@s.

7.1 Une chaîne de traitement cyclique

7.1.1 Le chemin de la production d’un catalogue encodé

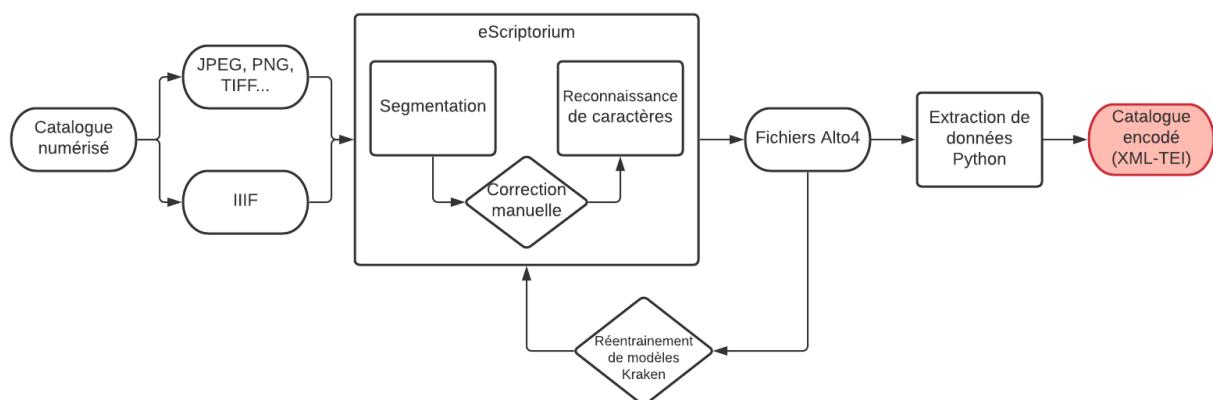


FIGURE 7.1 – Chaîne de traitement complète

L’image 7.1 résume bien la chaîne de traitement mise en place pour produire des catalogues encodés en XML-TEI depuis leur version numérisée. Celle-ci, en entrée de la chaîne de production, peut être soit dans un format basique de type JPEG, JPG, PNG, ou même TIFF, ou bien en IIIF¹. En effet, tous ces formats sont traitables dans eScriptorium,

1. Dans le cas de IIIF, il ne s’agit pas d’un ajout de fichier mais d’un ajout d’un url contenant

ce qui permet de les utiliser au sein de notre chaîne. Une fois que le catalogue numérisé y est entré, il s'agit d'appliquer le modèle de segmentation, qu'il est nécessaire par la suite de corriger, afin de vérifier les potentielles erreurs de détection de lignes - et parfois de zones. On réalise la reconnaissance de caractères avec le modèle correspondant. Il faut ici aussi relire le résultat obtenu et corriger manuellement les erreurs.

Ce travail de correction des résultats issus de l'application des modèles de segmentation et reconnaissance de caractères est nécessaire pour le moment. En effet, au début de l'utilisation de cette chaîne de traitement, on emploie les modèles **Chaource**. Comme signalé précédemment, ceux-ci ne permettent pas d'obtenir des résultats de très bonne qualité. Cette opération de vérification permet donc de récupérer une transcription acceptable et utilisable par la suite. Celle-ci donne également la possibilité de s'assurer de l'aspect des fichiers alto obtenus en sortie. Ils peuvent alors faire office de nouveaux jeux de données pour entraîner des nouveaux modèles plus performants, puisqu'il s'agit là de transcriptions et segmentations correctes. Ainsi, même si cette étape de correction et vérification est laborieuse sur le moment, elle permet de développer de nouveaux modèles, qui seront potentiellement de meilleure qualité et donc remplaceront **Chaource** dans la chaîne de traitement. On évolue alors petit à petit vers des résultats qui nécessitent de moins en moins de corrections manuelles, jusqu'à pouvoir réaliser une chaîne de traitement entièrement automatique en passant par Kraken² directement.

Par la suite, on applique sur la transcription, exportée d'eScriptorium sous la forme de fichiers alto, le prototype python d'extraction de données. Comme signalé précédemment, il est possible d'avoir des problèmes face à des fichiers alto mal formés, qui sont signalés dans le terminal. Il est alors nécessaire de les corriger puis de relancer le programme jusqu'à ne plus obtenir d'erreurs. On récupère alors le catalogue encodé en XML-TEI, qu'il faut encore corriger à l'aide de l'ODD.

Une dernière étape, non mentionnée dans le schéma, existe. Celle-ci consiste à appliquer une feuille de transformation XSLT sur le fichier XML-TEI obtenu afin de récupérer les données textuelles encodées sous la forme d'un csv. Celui-ci répertorie les différentes informations et permet une correction plus rapide et plus simple. Cela permet également de contourner le problème de l'utilisation d'Oxygen, logiciel payant, dans le cadre de l'élaboration d'une chaîne de traitement libre et ouverte.

7.1.2 De nouvelles données pour de nouveaux modèles d'HTR

Ainsi, la chaîne de traitement construite est cyclique. Les données, fichiers alto contenant la segmentation et la transcription, produites par le biais de la chaîne, alimentent un nouveau *dataset* permettant d'entraîner de nouveaux modèles d'HTR. Dans l'idée, ceux-ci remplacent leurs prédecesseurs et sont appliqués sur de nouveaux catalogues, qui eux même viendront grossir les données d'entraînement. Ce cercle continuera jusqu'à obtenir des modèles suffisamment fiables pour ne plus avoir à intervenir manuellement dans la chaîne de traitement.

le « manifest IIIF » du catalogue. Cela correspond à un document json qui répertorie à la fois des métadonnées précises d'un document - ici catalogue - et les images qui le composent. Ainsi, passer par IIIF permet d'obtenir des images de bonne qualité en entrée et de récupérer un certain nombre d'informations concernant le catalogue. Pour un exemple de manifest IIIF, celui du catalogue du Salon Rose Croix 1893 : <https://gallica.bnf.fr/iiif/ark:/12148/bpt6k54703109/manifest.json>.

2. En vue de cela, un morceau de code est ajouté dans la chaîne de traitement afin de réaliser l'HTR automatiquement par lignes de commandes lorsque cela sera possible : https://github.com/Juliettejns/extractionCatalogs/tree/main/fonctions/automatisation_kraken.

Au fur et à mesure de la production de catalogues, plusieurs modèles de transcriptions ont été produits. Le premier, **Fourme**, est issu d'un *dataset* contenant à la fois les premières données préparées, au nombre de 274 pages, ainsi que 100 pages de catalogues produits par la chaîne. Ces nouvelles pages proviennent de deux catalogues, un *Catalogue des artistes refusés* de 1863 et un *Catalogue du Salon Rose Croix* de 1893. Il s'agit donc dans les deux cas de nouvelles pages issues de catalogues d'exposition de la deuxième moitié du XIX^{ème} siècle. Ainsi le jeu d'entraînement est composé de 375 pages dont la majorité provient des catalogues d'exposition - 220 pages. L'entraînement du modèle n'a pas été réalisé jusqu'au bout, puisqu'il s'agissait essentiellement de déterminer si le long travail de correction dans le but de créer de la donnée réutilisable portait ses fruits. Cette décision de ne pas aller au bout du travail est dû au temps de production d'un modèle : un entraînement peut facilement durer plusieurs jours³ et il était nécessaire d'obtenir un premier aperçu de l'efficacité d'une chaîne de traitement cyclique rapidement. Ainsi, **Fourme** correspond au 13^{ème} modèle produit lors de l'entraînement⁴, ayant une *accuracy* de 96,46%.

Ce résultat s'avère assez bas, en comparaison de celui de **Chaource**, qui est de 97,19% sur le même jeu de données *test*. Cependant, lorsque les résultats de **Fourme** et **Chaource** sont comparés et mis côte à côte sur les données à produire, il est possible de remarquer une évolution nettement positive : les erreurs les plus communes, qu'il a fallu le plus corriger avec **Chaource**, ne sont presque plus présentes avec **Fourme**⁵. Ainsi, l'hypothèse d'un problème au niveau du *test* fixé a été émise. Ici, on a travaillé avec le *test* composé de 30 pages, présentant dans des proportions similaires des pages d'annuaires, de catalogues d'exposition et de ventes de manuscrits. Cela faisait sens lorsque le jeu d'entraînement avait une structure similaire. Or, maintenant, le *dataset* est composé à presque 60% de catalogues d'exposition. On peut donc supposer que le jeu de données a trop varié, et ne correspond plus forcément au *test*, ce qui a eu un impact sur les capacités des modèles produits à traiter les données maintenant en minorité. Une autre série de tests a été lancée en utilisant le test fixé ne comportant que des pages de catalogues d'exposition. Le résultat reste encore une fois très bas en comparaison de ce qui était attendu et de la qualité des résultats obtenus. Ainsi, on peut supposer que ces deux tests, réalisés lors des tout premiers jeux de données, ne correspondent plus du tout aux derniers modèles produits. Afin de vérifier cette hypothèse, et par là même leur qualité, il serait nécessaire de réaliser une nouvelle batterie de tests qui utiliserait un test fixe issus des pages les plus présentes dans le jeu de données actuel.

Un dernier modèle, **Gruyère**, a été entraîné à partir d'un jeu de données de 565 pages, soit les 375 pages du *dataset* de **Fourme** associées à des pages de catalogues d'exposition des Beaux-Arts de Rouen du XIX^{ème} siècle et l'intégralité du *Catalogue des Indépendants* de 1935. Cela permet d'obtenir un modèle ayant un taux d'*accuracy* de 91,98% à partir d'un entraînement se basant sur le test fixé de 30 pages. Encore une fois, les résultats sont particulièrement faibles en comparaison de la qualité de données fournies. De même que pour **Fourme**, lorsque le modèle est appliqué sur les données *test* ont obtient de bien meilleurs résultats à l'œil nu mais le score ne suit pas. Ainsi, ce dernier modèle semble être le meilleur obtenu jusqu'à présent et ses résultats chiffrés ne

3. Dans ce cas-ci, l'entraînement a été interrompu au bout de 3 jours.

4. Il a été sélectionné en tant que modèle ayant le meilleur taux d'*accuracy* au moment où les résultats ont commencé à stagner.

5. C'est par exemple le cas des majuscules, qui étaient très peu reconnues avec **Chaource**, mais aussi de certains nombres, comme 3 et 5 qui étaient bien plus confondus et des caractères en gras ou italique.

correspondent pas à ce qui est observé. C'est donc **Gruyère** qui est utilisé dans la chaîne de traitement.

7.2 Utilisations des données issues de la chaîne de traitement

7.2.1 Les données produites : catalogues encodés et exemple d'utilisation

Dans le cadre de mon stage, j'ai été chargée d'encoder un certain nombre de catalogues. Cela a permis d'alimenter la base de données mais également de contrôler et corriger la qualité des résultats de la chaîne de traitement créée. Grâce à Caroline Corbières puis à Ljudmila Petkovic, une grande quantité de catalogues a déjà été traité et inséré dans Basart. Ainsi, la plupart des documents qu'il reste à encoder sont un peu plus compliqués. L'application de la chaîne de traitement sur ces documents a donc nécessité de nombreux ajustements en fonction de la forme de chaque catalogue. J'ai alors travaillé sur des catalogues uniques issus de séries déjà traitées, à l'instar des salons des Indépendants, des artistes refusés, d'Automne ou encore Rose Croix. L'encodage de catalogues de l'exposition annuelle des Beaux-Arts de Rouen a également été entamé, dans le but d'observer la qualité de production d'une série complète de catalogues par la chaîne de traitement.

Les données de ces documents vont être versées dans Basart et seront donc interrogables par ce biais de différentes manières⁶. Un exemple d'utilisation de ces données a été réalisé grâce à un outil développé dans le cadre d'e-ditiones⁷. Celui-ci permet de segmenter, lemmatiser, normaliser un texte contenu dans un fichier XML-TEI et de récupérer ce résultat sous la forme d'un fichier csv. La première opération consiste à diviser le texte en *tokens* qui peuvent être définis comme des groupes de caractères ayant un sens ensemble, à l'instar d'un mot, d'une ponctuation, etc⁸. La lemmatisation correspond à obtenir le lemme de ce *token*, c'est à dire la forme canonique de l'élément. Par exemple, le mot « peint » est un token. Il s'agit du lemme « peindre » conjugué à la 3ème personne du singulier du présent de l'indicatif, ce qui correspond à la catégorie grammaticale du mot. Celle-ci se divise dans notre programme en deux éléments, le *pos*, *part of speech*, qui signale s'il s'agit là d'un verbe, d'un adjectif, d'un nom propre ou commun et l'étiquette morpho-syntaxique, *msd* qui donne des informations plus poussées sur le mot (ici, le mode, le temps, la personne, etc.). La normalisation consiste à obtenir, pour les textes en ancien français, une version moderne du *token*. Enfin, un dernier stade consiste à appliquer un NER, *Named entities recognition*, ou reconnaissance d'entités nommées, qui, comme son nom l'indique, permet de repérer les éléments correspondant à des noms, prénoms, lieux, dates dans un texte.

6. Un exemple, reposant sur la question de la place des femmes dans les salons des Indépendants, a été présenté dans le mémoire de Caroline Corbières.

7. Ce projet est disponible dans un dépôt github (<https://github.com/e-ditiones/Annotator>) J'ai participé à son développement et l'ai ensuite utilisé sur les données que je venais de produire. L'article déposé dans ce contexte est disponible en annexe (C.1).

8. HALFELD, « Compilers », Diaporama du cours de lexicométrique donné à l'université de Tours le 13 juillet 2008, <https://www.univ-orleans.fr/lifo/Members/Mirian.Halfeld/Cours/TLComp/13-0708-LexA.pdf>

Cette chaîne de traitement a donc partiellement été utilisée et appliquée sur les trois catalogues produits et encodés en XML-TEI à ce moment là : les *catalogues des œuvres exposées* du Salon des artistes refusés de 1863, du Salon Rose Croix de 1893 et du Salon des Indépendants de 1935. Actuellement, Béatrice Joyeux-Prunel s'intéresse particulièrement aux thèmes des titres d'œuvres. Nous nous sommes donc concentrées ici sur ceux-ci. Une fois le fichier csv de chaque catalogue contenant les *tokens* de chaque titres d'œuvres, leur lemme, *pos*, *msd* et étiquettes *NER* associés obtenus, je les ai exploités sous la forme de nuages de mots (les images 7.2, 7.3 et 7.4). Plus le mot est grand, plus sa fréquence au sein des titres des œuvres du catalogue est élevée.

7.2.2 Analyse des résultats obtenus : quantifier l'histoire de l'art



FIGURE 7.2 – Mots fréquents, *Catalogue [...] Refusés*, 1863

FIGURE 7.3 – Mots fréquents, Catalogue [...] Rose Croix, 1893



FIGURE 7.4 – Mots fréquents, *Catalogue [...] Indépendants*, 1940

Les mots qui composent un titre donnent des indications sur les sujets abordés dans les peintures présentées. Cela permet d'obtenir des informations sur les genres et types d'œuvres présents dans chaque exposition. Par exemple, ici, il est possible de voir la grande différence entre le Salon Rose Croix et les deux autres expositions. Il s'agit de la première exposition réalisée par l' « Ordre de la Rose-Croix catholique et esthétique du Temple et du Graal », mouvement ésotérique formé par Joséphin Péladan en réaction à l'essor du matérialisme, qui prône un renouveau de la spiritualité. Ces idées sont très bien visibles dans les différents mots les plus représentés, avec notamment le développement de tout un vocabulaire concernant l'aspect occulte, tel que « vision », « mystique », « muse ». À cela s'associent des éléments issus d'autres thèmes principaux du mouvement, tel que le catholicisme (« saint », « christ ») ou encore des personnages de haute importance pour le groupe, comme Balzac et Wagner⁹. Les principaux thèmes abordés sont également visibles pour les deux autres catalogues : pour le Salon des Refusés, on remarque l'importante

9. Alain GALOIN, *Le Salon de la Rose-Croix*, fr, URL : <https://histoire-image.org/fr/etudes/salon-rose-croix> (visité le 30/08/2021).

proportion de « portrait » et d’éléments associés, tels que « mademoiselle », « madame », « femme »... Dans le cas du salon des Indépendants, il semble être plus porté sur le thème de la nature, de par la quantité de mots faisant référence à la « nature », tels que « paysage », « fleur », etc. Dans les deux cas, l’autre type de thème reste visible dans le nuage de mots cependant il y a cependant une inversion des sujets traités. On peut également remarquer les types d’œuvres les plus importants pour chaque catalogue : « étude » pour le Salon Rose Croix, « portrait » pour le salon des Refusés et « peinture » pour le Salon des Indépendants, qui semblent traduire une évolution dans la façon de nommer les tableaux.

Face au peu de catalogues disponibles et à la différence entre les expositions étudiées, il est assez difficile de se forger une véritable opinion sur l’évolution des thèmes et les façons de titrer les peintures au cours du XIX^{ème} et du XX^{ème} sans tirer de conclusions hâtives. Cependant, cela permet de mesurer les capacités qu’offre le fichier XML-TEI en tant que format de stockage et les possibilités d’étude des données traitées. Il serait par exemple intéressant d’étudier l’évolution de ces termes sur un plus gros corpus, afin d’obtenir les lemmes les plus fréquents et leur développement sur plusieurs décennies. De même, les étiquettes NER, non employées par manque de temps, peuvent contenir un surplus d’informations quant à l’étude des personnalités¹⁰ ou lieux les plus représentés par exemple. Ainsi, malgré un jeu de données réduit qui empêche de tirer de véritables conclusions, ce type de visualisations permet de représenter de façon quantitative l’histoire de l’art.

7.3 Perspectives : limites et améliorations

7.3.1 Le problème de la prise en main

Ainsi, la chaîne de traitement réalisée est bien fonctionnelle, puisqu’elle permet d’obtenir un encodage assez bon ne nécessitant que très peu de corrections, dont la plupart sont corrigables en ajoutant des éléments au code python. Néanmoins, celle-ci ne permet pas une prise en main dite *user-friendly*, c’est-à-dire une utilisation facile pour ses usagers. En effet, la chaîne de traitement utilisée actuellement par Artl@s pour l’encodage des catalogues est centrée sur l’accessibilité de celle-ci, à un certain degré, pour les néophytes en informatique, ce qui n’est absolument pas le cas de celle construite ici.

De part sa construction autour d’un prototype python d’extraction de données, la chaîne n’est en effet que très peu accessible et utilisable pour quelqu’un n’étant pas initié à ce domaine. Si une maîtrise de python n’est pas forcément nécessaire pour utiliser le programme, cela peut s’avérer utile s’il s’agit d’ajouter une fonctionnalité permettant de déterminer le type du catalogue traité. Il est cependant primordial de comprendre et savoir réaliser des expressions régulières pour traiter des catalogues par le biais de cette chaîne de traitement. Comme expliqué précédemment¹¹, il est obligatoire avant chaque utilisation de vérifier que les expressions régulières employées correspondent au catalogue à traiter. De plus, malgré les précautions prises pour rendre cette étape facile et compréhensible, toutes les structures d’entrées de catalogues n’ont pas été transposées en regex. Il est donc possible de devoir créer soi même de nouvelles expressions régulières correspondant aux

10. Le corpus étant particulièrement réduit, cela avait peu d’intérêt ici. On pourrait cependant réaliser ce travail sur tout le corpus d’une série (tous les catalogues des Indépendants par exemple) ou sur tous les catalogues contenus dans Basart.

11. Voir 6.2.2

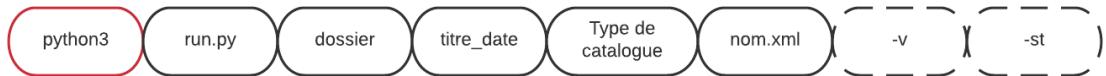


FIGURE 7.5 – Détail de la commande pour lancer le prototype

catalogues. L’unique solution est de créer le plus d’expressions régulières possibles en se référant aux différents catalogues disponibles dans Artl@s et par la suite de proposer une façon de sélectionner le type de séparateurs similaires aux choix du type d’entrée¹².

Afin de lancer les commandes plus facilement, le programme a été restructuré sous forme d’un CLI. Grâce à la librairie python click¹³, il a été possible de rendre le lancement du programme d’extraction de données plus facile et accessible. Ainsi, il est possible de réaliser cela grâce à une simple et unique commande lancée dans le terminal, qui répertorie les différents éléments nécessaires à la création du fichier XML-TEI. Cela permet aussi de décider d’activer, ou non, certaines options dans le traitement des fichiers, par exemple la vérification de la structure des fichiers alto. Il s’agit également de décider du format des documents en entrée : des fichiers alto issus d’eScriptorium et contenant la transcription ou bien des images numérisées de pages d’un catalogue qui seront alors directement transcrites dans le programme par le biais de Kraken. Ainsi, en plus de rendre le lancement du programme plus facile, la structure en CLI permet de préparer la chaîne à son automatisation complète, en proposant d’office une transcription automatique au sein du prototype. Une autre solution aurait été de transposer le code du prototype dans des *jupyter-notebook*, outil OpenSource permettant d’écrire du code au sein de fichiers partageables ouvrables dans un navigateur. Ce logiciel a l’avantage d’être intuitif pour les débutants - il est notamment particulièrement utilisé pour des cours - et de permettre une collaboration hors de Github. Cependant le code, construit ici via l’IDE pycharm, fonctionne sous la forme de modules¹⁴ et une migration sous la forme de fichiers uniques aurait été assez compliquée et longue. Ainsi, la meilleure solution actuelle était de passer par un CLI.

Un autre problème empêchant une prise en main aisée de la chaîne de traitement, il est nécessaire de maîtriser XML et XML-TEI pour le moment. En effet, une grosse partie de corrections et de vérifications des fichiers ALTO et XML-TEI produits est obligatoire afin d’obtenir une sortie correcte au programme. Pour ce qui est de la correction des entrées encodées, il est cependant possible de transformer les données en XML en csv grâce à une feuille de transformation XSLT. Cela permet de corriger plus facilement les informations. L’ajout des métadonnées est également un problème, puisqu’il est nécessaire de réaliser cela directement en XML. Une première solution a été de réaliser un CSS, *Cascading Style Sheet*. Traduisible par feuille de style en cascade, il s’agit d’un langage permettant de présenter des documents HTML ou XML en lui appliquant des règles de mise en forme (titrages, couleurs, polices, bordures, etc.). Ici, on associe une feuille de style CSS au document XML-TEI final afin d’obtenir un document composé de cases à remplir plutôt que de simples balises. Cela permet également de corriger les entrées encodées directement dans le XML. Une autre solution serait de compléter les métadonnées dans un document csv puis de le transformer en XML via une feuille XSLT. C’est ce qui a été

12. Nulle, Simple, Double ou Triple : Voir 6.2.1

13. La documentation est disponible ici : <https://click.palletsprojects.com/en/8.0.x/>

14. Voir 6.1.1.

privilégié dans le cadre de la chaîne de traitement actuelle d'Artl@s¹⁵. L'idée d'utiliser des feuilles CSS permet de rester en XML-TEI lors des corrections, ce qui est un point fort non négligeable. Cependant, cela entraîne également le besoin d'un logiciel pour le XML suffisamment fonctionnel, soit Oxygen, qui est payant.

Tout cela confirme une chaîne de traitement peu maniable et nécessitant des compétences en informatique pour l'utiliser. Plus qu'une véritable chaîne fonctionnelle, il s'agit là d'un prototype permettant de déterminer la faisabilité d'un *workflow* le plus *Open Source* possible ainsi que les avantages - tout comme les inconvénients - que cela procurerait.

7.3.2 Une chaîne ouverte et libre ?

Peut-on donc considérer le prototype de chaîne de traitement obtenu comme ouvert et libre ? Par ces adjectifs on entend une chaîne de traitement qui suit les critères de la Science Ouverte et lui garantie une complète reproductibilité des processus.

La transcription est réalisée au sein d'un outil ouvert, libre et dont le code est complètement accessible en ligne, permettant également la collaboration entre chercheurs quant à son élaboration. Ainsi, contrairement à Transkribus, outil payant et peu transparent sur son fonctionnement, l'utilisation de Kraken permet de connaître exactement la manière dont les données ont été traitées. De plus, celui-ci offre la possibilité de récupérer les modèles créés et de les rendre accessible à tous sur github. L'obligation de réaliser une chaîne de traitement semi-automatique, dû aux mauvais résultats d'HTR, entraîne cependant l'utilisation d'eScriptorium, interface graphique de Kraken. Si celle-ci est particulièrement pratique, permettant de visualiser et corriger manuellement la segmentation comme la transcription des données, elle nécessite d'accéder à un serveur - dans notre cas celui de l'INRIA. Il est cependant possible d'installer une version non-web directement sur son propre ordinateur, afin de ne pas en être dépendant. La migration réalisée de Transkribus vers Kraken permet également de rendre le processus de création des modèles plus transparents. En effet, ceux-ci sont maintenant récupérables et partageables, par le biais de dépôts github gratuits et ouverts. Y sont associés l'intégralité des jeux de données d'entraînement ainsi que leurs méthodes de production et d'utilisation.

Le prototype python d'extraction de données a quant à lui été réalisé dans un langage *Open Source*. Son code est intégralement disponible en ligne, de façon libre et gratuite. Grâce à son hébergement sur la plateforme Github, la collaboration entre chercheurs en vue de son amélioration est aussi possible. Ce programme a été développé en utilisant le logiciel Pycharm, une IDE¹⁶. Si il ne s'agit pas d'un logiciel *Open Source*, il est cependant possible de trouver d'autres environnements de développement intégré ouvert et libre. De plus, l'outil en question n'est pas obligatoire pour faire fonctionner la chaîne de traitement, celle-ci étant lancée directement dans le terminal.

L'unique logiciel réellement payant de notre chaîne de traitement est donc Oxygen. Il est nécessaire, pour utiliser cet éditeur de XML, d'acheter une licence annuelle. Utiliser Oxygen a donc un certain coût et n'est pas à disposition de tous, malgré la possibilité de

15. La feuille de transformation XSLT a été réalisée par Auriane Quoix dans le cadre de son stage à Artl@s en 2020. Elle est disponible ici : https://github.com/carolinecorbieres/ArtlasCatalogues/tree/master/0_Headers

16. Une IDE est un environnement de développement intégré. Il s'agit d'un ensemble d'outils permettant de coder plus facilement. Dans sa version gratuite, l'IDE permet de réaliser le programme et de le tester en même temps grâce à un terminal intégré, ce qui permet également d'ajouter directement les modifications réalisées dans Github. Dans sa version payante, elle offre la possibilité d'utiliser un unique logiciel pour plusieurs langages. (<https://www.jetbrains.com/fr-fr/pycharm/>)

télécharger une version d'essai de 30 jours. Cet outil a cependant fait ses preuves dans le monde de la recherche en humanités numériques et est particulièrement utilisé. Comme présenté précédemment, il est cependant tout à fait possible de passer par des csv pour corriger les données XML-TEI. La correction des fichiers ALTO, également réalisée dans Oxygen, peut s'avérer plus compliquée. Dans ce cas là, il peut être intéressant de réfléchir à des solutions d'éditeurs XML gratuits et libres, à l'instar de Xerlin¹⁷.

Ainsi, notre chaîne de traitement se base sur des logiciels et outils gratuits, libres et ouverts mais est également elle même une chaîne *OpenSource*. Elle est entièrement reproductible, transparente et permet la collaboration entre chercheurs, par le biais de l'accès libre gratuit et complet des données et processus. Elle s'inscrit alors complètement dans la démarche de la Science Ouverte et suit la volonté d'Artl@s d'obtenir une chaîne ouverte, à l'instar de sa base de données « libre d'accès, disponible à tous » .

17. <http://www.xerlin.org/>

Conclusion

Ce mémoire s'attache à décrire la mise en place d'un prototype de chaîne de traitement *OpenSource* de catalogues d'exposition. Face à un *workflow* fonctionnel et déjà utilisé au sein d'Artl@s, il développe une réflexion autour de l'utilité de la Science Ouverte en prenant appui sur l'exemple de la production de données pour Basart. Dans un premier temps, il présente une étude des enjeux et problématiques que posent le traitement des catalogues. Puis il s'intéresse à l'extraction d'informations issues d'une image avant de réfléchir à la structuration et l'annotation de ces données obtenues.

Une première partie permet donc de présenter les sources sur lesquelles est basé mon travail. Elle décrit tout d'abord l'histoire et l'utilité des catalogues papier et met en exergue l'intérêt d'un traitement numérique des documents semi-structurés, véritables réservoirs d'informations factuelles. On s'intéresse par la suite à la gestion informatique de ces documents. Celle-ci s'avère compliquée, face à la structure condensée des données qui, si elle est compréhensible aux yeux d'un humain, ne l'est pas pour une machine. Ce problème occupe plusieurs équipes de recherche, dont Artl@s, ce qui permet de présenter le cadre du projet tout en développant des solutions pour contourner cette situation.

Dans un second temps, je me suis intéressée à la question de la transcription automatique à partir d'images. Il s'agissait, premièrement, de trouver une alternative *OpenSource* à cette étape, réalisée par le biais du logiciel Transkribus dans la chaîne de traitement d'Artl@s. Pour ce faire, une réflexion a été menée autour des différents logiciels d'HTR ouverts et des formats de données existants. Par la suite, tout un travail a été réalisé sur la réalisation de jeux de données et l'entraînement de modèles de segmentation et de reconnaissance de caractères, ce qui a permis également de présenter les moyens de partage et stockage de ces informations, dans le cadre de la Science Ouverte. Enfin, un dernier point a décrit les autres informations récupérables présentes dans les images et leur utilité dans l'exploitation des catalogues.

Dans une dernière partie, je me suis intéressée à l'annotation de ces données récupérées à partir des images de catalogues. J'ai ainsi développé une réflexion autour de l'élaboration d'un prototype de programme python permettant de récupérer les informations textuelles des documents et de les structurer sous la forme d'un document XML-TEI interrogeable. Cela a permis d'obtenir une chaîne de traitement entièrement ouverte et libre capable de produire des données structurées à partir de catalogues d'exposition numérisés.

Ainsi, l'imbrication des différentes briques produites au cours de ce travail a permis d'obtenir une chaîne de traitement fonctionnelle et ouverte capable de passer d'une image numérisée à un fichier XML-TEI organisant des données structurées condensées et précises. Celle-ci est encore à un stade semi-automatique, suite à un niveau d'HTR qui ne permet pas sa complète automatisation. Cependant, grâce à l'entraînement de nouveaux modèles de segmentation et de transcription, on s'approche peu à peu de cet idéal, qui

permettra de réaliser cette chaîne sans correction manuelle, par le biais d'une simple commande. Elle reste tout de même un simple prototype, ne permettant pas de gérer tous les types de catalogues. Un gros travail de développement et d'accessibilité sur le programme python d'extraction et de structuration de données est donc nécessaire à sa complète utilisation. Il pourrait par exemple être possible, en restructurant un peu ce programme, de réaliser une application en ligne. Des images représentant les différents types de catalogues gérés par le prototype seraient présentées à l'utilisateur, qui en sélectionnerait un en même temps qu'il rentre le document qu'il souhaite structurer. Cela permettrait à la fois d'améliorer la prise en main du programme mais également de mettre à disposition du grand public un outil de transcription et de structuration automatique de documents semi-structurés. Cette idée s'intègre pleinement dans le mouvement de la Science Ouverte, permettant à toute la communauté scientifique de profiter du résultat de ce travail.

Ce mémoire a donc démontré que passer à la Science Ouverte est possible dans un projet de recherche. Cependant, cela nécessite de transformer les méthodes de travail afin d'appliquer ces principes et de permettre la réutilisation complète des données et protocoles. Ces idées doivent donc être présentes et prises en compte à chaque étape d'un projet, afin d'obtenir un résultat entièrement libre, ouvert et gratuit. Cela s'accompagne néanmoins d'inconvénients, à l'instar d'une prise en main plus compliquée des outils et de l'élaboration d'un simple prototype, qui nécessite un développement important par la suite, à l'inverse de la chaîne de traitement actuelle d'Artl@s. Pourtant, appliquer les principes de l'*Open Science* ouvre la porte à de nombreux avantages, qui ont été détaillés au cours de ce mémoire. L'utilisation d'outils ouverts, libres et gratuits permet de ne pas être dépendants de logiciels qui peuvent avoir un coût, à l'instar de Transkribus, ou dont le développement peut être interrompu. De plus, mener une réflexion sur le partage des données - leur structure, leur standardisation, leur format, leur réutilisation, leur document, leur reproduction - favorise leur interopérabilité entre les projets, comme c'est le cas avec SegmOnto, et le travail de groupe au sein de la communauté scientifique. La Science Ouverte développe donc à la fois l'autonomie et la collaboration entre chercheurs ainsi que la pérennisation des données de la recherche.

Table des matières

Résumé	iii
Remerciements	v
Bibliographie	vii
Liste des Acronymes	xiii
Introduction	1
I La numérisation des catalogues : problèmes et enjeux	5
1 Le catalogue : une source primaire pour les historiens	7
1.1 Définition globale	7
1.1.1 Histoire du catalogue papier	7
1.1.2 Étymologie	8
1.1.3 De l'utilisation des catalogues	9
1.2 Le catalogue d'exposition	10
1.2.1 Un objet au centre d'un événement : l'exposition	10
1.2.2 Le catalogue d'exposition : une structure normalisée dans le temps et l'espace	13
1.2.3 Une source primaire pour la recherche en histoire de l'art	14
1.3 Le catalogue de ventes de manuscrits	15
1.3.1 Au coeur d'un marché du manuscrit encore peu étudié	15
1.3.2 Le catalogue de vente de manuscrit : inventaire et commerce	16
1.3.3 Une source primaire pour le chercheur	17
2 Humanités numériques et catalogues	19
2.1 Un enjeu ancien au sein des humanités numériques	19
2.1.1 Enjeux et problèmes du traitement numérique des documents semi-structurés	19
2.1.2 Bref historique de la gestion numérique des catalogues	21
2.2 Un travail ancré dans un projet de plusieurs années...	23
2.2.1 ...Au sein d'une organisation pluri-institutionnelle...	23
2.2.2 ...Menant à la création d'une première chaîne de traitement automatique de leurs catalogues	25

II Extraction de l'information issue de l'image	29
3 Pourquoi, comment et quoi OCRiser ?	31
3.1 Qu'est-ce que la transcription automatique ?	31
3.1.1 Vocabulaire et étapes d'une transcription automatique	31
3.1.2 L'entraînement d'un modèle	33
3.2 Les données utilisées	34
3.2.1 Quelles données de travail ?	34
3.2.2 Quel format pour ces données ?	36
3.3 Les logiciels utilisés	38
3.3.1 Le choix d'un logiciel OpenSource	38
3.3.2 De Transkribus à eScriptorium : Organisation de la migration des données	42
4 L'entraînement d'un modèle commun de reconnaissance de texte pour les catalogues	45
4.1 Réalisation de modèles de segmentation.	45
4.1.1 L'initiative SegmOnto : un vocabulaire commun pour le nommage des zones	45
4.1.2 Préparation des données : application et adaptation de ce vocabulaire aux catalogues	48
4.1.3 Une preuve de concept pour le nommage des zones	51
4.2 Réalisation de modèles de reconnaissance de caractères	52
4.2.1 La création de jeux de données représentatifs du corpus de travail .	52
4.2.2 Comment évaluer ces modèles : outils et méthodes	54
4.3 Réflexions autour des modèles et jeux de données produits	56
4.3.1 Utiliser les modèles entraînés : mise en application	56
4.3.2 Vers un accès libre et ouvert de ces modèles et jeux de données . . .	58
5 Aller plus loin dans la récupération d'informations	61
5.1 Les illustrations	61
5.1.1 Panorama des images dans les catalogues	61
5.1.2 Comment récupérer les images ?	62
5.2 L'emphase typographique	64
5.2.1 L'emphase typographique dans les catalogues	64
5.2.2 État de l'art : récupérer l'information typographique en HTR . . .	65
5.2.3 La création d'un modèle de reconnaissance de l'emphase typographique	66
III Annotation de l'information récupérée	69
6 De l'ALTO à la TEI : encodage automatique de données sous la forme d'une application	71
6.1 La construction d'un prototype	71
6.1.1 Le choix d'une application python	71
6.1.2 La structuration du programme	73
6.2 L'extraction de données depuis les fichiers alto	74
6.2.1 Recensement des différentes typologies de catalogues	74

6.2.2	Expressions régulières et structure des entrées de catalogues	76
6.3	En sortie : résultats et perfectionnement du prototype	78
6.3.1	Présentation du fichier TEI obtenu	78
6.3.2	Améliorations et limites des résultats	80
7	De l'image à la TEI : Un prototype de chaîne de traitement complète	83
7.1	Une chaîne de traitement cyclique	83
7.1.1	Le chemin de la production d'un catalogue encodé	83
7.1.2	De nouvelles données pour de nouveaux modèles d'HTR	84
7.2	Utilisations des données issues de la chaîne de traitement	86
7.2.1	Les données produites : catalogues encodés et exemple d'utilisation	86
7.2.2	Analyse des résultats obtenus : quantifier l'histoire de l'art	87
7.3	Perspectives : limites et améliorations	88
7.3.1	Le problème de la prise en main	88
7.3.2	Une chaîne ouverte et libre ?	90
Conclusion		93
Table des figures		97
Liste des tableaux		101
Annexes		101
A Données		103
A.1	Sources : les catalogues	103
A.2	Jeux de données	103
A.3	Descriptions de formats	106
B HTR		109
B.1	Segmentation	109
B.1.1	Des exemples d'entrées	109
B.1.2	Des exemples de segmentation de pages entières	110
B.1.3	Résultats de l'application des modèles de segmentation sur des données tests	112
B.2	Reconnaissance de caractères	115
C Rapports et Articles		117
C.1	Articles	117
C.2	Rapports	143

Table des figures

1.1	Un exemple typique de page de catalogue d'exposition (<i>Catalogue [...] exposés</i> , Salon des Indépendants, 1913, p. 79)	13
1.2	Un exemple typique de page de catalogue de ventes de manuscrits (<i>Catalogue [...] manuscrits</i> , Charavay, 1846, p. 4)	16
2.1	<i>Catalogue [...] Nancy</i> , 1843, p.3	20
2.2	<i>Catalogue de feu M. de Bruyère</i> , Chalabre, 1833, p.102	20
2.3	<i>Bibliographie de France</i> , 1822, p.6	20
2.4	Chaîne de traitement actuelle d'Artl@s	25
3.1	Fonctionnement d'un OCR	32
3.2	Chaîne de production de la vérité terrain	35
3.3	Capture d'écran de l'interface eScriptorium	41
3.4	Chaîne de production complète des données de la vérité terrain	42
3.5	<i>Catalogue [...] Courbet</i> , 1882, p.2	43
4.1	<i>Catalogue Bovet</i> , 1887, p.79	46
4.2	<i>Catalogue Bovet</i> , 1887, p.79, sortie eScriptorium	50
4.3	<i>Lettre du Sieur de Balzac</i> , 1624 p.21	50
4.4	Répartition des zones dans le dataset Chaource	52
4.5	Les types d'erreurs	55
4.6	Exemple du calcul d'une distance de Levenshtein	55
5.1	<i>Catalogue des œuvres exposées</i> , Palais du Luxembourg, 1915, p.3	62
5.2	<i>Catalogue de manuscrits</i> , Bovet, 1887, p.18	62
5.3	<i>Catalogue des œuvres exposées</i> , Beaux-Arts de Nancy, 1892, p.004	65
5.4	<i>Catalogue de manuscrits</i> , Charavay, 1845, p.15	65
5.5	Exemple de transcription d'un catalogue	66
5.6	Catalogue de vente de manuscrits, Charavay, 1843, p.18	67
6.1	Fonctionnement schématique du prototype d'encodage	73
6.2	Schéma récapitulatif des types de catalogues	74
6.3	<i>Catalogue de l'exposition annuelle du musée de Rouen</i> , 1853, p.12	75
6.4	<i>Catalogue d'exposition des Beaux Arts de Nancy</i> , 1849, p.11	75
6.5	75
6.6	<i>Catalogue [...] Indépendants</i> , 1913, p.78	77
6.7	Exemple de regex	77
6.8	Exemple d'entrée encodée	79
6.9	Elements récupérés, <i>Catalogue [...] exposées</i> , Beaux Arts de Rouen, 1869, p.4	82

6.10	Elements récupérables, <i>Catalogue [...] exposées</i> , Beaux Arts de Rouen, 1869, p.4	82
7.1	Chaîne de traitement complète	83
7.2	Mots fréquents, <i>Catalogue [...] Refusés</i> , 1863	87
7.3	Mots fréquents, <i>Catalogue [...] Rose Croix</i> , 1893	87
7.4	Mots fréquents, <i>Catalogue [...] Indépendants</i> , 1940	87
7.5	Détail de la commande pour lancer le prototype	89

Liste des tableaux

3.1	Description numérique du corpus	35
3.2	Comparaison de différents OCR <i>Open Source</i>	38
4.1	Terminologie SegmOnto	47
4.2	Terminologie utilisée sur les données	49
5.1	Description numérique des pages du corpus de travail	67
5.2	Le <i>dataset BIR</i> (Issu de l'article)	67

Annexe A

Données

A.1 Sources : les catalogues

1	YOUNG Cybele with two Nymphs, portraits	Maria Cefway
2	A beggar boy	J. Rijfng
3	Basket of flowers	J. Edwards
4	Portrait of a gentleman	T. Beach
5	Portrait of a gentleman	J. Opie, R. A. Elect
6	Abraham and Isaac	W. Tate
7	Portrait of a lady and three children	Sir J. Reynolds, R. A.
8	Portrait of two horses	J. Boultnre
9	Portrait of a gentleman	W. R. Bigg
10	Portrait of a gentleman	M. H. Keymer

Type 2 : *Exhibition of the Royal Academy*, 1785, p. 1

CURPHEY (Grace), née en Ecosse. — 126, boulevard du Montparnasse, 14 ^e .
727 Les hortensias.
728 Les violettes.
729 Une étude.

Catalogue [...] indépendants, 1935, p.104

De Lotto Annibale	(ITALIA)
17 <i>Il Lavoro</i> (bronzo).	
18 <i>Il Risparmio</i> »	
(appartengono alla Cassa di Risparmio di Venezia).	

Esposizione Inter-nazionale... di Venezia, 1910

17. — Le Chasseur badois.

Signé à gauche : ..59. Gustave Courbet.
T. — H. 1.18. L. 1.75.
Appartient à M. D... M... .

Type 3 : *Catalogue [...] Courbet*, 1882, p. 39

Francisco BORES (1898—)

25. Composição sobre fundo rosa — 1945 — 73x60.
26. Natureza morta com garrafa — 1943 — 73x60.
27. Naturezà morta com doces — 1946 — 92x73.

Bienal de São Paulo, 1951, p. 54

211. **Chauvines** (Marie-Joseph-Louis d'Albret d'Ailly, duc de), habile chimiste. L. aut. sig., à M. Cochu. 27 juin 1789. 4 p. in-4. 2 50
212. **Chevalier** (Michel), Saint-Simonien, savant ingénieur. L. aut. sig. Paris, 3 juin 1840. 4 p. pl. in-8. 2 23
213. **Chevrenne** (le due de), fils du duc de Luynes, membre de l'Académie des sciences. L. aut. sig., au baron Destouches. Paris, 3 février 1817. 1 p. et demie in-4. Cachet. 2 23
214. **Chicoyneau**, premier chirurgien du roi, membre de l'Académie des sciences. 1^o Mémoire au roi, aut. sig. (à la 3^e personne), en faveur de son fils. 4 p. in-4; — 2^o supplique au roi Louis XV,

Catalogue [...] manuscrits, Laverdet, 1856, p.21

A.2 Jeux de données

Nom	Type	Provenance	Date	No. de pages
<i>Annuaire-almanach du commerce, de l'industrie...</i>	Annuaire		1898	150
<i>Exposition des oeuvres de M. Courbet à l'École des Beaux Arts</i>	Exposition	monographique	1882	24
<i>Catalogue des œuvres exposées</i>	Exposition	Salon des Indépendants	1892	5
<i>Catalogue des œuvres exposées</i>	Exposition	Salon des Indépendants	1913	7
<i>Catalogue des œuvres exposées</i>	Exposition	Salon des Indépendants	1923	5
<i>Catalogue des peintures, sculptures et miniatures</i>	Exposition	Palais du Luxembourg	1818	8
<i>Catalogue des peintures, sculptures et miniatures</i>	Exposition	Palais du Luxembourg	1867	5
<i>Catalogue des peintures, sculptures et miniatures</i>	Exposition	Palais du Luxembourg	1915	10
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exposition	Beaux-Arts de Nancy	1843	5
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exposition	Beaux-Arts de Nancy	1849	5
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exposition	Beaux-Arts de Nancy	1892	5
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	Biennale de Paris	1961	9
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	Biennale de Paris	1965	5
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	Biennale de Paris	1969	6
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	Beaux-Arts de Rouen	1853	7
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	Beaux-Arts de Rouen	1869	7
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	Beaux-Arts de Rouen	1888	7
<i>Catalogue de la Biennale de Sao Paulo</i>	Exposition	Biennale de Sao Paulo	1951	7
<i>Catalogue de la Biennale de Sao Paulo</i>	Exposition	Biennale de Sao Paulo	1972	5
<i>Catalogue de l'exposition de la société des amis des arts de Strasbourg</i>	Exposition	Société	1884	15
<i>Catalogue de la Biennale de Venise</i>	Exposition	Biennale de Venise	1895	5
<i>Catalogue de la Biennale de Venise</i>	Exposition	Biennale de Venise	1905	5
<i>Catalogue de la Biennale de Venise</i>	Exposition	Biennale de Venise	1920	5
<i>Revue des Autographes</i>	Manuscrits	Charavay	1870	18
<i>Revue des Autographes</i>	Manuscrits	Charavay	1871	32
<i>Revue des Autographes</i>	Manuscrits	Charavay	1873	5
<i>Revue des Autographes</i>	Manuscrits	Charavay	1877	16
<i>Revue des Autographes</i>	Manuscrits	Charavay	1880	16
<i>Revue des Autographes</i>	Manuscrits	Charavay	1881	16
<i>Revue des Autographes</i>	Manuscrits	Charavay	1883	16
<i>Revue des Autographes</i>	Manuscrits	Charavay	1885	14
<i>Catalogue de Ventes de manuscrits Charavay</i>	Manuscrits	Charavay	1845	16
<i>Catalogue de ventes de manuscrits Laverdet</i>	Manuscrits	Laverdet	1856	16
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscrits	Charavay	1857	12
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscrits	Charavay	1859	22
<i>Catalogue de vente de manuscrits Bovet</i>	Manuscrits	Bovet	1887	28
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscrits	Charavay	1896	27
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscrits	Charavay	1899	22
<i>Catalogue de vente de manuscrits Kra</i>	Manuscrits	Kra	1912	9
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscrits	Charavay	1919	32
<i>Catalogue de vente de manuscrits Bodin</i>	Manuscrits	Bodin	2009	18
<i>Catalogue de vente de manuscrits Bounure</i>	Manuscrits	Bounure	2009	5
<i>Catalogue de vente de manuscrits Bodin</i>	Manuscrits	Bodin	2017	12
<i>Catalogue de vente de manuscrits Aristophil</i>	Manuscrits	Aristophil	2018	21

Tableau des données disponibles

Nom	Type	Date	No. de pages	No. de Colonnes	Autres Zones
<i>Annuaire-almanach du commerce, de l'industrie...</i>	Annuaire	1898	50	2	Numbering
<i>Exposition des œuvres de M. Courbet à l'École des Beaux Arts</i>	Exposition	1882	19	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exposition	1892	5	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exposition	1913	7	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exposition	1923	5	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exposition	1818	2	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exposition	1867	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exposition	1843	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exposition	1849	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Paris</i>	Exposition	1892	5	1	Numbering
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	1961	5	1	Running Title
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	1965	4	1	Numbering
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exposition	1969	5	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	1853	5	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	1869	7	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exposition	1888	7	1	Numbering
<i>Catalogue de la Biennale de Sao Paulo</i>	Exposition	1951	6	1	Numbering
<i>Catalogue de la Biennale de Sao Paulo</i>	Exposition	1972	5	1	Numbering
<i>Catalogue de l'exposition de la société des amis des arts de Strasbourg</i>	Exposition	1884	15	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exposition	1895	5	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exposition	1905	3	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exposition	1920	5	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1870	5	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1871	6	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1873	4	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1877	4	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1880	2	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1881	2	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1883	5	1	Numbering
<i>Revue des Autographes</i>	Manuscrits	1885	2	1	Numbering
<i>Catalogue de ventes de manuscrits Charaway</i>	Manuscrits	1845	6	1	Numbering
<i>Catalogue de ventes de manuscrits Laverdet</i>	Manuscrits	1856	4	1	Numbering
<i>Catalogue de vente de manuscrits Charaway</i>	Manuscrits	1857	6	1	Numbering
<i>Catalogue de vente de manuscrits Charaway</i>	Manuscrits	1866	7	1	Numbering
<i>Catalogue de vente de manuscrits Bovet</i>	Manuscrits	1887	14	1	Stamp
<i>Catalogue de vente de manuscrits Charaway</i>	Manuscrits	1857	7	1	Numbering
<i>Catalogue de vente de manuscrits Charaway</i>	Manuscrits	1899	6	1	Numbering
<i>Catalogue de vente de manuscrits Kra</i>	Manuscrits	1912	9	2	Numbering
<i>Catalogue de vente de manuscrits Charaway</i>	Manuscrits	1919	8	1	Numbering

Tableau des données utilisées pour l'entraînement de modèles

Nom	Type	Date	Provenance	No de pages	Réutilisé en jeu d'entraînement ?
<i>Catalogue des œuvres exposées</i>	Exposition	1893	Salon Rose Croix	23	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1863	Salon des artistes refusés	67	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1935	Société des Indépendants	219	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1940	Salon d'Automne	96	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1853	Beaux Arts de Rouen	2	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1856	Beaux Arts de Rouen	46	✗
<i>Catalogue des œuvres exposées</i>	Exposition	1860	Beaux Arts de Rouen	68	✗
<i>Catalogue des œuvres exposées</i>	Exposition	1869	Beaux Arts de Rouen	3	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1872	Beaux Arts de Rouen	5	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1878	Beaux Arts de Rouen	3	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1888	Beaux Arts de Rouen	2	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1891	Beaux Arts de Rouen	5	✓
<i>Catalogue des œuvres exposées</i>	Exposition	1895	Beaux Arts de Rouen	6	✓

Tableau des données produites par la chaîne de traitement

A.3 Descriptions de formats

Ces descriptions des formats Alto et Page XML sont disponible en version document dans le dépôt de travail réalisé en collaboration avec Claire Jahan, stagiaire Artl@s¹.

1. https://github.com/Heresta/BAO_Stage_DH_ENS_2021/tree/main/documentationFormatsexistants

```

<!--
  Exemple de format ALTO2
  Documentation: https://github.com/altoxml/documentation/blob/master/v2/ALTO_changes_2_1.pdf
-->
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.loc.gov/standards/alto/ns-v2#"
      xmlns:page="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xsi:schemaLocation="http://www.loc.gov/standards/alto/ns-v2#
      http://www.loc.gov/standards/alto/alto.xsd">
  <!-- Balise racine et namespaces -->
  <Description>
    <!-- Informations sur les fichiers et le processus de réalisation du document -->
    <MeasurementUnit>pixel</MeasurementUnit>
    <DCRProcessing ID="Id0cr">
      <ocrProcessingStep>
        <processingDateTime>2020-11-01T14:47:50.282+01:00</processingDateTime>
        <processingSoftware>
          <softwareCreator>READ COOP</softwareCreator>
          <softwareName>Transkribus</softwareName>
        </processingSoftware>
      </ocrProcessingStep>
    </DCRProcessing>
  </Description>
  <Layout>
    <!-- Englobe tout les éléments sur la page -->
    <Page ID="Page1" PHYSICAL_IMG_NR="1" HEIGHT="2209" WIDTH="1074">
      <!-- Identifiant de la page, nombre d'images, taille (largeur et longueur) -->
      <TopMargin HEIGHT="0" WIDTH="1074" VPOS="0" HPOS="0"/>
      <LeftMargin HEIGHT="2209" WIDTH="0" VPOS="0" HPOS="0"/>
      <RightMargin HEIGHT="0" WIDTH="0" VPOS="0" HPOS="1074"/>
      <BottomMargin HEIGHT="0" WIDTH="1074" VPOS="2209" HPOS="0"/>
    </Page>
    <!-- Taille de l'imprimé = taille de la page -->
    <TextBlock ID="r_1_1" HEIGHT="103" WIDTH="933" VPOS="1" HPOS="43">
      <!-- Un TextBlock correspond à ce que Segm0nto appelle une région
          Chaque TextBlock a un identifiant précis, une taille (largeur et longueur) et
          des coordonnées x,y pour la situer dans la page selon un repère orthonormé. -->
      <Shape>
        <Polygon POINTS="43,1 976,1 976,104 43,104"/>
        <!-- Shape donne les coordonnées de la forme de la TextRegion -->
      </Shape>
      <TextLine ID="tl_1" BASELINE="114" HEIGHT="115" WIDTH="945" VPOS="-1" HPOS="36">
        <!-- TextLine correspond à une portion de la zone avec du texte
            Comme TextRegion elle a un identifiant, une taille (largeur et longueur),
            des coordonnées et une coordonnée de la ligne correspondant à la base du
            texte (baseline) -->
        <String ID="string_tl_1" HEIGHT="115" WIDTH="945" VPOS="-1" HPOS="36" CONTENT="<b>LIBRAIRIE ANCIENNE</b>"/>
        <!-- String correspond à la ligne de texte
            la balise a un identifiant, une taille, des coordonnées (similaires à celle de
            TextLine) et un contenu correspondant à la transcription -->
        <!-- [...] -->
      </TextLine>
    </TextBlock>
  </PrintSpace>
  </Page>
</Layout>
</alto>

```

Description du format alto

```

<!-- Description d'un document PageXML selon la documentation :
    https://github.com/PRIMA-Research-Lab/PAGE-XML/blob/master/PAGE-release/gts/pagecontent/2009-03-16/pagecontent.xsd -->
<!DOCTYPE gts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15
Metadata.xsd">
<!-- Metadata integrant les données sur le document: créateur, date de
creation, date de la dernière modification et commentaires -->
<Creator>Transkribus</Creator>
<Created>2019-11-28T15:07:37.583+01:00</Created>
<LastChange>2020-04-03T14:06:54.509+02:00</LastChange>
<Comments> Measurement unit: pixel PrimaryLanguage: English Language: FrenchStandard Producer: Transkribus </Comments>
<TranskribusMetadata docId="361091" pageId="14061559" pageNr="8" tid="26956060" status="DONE" userId="54586" imgUrl="https://files.transkribus.eu/Get? id=TXCAEAAANIBGABYJBKQLOMPEV&fileType=view" xmlUrl="https://files.transkribus.eu/Get?id=ZEOMABZQGTQ5ZXUECPBORXII" imageId="6554697"/>
</Metadata>
<!-- La balise Page contient la structure entière de la page.
    Ses attributs donnent le nom de l'image décrite et sa taille -->
<Page imageFilename="1845_05_14_CHA_type_0008.jpg" imageWidth="5866" imageHeight="9233">
<!-- La balise PrintSpace contient la structure entière de la page.
    Ses attributs donnent les coordonnées de la page -->
<PrintSpace>
    <!-- donne les coordonnées de la page -->
    <Coords points="97,-34 5963,-34 5963,9199 97,9199"/>
</PrintSpace>
<ReadingOrder>
    <!-- Ordre de lecture sous la forme d'un index associé à l'identifiant
        des différentes régions sous la forme d'un pointeur -->
<OrderedGroup id="ro_1604246200038" caption="Regions reading order">
    <RegionRef indexed="0" regionRef="r_1_1"/>
    <!-- [...] -->
</OrderedGroup>
</ReadingOrder>
<TextRegion type="paragraph" id="r_1_1" custom="readingOrder {index:0;}">
    <!-- TextRegion décrit une zone large, regroupement de plusieurs lignes.
        Elle a un type, un identifiant et un attribut custom,
        signale sa place dans l'index de lecture -->
    <Coords points="1849,726 4150,726 4150,1037 1849,1037"/>
    <!-- Coordonnées sous la forme x1,y1 de la région dans la page -->
<TextLine id="tl_1" primaryLanguage="English" custom="readingOrder {index:0;}">
    <!-- TextLine décrit une portion de la page avec du texte.
        Elle a un identifiant, une langue principale (ici anglais étrangement)
        et un ordre de lecture selon l'index établi précédemment -->
    <Coords points="1850,727 4149,727 4149,1036 1850,1036"/>
    <!-- Coordonnées du mot -->
    <Word id="w_2aabb1b2b1b1b1" language="English" custom="readingOrder {index:0;}">
        <!-- Word décrit un mot en particulier dans la ligne.
            Il a un identifiant, un langage (ici encore Anglais étrangement)
            et un ordre de lecture dans l'index -->
        <Coords points="1851,727 4149,727 4149,1036 1851,1036"/>
    <!-- Coordonnées du mot -->
<TextEquiv>
    <Unicode>GÄTM.IGNEim</Unicode>
    <!-- Mot transcrit par Transkribus -->
</TextEquiv>
<TextStyle fontFamily="Times New Roman" fontSize="27.0"/>
<!-- Informations sur le style du mot: police et taille -->
</Word>
<TextEquiv>
    <!-- texte corrige à la main de la transcription dans la
        balise Word, située dans le Textline -->
    <Unicode>CATALOGUE</Unicode>
</TextEquiv>
<TextLine>
    <TextEquiv>
        <Unicode>CATALOGUE</Unicode>
        <!-- Mot transcrit par Transkribus -->
    </TextEquiv>
    <!-- Detail sur les lignes qui séparent un élément d'un autre,
        avec identifiant, ordre de lecture et coordonnées -->
<SeparatorRegion id="r_4" custom="readingOrder {index:15;}">
    <Coords points="4604,0 5084,0 5084,12 4604,12"/>
</SeparatorRegion>
    <!-- [...] -->
</TextLine>
</TextRegion>
<!-- [...] -->
</Page>
</PcGts>

```

Description du format page

Annexe B

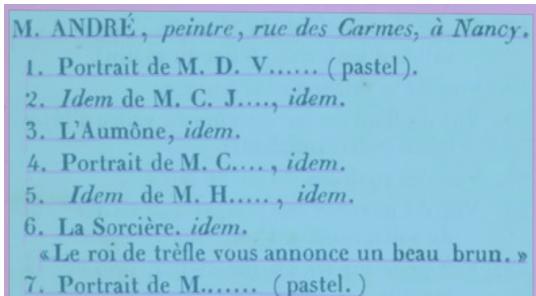
HTR

B.1 Segmentation

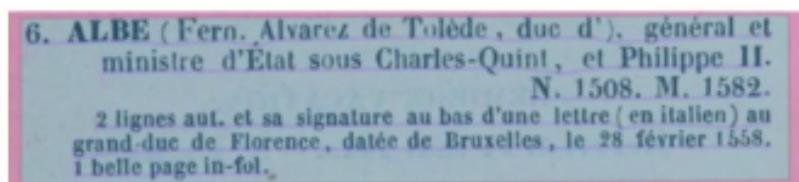
B.1.1 Des exemples d'entrées



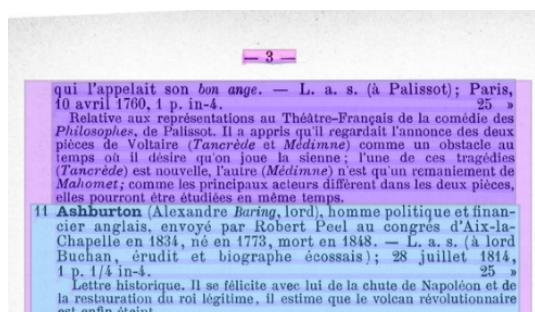
Exemple d'entrée d'annuaire



Exemple d'entrée de catalogue d'exposition



Exemple d'entrée de catalogue de vente de manuscrits



11 **Ashburton** (Alexandre Baring, lord), homme politique et financier anglais, envoyé par Robert Peel au congrès d'Aix-la-Chapelle en 1834, né en 1773, mort en 1848. — L. a. s. (à lord Buchan, érudit et biographe écossais); 28 juillet 1814, 1 p. 1/4 in-4.

Lettre historique. Il se félicite avec lui de la chute de Napoléon et de la restauration du roi légitime. Il estime que le volcan révolutionnaire est enfin éteint.

Exemple d'EntryEnd (en violet)

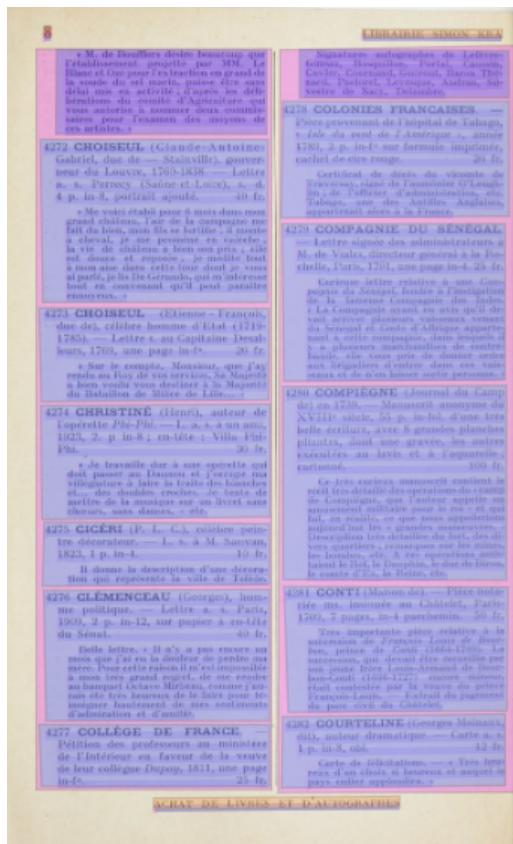
B.1.2 Des exemples de segmentation de pages entières

Zone	Quantité
<i>Decoration</i>	3
<i>Entry</i>	1500
<i>EntryEnd</i>	124
<i>Figure</i>	41
<i>Main</i>	301
<i>Numbering</i>	235
<i>Running Title</i>	65
<i>Stamp</i>	6
<i>Title</i>	9

Répartition numériques des zones du *dataset Chaource*

Couleur	Zone
Rose	<i>Main</i>
Bleu Clair	<i>Entry</i>
Bleu Foncé	<i>EntryEnd</i>
Rouge	<i>Numbering</i>
Orange	<i>Running Title</i>
Violet	<i>Figure</i>
Jaune	<i>Title</i>

Légende pour la segmentation



Catalogue[.] manuscrits, Kra,
1912, p.8

Revue des Autographes, Charavay, 1870, p.18

23. — Cheval de chasse sellé et boule-dogue en forêt, épisode de chasse à courre.

Signé à gauche: ..63. Gustave Courbet.
Salon de 1863.
Exposition particulière de 1867.
Au catalogue du Salon de 1863, ce tableau portait le titre de la *Chasse au renard*. L'auteur lui ayant fait subir quelques modifications après coup, un changement dans le titre devint nécessaire : le tableau reparut à l'exposition particulière de 1867, sous cette désignation que nous relevons sur le catalogue : « *Le Cheval du pi-queur*, épisode de chasse à courre.

T. — H. 1.12. L. 1.35.

Appartient à M. Recipon, député.

24. — La Femme à la vague.

Signé à gauche: ..68. G. Courbet.
T. — H. 0.63. L. 0.53.

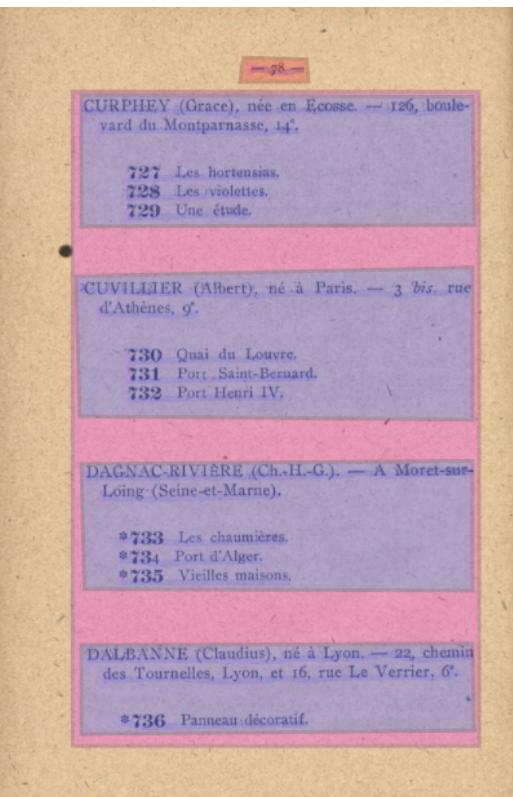
Appartient à M. Faure, de l'Opéra.

5. — La Dame aux bijoux, étude.

Signé à droite: Gustave Courbet, ..67.
Exposition particulière de 1867.
T. — H. 0.81. L. 0.64.

Appartient à M. Castagnary.

Catalogue [...] Courbet, 1884, p.48



Catalogue [...] Indépendants, 1913,
p.78

Desenho 3, 72, 69,5 x 50, papel.	1.000,00
Desenho 4, 72, 69,5 x 50, papel.	1.000,00
Desenho 5, 72, 69,5 x 50, papel.	1.000,00
14 Dias de Fevereiro, 72, 69,5 x 50, papel	

PIZA FILHO, (1949 — S. Paulo) — bp-72

Com Lápis de Côr, 71, 100 x 80	1.500,00
Da Janela, 71, 100 x 80	1.500,00
Vila dos Remédios, 72, 100 x 80	1.500,00

PRADO, Marcos da Silva (1950 — S. Paulo) bp-72

Círculo I, 72, 1,50 x 40, papel.	500,00
Arcos, 72, 1,50 x 40, papel.	500,00

QUEIRÓZ, José Carlos (S. José dos Campos) — masi

Encontro Cómico, 37 x 58.	500,00
Paisagem Cómica, 58 x 37.	500,00

REIF, Mariana (S. Paulo) — masi/bp-72

Pintura 1,72, 80 x 90.	1.600,00
Pintura 3, 72, 80 x 80.	1.600,00
Pintura 4, 72, 60 x 90.	2.000,00
Pintura 5, 72, 60 x 90	2.000,00

RETROZ, — S. Paulo) — bp-72

Monocromático 3.	2.000,00
Monocromático 4.	2.000,00

54

Catalogue [...] São Paulo, 1972,
p.54

1019

M.36 MALAKOFF (Avenue de) 43,84

1-3 Second, Paris, av. Malakoff, 1.	161 Feubert (Vve), Paris, av. Gde-Armée, 78.
5 Entrée av. Esplanade, 4.	2 Lhermeront, Paris, boul. Voltaire, 75.
12 Aldrophe (Vve), Paris, av. Malakoff, 7.	22 à 20 Crô des Omnibus, Paris,r. St-Honoré,155.
9 Legrand (Vve), Paris, av. Malakoff, 9.	24 Entrée r. Louvois, 62.
11 La Hamayde (de), gér. Bouts, Paris, r. Louvois.	26 à 30 Crémant, Paris, r. Louvois, 62.
13 Mouillé, Paris, r. Boissière, 26.	32 Berbier, Paris, r. Magdebourg, 22.
17 Pradeau, Paris, av. Trocadéro, 12.	34 Vassal, Paris, r. Lauriston, 93.
21 Stalín, Paris, av. Malakoff, 21.	36 Biron,gér.Beriaux, Paris,av. Malakoff, 40.
23-25 Berlaut, Paris, av. Malakoff, 27.	38-40 Berlaut, Paris, av. Malakoff, 40.
27 Yerles (Vve), Verhaeghe, Paris, r. Louvois, 18.	42 Berlaut (R.), Paris, r. Paix, 27.
29 Juillemin, Montreuil s/B., boul. Hôtel de Ville, 60 (Seine).	46-48 Entrée r. Lauriston, 108.
31 Montgolfier, Paris, av. Malakoff, 31.	50 Clermont, Paris, quai Debilly, 34.
35 Lainé, Paris, av. Kleber, 73.	52 à 56 Goncourt (St-Honoré), 34.
35 Maubouf, Paris, r. Victor Massé, 31.	58 Horowitz, Paris, av. Malakoff, 58.
37 Angerville, Paris, av. Malakoff, 37.	60 Rude, Paris, r. Aznale, 15.
39 Entrée r. Lauriston, 95.	66-69* Sallemeubles « St-Honoré d'Eylau », gér. Ruef, r. Aznale, 16.
41-43 Entrée r. St-Didier, 27.	s. n° Cotin, Paris, r. Royale, 9.
45-47 Entrée b. Paris, r. Malakoff, 4.	s. n° Roger (B*), Paris, r. La Bodie, 111.
49-51 Entrée b. Paris, r. Malakoff, 15.	86 à 92 Leblanc, Paris, av. Malakoff, 88.
s. n. Naud & Cie, Paris, r. Mogador prolongée, 4.	94 Pastureau, Paris, av. Malakoff, 94.
57 Cotta, Paris, av. Malakoff, 57.	96-98 Lévy, père, Paris, av. Malakoff, 98.
59 Nobel, Paris, av. Malakoff, 59.	100 Cohen, Paris, av. Malakoff, 100.
61 Houlang, Paris, av. Malakoff, 61.	102 Yerles (Vve), Paris, av. Malakoff, 102.
63 Dubout, Paris, av. Malakoff, 63.	104 Pissi-Wil (C* de), Paris, fr-St-Honoré, 31.
65 Crasmix, gér. Vincent, Paris, av. Villiers, 63.	110 Salignac-Fénoloz (Cte de), Paris, av. Malakoff, 110.
67 Ball, Paris, av. Victor Hugo, 51.	112-114 Bonnemains (de), Paris, r. Lapérouse, 2.
69 à 77 Picard (Vve), Paris, av. Malakoff, 38.	116 Castellane, Paris, av. Bosquet, 40.
79-81 Picard, Paris, av. Malakoff, 62.	124 Crozat de Lesser (B*), Paris, r. Volney, 12.
83 Bar (de), Paris, av. Malakoff, 45.	126-128 Lévy-Michel, Paris, av. Malakoff, 126.
85 Franck, Paris, av. Henri Martin, 2.	130-132 Abadie, Paris, av. Malakoff, 132.
89 à 99 S* Fougner Italiens *, Paris, boul. des Italiens, 19.	134 Plisson, Paris, av. Malakoff, 134.
105 Evans, Paris, av. Malakoff, 105.	s. n° des voitures, Paris, pl. Théâtre-Français, 1.
117 Bouasse-Lobel (M**), Parc-St-Maur, av. Nord, 33.	138 Méry-Picard, Paris, r. Pergolise, 20.
119 Grout, Paris, av. Malakoff, 119.	140 Malepsin, Paris, av. Malepsin, 140.
121 Labourdette, Paris, av. Malakoff, 121.	142 s. n° Méry-Picard, Paris, r. Pergolise, 20.
123 à 129 Kellner, Paris, av. Malakoff, 123.	144 Houdin, Paris, r. Louis-le-Grand, 9.
131-139 Grout, Paris, r. Montalivet, 12.	146 Gheet (M* de), Paris, av. Boisde Boulogne, 3.
135 Fortune, Paris, r. St-Fiacre, 12.	148-150 Duplan, Paris, r. Pyramides, 2.
137 Gatinet (de), Paris, av. Malakoff, 137.	152 Potier (Vve), Paris, boul. Madeleine, 21.
139 Poupart (C* de), Paris, boul. Hausmann, 149.	154 Poissonnier (de), Paris, r. Clémire, 28.
141 Prestrot, Paris, g. lerie Montpensier, 3-4.	156 Potier (Vve), Paris, boul. Madeleine, 21.
143 Philippon, Paris, av. Malakoff, 143.	158 Peyre, Neuilly s/S., r. Montresier, 3.
145 Baudenier (Vve), Paris, r. Moslay, 38.	160-162 Entrée av. Gde-Armée, 87.
147 Courier, Paris, r. Grenelle, 49.	166 Entrée av. Gde-Armée, 89.
149 Entrée villa Rechin, 1.	
155 Robert (Vve), Paris, av. Gde-Armée, 78.	
157 Commeau (de), Paris, av. Malakoff, 157.	
159 Neui et Cie, Paris, r. Mogador prolongée, 4.	

M.38 MALAQUAIS (Quai)

1 Firmin-Didot, Paris, r. Jacob, 56.
3 Ledreide la Clarizière, Paris, q. Malaquais, 3.
5 Pissi (M*), Versailles, r. Béthune, 14 (S-O.).

Annuaire [...] Propriétaires, 1898,
p.1019

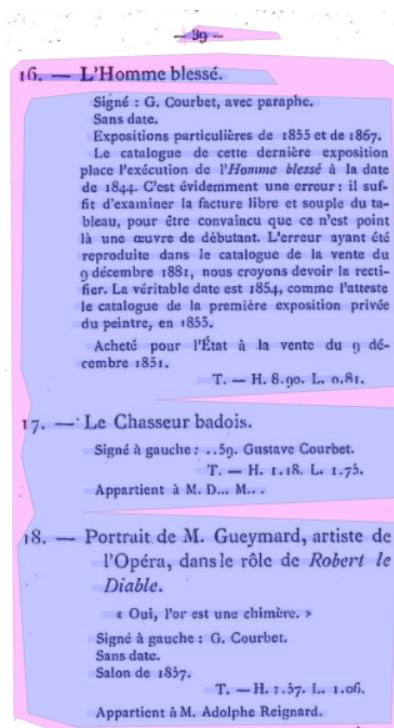
B.1.3 Résultats de l'application des modèles de segmentation sur des données tests

Modèle Abondance (sur un dataset de 30 pages)

— 21 —	650
167. Cambrenne (le général). Né à <i>Saint-Sébastien (Loire-Inférieure)</i> . L. aut. sig., à M. Franchetau, 2 Nantes. 8 septembre 1819. 1 p. in-4.	13. Dusouchet , Boulogne s/S., r. Sully, 32.
168. Carron (abbé Toussaint-Julien). Né à <i>Rennes</i> . L. aut. sig., au vicomte de Montmorency. Paris, 26 avril 1816. 1 p. in-4. 2 30	15. Worth , Paris, r. Lagrange, 10.
169. Castelman (Michel de), seigneur de <i>La Maurissière</i> , ambassadeur. Né à <i>La Maurissière</i> , M... 1692. Pièce notariée, 2 fois signée, 25 août 1586. 4 p. pl. et demi in-fol. 2 30	17. Pannier , Paris, r. Calais, 11.
200. Catinat (M. de), maréchal de France. Quitt. sig. (sur papier) à la somme de 500 livres. 18 mars 1704. 4 30	19. Hénault , Paris, r. Calais, 13.
201. Cauchy (de chevalier), secrétaire-archiviste du sénat-conservateur, littérateur. Né à <i>Rouen</i> en 1733. L. aut. sig. Paris, 29 avril 1806. 2 p. in-4.	15. Guitret , Paris, r. Calais, 15.
202. Causans (Jos.-A.-Vine, de <i>Mardon de</i>), mathématicien, gouverneur du prince de Conti, auteur de : <i>Spectacle de l'homme</i> , etc. Né à <i>Argenton</i> , L. aut. sig., à M. de Malesherbes. Paris, 22 décembre 1712. 2 p. in-4.	17. Chinchet , Paris, r. Amsterdam, 35.
203. Chalgrin architecte de l'arc de triomphe de l'Étoile. L. aut. sig. 21 octobre 1807. 2 p. in-8.	19. Milly (M^e de) , Paris, r. Calais, 19.
204. Châlon-sur-Marne reprend ses anciennes armoiries . L. aut. sig. de M. de Chamoni, maire de Châlons-sur-Marne, à M. Geoffroy, référendaire à la chancellerie, pour le prie de se charger de faire obtenir à la ville de Châlons la confirmation de ses anciennes armoiries, etc. 1 ^{re} déc. 1814. 1 p. in-4.	21. Maison du Sacré-Cœur , Paris, r. Calais
205. Champeion (Edouard), mem. de l'Académie des sciences. L. a. s., au roi. Bourges, 18 juin 1819. 3 gr. p. pl. in-fol. 5 *	23. Entrée r. Viatinille , 24.
207. Chapelle (M. d'agronome). Né à <i>Lyon</i> . L. aut. sig. 31 décembre 1821. 1 p. et demi in-4.	24. Châlon , Paris, av. Terreaux, 86.
208. Chapt de Rustaqne (Louis-Jacques de), archevêque de Tours. L. aut. sig., à M. Amelot. 27 octobre 1739. 1 p. in-fol. Gou- rigue.	4. Rennard , Paris, r. Amsterdam, 65.
209. Chapital (le comte), savant chimiste, membre de l'Institut. Né à <i>Lazare (Lozère)</i> . L. aut. sig., à M. Cadet de Gassicourt. Paris, 5 nov. 1812. 1 p. in-4.	6. Andronozoff , gr. ^e Beauzin, Paris, Ch. Anatin, 25.
210. Chastellain (du marquis de), poète, littérateur, membre de l'Académie française, auteur de : <i>Le poète philosophe</i> . L. aut. sig. papier. 1785. 4 p. pl. in-4.	10-12. Chalon-Sampson (M^e) , Paris, r. Calais,
211. Chauliac (Marie-Joseph-Louis d'Albret d'Ailly, due de), baron emmuni. L. aut. sig., à M. Corbin. 27 juin 1789. 1 p. in-4. 2 50	19. 20.
212. Chevallier (Michel), Saint-Simonien, savant ingénieur. L. aut. sig. Paris, 3 juin 1840. 1 p. pl. in-8.	14. Blech , Paris, r. Calais, 14.
213. Chevreuse (le duc de), fils du duc de Luynes, membre de l'Académie des sciences. L. aut. sig., au baron Desouches. Paris, 3 février 1817. 1 p. et demi in-4. Cachet.	16. Lalaurie , Paris, r. Calais, 16.
214. Chéoyean , premier chirurgien du roi, membre de l'Académie des sciences. 1 ^{re} Mémoire au roi, aut. sig. (à la 2 ^e personne), en faveur de son fils. 1 p. in-4; — 2 ^e supplique au roi Louis XV.	18. Henneaut , Paris, r. Calais, 18.
	20. Tripain (Vve) , Paris, r. Calais, 20.
	22. Grégoire , Paris, r. Calais, 22.
	24. Lecour (M^e) , Boulogne s/S., r. Est, 64.
	26. Genthier , Paris, r. Calais, 26.
	C. 1.6 CALLOT (Rue) ⁴¹
	1. Cartier , Paris, r. Masset, 1.
	3. Hodé , gr. ^e Odil, Paris, r. Callot, 5.
	7. Deschandelles , Paris, r. St-Honoré, 265.
	C. 1.7 CALMELS (Impasse) ⁴²
	3. Lézamy (Vve) , Paris, imp. Calmel, 3.
	5. Moulin , Paris, imp. Calmel, 5.
	7. Pommier , Paris, imp. Calmel, 7.
	2. (Terminal)
	4. Rion (M^e) , Paris, imp. Calmel, 4.
	6. Thibaud , Paris, imp. Calmel, 6.
	8. Entrée r. Palé-Nord , 18.
	24 ^{ea} . Loulême (Vve) , Paris, imp. Calmel, 24 ^{ea} .
	C. 1.8 CALMELS (Rue) ⁴³
	1-3. Thomas, Légalis , r. Voltaire, 73.
	5. Pontrel , Paris, r. Calmel, 5.
	7. Entrée r. Ordener , 118 ^{ea} .
	9. Devavanne , Paris, r. Calmel, 9.
	11. Cartey , Paris, r. Calmel, 11.
	13. Grégoire (M^e) , Paris, r. Calmel, 1.
	15. Raimond , Paris, r. Calmel, 15.
	17. Blech , Paris, r. Calmel, 17.
	19. Pelti (A.) , Paris, r. Lamartine, 6.
	25. Bayle , Paris, r. Calmel, 25.
	27. Garnier (M^e) , Paris, r. Calmel, 27.
	29. Grégoire (M^e) , Paris, r. Calmel, 29.
	31-33. Fripe (Vve) , Paris, r. Calmel, 31.
	35. Parmentier , Paris, r. Calmel, 35.
	37. Jaucourt , Paris, r. Tour, 91.

Catalogue [...] manuscrits, Laverdet, 1856, p.21

Annuaire [...] Propriétaires, 1898, p.650



Catalogue [...] Courbet, 1882, p.39

Modèle Beaufort (sur un dataset de 150 pages)

	21		650
107. Cambrouze (le général). Né à Saint-Sébastien (Loire-Inférieure). L. aut. sig., à M. Francheteau, à Nantes. 8 septembre 1819. 4 p. in-4. 3 * 198.		13. Dussuchet , Boulogne s/S., r. Sally, 32.	7. Haré , Paris, quai Louvre, 30.
108. Caron (abbé Toussaint-Jullien). Né à Rennes. L. aut. sig., au vicomte de Montmorency. Paris, 26 avril 1818. 4 p. in-4. 2 * 199.		15. Ricard , Paris, r. Tour, 78.	9. Worth , Paris, r. Lagrange, 10.
109. Castelnau (Michel de), seigneur de <i>Le Maurissier</i> , ambassadeur. Né à <i>La Maurissiere</i> , M. l'abbé. Pièce notariée, 2 fois signée. Paris, 1820. 4 p. in-4. 2 * 200.		17. Laporte , Paris, r. Sébastopol, 11.	11. Esquenazi , Paris, r. Calais, 13.
200. Catinat (Nicolas de), maréchal de France. Quitt. sig. (sur parchemin) de la somme de 500 livres. 18 mars 1704. 4 * 201.		21. Lalosse (de), Paris, r. Paul-Louis Courrier, 13.	15. Girard , Paris, r. Craté, 15.
201. Cauchy (le chevalier), secrétaire-archiviste du sénat-conservateur, littérateur. Né à Rouen en 1753. L. aut. sig. Paris, 29 avril 1806. 2 p. in-4. 2 *		22. Ribot , Paris, hôtel Pereire, 206.	17. Cinchant (Vve), Paris, r. Amsterdam, 55.
202. Causans (Jos.-L. Vime, de <i>Manteau-de</i>), mathématicien, gouverneur du prince de Condé, auteur des <i>Spectacles de l'Amour</i> , où Né à Paris. L. aut. sig., à M. de Malesherbes. Paris, 22 décembre 1739. 2 gr. p. in-fol. 3 *		23. Parmain , Riom [Puy-de-Dôme].	19. Milly (Mme de), Paris, r. Calais, 19.
203. Chaligny , architecte de l'Arc de triomphe de l'Étoile. L. aut. sig. 21 octobre 1807. 2 p. pl. in-8. 2 *		27. Pecher , Paris, r. Craté, 15.	21. Mission du Sacré-Cœur , Paris, r. Craté 21.
204. Châlon-sur-Marne prend ses anciennes armes. L. aut. sig. de M. de Chamorin, maire de Châlons-sur-Marne, à M. Geoffroy, référendaire à la chancellerie, pour prier de se charger de faire obtenir à la ville Châlons la confirmation des anciennes armoires, etc. 4 ^e déc. 1814. 4 p. pl. in-4. 6 *		29. Durst-Wild , frères, Paris, r. Caire, 39.	23. Rouyer (Vimille), 24.
205. Chamberlain (Edouard), mem. de l'Académie des sciences. L. u. s., au roi. Bonhiver, 18 juin 1819. 3 gr. p. pl. in-fol. 3 *		31. Bertrand , Paris, av. Opéra, 4.	2. Chalon , Paris, av. Ternes, 86.
Demande de secours pour la formation des établissements qui manquent en France, pour la prospérité de l'agriculture, commerce et industrie du royaume.		32. Boussinger , Paris, r. Bonaparte, 11.	4. Renard , Paris, r. Austerlitz, 65.
206. Chapelin (Jean), évêque de Bordeaux, membre de l'Assemblée constituante, garde-des-sceaux. Né à Rennes. L. aut. sig., au baron... Bordeaux, 28 const... 4 p. pl. in-4. 3 *		33. Cousin (Vve), Paris, r. Montalivet, 3.	6. Androsoff , gré. Bourrain, Paris, Ch. Austin, 25.
207. Chancy (Antoine), agronome. Né à Lyon. L. aut. sig. 31 décembre 1821. 4 p. et demi in-4. 2 *		35. Clermonte , Paris, r. Renne, 151 ^{re} .	10-12 Barbedienne-Sampson (Mme) , Paris, r. Calais, 10.
208. Chapt de Rustigne (Louis-Jacques de), archevêque de Tours. L. aut. sig., à M. Amelot, 27 octobre 1739. 4 p. pl. in-fol. Curieuse. 4 *		37. Bouchemain , gré. Menutin, Paris, r. Ulm, 38.	14. Bloch , Paris, r. Calais, 14.
209. Chaptal (le comte), savant chimiste, membre de l'Institut. Né à Lavaur (Lozère). L. aut. sig., à M. Caillie de Gassanoff. Paris, 5 juil. 1812. 4 p. pl. in-4. 2 *		39. Durst-Wild , frères, Paris, r. Caire, 39.	16. Labarque , Paris, r. Calais, 16.
210. Chastellier (de), chirurgien des hôpitaux, littérateur, membre de l'Académie française, auteur de : <i>Traité de la fièvre publique</i> . L. aut. sig. Versailles, 7 mai 1785. 4 p. pl. in-4. 3 *		41. Berizon , Paris, av. Opéra, 4.	18. Hennebert , Paris, r. Calais, 18.
211. Chautiez (Marie-Joseph-Louis d'Albert d'Ailly, duc de), baron chimiste. L. aut. sig., à M. Cochet, 27 juil. 1785. 4 p. in-4. 2 *		42. Decaux , Paris, r. N.-D. des Champs, 107.	20. Tigoni (Vive de), Paris, r. Calais, 20.
212. Chavallier (Michel), Saint-Simonien, savant ingénieur. L. aut. sig. Paris, 3 juil. 1840. 4 p. pl. in-8. 2 *		43. Rebot , Paris, r. Louvre, 6 et 8.	22. Thiébault , Paris, r. Craté, 11.
213. Chevreuse (le duc de), fils du duc de Luynes, membre de l'Académie des sciences. L. aut. sig., un baron Desfontaines, Paris, 6 février 1817. 4 p. et demi in-4. 2 *		44. Candot (M ^{me}), Paris, r. Volney, 4.	24. Lecour (M^{me}) , Boulogne s/S., r. Est, 64.
214. Chicoyenne , premier chirurgien du roi, membre de l'Académie des sciences. 1 ^{re} Mesure au roi, aut. sig. à la 2 ^{re} personne, en faveur de son fils. 4 p. in-4. — 2 ^{re} supplique au roi Louis XV. 2 *		45. Leclerc , Paris, r. St-Laurent, 10.	26. Gonthier , Paris, r. Calais, 26.
		46. Clément de Tonnelle (Vve) , ger. Naquet, Paris, r. Craté, 17.	
		47. Guillet , Paris, r. Caire, 28.	
		48. Hattier , Paris, r. Craté, 14.	
		49. Entrepas , Paris, r. Caire, 11.	
		50. Lefrère , Paris, r. Aubet, 19.	
		51. Lejeune , Paris, r. Assas, 72.	
		52. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		53. Richard , Paris, r. Assas, 72.	
		54. Richard , Paris, r. Craté, 15.	
		55. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		56. Guillet , Paris, r. Caire, 20.	
		57. Hattier , Paris, r. Craté, 14.	
		58. Entrepas , Paris, r. Caire, 11.	
		59. Lejeune , Paris, r. Assas, 72.	
		60. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		61. Richard , Paris, r. Craté, 15.	
		62. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		63. Guillet , Paris, r. Caire, 28.	
		64. Hattier , Paris, r. Craté, 14.	
		65. Entrepas , Paris, r. Caire, 11.	
		66. Lejeune , Paris, r. Assas, 72.	
		67. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		68. Richard , Paris, r. Craté, 15.	
		69. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		70. Guillet , Paris, r. Caire, 28.	
		71. Hattier , Paris, r. Craté, 14.	
		72. Entrepas , Paris, r. Caire, 11.	
		73. Lejeune , Paris, r. Assas, 72.	
		74. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		75. Richard , Paris, r. Craté, 15.	
		76. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		77. Guillet , Paris, r. Caire, 28.	
		78. Hattier , Paris, r. Craté, 14.	
		79. Entrepas , Paris, r. Caire, 11.	
		80. Lejeune , Paris, r. Assas, 72.	
		81. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		82. Richard , Paris, r. Craté, 15.	
		83. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		84. Guillet , Paris, r. Caire, 28.	
		85. Hattier , Paris, r. Craté, 14.	
		86. Entrepas , Paris, r. Caire, 11.	
		87. Lejeune , Paris, r. Assas, 72.	
		88. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		89. Richard , Paris, r. Craté, 15.	
		90. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		91. Guillet , Paris, r. Caire, 28.	
		92. Hattier , Paris, r. Craté, 14.	
		93. Entrepas , Paris, r. Caire, 11.	
		94. Lejeune , Paris, r. Assas, 72.	
		95. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		96. Richard , Paris, r. Craté, 15.	
		97. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		98. Guillet , Paris, r. Caire, 28.	
		99. Hattier , Paris, r. Craté, 14.	
		100. Entrepas , Paris, r. Caire, 11.	
		101. Lejeune , Paris, r. Assas, 72.	
		102. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		103. Richard , Paris, r. Craté, 15.	
		104. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		105. Guillet , Paris, r. Caire, 28.	
		106. Hattier , Paris, r. Craté, 14.	
		107. Entrepas , Paris, r. Caire, 11.	
		108. Lejeune , Paris, r. Assas, 72.	
		109. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		110. Richard , Paris, r. Craté, 15.	
		111. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		112. Guillet , Paris, r. Caire, 28.	
		113. Hattier , Paris, r. Craté, 14.	
		114. Entrepas , Paris, r. Caire, 11.	
		115. Lejeune , Paris, r. Assas, 72.	
		116. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		117. Richard , Paris, r. Craté, 15.	
		118. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		119. Guillet , Paris, r. Caire, 28.	
		120. Hattier , Paris, r. Craté, 14.	
		121. Entrepas , Paris, r. Caire, 11.	
		122. Lejeune , Paris, r. Assas, 72.	
		123. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		124. Richard , Paris, r. Craté, 15.	
		125. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		126. Guillet , Paris, r. Caire, 28.	
		127. Hattier , Paris, r. Craté, 14.	
		128. Entrepas , Paris, r. Caire, 11.	
		129. Lejeune , Paris, r. Assas, 72.	
		130. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		131. Richard , Paris, r. Craté, 15.	
		132. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		133. Guillet , Paris, r. Caire, 28.	
		134. Hattier , Paris, r. Craté, 14.	
		135. Entrepas , Paris, r. Caire, 11.	
		136. Lejeune , Paris, r. Assas, 72.	
		137. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		138. Richard , Paris, r. Craté, 15.	
		139. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		140. Guillet , Paris, r. Caire, 28.	
		141. Hattier , Paris, r. Craté, 14.	
		142. Entrepas , Paris, r. Caire, 11.	
		143. Lejeune , Paris, r. Assas, 72.	
		144. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		145. Richard , Paris, r. Craté, 15.	
		146. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		147. Guillet , Paris, r. Caire, 28.	
		148. Hattier , Paris, r. Craté, 14.	
		149. Entrepas , Paris, r. Caire, 11.	
		150. Lejeune , Paris, r. Assas, 72.	
		151. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		152. Richard , Paris, r. Craté, 15.	
		153. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		154. Guillet , Paris, r. Caire, 28.	
		155. Hattier , Paris, r. Craté, 14.	
		156. Entrepas , Paris, r. Caire, 11.	
		157. Lejeune , Paris, r. Assas, 72.	
		158. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		159. Richard , Paris, r. Craté, 15.	
		160. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		161. Guillet , Paris, r. Caire, 28.	
		162. Hattier , Paris, r. Craté, 14.	
		163. Entrepas , Paris, r. Caire, 11.	
		164. Lejeune , Paris, r. Assas, 72.	
		165. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		166. Richard , Paris, r. Craté, 15.	
		167. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		168. Guillet , Paris, r. Caire, 28.	
		169. Hattier , Paris, r. Craté, 14.	
		170. Entrepas , Paris, r. Caire, 11.	
		171. Lejeune , Paris, r. Assas, 72.	
		172. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		173. Richard , Paris, r. Craté, 15.	
		174. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		175. Guillet , Paris, r. Caire, 28.	
		176. Hattier , Paris, r. Craté, 14.	
		177. Entrepas , Paris, r. Caire, 11.	
		178. Lejeune , Paris, r. Assas, 72.	
		179. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		180. Richard , Paris, r. Craté, 15.	
		181. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		182. Guillet , Paris, r. Caire, 28.	
		183. Hattier , Paris, r. Craté, 14.	
		184. Entrepas , Paris, r. Caire, 11.	
		185. Lejeune , Paris, r. Assas, 72.	
		186. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		187. Richard , Paris, r. Craté, 15.	
		188. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		189. Guillet , Paris, r. Caire, 28.	
		190. Hattier , Paris, r. Craté, 14.	
		191. Entrepas , Paris, r. Caire, 11.	
		192. Lejeune , Paris, r. Assas, 72.	
		193. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		194. Richard , Paris, r. Craté, 15.	
		195. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		196. Guillet , Paris, r. Caire, 28.	
		197. Hattier , Paris, r. Craté, 14.	
		198. Entrepas , Paris, r. Caire, 11.	
		199. Lejeune , Paris, r. Assas, 72.	
		200. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		201. Richard , Paris, r. Craté, 15.	
		202. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		203. Guillet , Paris, r. Caire, 28.	
		204. Hattier , Paris, r. Craté, 14.	
		205. Entrepas , Paris, r. Caire, 11.	
		206. Lejeune , Paris, r. Assas, 72.	
		207. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		208. Richard , Paris, r. Craté, 15.	
		209. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		210. Guillet , Paris, r. Caire, 28.	
		211. Hattier , Paris, r. Craté, 14.	
		212. Entrepas , Paris, r. Caire, 11.	
		213. Lejeune , Paris, r. Assas, 72.	
		214. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		215. Richard , Paris, r. Craté, 15.	
		216. Tavares , ger. Belliard, Paris, r. Le Peletier, 18.	
		217. Guillet , Paris, r. Caire, 28.	
		218. Hattier , Paris, r. Craté, 14.	
		219. Entrepas , Paris, r. Caire, 11.	
		220. Lejeune , Paris, r. Assas, 72.	
		221. Trépinneaux , ger. Muier, Paris, r. Caire, 20.	
		222. Richard , Paris, r. Craté, 15.	
		223. Tavares</b	

Modèle Chaource (sur un dataset de 274 pages)

— 21 —

197. Cambronne (le général). Né à <i>Saint-Sébastien</i> (Loire-Inférieure). L. aut. sig., à M. Francheteau, à Nantes. 8 septembre 1819.
198. Caron (labbé Toussaint-Jullien). Né à <i>Reuilly</i> . L. aut. sig. au comte de Montmazyence. Paris, 26 avril 1816. 1 p. in-4. 30
199. Castelnau (Michel de), seigneur de <i>La Neuvière</i> , ambassadeur. Né à <i>La Neuvière</i> . M.. 1292. Pièce notariée, 2 fois signée, 25 août 1586. 1 p. pl. et demie in-fol. 3 *
200. Catinat (Nicolas de), maréchal de France. Quitt. sig. par partie de la somme de 100 livres. 18 octobre 1704. 4 50
201. Cauchié (de), chirurgien, secrétaire du sénat-conservateur, littérateur. Né à <i>Rouen</i> en 1755. L. aut. sig. Paris, 29 avril 1806. 2 p. in-4. 2 *
202. Cansane (Jos.-J.-Vine de <i>Mautou de</i> , mathématicien, gouverneur du prince de Coni, auteur de <i>Spéciale de l'homme</i> , etc. Né à <i>Aragon</i> . L. aut. sig., à M. de Malederbe. Paris, 22 octobre 1789. 2 p. pl. in-fol. 3 *
203. Chalons-sur-Marne reprend ses anciennes armes. L. aut. sig. 21 octobre 1807. 2 p. pl. in-8. 2 *
204. Châlons-sur-Marne reprend ses anciennes armes. L. aut. sig. 21 octobre 1807. 2 p. pl. in-8. 2 *
205. Chamberlain (Edouard), membre de l'Académie des sciences. L. a. s., au roi. Bonjour, 18 juin 1819. 3 gr. p. pl. in-fol. 3 *
Demande de secours pour la formation des établissements qui manquent en France, pour la prospérité de l'agriculture, commerce et industrie du royaume.
206. Champion (Gér.), évêque de Bordeaux, membre de l'Assemblée constitutive, garde-des-sceaux. Né à <i>Reuilly</i> . L. aut. sig. au baron de Bordeaux. 28 octobre 1 p. pl. in-4. 3 *
207. Chancy (Antoine), agronome. Né à <i>Lyon</i> . L. aut. sig. 31 décembre 1821. 1 p. pl. et demie in-4. 2 *
208. Chap de Bastignac (Louis-Jacques de), archevêque de <i>Tours</i> . L. aut. sig., à M. Amelot, 27 octobre 1739. 1 p. in-fol. Curieuse.
209. Chaptal (de camille), savant chimiste, membre de l'Institut. Né à <i>Louzun-Louzun</i> . L. aut. sig., à M. Cadet de Gassicourt. Paris, 22 octobre 1812. 1 p. pl. in-4. 2 75
210. Charteloux (de), poète, littérateur, membre de l'Académie française, auteur de <i>Traité de la félicité publique</i> . L. aut. sig. Versailles, 7 mai 1785. 1 p. pl. in-4. 3 50
211. Chauvelin (Marie-Joseph-Louis d'Albret d'Ally), duc de, habité chimiste. L. aut. sig., à M. Cochon, 27 juillet 1783. 1 p. in-4. 2 50
212. Chevalier (Bichelli), Saint-Simonien, savant ingénieur. L. aut. sig. Paris, 3 juin 1840. 1 p. pl. in-8. 2 25
213. Chrevrouse (de), docteur en médecine, membre de l'Académie de Lyon. L. aut. sig., à son frère, Jean-Baptiste. Paris, 20 octobre 1817. 1 p. et demie in-4. 2 25
214. Chirayens , premier chirurgien du roi, membre de l'Académie des sciences. 1 ^e Mémoire au roi, aut. sig. (à la 2 ^e personne), en faveur de son fils. 1 p. in-4; — 2 ^e supplique au roi Louis XV.

Catalogue [...] manuscrits, Laverdet, 1856, p.21

— 99 —

16. — L'Homme blessé.

Siglé : G. Courbet, avec paraphé.

Sans date.

Expositions particulières de 1855 et de 1867. Le catalogue de cette dernière exposition place l'exécution de *L'Homme blessé* à la date de 1844. C'est évidemment une erreur : il suffit d'examiner la facture libre et souple du tableau, pour être convaincu que ce n'est point là une œuvre de débutant. L'erreur ayant été reproduite dans le catalogue de la vente du 9 décembre 1851, nous croyons devoir la rectifier. La véritable date est 1854, comme l'atteste le catalogue de la première exposition privée du peintre, en 1855.

Acheté pour l'Etat à la vente du 9 décembre 1851.

T. — H. 8.90. L. 8.81.

17. — Le Chasseur badois.

Siglé à gauche : ...sg. Gustave Courbet.
T. — H. 1.18. L. 1.75.

Appartient à M. D... M..

18. — Portrait de M. Guermard, artiste de l'Opéra, dans le rôle de *Robert le Diable*.

* Oui, l'or est une chimère. *

Siglé à gauche : G. Courbet.

Sans date.

Salon de 1857.

T. — H. 1.57. L. 1.06.

Appartient à M. Adolphe Reignard.

Catalogue [...] Courbet, 1882, p.39

650

13. Dassouhet , Boulogne s/S., r. Sally, 32.
15. Ricard , Paris, r. Tour, 78.
17. Perte , Neuilly s/S., r. Ancelle, 9.
19. Longuet (Vve), Paris, r. Joubert, 26.
21. Lalesse (de), Paris, r. Paul-Louis Courier, 13.
23. Billet , Paris, boulevard Poissonnière, 206.
25. Pommeret , Paris, r. Porte-Dôme.
27. Decloux , Paris, r. Caïre, 27.
29. Besson , Paris, r. Bonaparte, 11.
31. Cousin (Vve), Paris, r. Montalivet, 3.
33. Bourrier , Paris, r. St-Michel, 4.
35. Clemente , Paris, r. Rennes, 151 ^{er} .
37. Bouffard , Paris, r. Meudon, Paris, r. Umn, 38.
39. Dard-Wild , frères, Paris, r. Caïre, 39.
41. Berçoux , Paris, av. Opéra, 4.
43. Decous , Paris, r. N.-D. des Champs, 167.
45. Reldot , Paris, r. Louvre, 6 et 8.
47. Cordier (Mme), gér. Lavarene, Paris, r. Berlin, 15.
49. Entrée r. Forges, 10.
51. Van der Nett , Paris, r. Caïre, 51.
53. Dumeuret (Herr), gér. Cavalier, Paris, r. Mazagran, 3.
52. Entrée boulevard Schœnholz, 113.
54. Jeanvier , gér. Léonard, Paris, r. Caïre, 4.
56. Levavasseur , Paris, r. Orléans, 1.
58. Mauré (Mme), Paris, r. Volney, 4.
60. Lecara , Paris, r. St-Laurent, 7.
62. Clerc de Tercyville (Vve), gér. Nequel, Paris, r. Lançay, 17.
64. Hattier , Paris, r. Caïre, 14.
66. Leblanc , Paris, r. Caïre, 14.
68. Lefèvre , Paris, r. Auber, 19.
70. Tréguier , Paris, gér. Mader, Paris, r. Caïre, 20.
72. Richard , Paris, r. Assas, 72.
74. Dussoix , Boulogne s/S., r. Sally, 32.
76. Dussoix , Paris, r. Béthune, Paris, r. Le Peletier, 18.
78. Gobelin , Paris, r. Caïre, 28.
80. Lalanne , Paris, r. Caïre, 30.
82. Delavaye , Paris, boulevard Tempé, 18.
84-36. Hébert (F.), Paris, N.-D. Victoires, 14.
88. Rulé (Bé), Paris, r. Cambon, 43.
90. Clain (Vve), gér. Marie, r. Louvre, 3.
92. Guenier (Vve), Paris, r. Charençon, 128.
94. Bordet , Paris, r. Assas, 97.
96. Meda (Bé), Paris, r. Cambon, 46.
98. Entrée pl. Caïre, 2.

14. — CALLOT (Rue)

1. Carlier , Paris, r. Musset, 1.
3-5. Hode , gér. Odile, Paris, r. Calot, 3.
7. Duchandiers , Paris, r. St-Honoré, 265.
12. — CALMELS (Impasse)
3. Lebaucq (Vve), Paris, imp. Calmels, 3.
5. Moulin , Paris, imp. Calmels, 5.
7. Thiébaut , Paris, imp. Calmels, 7.
2. Ternu , Paris, imp. Calmels, 2.
4. Rion (M ^e), Paris, imp. Calmels, 4.
6. Lamouche , Paris, imp. Calmels, 6.
18. Katrén r. Pôle Nord, 18.
24 ^{da} Loulmet (Vve), Paris, imp. Calmels, 24 ^{da} .
24 ^{da} Loulmet (Vve), Paris, imp. Calmels, 24 ^{da} .

14. — CALMELS (Rue)

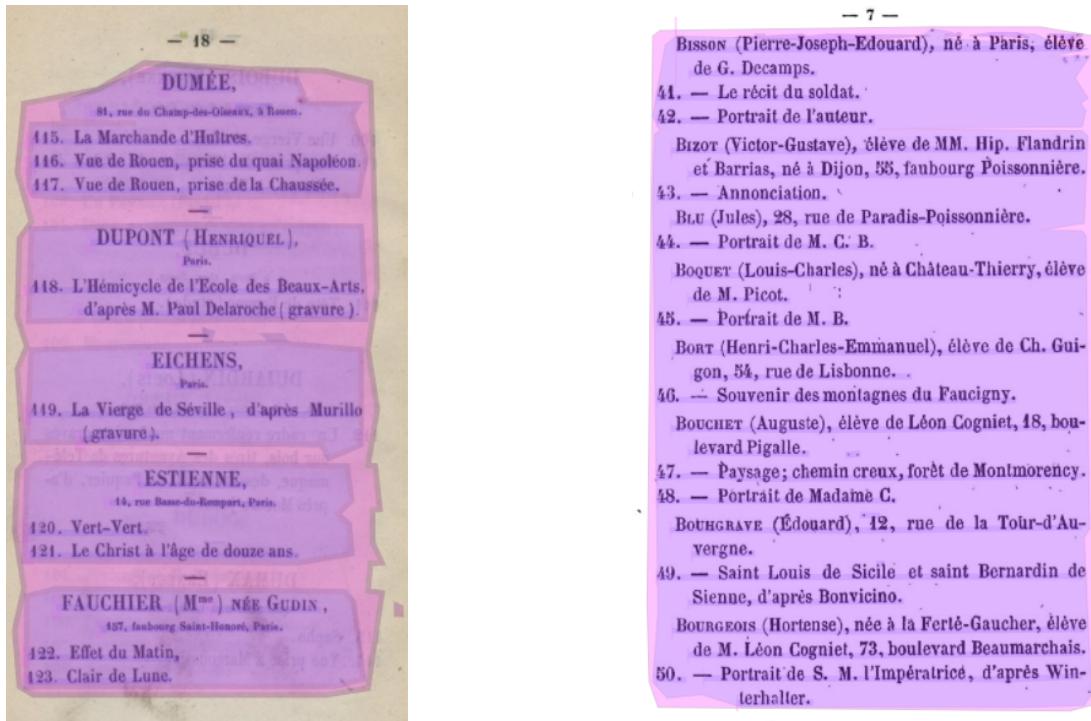
1-3. Thomas , Levallais, r. Voltaire, 73.
5. Pontelle , Paris, r. Calmels, 5.
7. Katrén r. Ordener, 118 ^{da} .
9. Devanenes , Paris, r. Calmels, 9.
11. Porter , Paris, r. Calmels, 11.
13. Imbert (Mme), Paris, r. Calmels, 1.
15. Raimondi , Paris, r. Calmels, 15.
17. Riecke , Paris, r. Calmels, 17.
19. Petit (A.), Paris, r. Lamartine, 6.
25. Bayle , Paris, r. Calmels, 25.
27. Garguier (M ^e), Paris, r. Calmels, 27.
29. Thibaut , Paris, r. Calmels, 29.
31-33 Foige (Vve), Paris, r. Calmels, 31.
35. Parmentier , Paris, r. Calmels, 35.
37. Jaucourt , Paris, r. Tour, 91.

Annuaire [...] Propriétaires, 1898,
p.650

— 104 —

GUERIN (Charles), né à Sens. — 1, rue Leclerc, 1 ^{er} .
1572 Nature morte. — 5.000 fr.
1573 Promenade. — 5.000 fr.
GUERIN (Pierre), né à Toulon. — A. La Serre, Pignans (Var).
1574 Jardin sur la Côte d'Azur, Cap Ferrat. — 700 fr.
1575 La Meije. — 700 fr.
GUERZONI (Stéphanie), née à Vienne. — Suisse. — 2, rue Antoine-Chantin, 1 ^{er} .
1576 Paysage. — 1.500 fr.
1577 Figure. — 2.500 fr.
GUI (Géo), né à Versailles. — 35, rue de Chatou, à Houilles (S.-et-O.).
1577 bis Avant d'éteindre (pastel). — 1.500 fr.
1577 ter Le rapin (pastel). — 1.500 fr.
GUIBERT-LASSALLE (André), né à Paris. — 43, rue Damrémont, 1 ^{er} .
1578 Paysage. — 1.100 fr.
1579 Nature morte : tête de cochon. — 750 fr.
GUICHARD (Mario), né à Saint-Etienne. — a. Les Fougères, chemin de Selaure, Saint-Etienne.
1580 Peinture.
1581 Peinture.
GUILLAUMET (Yvonne), née à Paris. — 47, rue de Passy, 10 ^{er} .
1582 Le Thuit : la basse-cour. — 1.200 fr.
1583 Le Thuit : le pavillon. — 1.100 fr.
GUINHALD (Bernard de), né à Saint-Calais (Sarthe). — a. Le Clos Berbère, n ^o , Cagnes-sur-Mer (Alpes-Maritimes).
1584 Gorges de l'Ariège. — 400 fr.
1585 Orgeix (Pyrénées-Orientales). — 400 fr.
GUITMAN (Albert), né à Nemirov — Français. — 81, rue Michel-Ange, 16 ^{er} .
1586 Le Vésinet : « Les Ibis ». — 1.000 fr.
1587 Vaches dans la prairie. — 2.000 fr.

Catalogue [...] indépendants, 1935,
p.104



Catalogue [...] Rouen, 1856, p.18

Catalogue [...] refusés, 1863, p.7

B.2 Reconnaissance de caractères

	dataset	Accuracy	CER
Abondance	30 pages sans repolygonisation	95,22%	4,8%
Beaufort	30 pages avec repolygonisation	67,88%	33,3%
Danablu	30 pages après correction des bugs de développement	95,22%	4,8%
Epoisse	30 pages dont les lignes sont natives d'eScriptorium	94,90%	5,10%

Mesures de la qualité des modèles produits(sur un test de 3 pages)

	dataset	30 pages		20 pages d'exposition	
		Accuracy	CER	Accuracy	CER
Chaource	100 pages	92,49%	7,51%	90,77%	9,23%
Fourme	375 pages	92,05%	7,95%	90,41%	9,59%
Gruyere	545 pages	91,86%	8,14%	90,92%	9,08%

Mesures de la qualité des modèles produits de plus de 100 pages

Zones		Lignes		Caractères		Résultats
Modèle	Dataset	Modèle	Dataset	Modèle	Dataset	CER
Chaource	Entraînement sur 274 pages	Chaource	Entraînement sur 274 pages	Beaufort	1er entraînement sur 30 pages sans repolygonisation	4,9 %
Chaource	Entraînement sur 274 pages	Chaource	Entraînement sur 274 pages	Fourme	Entraînement sur 30 pages avec segmentation lines blla	6,2 %
Chaource	Entraînement sur 274 pages	Modèle de base de Kraken		Beaufort	1er entraînement sur 30 pages sans repolygonisation	4,8 %
Chaource	Entraînement sur 274 pages	Modèle de base de Kraken		Fourme	Entraînement sur 30 pages avec segmentation lines blla	6,2 %
Chaource	Entraînement sur 274 pages	Chaource	Entraînement sur 274 pages	Chaource	Entraînement sur 100 pages	2,8 %
Chaource	Entraînement sur 274 pages	Chaource	Entraînement sur 274 pages	Gruyere	Entraînement sur 375 pages	4,6 %
						5,1 %

Comparaison des différents couples transcription-segmentation

Annexe C

Rapports et Articles

C.1 Articles

Les pages suivantes contiennent les propositions d'articles auxquelles j'ai participé dans le cadre de mon stage.

The BIR database – Identifying typographic emphasis in list-like historical documents

Anna Scius Bertrand*

name.surname@unige.ch

UniGE, Genève, Switzerland

HES-SO, Fribourg, Switzerland

EPHE-PSL

Paris, France

Simon Gabay

Ljudmila Petković

name.surname@unige.ch

UniGE

Genève, Switzerland

Juliette Janes

Caroline Corbières

Thibault Clérice

name.surname@chartes.psl.eu

ENC-PSL

Paris, France

ABSTRACT

Layout analysis and optical character recognition have become traditional tasks for processing historical prints, but are now insufficient. Additional information is found in typographic emphasis, such as bold and italic letters. They carry semantic meaning (titles, emphasis...) and also outline the structure of the page (entries, sub-parts...). Retrieving such data is therefore crucial for information extraction and automatic document structuring. In this paper, we introduce the Bold-Italic-Regular (BIR) database, which contains 285 pages of scanned, list-like historical prints that have been annotated at word level with bold and italic emphasis. Baseline results are provided for word detection and style classification using state-of-the-art deep neural network models, highlighting the challenges of the task.

CCS CONCEPTS

- Applied computing → Arts and humanities; Optical character recognition.

KEYWORDS

typographic information, historical print, object detection, style classification

ACM Reference Format:

Anna Scius Bertrand, Simon Gabay, Ljudmila Petković, Juliette Janes, Caroline Corbières, and Thibault Clérice. 2020. The BIR database – Identifying typographic emphasis in list-like historical documents. In *HIP'21: 6th International Workshop on Historical Document Imaging and Processing, September 05–07, 2021, Lausanne, CH*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3423603.3424002>

1 INTRODUCTION

Thanks to the recent improvements of Handwritten Text Recognition (HTR) engines [Kiessling 2019; Wick et al. 2018; ?] and the creation of user-friendly interfaces [Kiessling et al. 2019; Reul et al. 2019a], retrieving the text from an image has become an increasingly easy task. But along with the text itself, (old) prints offer additional graphic information that convey a semantic meaning that is still too often forgotten. It is the case of the layout (*e.g.* a paragraph starts with a carriage return and, potentially, an indentation), but also

of typographic emphasis (*e.g.* a title is in italic, such as a loan word). Identifying the latter is therefore crucial for digitisation tasks, may it be for traditional mining purposes (*e.g.* retrieving all the titles of works mentioned), but also automatic structuring of the text (*e.g.* encoding the logical structure thanks to physical and graphical hints found on the page).

The BIR database¹ introduced in this paper provides the research community with a benchmark dataset for developing and comparing methods for identifying **bold** and *italic* emphasis on scanned page images of list-like historical prints. In an experimental evaluation, we evaluate and compare state-of-the-art deep neural network models, including YOLOv5 [?] for word detection and MobileNetV2 [?] and Xception [?] for style classification. The obtained baseline results are promising but also demonstrate the difficulty of the task and the current limitations of the state of the art.

2 RELATED WORK

To be precise, “style” (*e.g.* italic, oblique...) or “weight” (*e.g.* bold, thin...) are attributes of possible “fonts” (*e.g.* Garamond, Arial...). When combined together we talk about a “typeface” (*e.g.* Garamond in bold vs Garamond in italic). These fonts are associated to “scripts” (*e.g.* latin for French or English, cyrillic for Russian or Serbian...). Very few researchers have addressed directly the question of typographic emphasis, but many have tackled the problem while dealing with a larger one – usually the identification of fonts or scripts, which are necessarily embodied in a typeface that can be in bold or italic.

The oldest approach to address typographic emphasis is the one of optical font recognition (OFR). Researchers have first tried to solve it using a font model base of several hundred known fonts and a multivariate Bayesian classifier [Zramdini and Ingold 1998]. Further studies have associated typographic emphasis with a texture. On the one hand, a simple weighted Euclidian distance has been used to classify fonts of synthetic data, after extracting features such as spatial frequency and orientation with multi-channel Gabor filters [Zhu et al. 2001]. On the other hand, local binary patterns have been used with a synthetically generated database of arabic text images [Nicolau et al. 2014]. Most recent approaches obviously use

¹The BIR database is available via <https://github.com/asciusb/BIR-database>.

neural networks, may they be LSTM [Tao et al. 2016] or CNN [Slimane et al. 2017].

In the past few years, in parallel to OFR, script recognition has benefited from numerous studies [Ubul et al. 2017], some of which may well be useful for the identification of typographic emphasis. It is particularly the case of experiments using the neural network implementation of OCR engines to identify multiple scripts at text-line level [Ul-Hasan et al. 2015], a strategy used successfully for recognizing semantico-typographic classes (dictionary entry, antiqua, fraktur, author's name and letter-spacing) in a real historical use case: Daniel Sander's *Wörterbuch der Deutschen Sprache* [Reul et al. 2019b].

3 THE BIR DATABASE

In the following, we comment on the general context of list-like historical prints, describe the specific document collection that constitutes the Bold-Italic-Regular (BIR) database, and provide details on how the scanned pages were annotated.

3.1 List-Like Historical Prints

During the past decade, list-like documents have drawn considerable attention from digital humanists, may they be art historians [Joyeux-Prunel and Marcel 2016], historians [di Lenardo et al. 2019], linguists [Salgado and Costa 2020], philologists [Gabay et al. 2020a]... Such documents have the particularity to use typographic emphasis to organise a rather dense information. Italic is used to identify professions in phone directories (cf. fig. 1), examples in dictionaries (cf. fig. 2) or title of works in manuscript catalogues (cf. fig. 3). Bold tends to have a more unified usage, and delimits the subject of the entry. Additional styles, such as small capitals, can also be found, as shown in the first example of the *Annuaire-Almanach*.

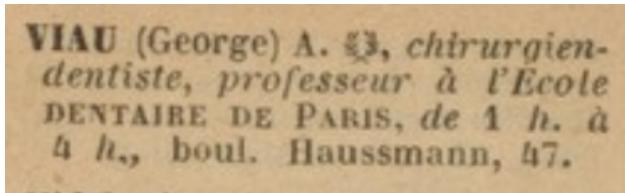


Figure 1: *Annuaire-almanach du commerce, de l'industrie, de la magistrature et de l'administration*, 1894, p. 1272, ark:/12148/bpt6k9732740w/f1498.

On top of additional mining options, typographic emphasis can be used to structure documents that we would like to encode. Bold is not only the topic of an entry, but the most efficient way to locate its beginning, and italic the only solution to locate specific passages. Without the information that some characters are in bold font, it is impossible to differentiate the first line (*24. La Rochefoucauld*) from the third (*2 gr...*) in the fig. 3. Similarly, it is impossible to distinguish the definition from the examples in the fig. 2, or

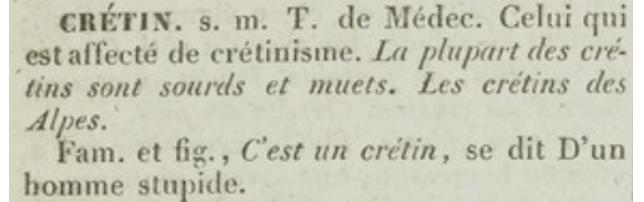


Figure 2: *Dictionnaire de l'Académie française - sixième édition*, Paris: Firmin Didot, 1835, vol. 1 (A-K.), p. 450, ark:/12148/bpt6k12804289/f490.



Figure 3: *Catalogues de lettres autographes, manuscrits, documents historiques, etc.*, Paris: Auguste Laverdet, N°1, avril 1856, p. 5, ark:/12148/bpt6k9687751c/f17.

the biographical sub-part (in green) from the philological one (in blue) in the fig. 4, and reconstitute with precision the entry in XML for further exploration [Gabay et al. 2020a].

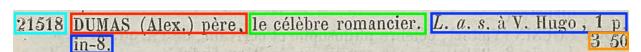


Figure 4: *Librairie autographe Ancienne et autographe Paris: Jacques Charavay, N°156, avril-mai 1867*, p. 6

```
<item n="21518" xml:id="CAT_001_e21518">
  <num type="lot">21518</num>
  <name type="author">Dumas (Alex.) père,</name>
  <trait>
    <p>le célèbre romancier</p>
  </trait>
  <desc xml:id="CAT_001_e21518_d1">L.a.s. à V. Hugo,
  1 p., in-8.</desc>
  <measure commodity="currency" unit="FRF">3 50</measure>
</item>
```

Figure 5: TEI modelling of fig. 4.

3.2 Document Collection

The BIR database contains a subset of 285 pages come from exhibition [Topalov et al. 2021] and sale catalogues [Gabay et al. 2021], for which we have strong evidences that using bold and italic information increase the precision of the extraction and the structuring [Gabay et al. 2020b]. Carefully selected excerpts have been taken in the documentation of the *Art@S* and the *Katabase* projects, *i.e.* mainly 19th French catalogues, which remains our primary target. Other types of documents (*e.g.* a 19th latin grammar) or in other languages than French (*e.g.* the São Art Paulo Biennale) have been added to diversify the data, as well as more recent catalogues

(20th and 21th c.). Tab. 1 lists the number of pages considered for the BIR database, according to different categories.

19 th prints	220	Exhibition catalogues	143
20 th prints	67	Manuscript catalogues	111
French documents	258	Other documents	33
Foreign documents	29		

Table 1: Number of pages according to different categories

3.3 Ground Truth Annotation

The words of the original database were segmented and classified according their style with the ABBYY FineReader software. The word segmentation was excellent but the style classification was far less successful. Each sentence was manually annotated with HTML tags to identify bold and italic words and the transcription was improved.

To conduct our experiments it was necessary to align the style information contained in the sentences with the locations of each word. Due to the manual correction of the transcription at the sentence level, the number of words contained in the sentence regularly differed from the number of words in the ABBYY FineReader output.

To solve this problem, a string edit distance algorithm was used to find the closest matching words, in order to correctly transfer the typographic emphasis. Finally, the ground truth file was manually inspected and corrected by adding, merging, and splitting word bounding boxes, and correcting the style information. Pages containing no text and those requiring too much manual editing were excluded. Table 2 lists some basic statistics of the resulting BIR database.

Documents	35
Pages	285
Words	88,019
- bold	2,106
- italic	5,745
- regular	80,168

Table 2: BIR database.

4 BASELINE RESULTS

Several experiments have been conducted on the BIR database to establish baseline results with methods from the current state of the art. Two tasks are considered, word detection and style classification.

4.1 Word Detection

Word detection aims at localizing words on a scanned page. We perform this task with a fully-convolutional object detection network, which takes a whole page image as input and provides bounding boxes around the detected words as output.

The YOLO [?] (You Only Look Once) model is considered, which has been a pioneering architecture for one-stage object detection and remained competitive over the years both in terms of speed and accuracy. The architecture consists of a convolutional *backbone*, followed by a feature pyramid *neck* that combines the extracted features at different scales, which are then processed by the *head* that computes two loss functions, one for bounding box regression and one for bounding box classification. In our case, three style classes are considered: bold, italic, and regular. Experiments are performed with the PyTorch-based YOLOv5 [?] version using the medium-sized YOLOv5m model with 21.4 million parameters pre-trained on the COCO [?] database.

Image preprocessing includes downscaling to a height of 1024 pixels, keeping the same aspect ratio, in order to fit the input images into the GPU memory. Furthermore, random scaling, translation, and rotation operations are applied during training to augment the number of training samples and improve the generalizability of the model.

4.2 Style Classification

Style classification aims at determining the style of an individual word image that has already been localized on the scanned page. The BIR database distinguishes bold, italic, and regular words.

Two fully-convolutional network architectures are used for establishing baseline results, namely MobileNetV2 [?] and Xception [?]. They were chosen with respect to their excellent classification performance on ImageNet [?] and to include both a smaller (MobileNetV2: 3.5 million parameters) and a larger (Xception: 22.9 million parameters) architecture. Experiments are performed with a Keras-based implementation and ImageNet pre-trained parameters. We drop the top layer, perform global average pooling to reduce the spatial dimensions, add a dense layer with 100 neurons, dropout, and rectified linear unit (ReLU) activation, and finish with a softmax classification layer. The architectures are illustrated in Figure 6.

Image preprocessing includes resizing to (160, 160) pixels and normalizing the RGB inputs to (-1,1) to match the pre-training condition.

As shown in Table 2, there is a significant class imbalance among the three classes, with *regular* being the majority class and *bold* and *italic* being the minority classes. We consider two strategies to alleviate this class imbalance:

- Balancing via over-sampling and under-sampling. We define a fixed amount of n training samples for each class, under-sample the majority class, randomly selecting n samples, and over-sample the minority classes, repeating the samples until n is reached.
- Using class weights for the loss function. We weight the loss of the minority classes with factor $\frac{m}{m'}$ where m is the number of majority samples and m' is the number of minority samples, in order to give more emphasis to the minority samples during training.



Figure 6: Network architecture for style classification using MobileNetV2 and Xception, respectively.

4.3 Experimental Setup

Two dataset splits are defined for experimental evaluation.

- **TVT.** In this standard setup, the 285 pages are randomly distributed into three distinct sets for training the neural networks (50%), validation of meta-parameters (25%), and testing of the final system performance (25%).
- **CV5.** In this cross-validation setup, the 35 documents are split into five distinct parts, each containing 7 documents. Five cross-validations are performed using three parts for training, one part for validation, and one part for testing, thus allowing to test the generalization capability of a trained network to an unseen document.

For word detection, the TVT setup is considered, training the YOLOv5m network with its default fine-tuning hyper-parameters 100 epochs on the training set until convergence. This base model is then further fine-tuned on an augmented dataset, randomly scaling, translating, and rotating each page hundred times. The augmented model is trained 20 epochs until convergence, which takes about 1 hour on two Titan RTX cards.

The detection performance is evaluated with respect to precision ($\# \text{correct detections} / \# \text{detections}$), recall ($\# \text{correct detections} / \# \text{words}$), and F1-score (harmonic mean of precision and recall), considering a detection as correct if its intersection over union (IoU) with the ground truth box is larger than 50%.

For style classification, an initial experiment is conducted with the MobileNetV2 for the TVT setup. Afterwards, more detailed experiments are conducted for the CV5 setup, comparing MobileNetV2 with the larger Xception architecture and the two strategies for coping with class imbalance, *class*

Table 3: Word detection results on the TVT test set.

	Precision	Recall	F1-Score
YOLOv5m	0.93	0.90	0.91

balancing and *class weights* respectively. The networks are trained with categorical cross-entropy loss and an adaptive Adam learning rate 50 epochs until convergence, which takes about 50 minutes on a Titan RTX card.

The best model is selected with respect to its accuracy on the validation set and then evaluated on the test set. Precision, recall, and F1-score are calculated for each style individually (bold, italic, regular), as well as their macro-average, i.e. the average of the evaluation measures without weighting with the number of samples from each class.

4.4 Results

Table 3 presents the detection performance achieved with the YOLO model on the TVT test set and Figure 7 illustrates some exemplary detection results. The object detection network is able to retrieve words with recall and precision over 90%, which demonstrates the feasibility of the approach but clearly leaves room for improvements. Typical errors include missing words (document in the middle) and errors due to ink bleed-through (document on the right).

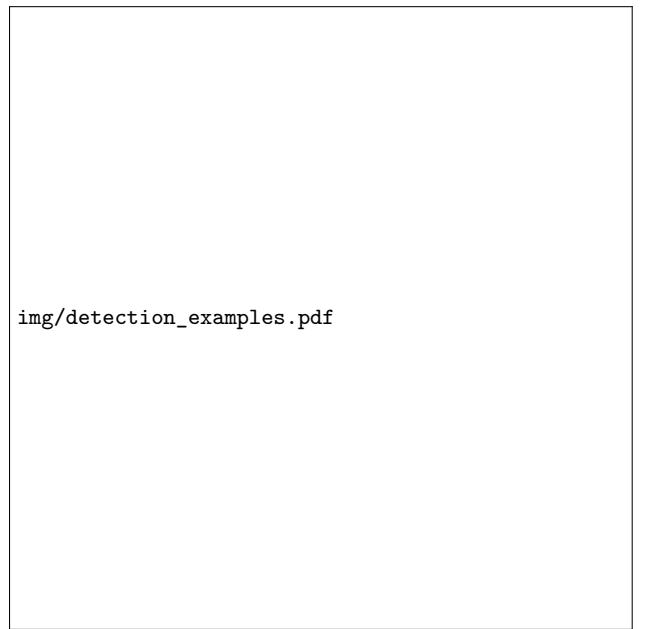
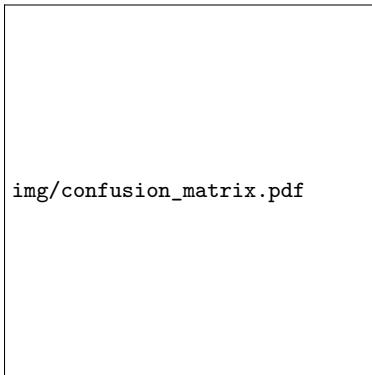


Figure 7: Word detection using YOLOv5m. Ground truth is marked in blue, detected words in green, missing words in red, and insertion errors in magenta.

Table 4: Style classification results with MobileNetV2 on the TTV test set.

	Precision	Recall	F1-Score
Regular	0.98	0.99	0.99
Bold	0.93	0.76	0.84
Italic	0.92	0.85	0.89
Macro-Average	0.95	0.87	0.90

Table 4 shows the style classification results with the MobileNetV2 model on the TTV test set. Overall, the achieved macro-average of 90% F1-Score is a promising result, especially when compared with ABBYY FineReader, which fails almost completely for style classification. While the retrieval of regular text is nearly perfect, only three out of four bold words are retrieved and the recall of italic words is also below 90%, illustrating the difficulty of the task even when training samples of the same document collection are available. Figure 8 provides more details on the misclassifications. There is no confusion between bold and italic but both means of typographic emphasis are frequently confused with regular text.

**Figure 8:** Style confusion matrix for MobileNetV2 on the TTV test set.

Finally, the transfer of style classification to unseen document collections is assessed in the CV5 cross-validation setup. Table 5 shows the macro-average of the F1-score for MobileNetV2 and Xception, as well as for the class balancing and class weight strategies when applied to MobileNetV2. In this scenario, the overall performance drops significantly, demonstrating how difficult it is to transfer knowledge on typewritten emphasis from one printed document collection to another. In four out of five cases the MobileNetV2 architecture achieves the best result. The basic strategies to cope with class imbalance are not able to improve the results.

Table 5: Style classification results on the CV5 test sets in terms of macro-average of the F1-score. The best result is highlighted in bold font for each cross-validation.

	CV1	CV2	CV3	CV4	CV5
MobileNetV2	0.80	0.77	0.76	0.71	0.60
Xception	0.93	0.70	0.74	0.70	0.59
Class Balance	0.76	0.64	0.69	0.67	0.54
Class Weights	0.71	0.59	0.74	0.69	0.57

5 CONCLUSION

The BIR database and its baseline results for word detection and style classification demonstrate the difficulty of identifying typewritten emphasis in historical prints. Given the importance of style emphasis for document analysis and understanding, we hope that our research database can contribute to the development of novel methods in this field. A promising line of research is related to building combined word localization, recognition, and style classification models. Furthermore, data augmentation, synthesis, and balancing strategies are expected to play a central role when developing more robust systems that generalize well to unseen document collections.

ACKNOWLEDGMENTS

This work benefited from funding from the Center of Excellence Jean Monnet IMAGO (École normale supérieure) and the support of Prof. Béatrice Joyeux-Prunel

REFERENCES

- Isabella di Lenardo, Raphaël Barman, Albane Descombes, and Frédéric Kaplan. 2019. Repopulating Paris: massive extraction of 4 Million addresses from city directories between 1839 and 1922. In *Digital Humanities 2019 Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO), Utrecht, The Netherlands. <https://dev.clariah.nl/files/dh2019/boa/0878.html>
- Simon Gabay, Ljudmila Petkovic, Alexandre Bartz, Matthias Gille Levenson, and Lucie Rondeau Du Noyer. 2021. Katabase: À la recherche des manuscrits vendus. In *Humanistica 2021. Humanistica*, Rennes, France. <https://hal.archives-ouvertes.fr/hal-03066108>
- Simon Gabay, Lucie Rondeau Du Noyer, Matthias Gille Levenson, Ljudmila Petkovic, and Alexandre Bartz. 2020b. Quantifying the Unknown: How many manuscripts of the marquise de Sévigné still exist?. In *Digital Humanities DH2020 (DH2020 Book of Abstracts)*. ADHO, Ottawa, Canada. <https://hal.archives-ouvertes.fr/hal-02889829>
- Simon Gabay, Lucie Rondeau Du Noyer, and Mohamed Khemakhem. 2020a. Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes. In *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica (Quaderni di Umanistica Digitale)*. Associazione per l'Informatica Umanistica e la Cultura Digitale, Milan, Italy, 113–118.
- Béatrice Joyeux-Prunel and Olivier Marcel. 2016. Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History. *Artl@s Bulletin* 4, 2 (2016), 80–104. <https://docs.lib.psu.edu/artlas/vol4/iss2/8>
- Benjamin Kiessling. 2019. Kraken - an Universal Text Recognizer for the Humanities. In *Digital Humanities 2019 Conference Abstracts* (2019-07). Alliance of Digital Humanities Organizations (ADHO),

- Utrecht, The Netherlands. <https://dev.clariah.nl/files/dh2019/boa/0673.html>
- B. Kiessling, R. Tissot, P. Stokes, and D. Stokl Ben Ezra. 2019. eScriotorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE Computer Society, Los Alamitos, CA, USA, 19–19. <https://doi.org/10.1109/ICDARW.2019.90032>
- Anguelos Nicolaou, Fouad Slimane, Volker Maergner, and Marcus Liwicki. 2014. Local Binary Patterns for Arabic Optical Font Recognition. In *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, Tours, France, 76–80. <https://doi.org/10.1109/DAS.2014.71>
- Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019a. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Paintings. *Applied Sciences* 9, 22 (2019), 4853. <https://doi.org/10.3390/app9224853>
- Christian Reul, Sebastian Göttel, Uwe Springmann, Christoph Wick, Kay-Michael Würzner, and Frank Puppe. 2019b. Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification: A Case Study on Daniel Sander's Wörterbuch der Deutschen Sprache. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATECH2019)*. Association for Computing Machinery, Brussels, Belgium, 33–38. <https://doi.org/10.1145/3322905.3322910>
- Ana Salgado and Rute Costa. 2020. O projeto Edição Digital dos Vocabulários da Academia das Ciências: o VOLP-1940. *Revista da Associação Portuguesa de Linguística* 7 (2020), 275–294. <https://doi.org/10.5281/zenodo.4453139>
- Fouad Slimane, Rolf Ingold, and Jean Hennebert. 2017. ICDAR2017 Competition on Multi-Font and Multi-Size Digitally Represented Arabic Text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. IEEE, Kyoto, Japan, 1466–1472. <https://doi.org/10.1109/ICDAR.2017.239> ISSN: 2379-2140.
- Dapeng Tao, Xu Lin, Lianwen Jin, and Xuelong Li. 2016. Principal Component 2-D Long Short-Term Memory for Font Recognition on Single Chinese Characters. *IEEE Transactions on Cybernetics* 46, 3 (mar 2016), 756–765. <https://doi.org/10.1109/TCYB.2015.2414920> Conference Name: IEEE Transactions on Cybernetics.
- Barbara Topalov, Simon Gabay, Caroline Corbières, Laurent Romary, Béatrice Joyeux-Prunel, and Lucie Rondeau du Noyer. 2021. Automating Art@s - extracting data from exhibition catalogues. In *EADH'21 Book of abstracts*. European Association for Digital Humanities (EADH), Krasnoyarsk, Russia.
- Kurban Ulbul, Gulzira Tursun, Alimjan Aysa, Donato Impedovo, Giuseppe Pirlo, and Tuergen Yibulayin. 2017. Script Identification of Multi-Script Documents: A Survey. *IEEE Access* 5 (2017), 6546–6559. <https://doi.org/10.1109/ACCESS.2017.2689159>
- Adnan Ul-Hasan, Muhammad Zeshan Afzal, Faisal Shafait, Marcus Liwicki, and Thomas M. Breuel. 2015. A sequence learning approach for multiple script identification. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Tunis, Tunisia, 1046–1050. <https://doi.org/10.1109/ICDAR.2015.7333921>
- Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *CoRR* abs/1807.02004 (2018). <http://dblp.uni-trier.de/db/journals/corr/corr1807.html#abs-1807-02004>
- Yong Zhu, Tieniu Tan, and Yunhong Wang. 2001. Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1192–1200. <https://doi.org/10.1109/34.954608> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Abdelwahab Zramdini and Rolf Ingold. 1998. Optical font recognition using typographical features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (aug 1998), 877–882. <https://doi.org/10.1109/34.709616> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

The BIR database – Identifying typographic emphasis in list-like historical documents

Anna Scius Bertrand*

name.surname@unige.ch

UniGE, Genève, Switzerland

HES-SO, Fribourg, Switzerland

EPHE-PSL

Paris, France

Simon Gabay

Ljudmila Petković

name.surname@unige.ch

UniGE

Genève, Switzerland

Juliette Janes

Caroline Corbières

Thibault Clérice

name.surname@chartes.psl.eu

ENC-PSL

Paris, France

ABSTRACT

Layout analysis and optical character recognition have become traditional tasks for processing historical prints, but are now insufficient. Additional information is found in typographic emphasis, such as bold and italic letters. They carry semantic meaning (titles, emphasis...) and also outline the structure of the page (entries, sub-parts...). Retrieving such data is therefore crucial for information extraction and automatic document structuring. In this paper, we introduce the Bold-Italic-Regular (BIR) database, which contains 285 pages of scanned, list-like historical prints that have been annotated at word level with bold and italic emphasis. Baseline results are provided for word detection and style classification using state-of-the-art deep neural network models, highlighting promising possibilities, such as near-human performance for isolated word classification, but also demonstrating limitations for the task at hand.

CCS CONCEPTS

- Applied computing → Arts and humanities; Optical character recognition.

KEYWORDS

typographic information, historical print, object detection, style classification

ACM Reference Format:

Anna Scius Bertrand, Simon Gabay, Ljudmila Petković, Juliette Janes, Caroline Corbières, and Thibault Clérice. 2020. The BIR database – Identifying typographic emphasis in list-like historical documents. In *HIP’21: 6th International Workshop on Historical Document Imaging and Processing, September 05–07, 2021, Lausanne, CH*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3423603.3424002>

1 INTRODUCTION

Thanks to the recent improvements of Handwritten Text Recognition (HTR) engines [Fischer et al. 2020; Wick et al. 2018] and the creation of user-friendly interfaces [Kiessling et al. 2019; Reul et al. 2019a], retrieving the text from an image has become an increasingly easy task. But along with the text itself, (old) prints offer additional graphic information that convey a semantic meaning that is still too often

forgotten. It is the case of the layout (*e.g.* a paragraph starts with a carriage return and, potentially, an indentation), but also of typographic emphasis (*e.g.* a title is in italic, such as a loan word). Identifying the latter is therefore crucial for digitisation tasks, may it be for traditional mining purposes (*e.g.* retrieving all the titles of works mentioned), but also automatic structuring of the text (*e.g.* encoding the logical structure thanks to physical and graphical hints found on the page).

The BIR database¹ introduced in this paper provides the research community with a benchmark dataset for developing and comparing methods for identifying **bold** and *italic* emphasis on scanned page images of list-like historical prints. In an experimental evaluation, we evaluate and compare state-of-the-art deep neural network models, including YOLOv5 [Jocher et al. 2021] for word detection and MobileNetV2 [Sandler et al. 2018] and Xception [Chollet 2017] for style classification. The obtained baseline results are promising but also demonstrate the difficulty of the task and the current limitations of the state of the art.

2 RELATED WORK

To be precise, “style” (*e.g.* italic, oblique...) or “weight” (*e.g.* bold, thin...) are attributes of possible “fonts” (*e.g.* Garamond, Arial...). When combined together we talk about a “typeface” (*e.g.* Garamond in bold vs Garamond in italic). These fonts are associated to “scripts” (*e.g.* latin for French or English, cyrillic for Russian or Serbian...). Very few researchers have addressed directly the question of typographic emphasis, but many have tackled the problem while dealing with a larger one – usually the identification of fonts or scripts, which are necessarily embodied in a typeface that can be in bold or italic.

The oldest approach to address typographic emphasis is the one of optical font recognition (OFR). Researchers have first tried to solve it using a font model base of several hundred known fonts and a multivariate Bayesian classifier [Zramdini and Ingold 1998]. Further studies have associated typographic emphasis with a texture. On the one hand, a simple weighted Euclidian distance has been used to classify fonts of synthetic data, after extracting features such as spatial frequency and orientation with multi-channel Gabor filters [Zhu et al. 2001].

¹The BIR database is available via <https://github.com/asciusb/BIR-database>.

On the other hand, local binary patterns have been used with a synthetically generated database of arabic text images [Nicolaou et al. 2014]. Most recent approaches obviously use neural networks, may they be LSTM [Tao et al. 2016] or CNN [Slimane et al. 2017].

In the past few years, in parallel to OFR, script recognition has benefited from numerous studies [Ubul et al. 2017], some of which may well be useful for the identification of typographic emphasis. It is particularly the case of experiments using the neural network implementation of OCR engines to identify multiple scripts at text-line level [Ul-Hasan et al. 2015], a strategy used successfully for recognizing semantico-typographic classes (dictionary entry, antiqua, fraktur, author's name and letter-spacing) in a real historical use case: Daniel Sander's *Wörterbuch der Deutschen Sprache* [Reul et al. 2019b].

3 THE BIR DATABASE

In the following, we comment on the general context of list-like historical prints, describe the specific document collection that constitutes the Bold-Italic-Regular (BIR) database, and provide details on how the scanned pages were annotated.

3.1 List-Like Historical Prints

During the past decade, list-like documents have drawn considerable attention from digital humanists, may they be art historians [Joyeux-Prunel and Marcel 2016], historians [di Lenardo et al. 2019], linguists [Salgado and Costa 2020], philologists [Gabay et al. 2020a]... Such documents have the particularity to use typographic emphasis to organise a rather dense information. Italic is used to identify professions in phone directories (cf. fig. 1), examples in dictionaries (cf. fig. 2) or title of works in manuscript catalogues (cf. fig. 3). Bold tends to have a more unified usage, and delimits the subject of the entry. Additional styles, such as small capitals, can also be found, as shown in the first example of the *Annuaire-Almanach*.

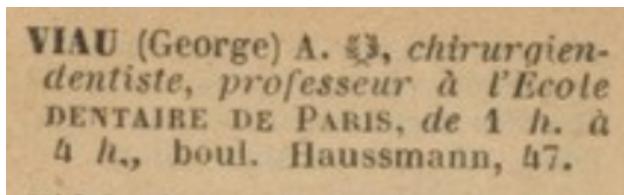


Figure 1: *Annuaire-almanach du commerce, de l'industrie, de la magistrature et de l'administration*, 1894, p. 1272, ark:/12148/bpt6k9732740w/f1498.

On top of additional mining options, typographic emphasis can be used to structure documents that we would like to encode. Bold is not only the topic of an entry, but the most efficient way to locate its beginning, and italic the only solution to locate specific passages. Without the information that some characters are in bold font, it is impossible to

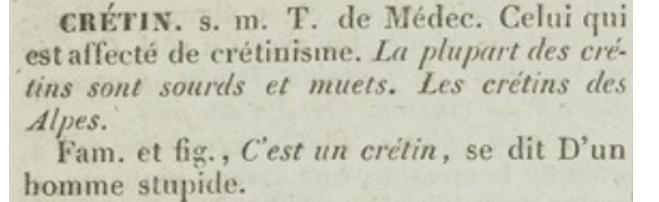


Figure 2: *Dictionnaire de l'Académie française - sixième édition*, Paris: Firmin Didot, 1835, vol. 1 (A-K.), p. 450, ark:/12148/bpt6k12804289/f490.

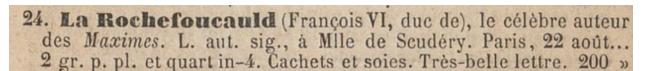


Figure 3: *Catalogues de lettres autographes, manuscrits, documents historiques, etc.*, Paris: Auguste Laverdet, N°1, avril 1856, p. 5, ark:/12148/bpt6k9687751c/f17.

differentiate the first line (24. *La Rochefoucauld*) from the third (2 gr...) in the fig. 3. Similarly, it is impossible to distinguish the definition from the examples in the fig. 2, or the biographical sub-part (in green) from the philological one (in blue) in the fig. 4, and reconstitute with precision the entry in XML for further exploration [Gabay et al. 2020a].

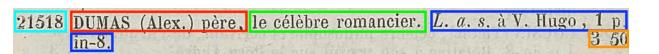


Figure 4: *Librairie autographe Ancienne et autographe Paris: Jacques Charavay, N°156, avril-mai 1867*, p. 6

```
<item n="21518" xml:id="CAT_001_e21518">
  <num type="lot">21518</num>
  <name type="author">Dumas (Alex.) père,</name>
  <trait>
    <p>le célèbre romancier</p>
  </trait>
  <desc xml:id="CAT_001_e21518_d1">L.a.s. à V. Hugo,
  1 p., in-8.</desc>
  <measure commodity="currency" unit="FRF">3 50</measure>
</item>
```

Figure 5: TEI modelling of fig. 4.

3.2 Document Collection

The BIR database contains a subset of 285 pages come from exhibition [Topalov et al. 2021] and sale catalogues [Gabay et al. 2021], for which we have strong evidences that using bold and italic information increase the precision of the extraction and the structuring [Gabay et al. 2020b]. Carefully selected excerpts have been taken in the documentation of the *Art@S* and the *Katabase* projects, i.e. mainly 19th French catalogues, which remains our primary target. Other types of

documents (*e.g.* a 19th Latin lexicon) or in other languages than French (*e.g.* the São Art Paulo Biennale) have been added to diversify the data, as well as more recent catalogues (20th and 21th c.). Tab. 1 lists the number of pages considered for the BIR database, according to different categories.

19 th prints	220	Exhibition catalogues	143
20 th prints	67	Manuscript catalogues	111
French documents	258	Other documents	33
Foreign documents	29		

Table 1: Number of pages according to different categories

3.3 Ground Truth Annotation

The words of the original database were segmented and classified according their style with the ABBYY FineReader software [Shapenko et al. 2018]. The word segmentation was excellent but the style classification was far less successful. Each sentence was manually annotated with HTML tags to identify bold and italic words and the transcription was improved.

To conduct our experiments it was necessary to align the style information contained in the sentences with the locations of each word. Due to the manual correction of the transcription at the sentence level, the number of words contained in the sentence regularly differed from the number of words in the ABBYY FineReader output.

To solve this problem, a string edit distance algorithm was used to find the closest matching words, in order to correctly transfer the typographic emphasis. Finally, the ground truth file was manually inspected and corrected by adding, merging, and splitting word bounding boxes, and correcting the style information. Pages containing no text and those requiring too much manual editing were excluded. Table 2 lists some basic statistics of the resulting BIR database.

Documents	35
Pages	285
Words	88,019
- bold	2,106
- italic	5,745
- regular	80,168

Table 2: BIR database.

4 BASELINE RESULTS

Several experiments have been conducted on the BIR database to establish baseline results with methods from the current state of the art. Two tasks are considered, word detection and style classification.

4.1 Word Detection

Word detection aims at localizing words on a scanned page. We perform this task with a fully-convolutional object detection network, which takes a whole page image as input and provides bounding boxes around the detected words as output.

The YOLO [Redmon et al. 2016] (You Only Look Once) model is considered, which has been a pioneering architecture for one-stage object detection and remained competitive over the years both in terms of speed and accuracy. The architecture consists of a convolutional *backbone*, followed by a feature pyramid *neck* that combines the extracted features at different scales, which are then processed by the *head* that computes two loss functions, one for bounding box regression and one for bounding box classification. In our case, three style classes are considered: bold, italic, and regular. Experiments are performed with the PyTorch-based YOLOv5 [Jocher et al. 2021] version using the medium-sized YOLOv5m model with 21.4 million parameters pre-trained on the COCO [Lin et al. 2014] database.

Image preprocessing includes downscaling to a height of 1024 pixels, keeping the same aspect ratio, in order to fit the input images into the GPU memory. Furthermore, random scaling, translation, and rotation operations are applied during training to augment the number of training samples and improve the generalizability of the model.

4.2 Style Classification

Style classification aims at determining the style of an individual word image that has already been localized on the scanned page. The BIR database distinguishes bold, italic, and regular words.

Two fully-convolutional network architectures are used for establishing baseline results, namely MobileNetV2 [Sandler et al. 2018] and Xception [Chollet 2017]. They were chosen with respect to their excellent classification performance on ImageNet [Deng et al. 2009] and to include both a smaller (MobileNetV2: 3.5 million parameters) and a larger (Xception: 22.9 million parameters) architecture. Experiments are performed with a Keras-based implementation and ImageNet pre-trained parameters. We drop the top layer, perform global average pooling to reduce the spatial dimensions, add a dense layer with 100 neurons, dropout, and rectified linear unit (ReLU) activation, and finish with a softmax classification layer. The architectures are illustrated in Figure 6.

Image preprocessing includes resizing to (160, 160) pixels and normalizing the RGB inputs to (-1,1) to match the pre-training condition.

As shown in Table 2, there is a significant class imbalance among the three classes, with *regular* being the majority class and *bold* and *italic* being the minority classes. We consider two strategies to alleviate this class imbalance:

- Balancing via over-sampling and under-sampling. We define a fixed amount of n training samples for each

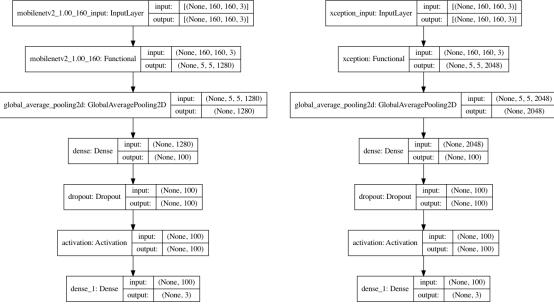


Figure 6: Network architecture for style classification using MobileNetV2 and Xception, respectively.

class, under-sample the majority class, randomly selecting n samples, and over-sample the minority classes, repeating the samples until n is reached.

- Using class weights for the loss function. We weight the loss of the minority classes with factor $\frac{m}{m'}$ where m is the number of majority samples and m' is the number of minority samples, in order to give more emphasis to the minority samples during training.

4.3 Experimental Setup

Two datasets splits are defined for experimental evaluation.

- **TVT.** In this standard setup, the 285 pages are randomly distributed into three distinct sets for training the neural networks (50%), validation of meta-parameters (25%), and testing of the final system performance (25%).
- **CV5.** In this cross-validation setup, the 35 documents are split into five distinct parts, each containing 7 documents. Five cross-validations are performed using three parts for training, one part for validation, and one part for testing, thus allowing to test the generalization capability of a trained network to an unseen document.

For word detection, the TVT setup is considered, training the YOLOv5m network with its default fine-tuning hyper-parameters 100 epochs on the training set until convergence. This base model is then further fine-tuned on an augmented dataset, randomly scaling, translating, and rotating each page hundred times. The augmented model is trained 20 epochs until convergence, which takes about 1 hour on two Titan RTX cards.

The detection performance is evaluated with respect to precision ($\# \text{correct detections} / \# \text{detections}$), recall ($\# \text{correct detections} / \# \text{words}$), and F1-score (harmonic mean of precision and recall), considering a detection as correct if its intersection over union (IoU) with the ground truth box is larger than 50%.

For style classification, an initial experiment is conducted with the MobileNetV2 for the TVT setup. Afterwards, more detailed experiments are conducted for the CV5 setup, comparing MobileNetV2 with the larger Xception architecture and the two strategies for coping with class imbalance, *class*

Table 3: Word detection results on the TVT test set.

	Precision	Recall	F1-Score
YOLOv5m	0.93	0.90	0.91

balancing and *class weights* respectively. The networks are trained with categorical cross-entropy loss and an adaptive Adam learning rate 50 epochs until convergence, which takes about 50 minutes on a Titan RTX card.

The best model is selected with respect to its accuracy on the validation set and then evaluated on the test set. Precision, recall, and F1-score are calculated for each style individually (bold, italic, regular), as well as their macro-average, i.e. the average of the evaluation measures without weighting with the number of samples from each class.

4.4 Results

Table 3 presents the detection performance achieved with the YOLO model on the TVT test set and Figure 7 illustrates some exemplary detection results. The object detection network is able to retrieve words with recall and precision over 90%, which demonstrates the feasibility of the approach but clearly leaves room for improvements. Typical errors include missing words (document in the middle) and errors due to ink bleed-through (document on the right).

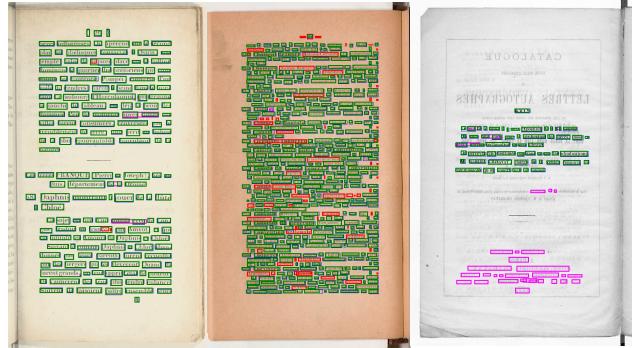


Figure 7: Word detection using YOLOv5m. Ground truth is marked in blue, detected words in green, missing words in red, and insertion errors in magenta.

Table 4 shows the style classification results with the MobileNetV2 model on the TVT test set. Overall, the achieved macro-average of 90% F1-Score is a promising result, especially when compared with ABBYY FineReader, which fails almost completely for style classification. While the retrieval of regular text is nearly perfect, only three out of four bold words are retrieved and the recall of italic words is also below 90%, illustrating the difficulty of the task even when training samples of the same document collection are available. Table 5 provides more details on the misclassifications. There

Table 4: Style classification results with MobileNetV2 on the TTV test set.

	Precision	Recall	F1-Score
Regular	0.98	0.99	0.99
Bold	0.93	0.76	0.84
Italic	0.92	0.85	0.89
Macro-Average	0.95	0.87	0.90

Table 5: Style confusion matrix for MobileNetV2 on the TTV test set.

	Regular	Bold	Italic
Regular	19,435	35	75
Bold	143	459	0
Italic	160	0	909

Table 6: Style classification results on the CV5 test sets in terms of macro-average of the F1-score. The best result is highlighted in bold font for each cross-validation.

	CV1	CV2	CV3	CV4	CV5
MobileNetV2	0.80	0.77	0.76	0.71	0.60
Xception	0.93	0.70	0.74	0.70	0.59
Class Balance	0.76	0.64	0.69	0.67	0.54
Class Weights	0.71	0.59	0.74	0.69	0.57

is no confusion between bold and italic but both means of typographic emphasis are frequently confused with regular text.

Finally, the transfer of style classification to unseen document collections is assessed in the CV5 cross-validation setup. Table 6 shows the macro-average of the F1-score for MobileNetV2 and Xception, as well as for the class balancing and class weight strategies when applied to MobileNetV2. In this scenario, the overall performance drops significantly, demonstrating how difficult it is to transfer knowledge on typewritten emphasis from one printed document collection to another. In four out of five cases the MobileNetV2 architecture achieves the best result. The basic strategies to cope with class imbalance are not able to improve the results.

It is also interesting to notice the relatively large variance of the results among the five cross-validations, highlighting the fact that each document collection has its specific properties and challenges.

4.5 Comparison with Human Performance

In a final experiment, we have investigated how the performance of the automatic system compares with human performance if the human is only presented with individual, cropped out word images without their context. For this purpose, we have randomly selected 1,000 words from the test set

Table 7: Style classification results on the 1,000 words test set in terms of F1-score.

	Regular	Bold	Italic
MobileNetV2	0.82	0.86	0.92
Human expert	0.85	0.87	0.95

Table 8: Style confusion matrix for MobileNetV2 on the 1,000 words test set.

	Regular	Bold	Italic
Regular	323	0	2
Bold	103	322	0
Italic	35	0	215

Table 9: Style confusion matrix for the human expert on the 1,000 words test set (excluding 27 non-annotated samples).

	Regular	Bold	Italic
Regular	303	9	4
Bold	93	331	0
Italic	17	0	216

of the TTV setup, according to a non-uniform distribution: 325 regular words, 425 bold words, and 250 italic words. A human expert, who has created ground truth at page level, was asked to classify the individual word images by putting them into three distinct folders. The human was made aware of the fact that the word distribution is not uniform, without providing any hints about their real distribution, to avoid a bias when distributing the words into the different folders.

To our surprise, the human expert refused to classify 27 images because they were near-empty or of low quality. These samples have been excluded from the evaluation of the human performance.

Tables 7-9 demonstrate that the automatic system reaches near-human performance for isolated word classification. While it is possible that neural networks may outperform humans for this task in the future, it also highlights that taking contextual information from the whole document page into account is expected to be a key element to improve the performance in the future.

5 CONCLUSION

The BIR database and its baseline results for word detection and style classification demonstrate the difficulty of identifying typewritten emphasis in historical prints. Given the importance of style emphasis for document analysis and

understanding, we hope that our research database can contribute to the development of novel methods in this field. A promising line of research is related to building combined word localization, recognition, and style classification models. Furthermore, data augmentation, synthesis, and balancing strategies are expected to play a central role when developing more robust systems that generalize well to unseen document collections. Going beyond isolated word classification by including more contextual information is expected to be of fundamental importance for improving the performance.

ACKNOWLEDGMENTS

This work benefited from funding from the Center of Excellence Jean Monnet IMAGO (École normale supérieure) and the support of Prof. Béatrice Joyeux-Prunel.

REFERENCES

- François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- Isabella di Lenardo, Raphaël Barman, Albane Descombes, and Frédéric Kaplan. 2019. Repopulating Paris: massive extraction of 4 Million addresses from city directories between 1839 and 1922. In *Digital Humanities 2019 Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO), Utrecht, The Netherlands. <https://dev.clariah.nl/files/dh2019/boa/0878.html>
- Andreas Fischer, Marcus Liwicki, and Rolf Ingold (Eds.). 2020. *Handwritten Historical Document Analysis, Recognition, and Retrieval — State of the Art and Future Trends*. World Scientific.
- Simon Gabay, Ljudmila Petkovic, Alexandre Bartz, Matthias Gilles Levenson, and Lucie Rondeau Du Noyer. 2021. Katabase: À la recherche des manuscrits vendus. In *Humanistica 2021*. Humanistica, Rennes, France. <https://hal.archives-ouvertes.fr/hal-03066108>
- Simon Gabay, Lucie Rondeau Du Noyer, Matthias Gilles Levenson, Ljudmila Petkovic, and Alexandre Bartz. 2020b. Quantifying the Unknown: How many manuscripts of the marquise de Sévigné still exist?. In *Digital Humanities DH2020 (DH2020 Book of Abstracts)*. ADHO, Ottawa, Canada. <https://hal.archives-ouvertes.fr/hal-02898929>
- Simon Gabay, Lucie Rondeau Du Noyer, and Mohamed Khemakhem. 2020a. Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographies. In *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica (Quaderni di Umanistica Digitale)*. Associazione per l'Informatica Umanistica e la Cultura Digitale, Milan, Italy, 113–118.
- Glenn Jocher et al. 2021. ultralytics/yolov5: v4.0 - nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration. <https://doi.org/10.5281/ZENODO.4418161>
- Béatrice Joyeux-Prunel and Olivier Marcel. 2016. Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History. *Artl@s Bulletin* 4, 2 (2016), 80–104. <https://docs.lib.psu.edu/artlas/vol4/iss2/8>
- B. Kiessling, R. Tissot, P. Stokes, and D. Stokl Ben Ezra. 2019. eScriptriorum: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE Computer Society, Los Alamitos, CA, USA, 19–19. <https://doi.org/10.1109/ICDARW.2019.90032>
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. 13th European Conf. on Computer Vision (ECCV)*. 740–755.
- Anguelos Nicolaou, Fouad Slimane, Volker Maergner, and Marcus Liwicki. 2014. Local Binary Patterns for Arabic Optical Font Recognition. In *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, Tours, France, 76–80. <https://doi.org/10.1109/DAS.2014.71>
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 779–788.
- Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019a. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences* 9, 22 (2019), 4853. <https://doi.org/10.3390/app9224853>
- Christian Reul, Sebastian Göttel, Uwe Springmann, Christoph Wick, Kay-Michael Würzner, and Frank Puppe. 2019b. Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification: A Case Study on Daniel Sander's Wörterbuch der Deutschen Sprache. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATECH2019)*. Association for Computing Machinery, Brussels, Belgium, 33–38. <https://doi.org/10.1145/3322905.3322910>
- Ana Salgado and Rute Costa. 2020. O projeto Edição Digital dos Vocabulários da Academia das Ciências: o VOLP-1940. *Revista da Associação Portuguesa de Linguística* 7 (2020), 275–294. <https://doi.org/10.5281/zenodo.4453139>
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.
- Andrey Shapenko, Vladimir Korovkin, and Benoit Leleux. 2018. AB-BYY: the digitization of language and text. *Emerald Emerging Markets Case Studies* (2018).
- Fouad Slimane, Rolf Ingold, and Jean Hennebert. 2017. ICDAR2017 Competition on Multi-Font and Multi-Size Digitally Represented Arabic Text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. IEEE, Kyoto, Japan, 1466–1472. <https://doi.org/10.1109/ICDAR.2017.239> ISSN: 2379-2140.
- Dapeng Tao, Xu Lin, Lianwen Jin, and Xuelong Li. 2016. Principal Component 2-D Long Short-Term Memory for Font Recognition on Single Chinese Characters. *IEEE Transactions on Cybernetics* 46, 3 (mar 2016), 756–765. <https://doi.org/10.1109/TCYB.2015.2414920> Conference Name: IEEE Transactions on Cybernetics.
- Barbara Topalov, Simon Gabay, Caroline Corbières, Laurent Romary, Béatrice Joyeux-Prunel, and Lucie Rondeau du Noyer. 2021. Automating Artl@s - extracting data from exhibition catalogues. In *EADH'21 Book of abstracts*. European Association for Digital Humanities (EADH), Krasnoyarsk, Russia.
- Kurban Ulubul, Gulzira Tursun, Alimjan Aysa, Donato Impedovo, Giuseppe Pirlo, and Tuergen Yibulayin. 2017. Script Identification of Multi-Script Documents: A Survey. *IEEE Access* 5 (2017), 6546–6559. <https://doi.org/10.1109/ACCESS.2017.2689159>
- Adnan Ul-Hasan, Muhammad Zeshan Afzal, Faisal Shahafat, Marcus Liwicki, and Thomas M. Breuel. 2015. A sequence learning approach for multiple script identification. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Tunis, Tunisia, 1046–1050. <https://doi.org/10.1109/ICDAR.2015.7333921>
- Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *CoRR* abs/1807.02004 (2018). <http://dblp.uni-trier.de/db/journals/corr/corr1807.html#abs-1807-02004>
- Yong Zhu, Tianlu Tan, and Yunhong Wang. 2001. Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1192–1200. <https://doi.org/10.1109/34.954608> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Abdelwahab Zramdini and Rolf Ingold. 1998. Optical font recognition using typographical features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (aug 1998), 877–882. <https://doi.org/10.1109/34.709616> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Towards automatic TEI encoding via layout analysis

Simon Gabay

Université de Genève (Switzerland)

Ariane Pinche, Claire Jahan, and Juliette Janes

École nationale des chartes | PSL (France)

The forefront of research on textual documents (may they be manuscripts and prints) is slowly moving from text recognition to automatic encoding. Transforming quickly images into XML-TEI documents is therefore the next important obstacle that needs to be tackled to offer enhanced mining options to digital libraries users.

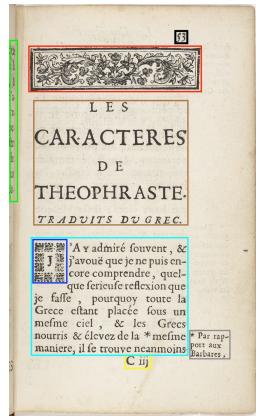


Figure 1: La Bruyère, *Les Caractères de Théophraste*, 1688, p.53. In black the page number, in red a headpiece, in brown the title, in cyan the text, in blue the drop capital, in yellow the signature, in grey a marginal note, in green noise.

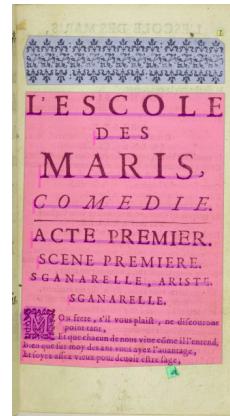


Figure 2: Example of a page with zones automatically recognised: Molière, *L'Ecole des Maris*, 1661, p.1. In yellow the page number, in purple the headpiece, in pink the text, in blue the drop capital, in green the signature.

If extensive research has been carried on complicated and highly standardised layouts such as manuscripts (Gabay, Rondeau Du Noyer, et al. 2020) or exhibition catalogues (Gabay, Topalov, et al. 2021) with dedicated tools like GROBID (Khemakhem 2020), more simple documents still await for an easily trainable and implementable solution. Most of the pages contain simple elements such as page numbers, marginal notes, signatures... (cf. fig. 1) which all have a different status than the actual body of the text, and therefore need to be disentangled from one another. For instance, finding all the mentions of Andromache in the play of the same name is a nonsense if we do not separate running titles (in which the name of the eponym heroins appear) from the dialogues.

Recent experiments show that a standardised description of the page and state-of-the-art models for layout analysis could offer a limited, yet efficient solution (cf. fig. 2) to numerous scholars for two reasons. On the one hand powerful open-source OCR engines are now available (Kiessling 2019) via user-friendly interfaces (Kiessling et al. 2019): it will become easier to train models tailored to one's need in a near future. On the other hand research institutions are now investing in OCR infrastructures (Gabay 2021; Romary et al. 2021), which will accelerate the production of data locally, without the direct support of GLAMs.

Using a common vocabulary to annotate zones (Gabay, Camps, et al. 2021), we have developed a generic workflow to analyse the layout, OCRise the text, and convert the ALTO output into minimally encoded TEI files (cf. table 1). This workflow is currently being tested on three different datasets: one of medieval manuscripts (cf. table 5), one of 17th c. literary prints (cf. table 4), and one of 19th c. catalogues (cf. table 6).

The key aspect of the workflow being zones detection, we have experimented two neural architectures for the layout analysis, and possible combinations of data to increase the efficiency. One model for each set has been trained, then two combinations: the medieval and the 17th c. data on the one hand, the 17th c. data and the 19th c. on the other hand (cf. table 2).

A second experiment as redimensioned the image taken as input in the VGSL specs (1 200→1 800)¹, which significantly improves the results for both main

Segmonto zone	TEI element
Damage	<damage>
Decoration	<figure>
DropCapital	<hi>
Figure	<figure>
Main	<p>
Margin	<note>
MusicNotation	<figure>
Numbering	<fw type="Numbering">
RunningTitle	<fw type="RunningTitle">
Seal	<figure>
Signatures	<fw type="Signatures">
Stamp	<figure>
Table	<table>
Title	<p>

Table 1: SegmOnto zones and their corresponding TEI element.

¹the architecture of the second training is therefore the following 1,1800,0,3 Cr7,7,64,2,2

	Medieval data	17 th c.	19 th c.	Combination 1	Combination 2
MIoU	0.6017	0.1965	0.2092	0.4883	0.1964
FWIoU	0.7445	0.7794	0.7157	0.7537	0.7322
MAcc	0.9805	0.9846	0.9613	0.9886	0.9834
Acc	0.9805	0.9846	0.9613	0.9886	0.9834

Table 2: Architecture 1: efficiency of complex models measured with Mean Intersection-Over-Union (MIoU), frequency-weighted intersection over union (FWIoU), Mean accuracy (MAcc) and accuracy (Acc). Combination 1 gathers Medieval data and 17th c., Combination 2 gathers 17th c. and 19th c. data.

categories (cf. table 3): intersection over union (*i.e.* zone detection) and accuracy (*i.e.* zone classification).

	Medieval data	17 th c.	19 th c.	Combination 1	Combination 2
MIoU	0.6127	0.1685	0.3662	0.5386	0.3269
FWIoU	0.7754	0.8002	0.7244	0.7845	0.7821
MAcc	0.9905	0.9886	0.9651	0.9917	0.9869
Acc	0.9905	0.9886	0.9651	0.9917	0.9869

Table 3: Architecture 2.

Lower scores for medieval data and 19th c. can easily explained by the relatively smaller size of the training set and the greater complexity of the layout. If the results obtained with combined datasets are hard to interpret without the existence of a test for layout analysis, scores do not suffer a significant drop: the combination seems to be relatively efficient.

Acknowledgements

This paper would not have been possible without the help of Thibault Clérice and Jean-Baptiste Camps (ENC), the SegmOnto technical and semantic working groups, the Kraken and eScriptorium development teams.

Gn32 Cr3,3,128,2,2 Gn32 Cr3,3,128 Gn32 Cr3,3,256 Gn32 Lbx32 Lby32 Cr1,1,32 Gn32
Lby32 Lbx32.

Appendices

Examples



Figure 3: *Manuscrit, BNF fr. 412, f°21, 13th c.*

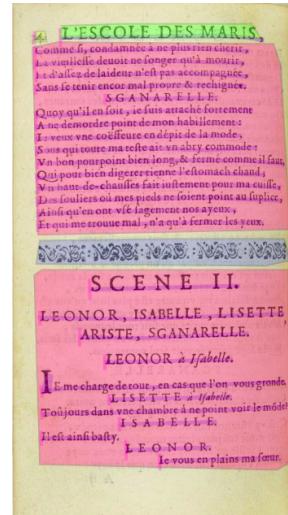


Figure 4: *Molière, L'Ecole des maris, Paris, G. de Luyne, 1661*

4417	
6. Eteinte r. Jours, Dammes (S.-et-O.).	8-10 Bourg, Paris, r. Anneau, 16.
5-7 Estete r. Dammesville, 50.	12 Milles, Varenne-Sur-Huire (Seine).
9. Estete r. Léon, 34.	14 André, Assise, r. Béon, 62 ^{me} (Seine).
11. Estete r. Léon, 39.	
13. Bercy, Paris, r. Laghouat, 24.	
14. Bercy, Paris, r. H ^{me} , 46.	
15 à 19 (Conseiller) Bérard, Paris, r. Orsay, 15.	
21. Théâtre (Vve), Paris, r. Laduyette, 121.	20-22 Béatrice, Paris, r. Béatrice, 27 (S.-et-O.).
23-25 Béatrice, Paris, r. Béatrice, 46.	23-25 Béatrice, Paris, r. Béatrice, 46.
27. Mercier, Théâtre (Sorbonne).	27. Mercier, Théâtre (Sorbonne).
2. Perret, Paris, r. Pigalle, 30.	3. Perret, Paris, r. Pigalle, 30.
4. Baudouin, Paris, r. Baudouin, 38.	5. Baudouin, Paris, r. Baudouin, 38.
6. Dubaïs, Paris, r. Galilé, 1.	7. Dubaïs, Paris, r. Galilé, 1.
8. Dubaïs, Paris, r. Galilé, 1.	9. Dubaïs, Paris, r. Galilé, 1.
10. Thomas, Perret, sille belleveu, 41 (Seine).	11. Dubaïs, Paris, r. Galilé, 1.
12. Quinson, Paris, r. Orsay, 12.	13-19 Chemins de fer de l'Est.
14. Dubaïs, Paris, r. Galilé, 1.	21-23 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
16. Durand, Paris, r. Breguet, 14.	24-26 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
18. Dubaïs, Paris, r. Galilé, 1.	27-29 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
20. Gouffier, Etiquage (Rue).	30-32 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
22. Paquet, pcr. Parmentier, Paris, r. Orsay, 22.	33-35 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
24. Paquet, pcr. Parmentier, Paris, r. Orsay, 22.	36-38 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
26. Grémard, Paris, r. Galilé, 1.	39-41 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
28. Grémard, Paris, r. Galilé, 1.	42-44 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
30. Grémard, Paris, r. Galilé, 1.	45-47 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
32. Courvoisier (M ^{me}), Paris, r. Ramey, 2.	48-50 Jules (Vve), et Berthe, Paris, r. Galilé, 1.
34. Courvoisier (M ^{me}), Paris, r. Ramey, 2.	51-53 Dubois, Paris, r. Gréville, 7.
36. Laubry, Melon, r. Ramey, 15 (S.-et-O.).	54-56 Dubois, Paris, r. Gréville, 7.
38. Brugui, Louange par Cresset (Creuse).	57-60 Chocet, Boissé St-Léger, r. Mercière, 3.
	61-64 Daillier, Paris, r. Ordener, 29.
	65-68 Daillier, Paris, r. Ordener, 29.
	69-72 Daillier, Paris, r. Ordener, 29.
	73-76 Daillier, Paris, r. Ordener, 29.
	77-80 Daillier, Paris, r. Ordener, 29.
	81-84 Daillier, Paris, r. Ordener, 29.
	85-88 Daillier, Paris, r. Ordener, 29.
	89-92 Daillier, Paris, r. Ordener, 29.
	93-96 Daillier, Paris, r. Ordener, 29.
	97-100 Daillier, Paris, r. Ordener, 29.
1. Estete r. Ravignan, 13.	101-104 Daillier, Paris, r. Ordener, 29.
3. Dubaïs, Paris, r. Ordener, 3.	105-108 Daillier, Paris, r. Ordener, 29.
3-7 Baudouin, Paris, r. Baudouin, 10.	109-112 Daillier, Paris, r. Ordener, 29.
2. Estete r. Ravignan, 15.	113-116 Daillier, Paris, r. Ordener, 29.
4. Le Chevrefeuille (M ^{me}), Paris, r. Nellié, 10.	117-120 Daillier, Paris, r. Ordener, 29.
10. Pellen, Paris, r. Brique, 2.	121-124 Daillier, Paris, r. Ordener, 29.

Figure 5: *Annuaire-almanach du commerce, de l'industrie..., 1898*

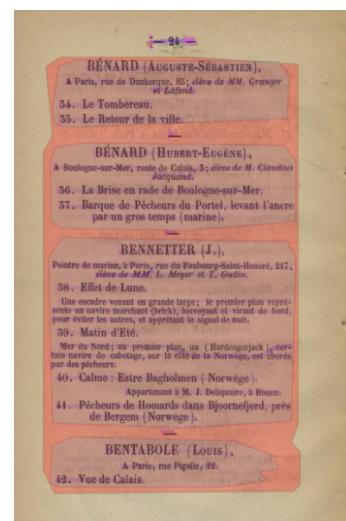


Figure 6: *Catalogue de l'exposition annuelle du musée de Rouen, 1860*

Datasets

Prints' title and author	Date	No. of pages	Prints' specificities
<i>Lettres du Sieur Balzac</i> , Jean-Louis Balzac	bpt1b8629420	1624	None
<i>Méduse</i> , Claude Boyer	bpt6k311844g	1697	No Damage zone
<i>Les caractères de Théophraste</i> , Jean de la Bruyère	bpt1b86070385	1688	No Damage zone
<i>Histoire amoureuse des Gaules</i> , Bussy-Rabutin	bpt1b8623309s	1665	No RunningTitle zone
<i>Le théâtre de P. Corneille</i> , Pierre Corneille	bpt6k10403751	1664	None
<i>Discours de la méthode</i> , René Descartes	bpt1b86069594	1637	None
<i>La princesse de Clèves</i> , Madame de La Fayette	bpt1b8610820b	1678	No Margin zone
<i>La Mariane</i> , Tristan L'Hermite	bpt6k1511072f	1639	No Title zone
<i>L'Ecole des femmes</i> , Molière	bpt1b8610785b	1663	No Damage and Margin zones
<i>George Dandin, ou le Mary confondu</i> , Molière	bpt1b8610793w	1669	No Margin zone
<i>Dom Garcie de Navarre</i> , Molière	+2258398909_2	1694	None
<i>Statira</i> , Nicolas Pradon	bpt6k8416272	1680	No Rubric line
<i>Athènaïs</i> , Nicolas Pradon	bpt6k857200c	1697	None
<i>Les plaideurs</i> , Jean Racine	bpt1b8610811c	1669	No Rubric line
<i>Oeuvres de Racine. Tome Premier</i> , Jean Racine	bpt6k8905809	1676	No Damage zone
<i>Oeuvres de Racine. Tome Second</i> , Jean Racine	bpt6k90581p	1676	None

Table 4: 17th century prints dataset description (always with zones Main, Decoration, DropCapital, Numbering, RunningTitle, Signatures, Stamp and always with lines Default and DropCapitalLine).

57

Manuscript ID	Date	No. of pages	No. of columns	Running Title	Drop Capital
Bnf, Arsenal 3516	13th	10	4	No	Yes
Bnf, ms fr. 22549	14th	3	3	No	Yes
Bnf, ms fr. 24428	13th	20	2	Yes	Yes
Bnf, ms fr. 412	13th	49	2	Yes	Yes
Bnf, ms fr. 844	13th	18	2	No	Yes
Cologny, bodmer, 168	13th	22	2	No	Yes
Vaticane, Reg. Lat., 1616	14th	41	1	No	Yes

Table 5: Medieval dataset description

Prints' title	Type	Date	No. of pages	No. of columns	
<i>Annuaire-almanach du commerce, de l'industrie...</i>	Annuaire	1898	50	2	Other zones
<i>Exposition des œuvres de M. Courbet à l'École des Beaux Arts</i>	Exhibition	1882	19	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exhibition	1892	5	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exhibition	1913	7	1	Numbering
<i>Catalogue des œuvres exposées, Société des Indépendants</i>	Exhibition	1923	5	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exhibition	1818	2	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exhibition	1867	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exhibition	1843	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exhibition	1849	5	1	Numbering
<i>Catalogue de l'exposition des Beaux-Arts de Nancy</i>	Exhibition	1892	5	1	Numbering
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1961	5	1	Running Title
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1965	4	1	Numbering Running Title
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1969	5	1	Numbering Running Title
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1853	5	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1869	7	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1888	7	1	Numbering
<i>Catalogue de la Biennale de Sao Paulo</i>	Exhibition	1951	6	1	Numbering
<i>Catalogue de la Biennale de Sao Paulo</i>	Exhibition	1972	5	1	Numbering
<i>Catalogue de l'exposition de la société des amis des arts de Strasbourg</i>	Exhibition	1884	15	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1895	5	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1905	3	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1920	5	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1870	5	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1871	6	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1873	4	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1877	4	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1880	2	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1881	2	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1883	5	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1885	2	1	Numbering
<i>Catalogue de ventes de manuscrits Charavay</i>	Manuscripts	1845	6	1	Numbering
<i>Catalogue de ventes de manuscrits Lavergne</i>	Manuscripts	1856	4	1	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1857	6	1	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1866	7	1	Numbering
<i>Catalogue de vente de manuscrits Boret</i>	Manuscripts	1887	14	1	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1857	7	1	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1899	6	1	Numbering
<i>Catalogue de vente de manuscrits Kra</i>	Manuscripts	1912	9	2	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1919	8	1	Numbering

Table 6: Catalogs dataset description

References

- Gabay, S. (2021). *FOrmes Numérisées et Détection Unifiée des Écritures*. <https://www.unige.ch/lettres/humanites-numeriques/index.php/?cID=188>.
- Gabay, S., J.-B. Camps, et al. (Sept. 2021). “SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)”. In: *Proceedings of the 1st International Workshop on Computational Paleography, IWCP@ICDAR 2021*. 1st International Workshop on Computational Paleography IWCP. Lecture Notes in Computer Science. Lausanne (Switzerland): Springer.
- Gabay, S., L. Rondeau Du Noyer, et al. (July 2020). “Quantifying the Unknown: How many manuscripts of the marquise de Sévigné still exist?” In: *Digital Humanities DH2020*. DH2020 Book of Abstracts. Ottawa, Canada: ADHO. URL: <https://hal.archives-ouvertes.fr/hal-02898929> (visited on 11/23/2020).
- Gabay, S., B. Topalov, et al. (Sept. 2021). “Automating Artl@s - extracting data from exhibition catalogues”. In: *Digital Humanities DH2020*. EADH2021 Book of Abstracts. Krasnoyarsk, Russia: ADHO. URL: REPLACE.
- Khemakhem, M. (2020). “Standard-based Lexical Models for Automatically Structured Dictionaries”. PhD thesis. Paris: INRIA.
- Kiessling, B. (July 2019). “Kraken - an Universal Text Recognizer for the Humanities”. In: *Digital Humanities 2019 Conference Abstracts*. Digital Humanities 2019 Conference. Utrecht, The Netherlands: Alliance of Digital Humanities Organizations (ADHO). URL: <https://dev.clariah.nl/files/dh2019/boa/0673.html> (visited on 03/16/2020).
- Kiessling, B. et al. (Sept. 2019). “eScriptorium: An Open Source Platform for Historical Document Analysis”. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). Vol. 2, pp. 19–19. DOI: 10.1109/ICDARW.2019.10032.
- Romary, L. et al. (2021). *CREMMA - Consortium pour la Reconnaissance d'Écriture Manuscrite des Matériaux Anciens*. <https://www.dim-map.fr/projets-soutenus/cremma/>.

Expanding the content model of annotationBlock

Alexandre Bartz¹, Simon Gabay², Juliette Janes³, and Laurent Romary⁴

¹Sorbonne Université (France)

²Université de Genève (Switzerland)

³PSL - Ecole nationale des chartes (France)

⁴INRIA - Almanach (France)

Linguistic annotation benefits from ISO specifications such as the Morphosyntactic Annotation Framework (MAF), whose recommendations have been added to the TEI P5¹ (ISO-24611 2012; Stührenberg 2012). Relying on feature structures (cf. ex. 1), these recommendations have however not been fully integrated to TEI stand off annotation (Bański et al. 2016) and it is currently impossible to encode feature structures within the `listAnnotation` and `annotationBlock` elements.

```
<seg>
  <w xml:id="s1w1">lōg</w>
  <w xml:id="s1w2">tems</w>
</seg>
<spanGrp type="wordForm">
  <span target="#s1w1" ana="#s1fs1"/>
  <span target="#s1w2" ana="#s1fs2"/>
</spanGrp>
<fs xml:id="s1fs1">
  <f name="lemma">
    <string>long</string>
  </f>
</fs>
<fs xml:id="s1fs2">
  <f name="lemma">
    <string>temps</string>
  </f>
</fs>
```

Example 1: TEI encoding following the MAF

¹<https://tei-c.org/release//doc/tei-p5-doc/en/html/FS.html>.

With the multiplication of annotation tools, the case is becoming more complex and is not limited to reference annotation (lemma, POS...) using the `fs` element anymore. For instance, normalisation tasks are becoming more and more common for medievalists (Stutzmann 2011) or modernists (Gabay and Barrault 2020) to encode various levels of transcription (*lōg tems* → *long tems*) or offer a fully version fully aligned with contemporary spelling (*lōg tems* → *longtemps*). New elements, such as `reg`, would therefore be extremely useful (cf. ex. 2).

```

<seg corresp="#s1">
  <w xml:id="s1w1">lōg</w>
  <w xml:id="s1w2">tems</w>
</seg>
[...]
<spanGrp type="wordForm">
  <span target="#s1w1" ana="#s1reg1"/>
  <span target="#s1w2" ana="#s1reg2"/>
</spanGrp>
<reg type="reg" xml:id="s1reg1">long</reg>
<reg type="reg" xml:id="s1reg2">tems</reg>
[...]
<spanGrp type="wordForm">
  <span target="#s1w1 #s1w2" ana="#s1norm1"/>
</spanGrp>
<reg type="norm" xml:id="s1norm1">longtemps</reg>

```

Example 2: Annotation of linguistic normalisation

Sadly, in case of multiple levels of embedded stand off data, the current version of the guidelines promotes a multiplication of `standOff` elements with a semantically unappropriated `seg` to store the necessary annotation (cf. ex. 3).

```

<standOff type="linguistic">
  <seg corresp="#s1">
    <spanGrp>[...]</spanGrp>
    <fs>[...]</fs>
  </seg>
</standOff>
<standOff type="norm">
  <seg corresp="#s1">
    <spanGrp>[...]</spanGrp>
    <reg>[...]</reg>
  </seg>
</standOff>

```

Example 3: Annotation of linguistic normalisation

Using the data of the *E-ditiones* project, we will make a case for a more appropriate use of `standOff` with one `listAnnotation` per annotation type and an extended version of the `annotationBlock` content model (cf. ex. 4). This latter, also an ISO recommendation (ISO-24624 2016), has originally been created to identify the reference features needed for transcribing spoken resources that are anchored on a single reference timeline, but also for integrating mechanisms to encompass most usual transcription conventions. It therefore needs to include not only the `fs` element for reference annotation, but also additional information about editorial transcription such as normalisation (`reg`) or any other philological intervention on the text (cf. the `model.pPart.transcriptional` class: `add`, `corr`, `damage`, `del`, `handShift`, `mod`, `orig`, `redo`, `reg`, `restore`, `retrace`, `secl`, `sic`, `supplied`, `surplus`, `unclear`, `undo`).

Data

Scripts and codes are available at <https://github.com/e-ditiones/Annotator>.

```

<TEI>
  <text>
    <body>
      <p>
        <seg xml:id="s1">
          <w xml:id="s1w1">lög</w>
          <w xml:id="s1w2">tems</w>
          <w xml:id="s1w3">a</w>
          <w xml:id="s1w4">geneve</w>
        </seg>
      </p>
    </body>
  </text>
<standOff>
  <listAnnotation type="linguistic">
    <annotationBlock corresp="#s1">
      <spanGrp type="wordForm">
        <span target="#s1w1" ana="#s1ing1"/>
        <span target="#s1w2" ana="#s1ing2"/>
        <span target="#s1w3" ana="#s1ing3"/>
        <span target="#s1w4" ana="#s1ing4"/>
      </spanGrp>
      <fs xml:id="s1ing1">
        <f name="lemma">
          <string>long</string>
        </f>
        <f name="pos">
          <symbol value="ADJqua"/>
        </f>
        <f name="nomb">
          <symbol value="s"/>
        </f>
        <f name="genre">
          <symbol value="m"/>
        </f>
        <f name="norm1">
          <string>long</string>
        </f>
      </fs>
      <fs xml:id="s1ing2">
        <f name="lemma">
          <string>tems</string>
        </f>
        <f name="pos">
          <symbol value="NOMcom"/>
        </f>
        <f name="nomb">
          <symbol value="s"/>
        </f>
        <f name="genre">
          <symbol value="m"/>
        </f>
      </fs>
    </annotationBlock>
  </listAnnotation>

```

```

        <f name="norm1">
            <string>temps</string>
        </f>
    </fs>
    <fs xml:id="s1ing3">
        <f name="lemma">
            <string>à</string>
        </f>
        <f name="pos">
            <symbol value="PRE"/>
        </f>
        <f name="norm1">
            <string>à</string>
        </f>
    </fs>
    <fs xml:id="s1ing4">
        <f name="lemma">
            <string>geneve</string>
        </f>
        <f name="pos">
            <symbol value="NOMpro"/>
        </f>
        <f name="norm1">
            <string>Genève</string>
        </f>
    </fs>
</annotationBlock>
</listAnnotation>
<listAnnotation type="normalisation">
    <annotationBlock corresp="#s1">
        <spanGrp type="formNorm">
            <span target="#s1w1 #s1w2" corresp="#s1norm1"/>
            <span target="#s1w2" corresp="#s1norm2"/>
            <span target="#s1w2" corresp="#s1norm3"/>
        </spanGrp>
        <reg xml:id="sinorm1">longtemps</reg>
        <reg xml:id="sinorm2">à</reg>
        <reg xml:id="sinorm3">Genève</reg>
    </annotationBlock>
</listAnnotation>
<listAnnotation type="NERD">
    <annotationBlock corresp="#s1">
        <spanGrp type="formNorm">
            <span target="#s1w4" ana="#s1ner1"/>
        </spanGrp>
        <fs xml:id="s1ner1">
            <f name="NER">
                <symbol value="B-loc.adm.town"/>
            </f>
            <f name="wikidata">
                <symbol value="Q71"/>
            </f>
        </fs>
    </annotationBlock>
</listAnnotation>
</standOff>
</TEI>
```

Example 4: TEI encoding following the MAF

References

- Bański, P. et al.** (2016). “Wake up, standOff!” In: *TEI Abstracts 2016*. TEI Conference 2016. Vienna, Austria. URL: <https://tei2016app.acdh.oeaw.ac.at/pages/show.html?document=banskiigaiffelopezmeoniromaryschmidtstadlerwitt.xml> (visited on 07/22/2021).
- Gabay, S. and L. Barrault** (June 2020). “Traduction automatique pour la normalisation du français du XVII e siècle”. In: *27ème Conférence sur le Traitement Automatique des Langues Naturelles*. TALN2020. Nancy, France: ATALA. URL: <https://hal.archives-ouvertes.fr/hal-02596669> (visited on 12/10/2020).
- ISO-24611** (2012). *Language Resource management — Transcription of Spoken Language — ISO 24611*. ISO.
- ISO-24624** (2016). *Language resource management — Morpho-syntactic annotation framework (MAF) — ISO 24624*. ISO.
- Stührenberg, M.** (Nov. 5, 2012). “The TEI and Current Standards for Structuring Linguistic Data”. In: *Journal of the Text Encoding Initiative* (Issue 3). ISSN: 2162-5603. DOI: 10.4000/jtei.523. URL: <https://journals.openedition.org/jtei/523> (visited on 07/22/2021).
- Stutzmann, D.** (2011). “Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?” In: *Kodikologie und Paläographie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*. Schriften des Instituts für Dokumentologie und Editorik 2. Ed. by **F. Fischer, C. Fritze, and G. Vogeler**, pp. 247–277. URL: <https://halshs.archives-ouvertes.fr/halshs-00596970> (visited on 01/09/2020).

C.2 Rapports

Les pages suivantes contiennent des rapports réalisés dans le cadre de mon stage. Ils permettent d'expliquer certains points, par le biais de manuels d'annotation ou de description.

Campagne d'annotation typographique

Juliette Janes

03-05-2021

1 Présentation du projet

Ce guide d'annotation a été réalisé dans le cadre de la création d'un dataset pour récupérer l'information typographique sur des imprimés du XIXème siècle. L'outil utilisé pour ce faire est un visualisateur d'images, <https://www.robots.ox.ac.uk/~vgg/software/via/via.html>, développé par le département des sciences de l'ingénieur d'Oxford. Il permet de charger une image, l'annoter et indiquer si il s'agit de gras, italique, gras-italique ou aucun, puis récupérer en sortie le csv correspondant. Ainsi, chaque information typographique est tokenisée et signalée. Le dataset ainsi obtenu sera renvoyé à Anna Scius-Bertrand afin d'entraîner un réseau de neurones pour la reconnaissance des typographies.

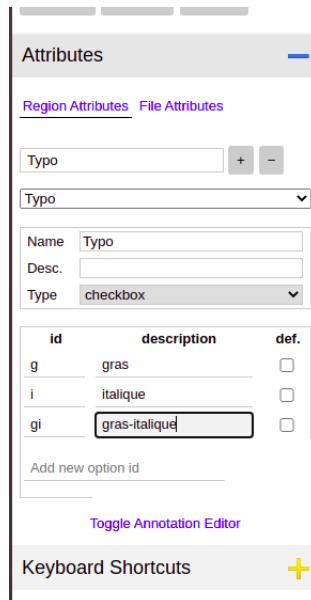


FIGURE 1 – Paramètres d'annotation à entrer dans le visualisateur d'images

2 Principes généraux pour l'annotation

L'idée est donc de tokeniser des passages des catalogues d'exposition et de ventes de manuscrits et d'y ajouter l'information typographique. Ainsi, chaque token devrait être bien encadré par des espaces de chaque côté.

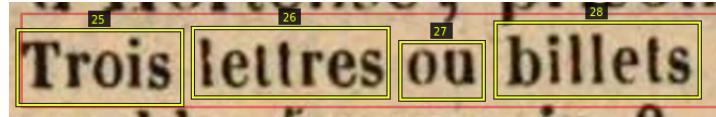


FIGURE 2 – Catalogue de vente de manuscrits, Charavay, 1843, p.18

2.1 Ponctuation

Lorsque la ponctuation est collée au mot, on l'intègre dans son token. Dans le cas où la ponctuation est encadrée par deux espaces, celle-ci est considérée comme un seul token.

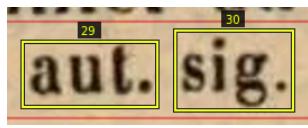


FIGURE 3 – Catalogue de vente de manuscrits, Charavay, 1843, p.18



FIGURE 4 – Manuel de synonymie, Doederlin, 1865, p.13

Dans le cas précis des tirets, lorsque ceux-ci lient deux mots entre eux, les deux mots sont considérés comme un seul token.

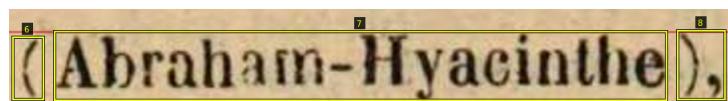


FIGURE 5 – Catalogue de vente de manuscrits, Charavay, 1843, p.14

2.2 Apostrophes

Un mot associé à une apostrophe, n'ayant donc pas d'espace les séparant, doit être considéré comme un seul token.



FIGURE 6 – Catalogue de vente de manuscrits, Charavay, 1843, p.14

2.3 Chiffres

Dans le cas des chiffres, souvent signalant des prix dans le cas des catalogues, ils sont considérés comme un seul token, même dans le cas d'un petit espace entre ceux-ci.

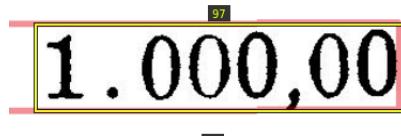


FIGURE 7 – Catalogue d'exposition, Sao Paulo, 1972, p.3

3 Ambiguïtés fréquentes

3.1 Typographies différentes dans un même token

Dans le cas où un token contient deux typographies différentes en son sein, on garde ce token. On lui associe la typographie italique ou grasse et on ne prend pas en compte la typographie normale.

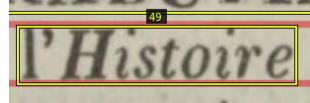


FIGURE 8 – Catalogue de ventes de manuscrits, JA, 1857, p.4

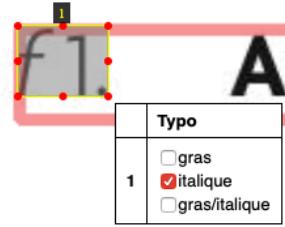


FIGURE 9 – Catalogue de ventes de manuscrits, Bodin, 2009, p.3

3.2 Décorations

Si des décos se situent sur la ligne à tokeniser, il faut les tokeniser également.



FIGURE 10 – Catalogue de ventes de manuscrits, Bovet, 1887, p.76

Manuel pour la segmentation des catalogues

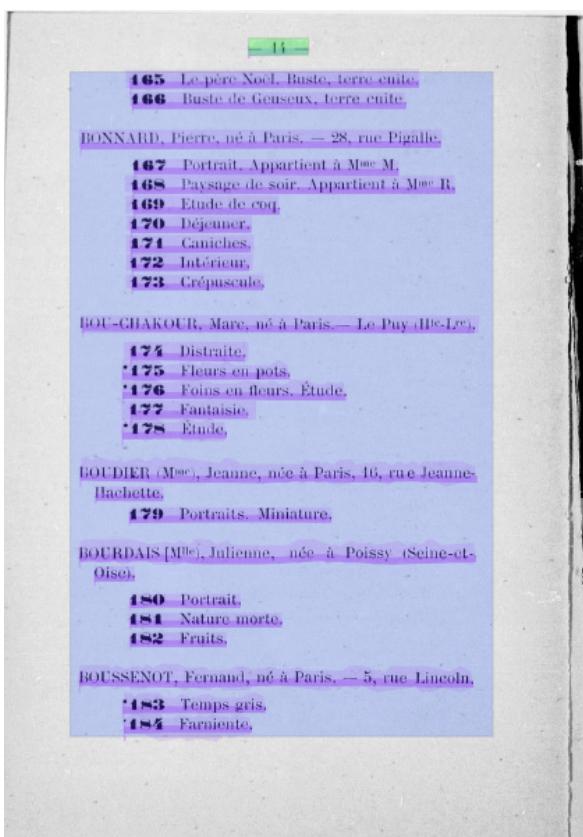
Juliette Janes

04-06-2021

Ce manuel a été réalisé dans le cadre de la réalisation d'un dataset de segmentation de catalogues sur eScriptorium.¹ Il permet de donner un aperçu des différents choix faits lors de la préparation des données.

1 Segmentation utilisant le vocabulaire de SegmOnto

1.1 Cas les plus fréquents



752	
58-60	Romainin, aux Deux-Moulins près Versailles (S.-et-O.).
62	Pochet, Paris, av. Clichy, 160.
64	Entrée r. Boursault, 19.
66	Gufé, Paris, r. Chercé-Midi, 2.
68	Lemaire (M ^{me}), Paris, r. Chercé-Midi, 9.
70	Belzacq, Paris, r. Printemps, 1.
72-78	Chemin-de-fer de l'Ouest.
80	Entrée r. Rome, 93.
82	Oder, Paris, r. Lopellbach, 9.
84	Marge, Paris, av. Wagner, 19.
86	Bourneuf, Paris, r. Damesme, 86.
88	C ^e G ^{re} des Oumilius, Paris, r. St-Honoré, 155
90	Loiraud, Paris, pl. Percière, 9.
92	C ^e Ass-Vies Le Phénix, Paris, r. Lafayette, 33.
94	Entrée r. Sausure, 2.
96	Entrée r. Sausure, 1.
98	Montagne, Paris, r. Damesme, 98.
100	Delurck, Paris, r. Rivoli, 116.
102	Sanson, Argenteuil (S.-et-O.).
104	Robin, Paris, r. Martyrs, 59.
106	Laville, Paris, r. Lyon, 22.
108	Boutillier du Retail (M ^{me}), La Belletière, par Viviane (Vienne).
110	Daniel, Paris, r. Damesme, 1.
112	Pernet (Vcl), Paris, r. Cardinet, 36.
114	Rouxeville, Paris, r. Crozatier, 50.
116	Cayron (Vcl), Paris, av. St-Ouen, 119.
118	Lécuyer (M ^{me}), Paris, r. La Barre, 44.
120	Hainaut, Boulogne s/S., r. Paris, 160.
D.6 DAMESME (Impasse) ²¹	
1	Moulin, Paris, imp. Damesme, 1.
3	Huel, Paris, imp. Damesme, 3.
5	Sellier, Paris, imp. Damesme, 5.
7	Brellez, Paris, imp. Damesme, 7.
9	Salat, Paris, imp. Damesme, 9.
11	Kirn r. Damesme, 39.
13	Nola, Paris, imp. Damesme, 4.
15	Filiâtre, Paris, imp. Damesme, 6.
17	Gilbert, Paris, imp. Damesme, 8.
19	Blanchefeuille, Paris, imp. Damesme, 10.
D.7 DAMESME (Rue) ²²	
1 à 5	Ville de Paris.
7-9	Baudran (Victor), Paris, r. Vilain, 28.
11	Burgien, Paris, r. Damesme, 37.
13	Goujetot, St-Mandé, av. Damesnil, 47.
15	St-Marc, Viry-Châtillon (S.-et-O.).
17	Entrée r. Baudran, 1.
19	Querrile (Vcl), Paris, r. Saintonge, 45.
21	Bourgoin, Paris, r. Damesme, 21.
23	Pincemail, Paris, boul. Arago, 15.
25	Mathieu, Paris, r. Damesme, 25.
27	Bourhon, Paris, r. Damesme, 27.
D.8 DAMIETTE (Rue de) ²³	
1	Launay (Vcl), gér. Lacau, Paris, r. Etienne-Marcel, 50.
3	Entrée r. Nil, 2.
5-7	Entrée r. Aboukir, 96.
2-4	Launay (Vcl), gér. Lacau, Paris, r. Etienne-Marcel, 50.
4 ^{me}	Entrée r. Caire, 51.
6-8	Entrée r. Caire, 53.

FIGURE 1 – Société des Artistes Indépendants, 1892, p.14

1. voir compte rendu segmentation 20-04-2021

FIGURE 2 – Annuaire des propriétaires de Paris, 1898, p.777

Les figures 1 et 2 présentent deux types de pages assez communes dans le dataset. La pagination, en haut de la page, en vert, est signalé par une zone **Numbering** tandis que le reste de la page est considéré comme une zone **Main**, en violet.

Lorsqu'un titre courant est présent sur la page, on le signale avec une zone **Running Title**, comme visible en jaune sur la figure 3. Dans le cas où le titre courant est situé sur la même ligne que la numérotation, comme c'est le cas ici, il faut supprimer la ligne et en recréer une pour chacune des zones, afin de bien les séparer.

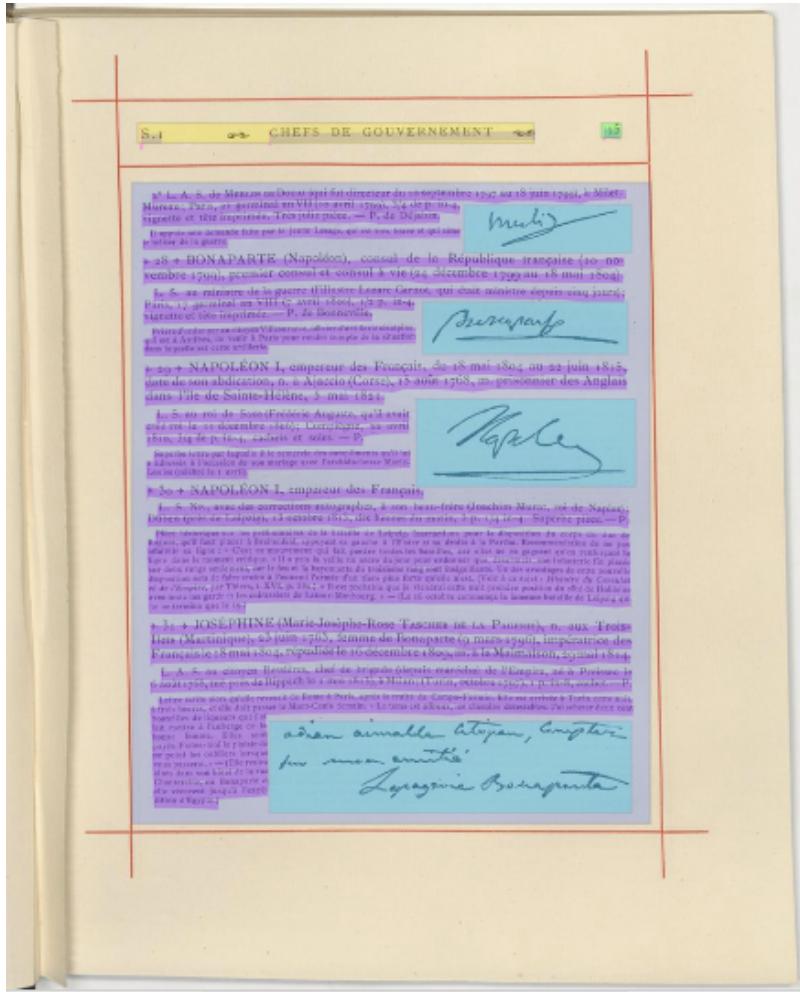


FIGURE 3 – *Catalogue de manuscrits*, Bovet, 1887, p.95

1.2 Cas particuliers

Lorsqu'il y a des images, à l'instar de la figure précédente, il faut les signaler comme étant des zones **Figure**, visibles ici en bleu, même si il s'agit là de signatures.

La zone **Title** correspond au titre d'une unité codicologique uniquement. Ainsi, la figure 4 présente une zone **Title**, en rouge, tandis que le titre *Pintura* de la page 5, qui correspond à un titre de partie dans le livre, est intégré directement à la zone **Main**, en violet.

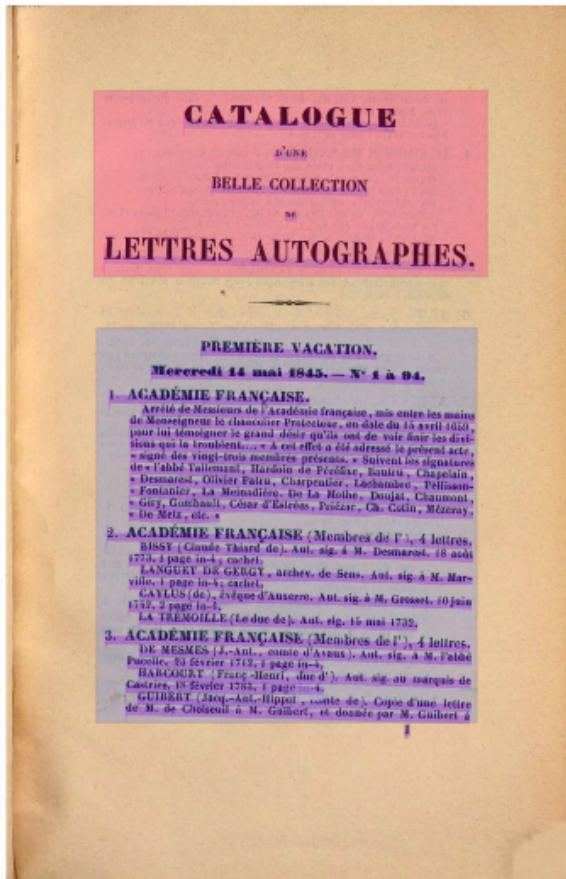


FIGURE 4 – *Catalogue de manuscrits*, Charavay, 1845, p.1

Hyatt Mayer, Conservador de Gravuras do "Metropolitan Museum of Art"; Una Johnson, Conservador de Gravuras e Desenhos do "Brooklyn Museum", e Dorothy Lytle, da Secção de Gravuras do "Museum of Modern Art".

Os pintores estão representados por obras cujo número varia de um a três, variação essa dependente do seu tamanho; os escultores por uma, e os gravadores por três. No caso daqueles artistas que apresentam mais de uma obra, o Júri tentou escolher peças que ilustrassem o desenvolvimento e variedade do seu estilo. De um modo geral, as obras integrantes da representação norte-americana foram concluídas no último decênio, mas, em alguns casos, houve necessidade de incluir-se obras mais antigas. A representação é composta tanto de artistas nascidos nos Estados Unidos como daqueles que nasceram no exterior, mas que fixaram sua residência e produziram uma parte considerável de sua obra.

RENÉ D'HARNONCOURT
Diretor do "Museum of Modern Art"
— New York

PINTURA

Ivan Le Lorraine ALBRIGHT (EE. UU. 1897 —)

1. Mulher — 1928. 84x56. Museum of Modern Art, New York

William BAZIOTES (EE. UU. 1911 —)

2. Natureza morta — máscaras — 1946 — 91x122. Philip C. Johnson, New York
3. O sonâmbulo — 1951. 122x102. The Kootz Gallery, New York

75

FIGURE 5 – *Biennale de São Paulo*, 1951, p.75

2 Nettoyage des lignes

Une fois les informations précédentes ajoutées à la segmentation, il est possible de s'atteler à un nettoyage rapide des lignes. Il y a assez peu de choses à corriger, cependant, il faut bien vérifier que les lignes sous les accolades ont été supprimées, ainsi que remettre au niveau de la ligne les points qui descendent trop haut ou trop bas. C'est parfois le cas pour les cadratins et les éléments en exposant, par exemple *Mme* et *Mlle*. Ces cas sont plus fréquents dans les annuaires, mais peuvent être rencontrés dans les autres documents.

Attention : Avant de passer à l'étape de l'ajout des entrées, il est nécessaire de lier les lignes à la zone **Main** en sélectionnant toutes les lignes puis en cliquant sur le bouton présenté ci dessous.



3 Ajout des entrées

Il faut par la suite ajouter à la segmentation les entrées de catalogues. Cet élément n'est pas présent dans le vocabulaire SegmOnto mais a été ajouté ici dans le but d'intégrer un second niveau de description des documents, afin d'améliorer l'extraction de données une fois les pages océsirées.

La zone **Entry** permet de désigner chaque entrée de catalogue. Les figures 6, 7 et 8 donnent une idée de l'aspect de ceux-ci pour chaque type de données présents dans le dataset.

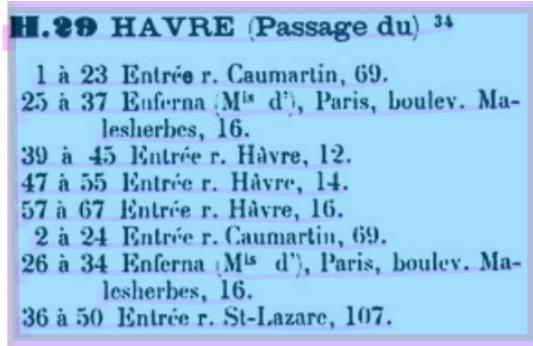


FIGURE 6 – *Annuaire des propriétaires de Paris*, 1898, p.938

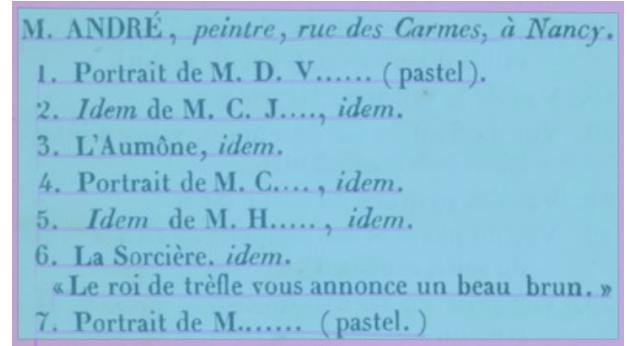


FIGURE 7 – *Catalogue des œuvres exposées à Nancy*, 1843, p.1

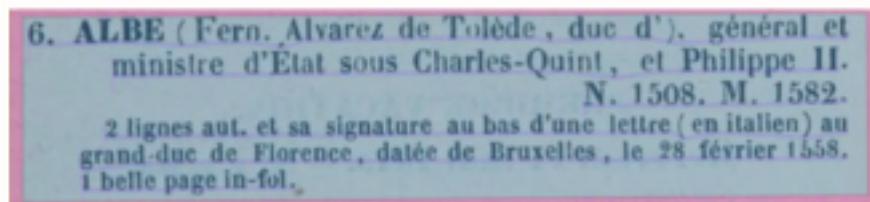


FIGURE 8 – *Catalogue de manuscrits, Charavay*, 1845, p.2

Les **Entry** peuvent contenir d'autres zones, essentiellement des zones **Figure**, comme visible dans la figure 9.

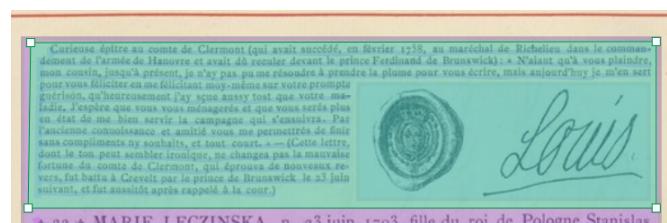


FIGURE 9 – *Catalogue de manuscrits, Bovet*, 1887, p.76

Dans le cas où une entrée est à cheval entre deux pages, il faut tagger le début de l'entrée comme **Entry** et les suivantes comme **EntryEnd**. Un exemple de cette zone est visible en violet sur la figure 10, le bleu correspondant à **Entry**.

qui l'appelait son *bon ange*. — L. a. s. (à Palissot); Paris,
10 avril 1760, 1 p. in-4. 25 »

Relative aux représentations au Théâtre-Français de la comédie des *Philosophes*, de Palissot. Il a appris qu'il regardait l'annonce des deux pièces de Voltaire (*Tancrède* et *Médimne*) comme un obstacle au temps où il désire qu'on joue la sienne; l'une de ces tragédies (*Tancrède*) est nouvelle, l'autre (*Médimne*) n'est qu'un remaniement de *Mahomet*; comme les principaux acteurs diffèrent dans les deux pièces, elles pourront être étudiées en même temps.

11 **Ashburton** (Alexandre Baring, lord), homme politique et financier anglais, envoyé par Robert Peel au congrès d'Aix-la-Chapelle en 1834, né en 1773, mort en 1848. — L. a. s. (à lord Buchan, érudit et biographe écossais); 28 juillet 1814, 1 p. 1/4 in-4. 25 »

Lettre historique. Il se félicite avec lui de la chute de Napoléon et de la restauration du roi légitime, il estime que le volcan révolutionnaire est enfin éteint.

FIGURE 10 – *Revue des Autographes*, 1885, p.3

Typologie des catalogues

Juliette Janes

juin 2021

Le but de ce manuel est de classer les différentes typologies existantes de catalogues de manuscrits et d'exposition dans l'idée de pouvoir en extraire les informations le plus facilement possible, en fonctionnant par groupe de catalogues similaires. Dans un premier temps, j'ai divisé, sur le modèle du papier *Automating Artl@s, extracting datas from exhibition catalogs*, les différents types d'entrées de catalogues. Je me suis ensuite concentrée sur les informations, emphases typographiques et ponctuations variants au sein de chacun de ces groupes.

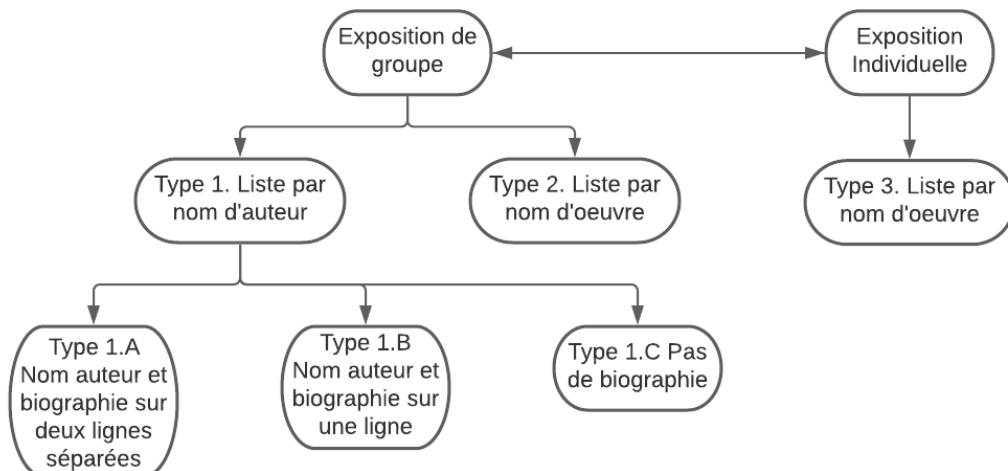


FIGURE 1 – Schéma récapitulatif des types de catalogues

1 Type 1 : Catalogues par liste d'auteur

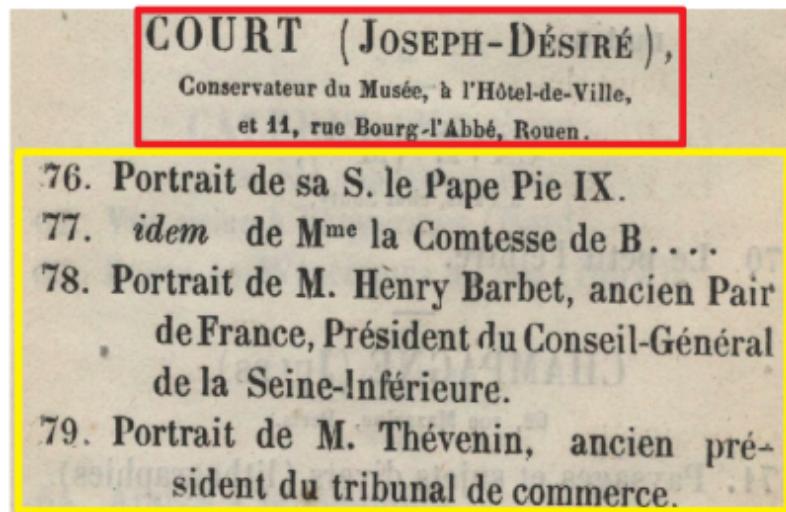


FIGURE 2 – Catalogue de l'exposition annuelle du musée de Rouen, 1853, p.12

La figure 2 est un exemple de base de la première catégorie, définie dans le papier mentionné. Il s'agit d'une entrée de catalogue d'exposition de groupe, listé par nom d'auteur. Une entrée correspond donc dans ce cas là, à un exposant, en rouge, pour lequel toutes les œuvres présentées dans l'exposition ont été listées, en jaune.

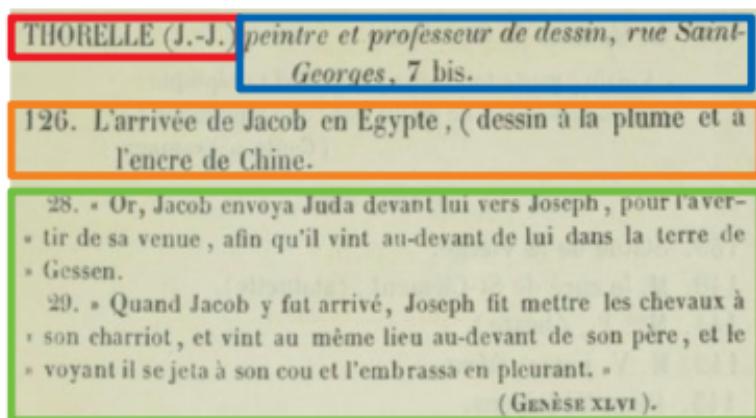


FIGURE 3 – Catalogue d'exposition des Beaux Arts de Nancy, 1849, p.11

La structure décrite peut être subdivisée en plusieurs parties, tel que présentées dans la figure 3. Si chaque entrée contient au moins le nom de l'auteur, en rouge, et le nom de l'œuvre, en orange, il peut aussi être indiqué des informations biographiques, en bleu, et de multiples informations sur l'œuvre en question, en vert. Chacun de ces éléments ont également un aspect qui peut être variable et que l'on va décrire.

1.1 Types d'organisation d'entrées

1.1.1 Type 1.A : Entrées sur trois parties

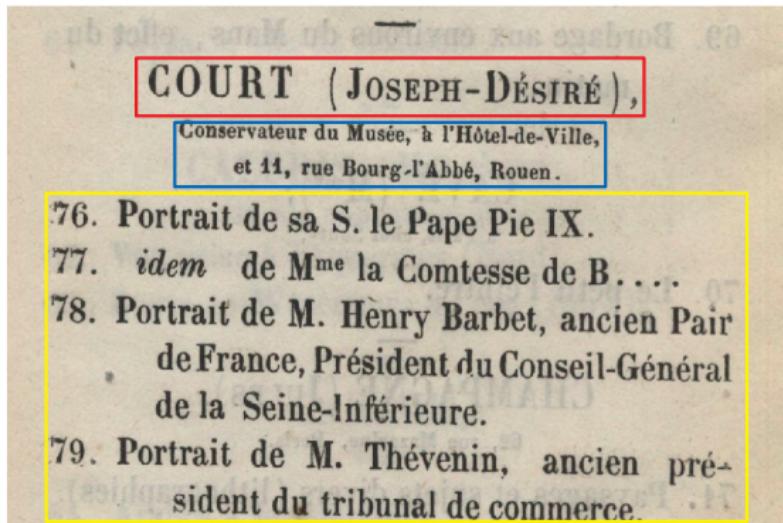


FIGURE 4 – Catalogue de l'exposition annuelle du musée de Rouen, 1853, p.12

La première partie donne le nom de l'auteur, la seconde des informations sur celui-ci, telles que son adresse de domiciliation, son école ou sa nationalité et la dernière est composée des œuvres de l'auteur présentées lors de l'exposition, sous la forme d'une liste numérotée.

1.2 type 1.B : Entrée sur Deux parties

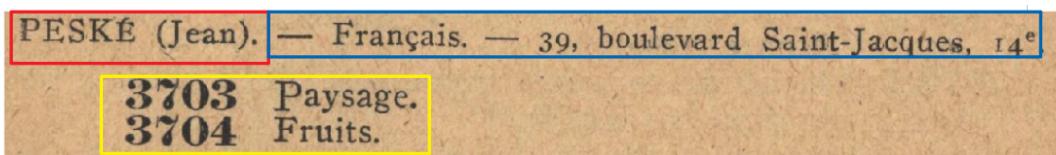


FIGURE 5 – Société des artistes indépendants, 1892, p.14

La figure 5 est un exemple typique de la seconde catégorie d'entrées que l'on peut trouver dans les catalogues d'exposition. Elle est composée de deux parties distinctes, la première présentant le nom de l'auteur, puis des informations complémentaires sans sauter de lignes et la seconde comportant la liste numérotée des œuvres exposées par l'auteur.

1.3 Structure des noms d'artistes

1.3.1 Structure basique

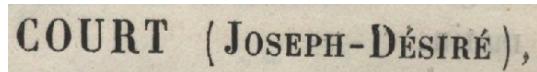


FIGURE 6 – Catalogue [...] Rouen, 1853, p.12

La figure 6 correspond à la structure la plus basique et la plus commune que peuvent prendre les nom d'artistes dans les catalogues d'exposition. Le nom y est en majuscule, suivi du prénom, en minuscule et entre parenthèses et d'une virgule. Cette structure peut varier un petit peu en fonction des catalogues, notamment au niveau de la ponctuation utilisée.

1.3.2 Variation de la structure basique : Ponctuation

Dans les figures ci-jointes, il est ainsi possible d'observer des structures assez similaires à la ponctuation changeante. Ainsi, la figure 7 est identique à la structure de base mais ne contient pas de virgule finale. Les figures suivantes n'ont plus le prénom entre parenthèses. Dans le cas de la figure 8, le nom est en majuscule, suivi d'une virgule, du prénom en minuscule et d'une autre virgule. Cette structure est légèrement modifiée dans les figures suivantes, où la virgule finale est transformée en point ou en tiret cadratin. Ce type de structure peut être également fréquemment associée à une emphase en gras du nom et du prénom, ou encore plus souvent, uniquement du nom. Ces exemples sont principalement issus de catalogues de la seconde moitié du XIXème siècle.



FIGURE 7 – *Catalogue [...] Nancy, 1849, p.11*

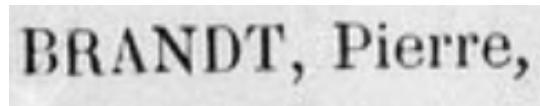


FIGURE 8 – *Exposition des artistes indépendants, Paris, 1892, p.15*

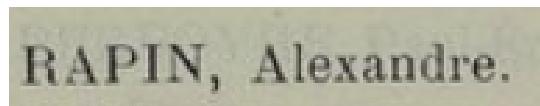


FIGURE 9 – *Catalogue d'exposition des beaux arts de Strasbourg, 1884, p.1*

Dagnan-Bouveret P. A. J. —

FIGURE 10 – *Catalogue[...], Venise, 1895*

1.3.3 Variation de la structure basique : Ordre

À l'instar de la figure 11, certains catalogues, plus tardifs et datant de la deuxième moitié du XXème siècle, présente le prénom de l'artiste puis son nom, le plus souvent en majuscules et en gras.

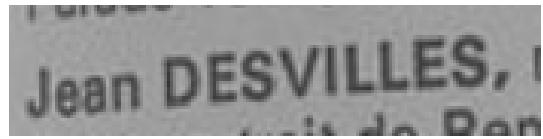


FIGURE 11 – Catalogue [...] Paris, 1965, p.175

1.3.4 Variation de la structure basique : Éléments mentionnés

Dernier type de variation, certains catalogues ne donnent pas le prénom des artistes. Dans le premier cas, pour les catalogues de la première moitié du XIXème siècle, seul le titre de civilité est mentionné, suivi d'un point.

La figure 13 présente un cas assez rare mais qui peut être rencontré. Certaines entrées peuvent ne pas faire mention du prénom de l'artiste, dans des catalogues où le prénom est mentionné. On peut donc supposer que celui-ci n'est pas connu.

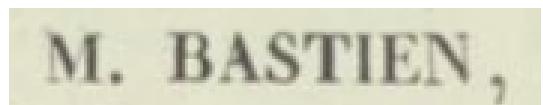


FIGURE 12 – Catalogue [...] Nancy, 1843, p.4



FIGURE 13 – Catalogue [...] Luxembourg, 1867, p.3

1.4 Structure des biographiques d'artistes

1.4.1 Structure sans saut de ligne

Les exemples suivants donnent une idée de l'aspect de base de l'élément biographique, lorsqu'il est présent dans l'entrée de catalogue. Celui-ci peut se situer soit à la suite du nom de l'artiste, comme pour les premières figures, soit sur la ligne en dessous, comme les derniers exemples.

Lorsque l'information biographique est située à la suite du nom de l'artiste, elle peut commencer par une virgule, à l'instar de la figure 14, et 16 ou un tiret cadratin, comme la figure 15. Elle se finit le plus souvent par un point et un saut de ligne. Cette partie est assez fréquemment mise en italique, et peut contenir l'adresse de l'artiste (qui commence le plus souvent par un chiffre), sa nationalité, sa ville, son maître et sa date de naissance. Dans certains cas, ces informations sont formalisées sous l'aspect d'un paragraphe rédigé, dont les informations seraient beaucoup plus compliquées à extraire, comme pour le catalogue de Venise. La plupart du temps, lorsqu'il s'agit plutôt d'une liste d'informations, séparées, dans le cas où l'information biographique a commencé par une virgule, par une virgule, ou, dans le cas où elle a commencé par un cadratin, par un cadratin. Dans certains cas, les dates et lieux de naissance, souvent associés, sont situés dans des parenthèses, qui peuvent commencer ou tenir lieu d'information biographique, comme pour les catalogues de São Paulo. C'est également le cas avec les nationalités.

RIGOLOT (Albert-Gabriel), né à Paris, élève de MM. Pelouse et Allongé; méd. de verm., Rouen; méd. d'arg., Rennes, Perpignan; méd. 2^e cl. (au blanc et noir), Paris.— Avenue d'Orléans, 52, Paris.

FIGURE 14 – Catalogue [...] Rouen, 1888

Dagnan-Bouveret P. A. J. — N. il 7 gennaio 1852 a Parigi. Fu discepolo di Gérôme ed espose per la prima volta al *Salon* del 1879 il quadro « *Un matrimonio mediante fotografie* », cui seguirono « *Benedizione d'un Pari* » (1882), « *La Vaccinazione* » (1883), « *Cavalli all'abbeveratoio* » (1884), « *Santa Vergine* » (1885) « *Pane benedetto* ». Le opere del Dagnan-Bouveret s'ispirano a una concezione delicata della vita, e spesso ad un sentimento di dolce religiosità. Egli è il poeta delle pie costumanze bretoni.

FIGURE 15 – Catalogue [...] Venise, 1895

VERNET (ÉMILE-JEAN-HORACE), né à Paris, en 1789. élève de Vincent, chevalier de la Légion-d'Honneur en 1814, officier en 1825, membre de l'Institut en 1826, directeur de l'Académie de France à Rome en 1828, commandeur de la Légion-d'Honneur en 1842, grand officier en 1862, mort en 1863.

FIGURE 16 – Catalogue [...] Luxembourg, 1867, p.3

Mary VIEIRA, née à São Paulo en 1927.

FIGURE 17 – Catalogue [...] Paris, 1961

SCHMIDT, Marcos Rodolfo (1943 — S. Paulo) — bp-72

FIGURE 18 – Catalogue [...] São Paulo, 1972

Willem de KOONING (Holanda, 1904 —)

FIGURE 19 – Catalogue [...] São Paulo, 1951

1.4.2 Structure avec saut de ligne

Dans d'autres cas, les informations biographiques sont mentionnées après avoir sauté une ligne. Chaque information peut être contenu dans une ligne, à l'instar du catalogue de Rouen. Le cas du catalogue de Nancy est particulièrement peu fréquent. Les informations biographiques tel que le maître et la ville de naissance sont ajoutées à la suite du nom de l'auteur tandis que l'adresse de l'artiste est située sur la ligne suivante en italique. C'est également le cas pour le catalogue de Strasbourg, où la ville de l'artiste est mentionné sur une ligne, en italique.

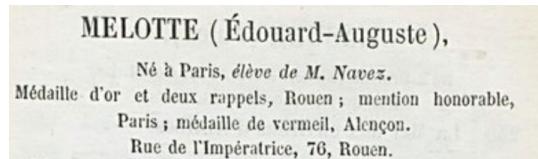


FIGURE 20 – Catalogue [...] Rouen, 1869

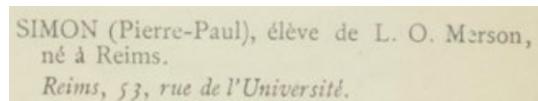


FIGURE 21 – Catalogue [...] Nancy, 1867

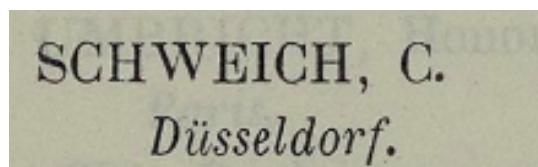


FIGURE 22 – Catalogue [...] Strasbourg, 1884

1.5 Structure des titres d'oeuvres et de leurs informations complémentaires

Cette partie présente les variations de structure des œuvres et de leurs informations complémentaires. Si la structure de la numérotation et du titre des peintures varient peu, les informations complémentaires peuvent être contenues entre parenthèses, après un tiret, des points de suspension...

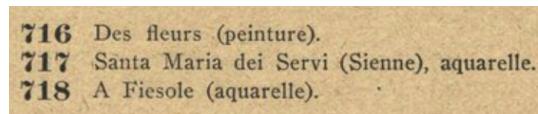


FIGURE 23 – Catalogue [...] Société des artistes indépendants, 1913

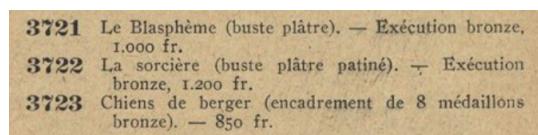


FIGURE 24 – Catalogue [...] Société des artistes indépendants, 1923

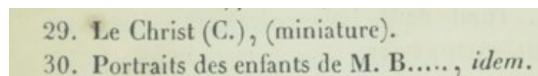


FIGURE 25 – Catalogue [...] Nancy, 1843

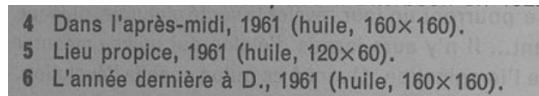


FIGURE 26 – *Catalogue[...] Paris, 1961*

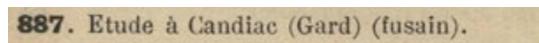


FIGURE 27 – *Catalogue[...] Rouen, 1888*

2 Type 2 : Catalogue par liste d'oeuvre

2.1 Structure



FIGURE 28 – *The exhibition of the Royal Academy, 1831*, p.7

Cette catégorie concerne les entrées de catalogues sur une seule ligne, à l'instar des catalogues anglais et canadiens. Elle se présente sous la forme d'une liste numéroté des œuvres avec, associé à chaque œuvre, le nom de son auteur, le plus souvent en italique.

3 Type 3 : Catalogue d'exposition individuelle

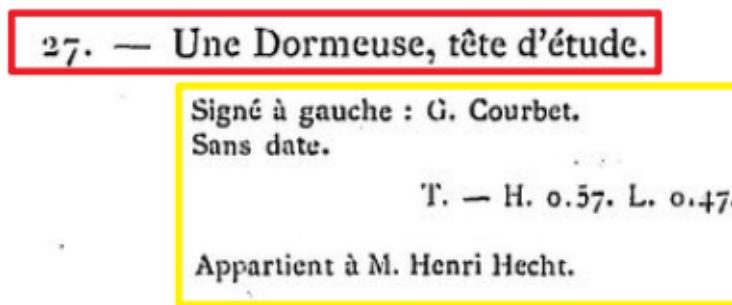


FIGURE 29 – *Catalogue d'exposition de Courbet, 1882*, p.42

Cette catégorie concerne les entrées de catalogues d'expositions monographiques. Comme il s'agit d'exposer un seul peintre, les entrées sont directement des œuvres numérotées avec plusieurs types d'informations ajoutées (Image 29). Les images suivantes présentent des exemples des différents informations complémentaires qui peuvent être associées au nom de l'œuvre : signature (en bleu), description de la peinture (en orange), dimensions (en violet), date de production (en jaune) ou encore propriétaire (en vert).

27. — Une Dormeuse, tête d'étude.

Signé à gauche : G. Courbet.

Sans date.

T. — H. 0.57. L. 0.47.

Appartient à M. Henri Hecht.

FIGURE 30 – Catalogue d'exposition de Courbet, 1882, p.42

33. — La Sorcière, copie d'après Franz Hals.

Signé à gauche : ..69. G. Courbet.

A droite, de la main du peintre, le monogramme de Franz Hals, avec la date 1645, et au-dessous : Aix-la-Chapelle.

Cette copie a été faite d'après le tableau original qui faisait partie de la galerie Suermondt.

T. — H. 0.85. L. 0.69.

FIGURE 31 – Catalogue d'exposition de Courbet, 1882, p.46

135. — Le Peintre à son chevalet.

Crayon noir.

Exposition particulière de 1855, où il était inscrit au catalogue avec la date de 1848.

Exposition particulière de 1867.

H. 0.5 . L. 0.33.

FIGURE 32 – Catalogue d'exposition de Courbet, 1882, p.86