

ÉCOLE NATIONALE DES CHARTES

Lucie Rondeau du Noyer

Élève de l'École normale supérieure de Paris

Diplômée de master

Agrégée d'histoire

ENCODER AUTOMATIQUEMENT DES CATALOGUES EN XML-TEI

PRINCIPES, ÉVALUATION ET APPLICATION À LA REVUE DES AUTOGRAPHES DE LA LIBRAIRIE CHARAVAY

Mémoire pour le diplôme de master

« Technologies numériques appliquées à l'histoire »

2019

Résumé

Dans le cadre d'un projet FNS consacré à l'écriture privée au XVII^e siècle, Simon Gabay, collaborateur scientifique en charge des humanités numériques à l'UniNe, a eu recours à de nombreux catalogues de vente d'autographes édités à Paris au XIX^e siècle. S'est ainsi imposée l'idée que l'exploitation systématique de ces catalogues constitue une étape incontournable pour la création d'une base de données des manuscrits français du XVII^e siècle, objectif du projet e-ditiones, conduit par Simon Gabay à l'UniNe depuis 2018.

Dans ce contexte, l'enjeu de mon stage était de proposer une chaîne de traitement complète, efficace et automatique permettant le passage d'un corpus numérisé de catalogues de vente d'autographes de la deuxième moitié du XIX^e siècle à une base de données aisément accessible et interrogeable.

Après un état de la recherche concernant les documents autographes et leur commercialisation depuis le XIX^e siècle, période de grande popularité des ventes et des collections d'autographes, le présent mémoire décrit les aspects techniques du travail effectué au cours de mon stage. Il détaille et évalue les outils numériques sélectionnés et mis en œuvre à chaque étape. Il démontre en particulier la pertinence et l'adaptabilité du logiciel GROBID-dictionaries pour le traitement de ressources non lexicales mais de forme encyclopédique.

En complément de la démarche technique, l'étude de la famille Charavay et le travail effectué sur les publications éditées par cette dynastie de libraires mettent en valeur la variété et la richesse des informations contenues dans les catalogues de ventes. L'exploitation systématique de ces sources précieuses et encore trop peu exploitées est aujourd'hui facilitée par les technologies numériques. Il est ainsi permis d'esquisser quelles sont les perspectives offertes dans de nombreux domaines de recherche par le traitement automatique et la mise en base de données des catalogues de vente.

Mots clés : catalogues de vente ; autographes ; manuscrits ; XML-TEI ; GROBID-dictionaries ; apprentissage supervisé ; *feature engineering* ; structuration automatique de données ; Charavay ; patrimoine écrit.

Informations bibliographiques : Lucie Rondeau du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay*, mémoire de master « Technologies numériques appliquées à l'histoire », dir, Thibault Clérice et Simon Gabay, École nationale des chartes, 2019.

Il y a deux catégories de marchands d'autographes les marchands sérieux, patentés, honorés, honorables, et les marchands d'occasion, courtiers marrons de la frivolité et du scandale, colportant sous le manteau et semant, à prix d'or, la profanation des souvenirs et le déshonneur des familles. Nous ne parlerons que des premiers [...].

Adolphe de Lescure, *Les autographes et le goût des autographes en France et à l'étranger : portraits, caractères, anecdotes, curiosités*

Remerciements

Je remercie en premier lieu Thibault Clérice, responsable pédagogique du master « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Tuteur de mon stage, il m'a donné l'opportunité d'effectuer un travail de fin d'études à l'Université de Neuchâtel et de mettre ainsi en pratique dans les meilleures conditions l'enseignement dispensé au cours de l'année 2018-2019.

Tous mes remerciements vont également à l'ensemble de l'Institut de littérature française de l'Université de Neuchâtel, particulièrement à Alain Corbellari, professeur ordinaire de littérature médiévale, et à Simon Gabay, collaborateur scientifique en charge des humanités numériques. Référent de mon stage, Simon Gabay a encadré et soutenu mon travail avec constance et enthousiasme. Je suis aussi très reconnaissante à Céline Künzi qui a grandement facilité mon arrivée et le déroulement de mon stage à l'UniNe (et a su m'apprendre de nombreux helvétismes).

Grâce à la disponibilité et à l'aide de Mohamed Khemakhem, doctorant en informatique à l'Inria et programmeur de GROBID-dictionaries, de nombreuses difficultés techniques ont pu être aplanies. Je tiens à lui exprimer ma reconnaissance, ainsi qu'à Laurent Romary, directeur de recherche à l'Inria.

Ce travail a reçu le soutien de la Fondation-Maison Borel où j'ai pu séjourner deux mois : je remercie la Fondation d'avoir retenu ma candidature. Je remercie également Pierrette Bertolucci et Anna Lucatelli pour toute l'aide apportée au cours de mon séjour neuchâtelois.

En juillet 2019, j'ai eu la chance de participer à l'école d'été « Décrire, reconstituer, explorer les bibliothèques de la première modernité » organisée par l'Equipex Biblissima et la Bibliothèque Mazarine. Je remercie ses initiateurs, Yann Sordet et Patrick Latour, ainsi que l'ensemble de ses participants. Cette expérience collective m'a permis de prolonger les perspectives ouvertes pendant mon année à l'École nationale des chartes et approfondies au cours de mon stage.

Merci à Camille ainsi qu'à Élise et Yannick, mes hôtes fribourgeois, qui m'ont accueillie si généreusement lors de mes premières semaines en Suisse. Merci enfin à Dominique, Emmanuel, Gabriel et Justine pour leur soutien sans faille.

Introduction

Alors qu'il préparait une édition critique de la correspondance de Descartes, Erik-Jan Bos, chercheur en philosophie à l'Université d'Utrecht, a localisé en Pennsylvanie, grâce à une simple requête sur un moteur de recherche, un document disparu depuis près de 150 ans¹. Dans les collections du Haverford College, se trouvait une lettre autographe de Descartes volée au XIX^e siècle à la bibliothèque de l'Institut de France par Guglielmo Libri, célèbre bibliophile italien qui dut quitter la France en 1848 suite à des accusations de vols dans les bibliothèques publiques qu'il était chargé d'inspecter². Cette découverte qui a retenu l'attention de la communauté scientifique indique combien les outils de recherche numériques sont un adjuvant précieux pour retracer l'histoire de documents fondamentaux et inédits.

Depuis deux décennies, la valorisation des archives privées est telle que les manuscrits et les autographes ont pu « devenir [des] objets de placements financiers³ », voire de spéculation. Les services d'archives publiques en France ont été confrontés à des problématiques nouvelles⁴ du fait de l'ampleur inédite de ce phénomène. Alors que les prix s'envolent, de quels moyens concrets disposent-ils pour empêcher le démembrement de fonds privés, s'opposer à la vente d'archives publiques qui n'auraient jamais dû apparaître en premier lieu sur le marché privé et se protéger du trafic international d'autographes ? À court terme, l'ouverture d'une instruction judiciaire au printemps 2014 contre Aristophil suivie de la faillite de cette « société par actions simplifiées, et à visée clairement spéculative⁵ » se traduit par un relatif tassement des prix des manuscrits et autographes sur le marché parisien. À plus long terme, il est cependant peu probable que le marché soit durablement enrayé si l'on en juge par les premières ventes de liquidation. Les 130 000 documents du fonds Aristophil dont la dispersion est prévue sur une période de six ans devraient être majoritairement acquis par des acteurs privés.

Au côté des archivistes, les chercheurs sont également concernés par le développement du marché privé des manuscrits. C'est particulièrement vrai pour celles et ceux d'entre eux qui cherchent à établir des éditions critiques. Une telle entreprise nécessite d'être en mesure de localiser les manuscrits, y compris ceux détenus par des collectionneurs

1. S. Ornes, « Q&A: Descartes' Decipherer », *Nature*, 483, 28 mars 2012, p. 540.

2. A. Maccioni Ruju, et M. Mostert, *The Life and Times of Guglielmo Libri (1802-1869) : Scientist, Patriot, Scholar, Journalist and Thief, A Nineteenth-Century Story*, Hilversum, Verloren Publishers, 1995.

3. P. Even, « Les archives : un marché ? », *Pouvoirs*, n° 153/ 2, avril 2015, p. 95-107.

4. P. Marcilloux, *Les ego-archives : traces documentaires et recherche de soi*, Rennes, Presses universitaires de Rennes, 2013.

5. P. Even, « Les archives : un marché ? », art. cit.

Introduction

privés. Il est aussi utile d'avoir connaissance de l'existence de fac-similés en cas d'impossibilité à consulter les originaux. Dans les deux cas, le recours à des catalogues de vendeurs d'autographes et de manuscrits est d'une aide certaine. L'exploitation systématique de ces catalogues par leur mise en série pourrait permettre en outre de mieux connaître l'histoire de la circulation du patrimoine écrit et la provenance des manuscrits autographes. Elle serait un outil au service l'authentification des documents pour les acteurs économiques du marché, les éditeurs de correspondances et les chercheurs spécialistes de l'histoire du patrimoine écrit.

Mon stage à l'Institut de littérature française de l'Université de Neuchâtel (UniNe) s'inscrivait dans la suite d'un projet financé par le FNS (Fonds national suisse de la recherche scientifique). Conduit entre 2015 et 2018 sous la direction de Marc Escola de l'Université de Lausanne et d'Alain Corbellari de l'Université de Neuchâtel et intitulé « L'écriture privée au XVII^e siècle : étude philologique des manuscrits de Mme de Sévigné⁶ », ce projet est né du constat de l'absence d'une étude systématique des manuscrits de la plus grande épistolière du XVII^e siècle. Alors que l'étude des manuscrits du XVIII^e siècle a connu un renouvellement notable au cours de ces dernières années⁷, bien des textes du XVII^e siècle attendent un traitement comparable. Remarquer que certains éditeurs scientifiques des textes littéraires de la période classique s'appuient peu sur les apports de la philologie romane est encore d'actualité⁸.

La surprenante absence d'un catalogue inventoriant les lettres autographes de Madame de Sévigné est d'autant plus dommageable qu'elle empêche de mener à bien certaines recherches, notamment sur l'orthographe du XVII^e siècle. Elle est également un frein à l'étude comparative des éditions successives de l'œuvre de Madame de Sévigné. Le projet « L'écriture privée au XVII^e siècle » visait donc à proposer une nouvelle édition des lettres de Madame de Sévigné et à conduire une analyse linguistique et éditoriale du corpus. Il a notamment résulté en la publication d'un catalogue électronique faisant état des sources dispersées relatives à ces lettres⁹. Cette démarche philologique s'est largement appuyée sur les catalogues de ventes de manuscrits et de lettres autographes, sources incontournables du fait de la forme épistolaire de l'œuvre de Madame de Sévigné.

6. Le descriptif complet du projet est consultable à l'adresse suivante : <http://p3.snf.ch/projects-157169>.

7. Un bon aperçu de ce renouvellement est donné par N. Ferrand dans l'introduction d'un numéro spécial de *Genesis* consacré à la question des brouillons du XVIII^e siècle : « L'Ancien et le Nouveau Régime des manuscrits de travail », *Genesis. Manuscrits – Recherche – Invention*, n° 34, avril 2012, p. 7-17.

8. S. Gabay, « Éditer le Grand Siècle au XIX^es. Remarques sur les choix (ortho)graphiques de quelques éditeurs », document de travail, 2018. hal-01907239

9. S. Gabay, « Sources sévignéennes », *Projet e-ditiones / Université de Neuchâtel*, 2018, disponible à l'adresse suivante : https://f.hypotheses.org/wpcontent/blogs.dir/5238/files/2018/11/Sources_sevigneennes.pdf.

Introduction

Depuis 2018 et dans le cadre d'un nouveau projet intitulé e-ditiones¹⁰, Simon Gabay, collaborateur scientifique de l'UniNe en charge des humanités numériques, poursuit l'exploitation des catalogues de vente d'autographes. Il s'agit désormais d'étendre le champ de recherche afin d'établir un catalogue recensant le plus grand nombre de manuscrits français du XVII^e siècle et éclairant l'histoire de leur transmission et de leur circulation jusqu'à nos jours.

Les catalogues de vente de manuscrits et autographes, matériau indispensable à la bonne conduite de ce projet, sont eux-mêmes des sources relativement difficiles d'accès. Produits par des libraires ou des experts marchands d'autographes depuis le début du XIX^e siècle, ils n'ont pas fait l'objet de collectes systématiques par les institutions patrimoniales avant la deuxième moitié du XX^e siècle. A la Bibliothèque nationale de France comme aux Archives nationales, les fichiers constitués à partir de certains de ces catalogues sont plus couramment exploités que les collections de ces imprimés qui n'ont pas fait l'objet de campagnes de numérisation spécifiques.

La problématique est alors la suivante : en l'absence d'un corpus constitué et aisément accessible de catalogues de vente de lettres autographes, quelles sont les technologies numériques permettant de faciliter la consultation et l'exploitation de ces sources ? Dans quelle mesure la production de documents encodés et mis à disposition des chercheurs peut-elle servir l'histoire de la circulation de l'écrit aussi bien que les recherches philologiques et ecdotiques ?

Durant mon stage à l'UniNe sous la tutelle de Simon Gabay, j'ai précisément travaillé à la mise au point d'une chaîne de traitement permettant de passer le plus efficacement et automatiquement possible de la version numérisée d'un ensemble de catalogues à une base de données facilement disponible et interrogeable. L'un des enjeux de cette démarche était la structuration automatique pour exploiter pleinement des documents dont les seules numérisation et mise en ligne apparaissent aujourd'hui comme insuffisantes.

Après avoir fait le point sur l'état de la recherche concernant les documents autographes et leur commercialisation depuis le XIX^e siècle, je présenterai les trois temps de cette démarche technique. Il s'agissait en premier lieu de mettre au point une chaîne de traitement complète, permettant de passer du catalogue numérisé à un document XML-TEI intégrable à une base de données et interrogeable *via* une interface. Pour plusieurs étapes de cette chaîne de traitement, il a ensuite été nécessaire de produire des données d'entraînement. Ces dernières ont servi à conduire une évaluation des outils retenus. Enfin, j'ai constitué pour leur mise en ligne un corpus test de soixante-quinze numéros de la *Revue des Autographes, des*

10. Le carnet de recherche de ce nouveau projet est disponible en ligne sur la plateforme Hypothèses, à l'adresse <https://editiones.hypotheses.org>.

Introduction

curiosités de l'histoire et de la biographie, périodique recensant les documents en vente à prix marqués dans la librairie de Gabriel et d'Eugène Charavay.

Première partie

Le marché des lettres autographes : historiographie et typologie des sources

1.1. Les catalogues de vente, une source majeure pour l'histoire du livre et l'histoire de l'art

Les catalogues, qu'ils soient produits par des acteurs privés ou commerciaux et quelle que soit leur fonction (consultation, récolement, préparation d'une vente), sont une source majeure pour la recherche en histoire de l'art comme en histoire du livre¹¹. Dans ces deux domaines de recherche, ils sont depuis longtemps des outils de travail et de référence.

Les catalogues permettent aux acteurs du marché de l'art et de la bibliophilie d'avoir accès à des informations précieuses relativement aux modalités et aux résultats des ventes passées. Ces éléments représentent un enjeu économique car ils orientent en partie les décisions d'achat et les prix de vente pratiqués. Les nouvelles technologies de l'information et de la communication permettent le développement à une échelle sans précédent de bases de données recensant des transactions anciennes.

Ces grandes bases de données sont conçues, alimentées et maintenues par des acteurs privés du marché de l'art (comme la société française spécialisée Artprice), des éditeurs scientifiques (notamment Brill¹²) ou des institutions culturelles et de recherche (parmi lesquelles se distingue le Getty Research Institute et sa série de Getty Provenance Index Databases). Dans la mesure où la plupart d'entre elles recensent des titres de catalogues imprimés sans indexation de leur contenu, elles restent généralistes et se prêtent difficilement à la recherche de lettres autographes.

Ainsi, dans la base « *The Book Sales Catalogues of the Dutch Republic 1599-1800* » de Brill qui vise à recenser tous les catalogues de vente de livres en Hollande aux XVII^e et XVIII^e siècles, moins de dix notices attestent de façon indubitable de la présence d'autographes. Il est également à noter que la consultation des documents de cette base, comme de nombreux autres outils de référence proposés par des entreprises, suppose un abonnement coûteux de la part des bibliothèques ou des institutions de recherche.

11. A. Charon et E. Parinet (dir.), *Les ventes des livres et leurs catalogues, XVII^e-XX^e siècle*, Paris, Publications de l'École nationale des chartes, 2000.

12. Brill maintient par exemple, dans le domaine de l'art, le répertoire numérique *Art Sales Catalogue Online* qui se fonde sur le *Répertoire des Catalogues de Ventes Publiques intéressant l'Art ou la Curiosité* de Frits Lugt. Dans le domaine de l'histoire du livre, Brill se charge de la publication de la base de données « *The Book Sales Catalogues of the Dutch Republic 1599-1800* », version numérique d'une collection de microfilms de catalogues de vente lancée par l'historien du livre Bert van Selm et poursuivie par J. A. Gruys et H. W. de Kooker.

1.1. Les catalogues de vente, une source majeure pour l'histoire du livre et l'histoire de l'art

1.1.1. Catalogues et histoire du livre : multiplication des éditions électroniques et des bases de données

Les catalogues de vente publique comptent parmi les sources fondamentales pour les historiens du livre. En attestent l'édition soignée de certains des plus anciens exemplaires connus¹³ et l'existence de corpus les recensant pour la période moderne¹⁴ durant laquelle se diversifient, à partir de la Renaissance, les types et les fonctions des catalogues, en lien avec l'essor du commerce du livre imprimé¹⁵. Alors que l'époque contemporaine marque un moment d'explosion du nombre des catalogues commerciaux, les études et les tentatives d'inventaires ne sont pas aussi nombreuses que pour les périodes précédentes. Pour preuve, le champ d'investigation de l'équipement d'excellence Biblissima, « observatoire du patrimoine écrit¹⁶ », financé depuis 2011 par le programme des Investissements d'avenir, se limite précisément à la fin du XVIII^e siècle.

Parmi les corpus de catalogues commerciaux modernes, la base de données « Esprit des Livres » est construite et hébergée par l'École nationale des chartes depuis 1998. Elle recense, grâce à un travail de repérage collectif, l'ensemble des catalogues de vente de livres antérieurs au XIX^e siècle conservés dans les bibliothèques parisiennes. À l'origine, elle mettait l'accent sur l'identification des éditions de catalogues, de leurs exemplaires, de leurs particularités et de leurs possesseurs. Il est intéressant de noter que son intégration à Biblissima a mené à la création d'un nouvel aspect de la base. L'objectif complémentaire est désormais d'identifier les manuscrits mentionnés dans les catalogues, afin d'intégrer les métadonnées les concernant aux notices du site « Esprit des Livres ».

Ce nouvel axe de la base de données confirme que le renouvellement de l'étude des catalogues, permis par les humanités numériques, rejoint la volonté de mettre en place des ressources fédérées pour étudier la circulation des manuscrits, aux échelles nationale (avec la

13. P. Delsaerd et Y. Sordet (dir.), *Lectures princières & commerce du livre : la bibliothèque de Charles III de Croÿ et sa mise en vente, 1614*. 2 vol., Enghien / Paris, Fondation d'Arenberg / Éditions des Cendres, 2017.

14. G. Mandelbrote, « La nouvelle édition de G. Pollard et A. Ehrman, *The Distribution of Books by Catalogue from the invention of printing to AD 1800* : Bilan des travaux préparatoires : catalogues français » dans *Les ventes des livres et leurs catalogues, XVII^e-XX^e siècle*, *op.cit.*

15. Y. Sordet, « Pour une histoire des catalogues de livre : matérialité, formes, usages » dans *De l'argile au nuage, une archéologie des catalogues : II^e millénaire av. J. C. - XXI^e siècle*, p. 21

16. A.-M. Turcan-Verkerk et P. Bertrand, « BIBLISSIMA : Bibliotheca bibliothecarum novissima, an observatory for the written cultural heritage of the Middle Age and the Renaissance », dans B. Saou-Dufrene et B. Barbier (éd.), *Heritage and Digital Humanities. How should training practices evolve?*, Berlin, Lit, 2014, p. 129-139.

1.1. Les catalogues de vente, une source majeure pour l'histoire du livre et l'histoire de l'art

base de données Bibale¹⁷) comme internationale (avec le projet *Mapping Manuscript Migrations*¹⁸). D'autres projets soutenus par Biblissima poursuivent le même objectif de reconstituer des collections et des bibliothèques aujourd'hui dispersées grâce à la numérisation et au moissonnage de catalogues. Les Archives nationales ont ainsi étudié les catalogues des bibliothèques ecclésiastiques saisies pendant la période révolutionnaire (1770-1797). Il s'agissait de les localiser dans les différentes séries du cadre de classement et de les numériser pour consultation dans la salle des inventaires virtuelle. Rassembler virtuellement des documents épars dans les fonds documentaires, assurer leur pérennité et permettre leur consultation en ligne sont des objectifs en partage avec notre approche des catalogues de vente de lettres autographes.

Toujours dans le cadre de Biblissima, une attention certaine a été portée à l'édition scientifique des sources catalographiques. Pour la période moderne, il est important de mentionner le travail d'édition des inventaires des bibliothèques de Mazarin et de Richelieu en cours à la Bibliothèque Mazarine. Ces documents sont des substituts aux catalogues inexistantes de ces deux collections. Une réflexion poussée sur les modalités d'édition électronique a précédé ces travaux d'édition. Elle a donné lieu à la création de l'environnement de travail Thecae dont la méthodologie et la marche à suivre ont été publiées en ligne¹⁹.

1.1.2. Catalogues et histoire de l'art : chantiers de numérisation et étude des circulations

Comme dans le domaine de l'histoire du livre, les recherches conduites sur le marché de l'art au prisme de la notion de circulation se sont multipliées ces dernières années. Le groupe de recherche Artl@s s'est ainsi lancé depuis 2009 dans la constitution de BasArt, une base de données mondiale des catalogues d'exposition²⁰. Béatrice Joyeux-Prunel, sa fondatrice et principale animatrice, s'éloigne de l'exploitation monographique et économique des

17. Pour plus de détails, se reporter à H. Wijsman, « The Bibale Database at the IRHT: A Digital Tool for Researching Manuscript Provenance », *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, n° 2017/1, p. 328-34, disponible à l'adresse suivante : https://repository.upenn.edu/mss_sims/vol1/iss2/10.

18. T. Burrows, E. Hyvönen, L. Ransom et H. Wijsman, « Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts », *Manuscript studies* 2019 (3.1), disponible à l'adresse suivante : https://repository.upenn.edu/mss_sims/vol3/iss1/13.

19. M. Bisson, A. Goloubkoff et E. Kuhry, « Éditer un inventaire XML-TEI P5 », Université de Caen / Pôle Document Numérique, mis en ligne le 25 février 2019.

20. S.A. Matei, « ARTL@S and BasArt : A Loose Coupling Strategy for Digital Humanities », *Artl@s Bulletin : Pour une histoire spatiale des arts et des lettres*, Volume 1, Issue 1, 2012.

1.1. Les catalogues de vente, une source majeure pour l'histoire du livre et l'histoire de l'art

catalogues d'exposition et revendique de les utiliser pour mettre en valeur les flux et les échanges artistiques existant à l'échelle globale depuis le XIX^e siècle²¹.

De vastes programmes de numérisation de catalogues ont été entrepris ces dernières années. Ils ont conduit à la mise en ligne de nombreux catalogues d'art et à la conception de modes de consultation facilitant la recherche : mise en base de données ou recherche plein texte.

En Allemagne, par exemple, deux projets de recherche ont permis la numérisation d'une part substantielle des catalogues d'art édités entre 1900 et 1945 et conservés à la Bibliothèque Universitaire d'Heidelberg. Le premier, intitulé « *Kunst – Auktionen – Provenienzen* » (« Art, Enchères, Provenances »), vise à reconstituer le marché de l'art en Allemagne dans la première partie du XX^e siècle « au miroir des catalogues de vente aux enchères entre 1901 et 1929²² ». Il s'agit de fournir aux chercheurs et aux acteurs culturels un outil facilitant les recherches de provenance. L'objectif est de centraliser un matériau archivistique dispersé dans de nombreuses bibliothèques et institutions muséales et d'interroger, y compris en mode plein texte, des sources jusqu'alors difficiles d'accès.

Le second programme, d'envergure internationale et appuyé par le Getty Research Institute, consiste en la numérisation de catalogues édités entre 1930 et 1945. Le principal objectif est l'identification des œuvres d'art vendues sous le régime nazi. Les études de provenance qui se sont développées et institutionnalisées au cours des années 1990 hors du champ académique, notamment pour retrouver les œuvres confisquées et dispersées pendant la période nazie, ont trouvé à s'intégrer aujourd'hui dans des structures universitaires²³. Parce qu'elles s'appuient sur un travail intensif des sources historiques, elles se révèlent complémentaires aux études du marché de l'art largement inspirées par la nouvelle sociologie économique et permettent, comme le programme Artl@s, d'échapper à une approche purement quantitative.

Si les recherches de provenance ont abouti à la mise en ligne de larges bases de données dans le domaine des beaux-arts comme dans le secteur de la vente des livres anciens et précieux, la place qu'y tiennent les documents manuscrits et autographes et les catalogues s'y rapportant demeure marginale. Dans une présentation de la méthodologie de l'Index de provenances constitué par les personnels du Getty Research Institute²⁴, il est explicitement

21. B. Joyeux-Prunel, « Circulation and the Art Market », *Journal for Art Market Studies*, 1-2, 2017.

22. La base est disponible en ligne à l'adresse suivante :

<https://digi.ub.uni-heidelberg.de/en/sammlungen/artsales.html>.

23. J. Gramlich, « Reflections on Provenance Research : Values – Politics – Art Markets », *Journal for Art Market Studies*, 1-2, 2017.

24. R. Cuadra et S. Michels, « Publishing German Sales, A Look under the Hood of the Getty Provenance Index », *IRIS: Behind the Scenes at the Getty*, 2013, disponible en ligne à l'adresse suivante : <https://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index>.

1.1. Les catalogues de vente, une source majeure pour l'histoire du livre et l'histoire de l'art

précisé que les catalogues purement bibliographiques ont été écartés lors de la construction de la base de données. Ce sont les objets d'art tels que les peintures, les sculptures et les dessins qui intéressent en premier lieu les instigateurs du projet. Il en va de même des entrées relatives aux documents manuscrits et autographes dans les catalogues composites : elles n'ont pas été intégrées aux bases de données finales.

En France, la bibliothèque de l'Institut National d'Histoire de l'Art (INHA) a également initié un chantier de numérisation de ses catalogues de vente provenant de la collection Jacques Doucet. Sur un total de 7 512 catalogues de vente disponibles en ligne, seuls 136 incluent des lettres et autographes. Il est également à noter que cette collection numérisée ne comporte aucun catalogue du XIX^e siècle, période du premier apogée du commerce d'autographes à Paris.

À l'issue de ce tour d'horizon des projets d'humanités numériques permettant l'exploitation des catalogues de vente d'art ou de livres, force est de constater qu'il n'existe pas de base de données spécifique des autographes ou des manuscrits du XVII^e siècle. Cette carence, qui est à mettre en relation avec la connaissance lacunaire que nous possédons du marché de l'autographe parisien au XIX^e siècle, justifie l'ambition du projet e-ditiones porté par Simon Gabay à l'UniNe.

1.2. Le commerce des lettres autographes, un secteur du marché de l'art encore peu étudié

1.2.1. Le poids historique de la parole de l'expert vendeur

Les travaux d'histoire consacrés à la vente et au marché des autographes sont peu nombreux et ont été majoritairement publiés dans des pays anglo-saxons²⁵. En règle générale, les collectionneurs sont plus étudiés, et donc mieux connus, que les autres acteurs du marché : vendeurs, libraires et experts.

En France, y compris dans les publications universitaires et scientifiques, la parole des marchands d'autographes reste peu disputée. L'ouvrage de référence *Les autographes*²⁶ en est un exemple. S'il contient des contributions d'universitaires renommés tels que François Moureau, il est publié sous la direction d'Alain Nicolas, libraire spécialisé, et reste avant tout

25. On se reportera notamment à A.N.L. Munby, *The Cult of the Autograph Letter in England*, Londres, Athlone press, 1962 et à L.J. Cappon « Walter R. Benjamin and the Autograph Trade at the Turn of the Century », *Proceedings of the Massachusetts Historical Society*, n° 78, 1966, p. 20-37.

26. A. Nicolas (dir.), *Les autographes*, Paris, Maisonneuve & Larose, 1988.

1.2. Le commerce des lettres autographes, un secteur du marché de l'art encore peu étudié

destiné à un public d'amateurs cherchant à se constituer ou à enrichir une collection.

Le marchand d'autographes parisien Thierry Bodin est aussi présent dans les publications et les colloques²⁷, signe de la persistance de la figure du marchand « expert », dans la lignée d'Étienne Charavay (1848-1899), archiviste paléographe et héritier de la librairie Jacques Charavay.

En 1998, Thierry Bodin remarquait que le commerce des autographes et des manuscrits littéraires était restreint et moins bien établi que celui des livres précieux dans la mesure où il ne reposait pas sur des cotes bien définies²⁸. Depuis le début du XXI^e siècle, le marché en France a cependant fluctué et connu une très forte valorisation des autographes. Cette nouveauté a attiré l'attention des chercheurs dans les domaines de l'économie de la culture²⁹ et de la sociologie critique qui identifient la valorisation des objets patrimoniaux comme l'un des traits dominants du capitalisme d'aujourd'hui³⁰.

1.2.2. « Le prix de l'écrit » : l'émergence d'un questionnement interdisciplinaire

Depuis le début de la décennie 2010, un questionnement interdisciplinaire relatif au « prix de l'écrit » s'est développé autour de la problématique suivante : « À quel prix s'échangent sur la longue durée les objets écrits non destinés d'emblée au commerce – actes notariés ou administratifs, manuscrits, documents d'archives, autographes, et autres vieux papiers³¹ ? ». Comme en témoigne la formule de Yann Potin, les documents autographes sont appelés à occuper une place centrale dans ce nouveau champ de recherche. Comme dans l'énumération citée, il est important de distinguer manuscrits et autographes : « [l]a différence réside dans le fait que l'autographe est un manuscrit qui porte les traces de l'intervention physique de l'auteur lui-même³². »

27. Entre autres articles universitaires et chapitres de livres, on peut se reporter à T. Bodin et J. Neefs, « Les autographes. Entretien », *Genesis (Manuscrits-Recherche-Invention)*, n° 7, 1995, p. 177-184 et T. Bodin, « Les grandes collections de manuscrits littéraires » dans *Les ventes des livres et leurs catalogues, XVII^e-XX^e siècle, op.cit.*

28. T. Bodin et J. Neefs, « Les autographes. Entretien », art. cit.

29. I. M. Mendoza, « L'économie du patrimoine écrit : le marché des autographes », thèse de doctorat, Université Paris 1 Panthéon-Sorbonne, 2010. Les principaux apports de la thèse sont synthétisés dans le document de travail rédigé par I. Mendoza Miranda, F. Gardes, X. Greffe et P.-C. Pradier, « Are Autographs Integrating the Global Art Market? The Case of Hedonic Prices for French Autographs (1960-2005) », [halshs-01025095/](https://halshs.archives-ouvertes.fr/halshs-01025095/)

30. L. Boltanski et A. Esquerre, « La « collection », une forme neuve du capitalisme la mise en valeur économique du passé et ses effets », *Les Temps Modernes* n° 679, octobre 2014, p. 5-72 et L. Boltanski et A. Esquerre, *Enrichissement : une critique de la marchandise*. Paris, Gallimard, 2017.

31. Y. Potin, « Le prix de l'écrit », *Genèses*, n° 105/4, novembre 2016, p. 3-7.

32. G. Docquier, « Le document autographe, une « non-réalité » pour l'historien ? », *Le Moyen Âge*, Tome CXVIII, n° 2, août 2012, p. 408.

1.2. Le commerce des lettres autographes, un secteur du marché de l'art encore peu étudié

Le travail de recensement des documents autographes et des manuscrits passant en vente apparaît comme une nécessité pour mener des travaux philologiques. À titre d'exemple, la revue *Recherches sur Diderot et sur l'Encyclopédie* possède une rubrique récurrente intitulée « Documents et autographes ». En plus de reposer sur le dépouillement de catalogues de vente, elle propose dans la mesure du possible des renvois à des catalogues antérieurs. Un rapprochement s'opère aujourd'hui entre les pratiques des philologues pour qui le document autographe est une source fondamentale et celles des historiens qui ont mis plus longtemps à considérer pour lui-même le phénomène d'autographie³³. Au vu de récents travaux publiés par des médiévistes, il est désormais difficile de soutenir que les autographes sont une « non-réalité » pour les historiens : une monographie publiée en Italie en 2014 étudie par exemple la signification et la valeur de l'écriture revendiquée comme autographe entre le XI^e et le XIII^e siècle³⁴.

Les historiens rejoignent donc des chantiers de recherche ouverts depuis plusieurs décennies par les historiens de la littérature dont les travaux ont porté sur le rôle des autographes dans la constitution de l'histoire littéraire³⁵, la représentation littéraire des collections d'autographes³⁶ et les figures d'écrivains collectionneurs de l'écrit³⁷. Ces réflexions prennent désormais un tour plus matériel³⁸ : l'histoire des collections de documents écrits au XIX^e siècle s'est particulièrement développée, en lien avec l'intérêt nouveau pour la question des circulations. Dans ce cadre, les catalogues de vente de lettres autographes, leur structuration et leur mise à disposition de la communauté scientifique ont toute leur pertinence.

1. 3. Les catalogues de vente de lettres autographes : une typologie

Parce qu'ils sont des documents commerciaux à la présentation normalisée, les

33. G. Docquier. « Le document autographe, une « non-réalité » pour l'historien ? », art. cit, p. 408.

34. M. Long, *Autografia ed epistolografia tra XI e XIII secolo : Per un'analisi delle testimonianze sulla « scrittura di propria mano »*, Milan, Ledizioni, 2014.

35. Sur l'importance des documents autographes dans la naissance de l'histoire littéraire universitaire, on consultera L. Fraisse, *Les fondements de l'histoire littéraire : de Saint-René Taillandier à Lanson*, Paris, H. Champion, 2002 et P.-J. Dufief, « Correspondances et histoire littéraire (1850-1900) » dans L. Fraisse (dir.), *L'histoire littéraire à l'aube du XXI^e siècle : Controverses et consensus*, Paris, Presses Universitaires de France, 2005.

36. D. Pety, *Poétique de la collection au XIX^e siècle : Du document de l'historien au bibelot de l'esthète*, Nanterre, Presses universitaires de Paris Nanterre, 2012.

37. M. Le Bail, « L'amour des livres la plume à la main: écrivains bibliophiles du XIX^{ème} siècle », Thèse de doctorat, Université Toulouse-Jean Jaurès, 2016 et M. Charreire, « Un marchand d'histoire au XIX^{ème} siècle », *Genèses*, n° 105/4, novembre 2016, p. 36-56.

38. N. Preiss (éd.), *Le XIX^e siècle à l'épreuve de la collection*, Reims, ÉPURE, 2018.

1. 3. Les catalogues de vente de lettres autographes : une typologie

catalogues de vente de lettres autographes semblent se prêter à un travail de mise en série. Avant de s'interroger sur les modalités techniques de leur traitement automatisé grâce aux technologies numériques, il convient d'établir qu'ils forment une base suffisamment homogène pour faire corpus.

La tâche n'a rien d'évident car il n'existe pas dans les grandes institutions patrimoniales françaises de fonds uniquement consacrés aux catalogues d'autographes. La sous-série AB XXXVIII des Archives nationales en témoigne. Intitulée « Collection des catalogues de vente d'autographes & livres anciens imprimés », elle couvre un vaste champ concernant des ventes de lettres autographes mais aussi d'éditions originales, de documents historiques non signés, d'objets d'art, etc. Le contenu mixte de beaucoup de catalogues empêche de classer thématiquement et de repérer à la première lecture les catalogues contenant des autographes. Répertorier dans un tel fonds tous les catalogues partiellement consacrés à des lettres autographes suppose la consultation directe des sources, actuellement non numérisées.

Dans le cadre de mon stage, seuls des catalogues dédiés exclusivement à la vente de lettres autographes ont été mobilisés. Le critère d'identification est nominal, c'est-à-dire qu'il repose sur la présence du mot « autographes » dans le titre du catalogue.

1.3.1. Catalogues d'enchères et catalogues à prix marqués : le reflet d'un marché dual

Une fois cette restriction posée, une certaine diversité demeure au sein des catalogues de vente de lettres autographes. Pour en établir un premier classement, il est pertinent de s'inspirer des remarques formulées par Nicole Masson à propos des catalogues de bibliothèques vendues à Paris dans la deuxième moitié du XVIII^e siècle³⁹. À partir d'une soixantaine d'imprimés déposés à la Bibliothèque nationale de France, elle propose « une sorte de double typologie », entre « vente en bloc de l'ensemble de la bibliothèque ou vente en détail des livres, à l'unité ou par petits lots » et « vente à l'amiable à prix marqués ou vente aux enchères ». Dans le domaine des autographes également, il est courant de vendre aux enchères les plus belles et larges collections d'autographes et de vendre au détail et à prix marqués des pièces plus éparses.

Qu'ils concernent des ventes aux enchères de collections ou du commerce de détail à prix marqués, les catalogues d'autographes présentent une organisation comparable. Les entrées décrivant les autographes à vendre sont ordonnées, la plupart du temps, par ordre

39. N. Masson, « Typologie des catalogues de vente », dans *Les ventes des livres et leurs catalogues, XVII^e-XX^e siècle, op. cit.*, p. 117-129.

1. 3. Les catalogues de vente de lettres autographes : une typologie

alphabétique ou par ordre thématico-alphabétique. Elles commencent systématiquement par un nom de personne ou d'organisation. Le caractère initial du nom est renforcé par une mise en valeur typographique elle aussi systématique, dont les modalités peuvent varier d'un catalogue à l'autre (gras, petites majuscules, etc.).

AUTOGRAPHES A PRIX MARQUÉS

MAISON Gabriel CHARAVAY

Dirigée par **Eugène CHARAVAY Fils**, expert en autographes,

8, QUAI DU LOUVRE, A PARIS.

Abréviations : L. a. s., lettre autographe et signée; Acad. fr., Académie française; Sousc., compliment terminant la lettre; P., page.

- 1 **Agassiz** (Louis), naturaliste Suisse, célèbre par ses travaux sur les glaciers. — L. a. s., en anglais, au professeur Silliman; Boston, 1847, 1 p. 1/2 in-4. Cachet. Relative à ses lectures sur les glaciers. 3 »
- 2 **Aguesseau** (H.-Cardin-J.-B. d'), philanthrope, de l'Acad. fr., petit-fils du chancelier, né à Paris. — L. a.s.; Fresnes, 1822, 1 p. in-4. 3 »

Figure 1 – Extrait d'une liste de lettres autographes à prix marqués
(*Revue des Autographes* n° 59)

Cette emphase portée sur le nom propre était l'idée que, dans la salle de ventes comme dans la librairie, c'est plus souvent le nom que le contenu du document qui fait vendre. Il est d'ailleurs remarquable que les noms placés à l'initiale ne correspondent pas toujours au scripteur ni au signataire de la lettre.

Il existe même, dans les catalogues de libraires ou les revues spécialisées, des encarts publicitaires consistant seulement en des listes d'auteurs dont il est possible de se procurer des cartes ou lettres autographes. Dans ce cas précis, le support vendu qui ne porte qu'une signature n'est pas décrit.

LETTRES AUTOGRAPHES A 2 Fr. PIÈCE, AU CHOIX.

Membres de l'Académie française. Guizot, Martin (H.), Ponsard, Sardou (V.), Sicard (l'abbé), Arnault (V.), Augier (Em.), Bausset (le cardinal), Sainte-Beuve, Berryer, Jouy (de), Mignet, Noailles (le duc de), Scribe, Vitet, Droz (Jos.), Michaud, Royer-Collard, Nizard (D.), Ampère (J.-J.), Falloux (de), Sandeau (J.), Dufaure, Dumas (J.-B.), Fontanes, François de Neuchâteau, Laya, Legouvé (Ern.), Pasquier (le duc).

Littérateurs. Deschamps (Em.), Monteil (A.-A.), Baschet (Arm.), Esquiros (Alph.), Sue (Eug.), Karr (Alph.), Reybaud (L.), Gozlan (L.), Quinet (Edg.), Chennedollé, Méry, Lamotte-Langon, Brucker (R.), Roger de Beauvoir, Gouffé (Arm.).

Musiciens. Berlioz, Adam (Ad.), Caraffa, Maillard (Aimé), Kreutzer (L.), Paër, Pradher, Rigel (H.-J.), Seligmann, Zimmermann, Anders (G.-F.), Henrion (P.), Panseron, Reicha, Labarre.

Acteurs, actrices, chanteurs, cantatrices. Déjazet, Agar (M^{lle}), Bouffé, Bloch (Rosine), Miolan-Carvalho (M^{me}), Pierson (Blanche); Viardot (Pauline), Laurent (Marie), Duprez (G.).

Divers. Les peintres Schnetz et Giraud; David d'Angers; Benj. Constant, Dupont de l'Eure, Cantagrel, Orfila; M^{mes} Desbordes-Valmore, Dorval.

Figure 2 – Liste de lettres autographes non assorties de leur description
(extrait de la *Revue des Autographes* n° 55)

1.3.2. Les revues-catalogues : des supports commerciaux hybrides

Les catalogues de documents autographes les plus connus, réputés et consultés sont ceux qui inventorient les trois plus riches collections constituées dans la deuxième moitié du XIX^e siècle par Benjamin Fillon (1819-1881)⁴⁰, Alfred Bovet (1841-1900)⁴¹ et Alfred Morrison (1821-1897)⁴². Ces catalogues « demeurent comme des ouvrages de référence, à la fois pour les fac-similés qui les illustrent et pour les descriptifs et commentaires développés

40. *Inventaire des autographes et des documents historiques composant la Collection de M. Benjamin Fillon...*, 5 volumes, Paris / Londres, E. Charavay / F. Naylor, 1877-1883.

41. É. Charavay et F. Calmettes, *Lettres autographes composant la collection de M. Alfred Bovet*, deuxième édition corrigée et augmentée, Paris, Charavay frères, 1887 [1884-1885].

42. *The collection of autograph letters and historical documents formed by Alfred Morrison. Second series, 1882-1893.* 3 vol., Londres, 1893.

1. 3. Les catalogues de vente de lettres autographes : une typologie

qui accompagnent les fiches⁴³. ».

Toutes les collections d'autographes dispersées n'ont pas fait l'objet d'un travail d'analyse et d'érudition. Lors de la plupart des ventes aux enchères, le catalogue est avant tout un « ouvrage de « consommation courante », [...] une sorte de prospectus commode qui permet à l'acheteur potentiel de se faire à l'avance une idée de la collection vendue, puis de suivre aisément le déroulement de la vente en y portant éventuellement des annotations⁴⁴. ».

À l'autre bout du spectre, il existe des brochures commerciales ou des revues périodiques destinées à promouvoir les autographes en vente à prix marqués dans les librairies spécialisées. De tels supports sont édités et publiés par les libraires eux-mêmes qui les adressent à leurs abonnés et leurs meilleurs clients. Si ces publications périodiques ne se présentent pas sous forme d'inventaires exhaustifs et mêlent aux listes d'autographes articles et publicités, elles méritent d'être intégrées à notre corpus pour rendre compte de toute la diversité du marché parisien au XIX^e siècle.

Pour compléter depuis Bruxelles sa collection de manuscrits et d'autographes et mener à bien son travail d'historien et d'archiviste du romantisme⁴⁵, l'érudit belge Charles de Spoelberch de Lovenjoul (1836-1907) ne se limitait pas à la consultation des catalogues de vente aux enchères les plus complets et les mieux édités. Il collectionnait également les bulletins et les périodiques plus modestes que lui adressaient les libraires parisiens⁴⁶. L'ensemble de catalogues légué à l'Institut de France en même temps que la collection de manuscrits littéraires de Lovenjoul montre que les collections les plus prestigieuses se sont en partie constituées à partir des brochures les plus modestes et publicitaires.

1.3.3. L'activité éditoriale de la famille Charavay : un exemple de diversification des catalogues

Active pendant plusieurs générations dans les mondes lyonnais et parisien de la librairie, la famille Charavay n'a pas fait à ce jour l'objet d'une étude monographique. Ses membres ont été davantage étudiés pour leur implication dans le milieu communiste ou leur

43. T. Bodin, « Les grandes collections de manuscrits littéraires » dans *Les ventes de livres et leurs catalogues*, *op. cit.*

44. N. Masson, « Typologie des catalogues de vente », dans *Les ventes des livres et leurs catalogues*, *op. cit.*

45. C. Faivre d'Arcier, *Lovenjoul (1836-1907): une vie, une collection*. Paris, Kimé, 2007.

46. C. Faivre d'Arcier, « Lovenjoul et ses catalogues au cœur d'un service de commissions » dans *Le Livre entre le commerce et l'histoire des idées : Les catalogues de libraires (XV^e-XIX^e siècle)*, Paris, Publications de l'École nationale des chartes, 2011.

1. 3. Les catalogues de vente de lettres autographes : une typologie

œuvre d'historien que pour leur activité commerciale. C'est Étienne Charavay, archiviste paléographe et biographe de Lafayette qui est de nos jours le plus cité et étudié, notamment pour son rôle d'expert pendant l'Affaire Dreyfus⁴⁷.

Si le détail de l'activité marchande des Charavay reste relativement mal connu, la prééminence et le sérieux de cette dynastie de libraires dans le commerce des autographes sont largement établis. En 1865, dans son ouvrage sur les autographes, Adolphe de Lescure déclare déjà : « Les affaires de ce genre [le commerce d'autographes], celles du moins que ne sauraient effleurer nos reproches, sont aujourd'hui heureusement concentrées entre les mains de MM. Laverdet et Charavay aîné⁴⁸. ».

L'activité de cette famille de libraires se trouve redoublée cette même année 1865 par l'établissement de la librairie de Gabriel Charavay (1818-1879). Ce dernier rachète le cabinet d'autographes du libraire Auguste Laverdet (1805-1865) alors installé 50 rue Saint-André-des-Arts. Son frère Jacques Charavay dit l'Aîné (1809-1867), ancien huissier, a lui fondé dès 1830 une librairie à Lyon qu'il a transportée à Paris vers 1843.

Selon la version en ligne du Maitron, dictionnaire biographique du mouvement ouvrier, c'est la volonté de rédiger des « notices sur des personnages de la Révolution » qui a amené Gabriel Charavay, ouvrier bonnetier de son état, à acquérir des brochures et des manuscrits de la période révolutionnaire⁴⁹, avant de devenir un expert des autographes et de s'établir comme marchand de livres et de documents.

Son activité commerciale est complétée par de nombreuses publications consacrées aux autographes. Sa maison édite et diffuse entre 1866 et 1936 un catalogue mensuel intitulé *Revue des Autographes, des curiosités de l'histoire et de la biographie*. De 1879 à 1892 son fils Eugène en assure la rédaction. Lui succèdent sa veuve entre 1892 et 1918, puis sa fille Gabrielle.

Pour promouvoir son activité, la librairie Jacques Charavay reprend la publication d'une autre revue initiée par Gabriel en 1862 : *L'Amateur d'autographes*. La réputation de cette revue mensuelle grandit encore quand elle passe sous la direction d'Étienne, fils, successeur de Jacques et ancien élève de l'École des chartes. Dans une brochure éditée à sa mort, son ami et collaborateur Maurice Tourneux (1849-1917) regrette l'irrégularité des livraisons de *L'Amateur d'autographes* qui s'explique par la fréquence des missions confiées à Étienne du fait de sa réputation d'expert en autographes⁵⁰. Comme en atteste sa bibliographie posthume et

47. T. Ribémont, « Les historiens chartistes au cœur de l'affaire Dreyfus », *Raisons politiques*, n° 18-2, 2005, p. 97-116.

48. A. de Lescure, *Les Autographes et le Goût des autographes*, Paris, J. Gay, 1865, p. 52.

49. Maitron en ligne, CHARAVAY frères [CHARAVAY Gabriel et CHARAVAY Jean], notice revue, corrigée et complétée par Jacques Grandjonc, version mise en ligne le 20 février 2009, dernière modification le 17 décembre 2018, disponible à l'adresse suivante : <http://maitron-en-ligne.univ-paris1.fr/spip.php?article28465>.

50. M. Tourneux, *Étienne Charavay : sa vie et ses travaux*, Paris, E. Charavay, 1900, p. 6.

1. 3. Les catalogues de vente de lettres autographes : une typologie

sélective rédigée par Tourneux⁵¹, Étienne Charavay s'est chargé de la rédaction de nombreux catalogues de vente pour préparer la dispersion des plus prestigieuses collections d'autographes de son temps.

Grâce à leur double statut d'experts et de libraires, les générations successives de la famille Charavay ont été en position de constituer tous les types de catalogues mentionnés précédemment, de la somme érudite monumentale à la simple brochure commerciale⁵². Les catalogues et les périodiques produits par les deux librairies Charavay sont en partie conservés par la Bibliothèque nationale de France et les Archives nationales. Ces sources sont couramment mais indirectement sollicitées par les philologues, les éditeurs scientifiques et les biographes. Au lieu de consulter directement les catalogues, ces chercheurs ont pris l'habitude de s'appuyer sur le « fichier Charavay » mis à disposition par chacune des deux institutions.

Le Département des Manuscrits de la Bibliothèque nationale de France a hérité en 1939 du « fichier Charavay » constitué au cours de sa vie par le célèbre bibliophile et érudit Seymour de Ricci (1881-1942)⁵³. Composé de 183 classeurs regroupant des notices découpées dans les catalogues Charavay et classées alphabétiquement, le fichier « permet de retrouver la trace de documents désormais épars⁵⁴ » mais sans référence ou accès au catalogue d'origine.

La sous-série AB XXXVIII des Archives nationales possède elle aussi un « fichier Charavay » établi au cours de la seconde moitié du XIX^e siècle en découpant des notices aujourd'hui conservées dans 51 boîtes. Contrairement au fichier consultable au Département des Manuscrits de la Bibliothèque nationale de France, la référence du catalogue d'origine figure sur la plupart des fiches disponibles aux Archives nationales. Ce fichier a été racheté au libraire Degrange en même temps que des milliers de brochures de vente d'autographes en 1961 et un index manuscrit en a été établi dès 1962.

La dispersion et la connaissance imprécise du contenu des catalogues de vente laissent ces sources partiellement inexploitées, ce qui laisse à penser que de nombreux fac-similés ou transcriptions de documents manuscrits disparus ou conservés dans des fonds privés peuvent s'y trouver à l'insu des chercheurs. L'intérêt d'un dépouillement approfondi des catalogues de vente d'autographes et de documents historiques peut aussi aider à une meilleure connaissance de fonds d'archives aujourd'hui dispersés et à la redécouverte de documents par les historiens qui n'en ont pas connaissance faute d'inventaires spécifiques.

51. M. Tourneux, *Étienne Charavay...*, *op. cit.*, p. 10-43.

52. Un panorama de l'ensemble des publications de la première génération des Charavay est disponible dans A. de Lescure, *Les Autographes et le Goût des autographes*, *op. cit.*, p.53.

53. J. Porcher, « A la Bibliothèque nationale : le legs Seymour de Ricci », *Bibliothèque de l'École des chartes* 105-1, 1944, p. 229-33.

54. *Ibidem*, p. 230.

1. 3. Les catalogues de vente de lettres autographes : une typologie

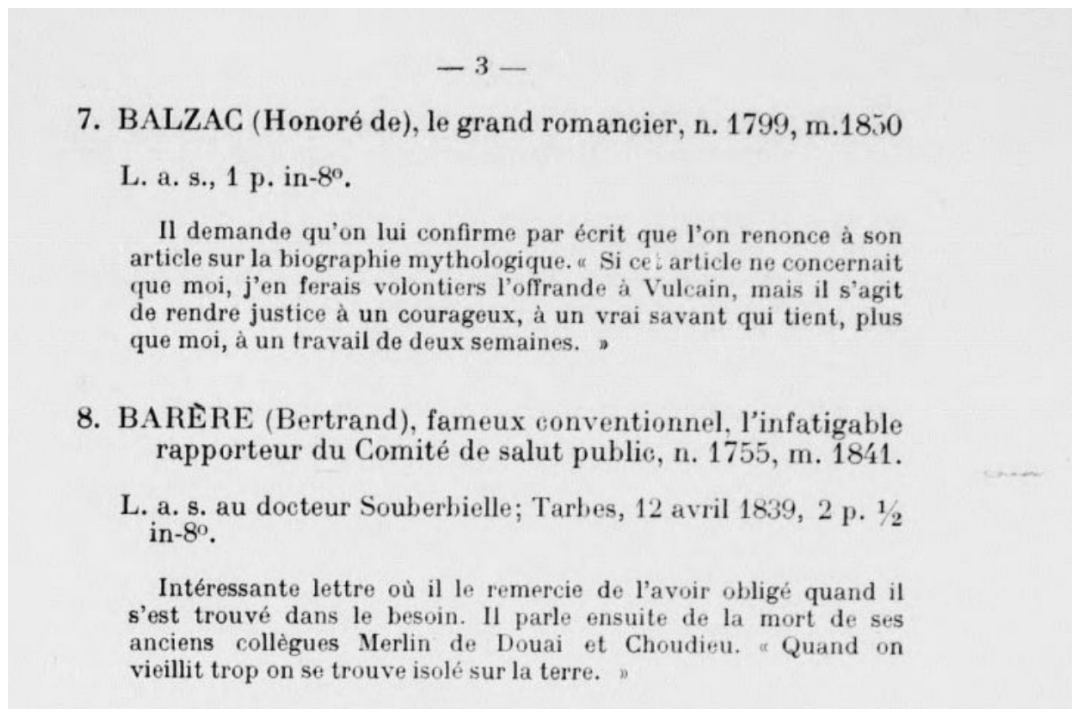


Figure 3 – Extrait d'un catalogue de vente aux enchères rédigé par Noël Charavay
(*Catalogue d'une intéressante collection de lettres autographes...* vente du 14 décembre 1908)

En 1965, Michel Nortier, conservateur à la Bibliothèque nationale de France, s'attelle à reconstituer l'histoire complexe des archives de la Chambre des Comptes de Paris⁵⁵. Pour localiser tout ou partie de ce fonds ayant connu plusieurs vagues de dispersion successives depuis le XVIII^e siècle, il préconise le recours systématique aux catalogues de vente d'autographes édités après 1840. Cette date est retenue car elle correspond d'abord à l'essor et à la professionnalisation des libraires vendeurs d'autographes, ensuite parce que le développement d'un marché privé de l'autographe a conduit à une dispersion accrue des fonds. S'il rappelle que le contenu des catalogues de vente doit être traité avec prudence, Michel Nortier va jusqu'à affirmer que leur dépouillement exhaustif « compenserait d'une certaine façon le sort funeste qui a mené à leur disparition, sinon à leur perte définitive, des documents qui avaient pourtant traversé sans inconvénient plusieurs siècles de notre histoire⁵⁶ ».

Des appels au dépouillement à grande échelle des catalogues de vente d'autographes sont aussi formulés dans les champs de la recherche littéraire et de la conservation du

55. M. Nortier, « Le sort des archives dispersées de la Chambre des Comptes de Paris », *Bibliothèque de l'Ecole des chartes*, n° 123/2, 1965, p. 460-537.

56. *Ibidem*, p. 487.

1. 3. Les catalogues de vente de lettres autographes : une typologie

patrimoine. L'inventorisation manuelle des catalogues est dès le XIX^e siècle conçue comme le principal antidote à la déprédation des dépôts publics encouragée par la subite extension du marché des autographes à partir de 1835⁵⁷. Elle aboutit à la publication des deux volumes du *Dictionnaire de pièces autographes volées aux bibliothèques publiques...* chez Panckoucke⁵⁸. Cette entreprise est cependant limitée du fait de la disproportion entre les moyens humains et techniques disponibles et l'immensité du matériau documentaire à traiter. Cent soixante-dix ans plus tard, les technologies numériques offrent une possibilité de remédier à la sous-exploitation des catalogues de vente en mettant en place des systèmes de structuration et d'indexation automatique de ces précieuses sources historiques.

57. A. de Lescure, *Les Autographes et le Goût des autographes*, op. cit., p. 8.

58. L. Lalanne et H. Bordier, *Dictionnaire de pièces autographes volées aux bibliothèques publiques de la France, précédé d'observations sur le commerce des autographes*, Paris, Panckoucke, 1851 et L. Lalanne et H. Bordier, *Dictionnaire de pièces autographes volées aux bibliothèques publiques de la France précédé d'observations sur le Commerce des autographes ... 3e et 4e livraisons*, Paris, Panckoucke, 1853.

Deuxième partie

Du catalogue à la base de données : méthodologie et évaluation d'une chaîne de traitement

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

Les débouchés scientifiques de la construction de bases de données recensant le contenu des catalogues de vente d'autographes peuvent être multiples. Proposer une plus grande accessibilité de ces sources serait utile pour l'établissement d'éditions critiques de textes autographes et manuscrits, en permettant par exemple d'avoir connaissance de documents inédits ou de fac-similés inexploités. La mise en ligne pourrait aussi être un point d'appui à des recherches historiques ou philologiques nouvelles. Il serait ainsi possible de mobiliser cette base pour étudier sur le long terme les pratiques commerciales des vendeurs d'autographes ou pour tenter de déterminer de nouveaux moyens d'authentification des documents manuscrits et autographes.

L'enjeu technique principal de mon stage était de structurer de manière la plus automatique possible le contenu des catalogues. Les outils de reconnaissance optique de caractères sont aujourd'hui extrêmement performants mais ne peuvent pas constituer à eux seuls un traitement informatique complet et satisfaisant de nos catalogues dans lesquels la structure de l'information est aussi significative que le contenu des notices.

L'attention portée à la reconnaissance et à la restitution par des moyens numériques de cette structure explique le choix d'encoder les métadonnées et le texte des catalogues de vente dans le standard XML-TEI⁵⁹. Développée depuis 1987 par un ensemble de membres de la communauté scientifique internationale, la *Text Encoding Initiative* vise à définir un langage commun pour partager et mutualiser les textes encodés numériquement. Les documents encodés dans ce format permettent de rendre compte des particularités structurelles du document de départ, car les balises ajoutées au texte ne sont pas, contrairement au langage HTML, de simples indications de mise en page. L'usage de ces balises, appelées « éléments » dans la terminologie de la TEI, est décrit en détail dans les « Recommandations pour l'encodage et l'échange de textes électroniques » disponibles en ligne⁶⁰. Chacun des éléments possède un sens précis et son contenu sémantique peut être encore affiné par l'ajout d'attributs dont l'usage est tout aussi réglementé.

Les documents au format XML-TEI ont l'avantage d'être facilement interrogeables par

59. Pour une présentation complète des principes et des avantages de ce format standardisé d'encodage des documents textuels, se reporter à L. Burnard, *What is the Text Encoding Initiative?*, Marseille, OpenEdition Press, 2014.

60. Intitulée P5, la version courante des recommandations de la TEI est disponible à l'adresse suivante : <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/index.html>.

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

différents biais : requêtes dans des langages eux-même standardisés (XQuery et XPath) ou consultation facilitée par le développement d'une interface graphique. Recourir à un langage informatique maintenu et utilisé par une communauté scientifique active est aussi une garantie de pérennité et simplifie l'échange d'informations ou l'intégration des catalogues de vente à d'autres projets et bases de données.

L'un des enjeux de mon stage était de proposer une chaîne de traitement efficace des catalogues de vente d'autographes numérisés. Schématisée ci-dessous, elle fait appel à plusieurs logiciels dont Transkribus, logiciel de reconnaissance automatique de texte, et GROBID-dictionaries. Nous détaillerons leurs propriétés, les raisons de leur utilisation et leurs apports à notre démarche dans la suite de ce deuxième chapitre. Afin de garantir la qualité et l'exploitabilité des fichiers transcrits et structurés automatiquement et d'en faciliter la correction manuelle, un ensemble de transformations grâce à des feuilles de style XSL et des schémas de contrôle de la structure des fichiers ont été mis en place.

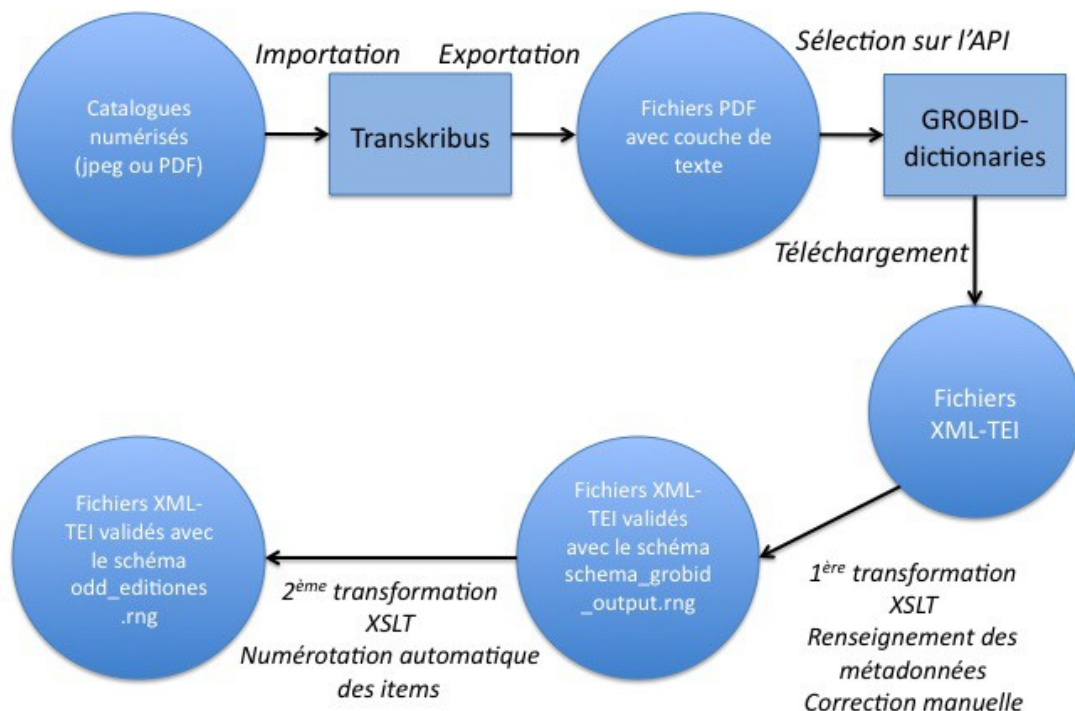


Figure 4 – Représentation schématique de la chaîne de traitement

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

2.1.1. Un traitement qui se concentre sur les entrées de catalogues

Le but de la chaîne de traitement présentée dans ce mémoire n'est pas de constituer des éditions scientifiques des catalogues et des revues-catalogues. En plus du fait qu'il existe un relatif retard de la réflexion sur l'encodage XML-TEI des périodiques, les projets en cours privilégient la visualisation des revues plutôt que la structuration de leur contenu. Or, dans le cas des catalogues produits par les libraires, il importe moins de mettre en avant la production éditoriale des éditeurs que de donner accès aux chercheurs à l'ensemble des ventes qu'ils ont proposées ou réalisées. En effet, la nature et le degré de structuration des informations présentes dans les catalogues de vente, en particulier dans les revues-catalogues à prix marqués, sont beaucoup trop disparates pour ne pas faire le choix d'écarter un certain nombre de pages de la chaîne de traitement automatique. C'est le cas notamment des pages de garde ou de celles qui présentent des articles ou des réclames.

La relative homogénéité du matériau à encoder est l'une des conditions d'efficacité de la structuration automatique des documents et l'un des prérequis à la construction d'une base de données dédiée. Le choix a été fait de concentrer les efforts sur les entrées de catalogues, données traitables automatiquement et uniformément. Sont exclus de la chaîne de traitement automatique les articles et listes non structurées par entrées. Cela n'empêche pas que notre chaîne de traitement s'applique à des catalogues de vente aux enchères aussi bien qu'à des catalogues à prix marqués.

Ce parti-pris ne signifie pas que toutes les informations présentes hors des entrées sont perdues. Dans le cas des catalogues de vente aux enchères, où les informations extérieures concernent le plus souvent les modalités de la vente, elles sont reportées manuellement et de façon détaillée dans le <teiHeader> de chaque fichier numérique dérivé d'un catalogue. Il en va de même pour les informations bibliographiques relatives à chaque catalogue imprimé.

2.1.2. Principes et choix de l'encodage manuel

Le traitement automatique des catalogues en vue de leur mise en base de données est solidaire d'un travail de réflexion sur la nature du texte encodé. Dans le cas présent, la décision a été prise d'encoder le contenu commun à tous les exemplaires d'un même catalogue plutôt que de rendre compte des spécificités de l'exemplaire numérisé auquel est appliquée la chaîne de traitement. Exprimé dans les termes du modèle conceptuel FRBR

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

(*Functional Requirements for Bibliographic Records* ou « spécifications fonctionnelles des notices bibliographiques »), cela revient à encoder les catalogues comme manifestation et non comme item.

Les annotations manuscrites, souvent des prix inscrits en marge dans les catalogues de vente aux enchères, sont en effet difficilement traitables automatiquement : la main change d'un catalogue à l'autre et chacun nécessiterait le développement de modèles spécifiques de reconnaissance optique de caractères. Quand elles existent, les annotations méritent également d'être recoupées et complétées par d'autres sources documentaires (autres exemplaires annotés, comptes-rendus de la vente, procès-verbal des commissaires-priseurs). Le but est donc d'encoder automatiquement les informations présentes dans tous les exemplaires d'un catalogue et de réserver, si besoin, la possibilité d'un travail de saisie manuelle pour les exemplaires annotés.

Du fait des difficultés à entraîner la transcription automatique des pages de garde (moindre structuration et moindre stabilité de la mise en page, présence d'ornements ou de variations typographiques qui sont autant de bruit pour les logiciels de reconnaissance automatique de texte), les informations présentes sur cette page sont consignées à la main dans l'élément <teiHeader>.

Parce que nous avons fait le choix d'encoder des manifestations d'œuvres et non des items, l'élément <sourceDesc> se compose en premier lieu d'un élément <bibl> contenant une référence bibliographique à la structure relativement souple. Ainsi, il est possible de décrire les différents catalogues, quelle que soit leur forme et sans trop de contraintes. La possibilité de multiplier les éléments <publisher> est importante. D'un point de vue de l'histoire des circulations de manuscrits et du marché des autographes, il est intéressant de savoir si les catalogues sont édités simultanément dans différents pays. Ce sont des indices précieux quant au degré et au rayon de publicisation d'une vente aux enchères.

```
<sourceDesc>
  <bibl>
    <title>Précieux autographes composant la Collection du Président Robert Schuman. Première partie</title>
    <publisher>Michel Castaing</publisher>
    <pubPlace>3 rue Furstenberg, Paris</pubPlace>
    <publisher>Pierre Cornuau</publisher>
    <pubPlace>22 rue Laffitte, Paris</pubPlace>
    <date>1965</date>
  </bibl>
```

Figure 5 – Extrait du <teiHeader> décrivant le catalogue de la vente préparant la dispersion de la collection d'autographes de Robert Schuman en 1965

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

Toujours dans le cas d'une vente aux enchères, le catalogue imprimé renvoie à une vente avec laquelle il ne saurait se confondre. Par exemple, la date de publication d'un catalogue et la date effective des vacations de la vente aux enchères diffèrent nécessairement. Cependant, les informations données par la page de garde concernent avant tout la vente effective, et c'est cette dernière qui est la plus instructive si l'on a pour objectif d'étudier la circulation des manuscrits et des autographes.

Le deuxième élément composant la section <sourceDesc> est donc l'élément <listEvent> qui permet de lister l'ensemble des événements en lien avec le catalogue. Après analyse du déroulé des ventes, les sources complémentaires susceptibles de compléter le catalogue sont de quatre types, qui correspondent à autant d'attributs associés à un élément <event>.

```
<listEvent>
  <event type="auction">
    <!-- ab? -->
    <p>
      <address>
        <addrLine>Hôtel des commissaires-priseurs</addrLine>
      </address>
      <persName type="auctioneer">Me R.-G. Boisgirard</persName>
      <persName type="expert">Michel Castaing</persName>
      <persName type="expert">Pierre Cornuau</persName>
      <persName type="collector" ref="#PE2_000012">Robert Schuman</persName>
      <date from="1965-03-04" to="1965-03-05">les 4 et 5 mars 1965.</date>
    </p>
  </event>
  <event type="report">
    <p>
      <bibl>«La Bibliothèque et la Collection Robert Schuman »,
        <title>Bulletin de la Société de l'Histoire du Protestantisme Français (1903-2015)</title>,
        Vol. 111 (Juillet-Août-Septembre 1965), pp. 250-255.
        <ptr target="https://www.jstor.org/stable/24292644"/>
      </bibl>
    </p>
  </event>
</listEvent>
```

Figure 6 – Extrait du <teiHeader> décrivant les modalités de la vente préparant la dispersion de la collection d'autographes de Robert Schuman en 1965

Quatre attributs différents peuvent être utilisés pour compléter l'élément <event> :

- En premier lieu, la **vente** (*sale*) elle-même. Un événement de ce type est décrit en listant les lieux, les organisations et les personnes intervenant dans cette vente. Pour garantir

2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés

l'uniformité de la saisie d'un catalogue à l'autre, le typage des éléments <persName> correspond à une liste fermée : *auctioneer* (commissaire-priseur), *collector* (collectionneur), *expert* (« expert » qui garantit l'authenticité des documents), *sales assistant* (dans les catalogues de vente aux enchères du XIX^e siècle, les libraires qui sont mentionnés sans référence à un statut d'expert) et *seller* (si la personne à l'origine de la vente est distincte du collectionneur) ;

- deuxièmement, les **comptes-rendus** (*record*) dont la vente a pu faire l'objet. C'est en effet souvent dans la presse spécialisée que les résultats des enchères ont été enregistrés et rendus publics pour les amateurs d'autographes. Consulter et reproduire une telle source permet d'évaluer, s'il existe, l'exactitude du prix mentionné de façon manuscrite sur les exemplaires de catalogues ;

- troisièmement, les **procès-verbaux** (*minutes*) produits par les commissaires-priseurs et remis aux autorités. Dans le cas des ventes aux enchères organisées à Paris, les procès-verbaux sont déposés aux Archives de Paris. Ces documents sont une autre source de connaissance des prix atteints lors d'une vente aux enchères. Ils permettent par ailleurs de mettre en perspective les catalogues, sources partielles, parfois manipulées pour des raisons commerciales ou ne rendant pas toujours compte de tous les lots et pièces proposés à la vente ;

- enfin, le dernier type d'événement relève de l'**identification** (*identification*). Un grand nombre de ventes aux enchères se conduisent sans que l'identité du vendeur ne soit révélée. L'identification a posteriori du propriétaire et du vendeur d'une collection dispersée est un événement à part entière. Les sources les plus fréquemment utilisées pour ce genre d'attributions sont les deux volumes du *Dictionnaire des pièces volées dans les bibliothèques publiques* de Lalanne et Bordier⁶¹, ainsi que le contenu éditorial de *L'Amateur d'autographes*.

Si plusieurs exemplaires du même catalogue ont été consultés pour compléter la version numérique du catalogue, la référence précise de chacun d'entre eux est enregistrée dans un élément <witness>, lui-même contenu dans un élément <listWit>.

Le reste des informations encodées correspond aux différentes entrées du catalogue. Elles sont traitées et structurées automatiquement grâce au logiciel GROBID-dictionaries et leur structure sera exposée ci-dessous (voir section 2.2.2.).

61. L. Lalanne et H. Bordier, *Dictionnaire de pièces autographes volées aux bibliothèques publiques de la France, précédé d'observations sur le commerce des autographes*, op. cit., 1851 et 1853.

2.2. Présentation détaillée et évaluation des outils retenus

2.2.1 La reconnaissance automatique de caractères : Transkribus

Un préalable au traitement des catalogues numérisés au cours de mon stage était que les images numériques correspondantes soient transformées en fichiers PDF recouverts d'une couche de texte. Pour effectuer cette transformation, mon référent de stage Simon Gabay avait décidé de recourir au logiciel Transkribus qui permet à la fois la reconnaissance optique de caractères (en anglais, *optical character recognition* ou OCR) et la reconnaissance de texte manuscrit (en anglais, *handwritten text recognition* ou HTR).

Quoique les catalogues étudiés soient des imprimés, recourir à la fonction HTR de Transkribus plutôt qu'à la fonction OCR s'est révélé plus efficace, du fait de la présence d'effets typographiques diversifiés et de la qualité parfois médiocre des numérisations. À titre d'exemple, les noms propres imprimés en gras en début de notice n'étaient pas correctement reconnus par le modèle OCR standard de Transkribus. Développer un modèle HTR spécifique a permis de contourner ce problème et de garantir l'exactitude de la retranscription des noms propres, élément impératif pour assurer la qualité de la base de données finale.

Chaque catalogue destiné à intégrer cette base de données était l'objet d'un traitement en deux temps dans Transkribus. Sa mise en page était automatiquement détectée par la fonction consacrée du logiciel puis la transcription était effectuée en s'appuyant sur un modèle HTR développé avant le début de mon stage. Le résultat était ensuite exporté au format PDF et couche de texte avant d'être structuré automatiquement par GROBID-dictionaries.

Si le traitement des catalogues par Transkribus s'est révélé très satisfaisant sur le plan de la retranscription, deux limites sont néanmoins à souligner. Tout d'abord, la version actuelle du logiciel ne permet pas d'indiquer automatiquement dans le résultat exporté quels sont les passages imprimés en gras et en italique. Le premier cas n'est pas problématique dans la mesure où la mise en gras du nom correspond à une volonté, déjà notée plus haut, de souligner ce qui fait vendre : le nom de l'auteur de l'autographe (ou plus rarement de son destinataire). La perte des informations en italique est plus gênante dans la mesure où elle correspond à une perte sèche d'informations et ne permet plus de distinguer les titres d'œuvres mentionnées du reste du contenu textuel.

2.2. Présentation détaillée et évaluation des outils retenus

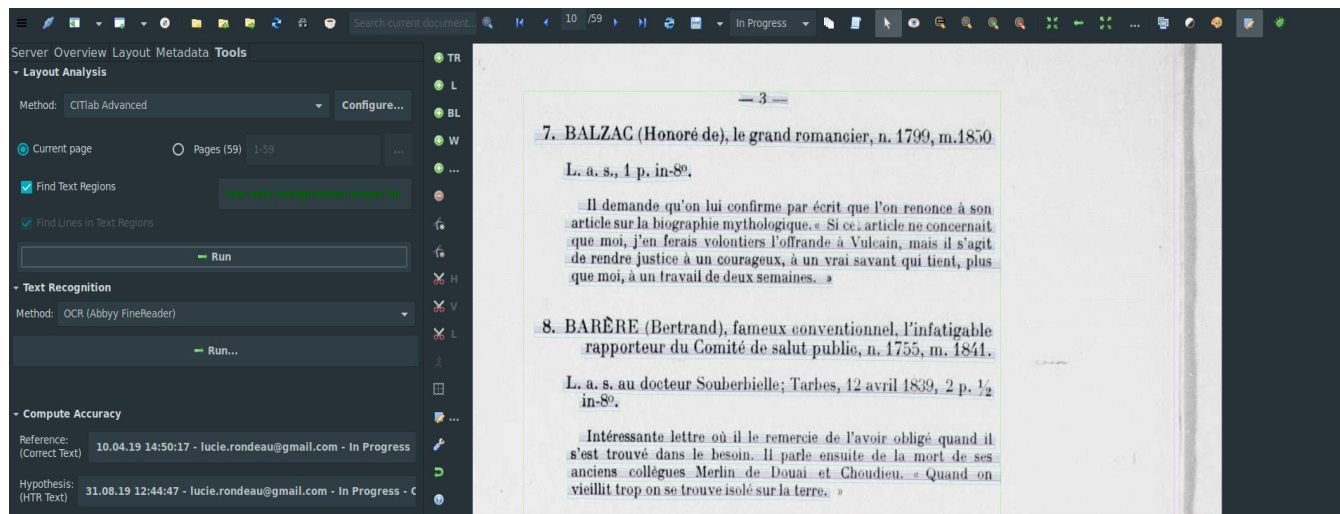


Figure 7 – Capture d'écran du logiciel Transkribus : l'analyse de la mise en page est la première étape de traitement des catalogues de vente

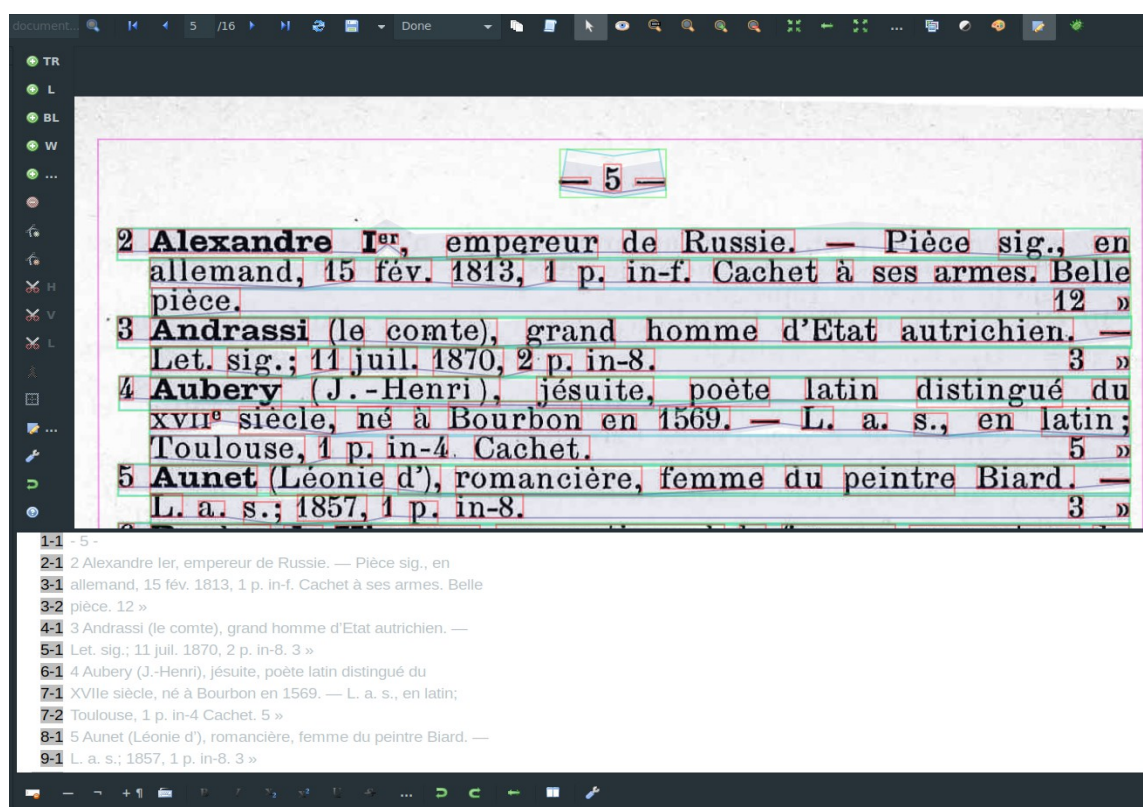


Figure 8 – Capture d'écran du logiciel Transkribus : après application du modèle HTR, le logiciel affiche la transcription réalisée en regard du catalogue numérisé

2.2. Présentation détaillée et évaluation des outils retenus

De plus, il est possible qu'à l'avenir, les services offerts par Transkribus (transcription automatiques des documents et production de modèles HTR) deviennent payants. Pour ne pas dépendre d'un logiciel dont le coût d'utilisation sur le moyen terme est inconnu, un nouveau modèle destiné à la transcription automatique des catalogues de vente a donc été développé par Simon Gabay sur le logiciel libre et *open source*⁶² Kraken⁶³.

2.2.2 La structuration automatique des données : GROBID-dictionaries

2.2.2.1 L'intérêt et les conditions d'une utilisation « métaphorique » du logiciel

Il n'existe pas à ce jour de logiciel spécifiquement dédié à la structuration automatique de documents catalographiques, en particulier si leur mise en page diffère d'une page à l'autre ou dans le temps. Pour pallier cette lacune, le laboratoire d'humanités numériques de l'École Polytechnique Fédérale de Lausanne défend la production d'outils de structuration les plus génériques possibles. Il a notamment mis à libre disposition de la communauté scientifique la solution dhSegment⁶⁴. Cependant, ses applications actuelles relèvent plus souvent de l'analyse de la mise en page de sources historiques que de la structuration détaillée des données textuelles.

La technologie dhSegment a notamment été utilisée en 2018 par Raphaël Barman pour extraire automatiquement les informations contenues dans quelques 1 900 catalogues de vente aux enchères conduites à l'Hôtel Drouot entre 1939 et 1945. Comme en atteste le modèle de données consultable en ligne⁶⁵, le but est d'identifier et d'extraire les différentes entrées du catalogue. Le texte descriptif de chaque entrée n'est pas plus précisément structuré et il est versé en l'état dans une base de données relationnelle. Or, le but de notre chaîne de traitement des catalogues est de distinguer finement les différents éléments de chaque entrée et de récupérer des documents structurés en XML-TEI. L'emploi d'un logiciel plus spécifique s'est donc imposé.

GROBID-dictionaries est l'un des sous-projets de GROBID, solution originellement développée pour extraire automatiquement les informations bibliographiques présentes dans

62. B. Kiessling, « *Kraken - a Universal Text Recognizer for the Humanities* » DH2019, Utrecht, Pays-Bas. Disponible à l'adresse : <https://dev.clariah.nl/files/dh2019/boa/0673.html>

63. S. Gabay (éd.), *19th fixed-price and auction catalogues: Ground Truth and Models for OCR*, Neuchâtel: Université de Neuchâtel, 2019. Ce jeu de données est disponible à l'adresse suivante : <https://github.com/OCRCat>

64. S. Ares Oliveira, B. Seguin et F. Kaplan, « dhSegment: A generic deep-learning approach for document segmentation », *CoRR*, avril 2018 (article révisé en août 2019).

65. Le modèle de données est disponible à l'adresse suivante : https://github.com/raphaelBarman/auca-schema/blob/master/sql_schema_description.md.

2.2. Présentation détaillée et évaluation des outils retenus

les articles scientifiques⁶⁶. Cherchant à transposer le modèle et le principe de GROBID aux informations lexicales, GROBID-dictionaries est une librairie java d'apprentissage supervisé. À partir de fichiers PDF, elle permet de traiter, d'extraire et de structurer automatiquement en XML-TEI les informations textuelles contenues dans des dictionnaires numérisés. Il existe d'autres sous-projets destinés à la reconnaissance et la désambiguïsation d'entités nommées (GROBID-NERD) ou à l'identification et à la normalisation des unités physiques (GROBID-quantities).

GROBID-dictionaries a été conçu pour être indépendant de tout modèle de structuration des entrées lexicales, ce qui explique le choix de l'apprentissage supervisé et non d'une approche par règles. Le logiciel peut donc être appliqué pour structurer automatiquement des annuaires numérisés⁶⁷ ou nos catalogues de vente d'autographes. Le champ d'application de GROBID-dictionaries peut être étendu à toute ressource numérisée possédant une forme encyclopédique.

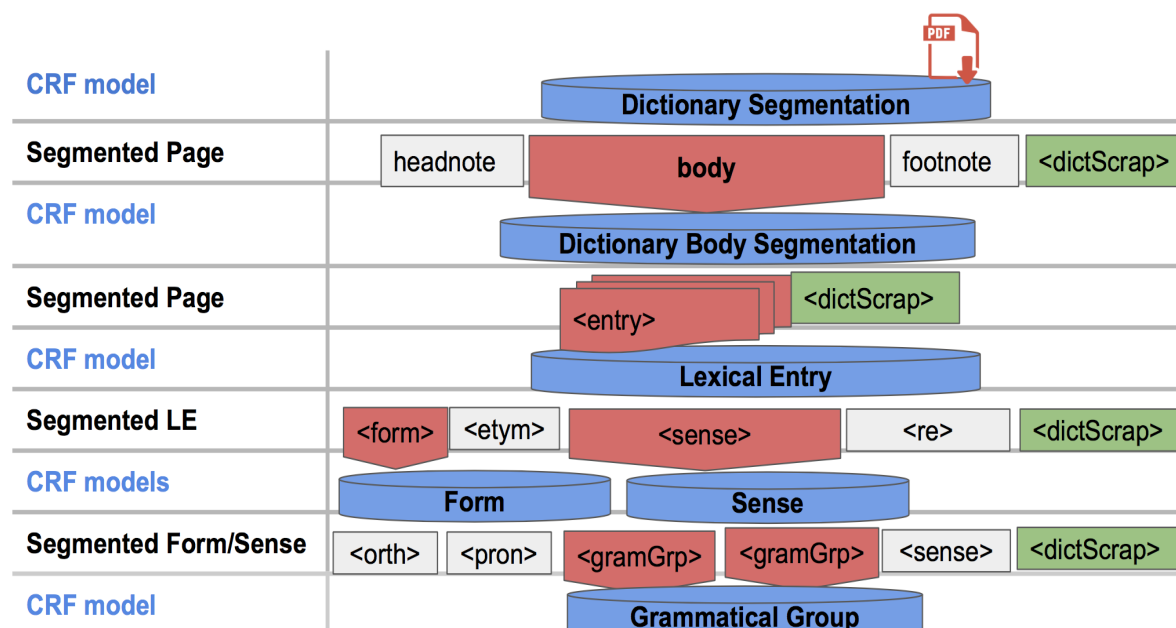


Figure 9 - Modèle de départ du logiciel GROBID-dictionaries

66. L. Romary et P. Lopez, « GROBID - Information Extraction from Scientific Publications », janvier 2015. hal-01673305

67. M. Khemakhem, C. Brando, L. Romary, F. Mélanie-Becquet et J.-L. Pinol, « Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories », JADH2018 « Leveraging Open Data », Septembre 2018, Tokyo. hal-01814189

2.2. Présentation détaillée et évaluation des outils retenus

Le logiciel s'appuie sur l'application en cascade de modèles dits CRF (*Conditional Random Fields*)⁶⁸. Ces « champs aléatoires conditionnels » permettent de construire des modèles probabilistes qui, à partir d'un corpus d'observations déjà annotées, associent à chaque nouvel élément une étiquette correspondante⁶⁹.

Le premier modèle intitulé « *Dictionary Segmentation* » vise à segmenter chaque page de dictionnaire en trois éléments : un entête qui correspond au numéro de page ou rappelle la section du dictionnaire (<headnote>), le corps du texte (<body>) et le pied de page (<footnote>).

A ce premier niveau, l'élément <dictScrap> est utilisé pour qualifier toute information textuelle qui ne trouverait pas sa place dans les catégories définies ci-dessus.

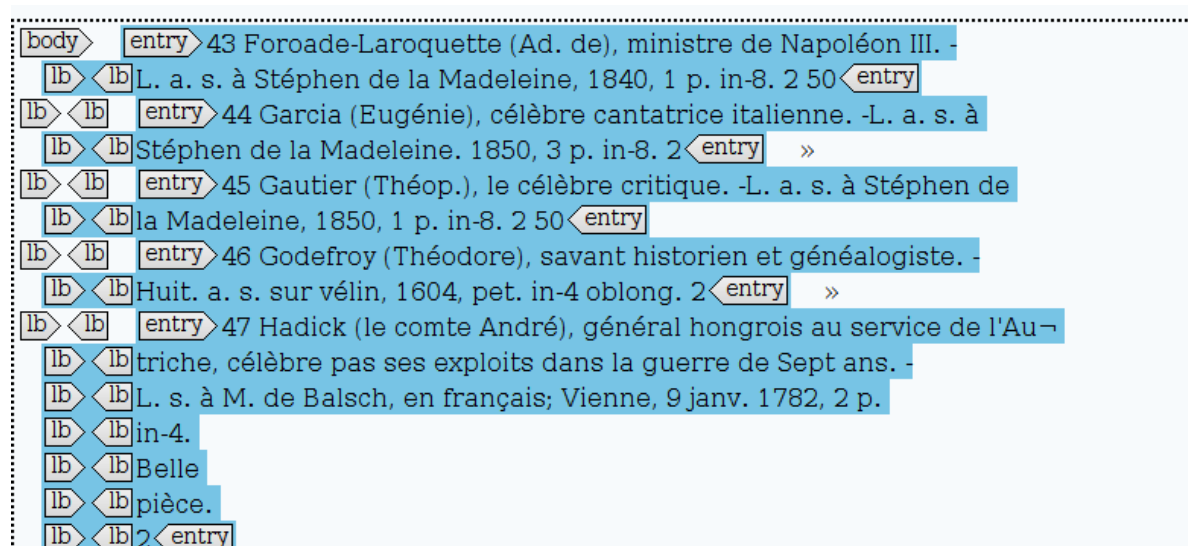


Figure 10 – Données d'entraînement où sont annotées manuellement les différentes entrées (extrait de la *Revue des Autographes* n° 25)

Au sein du corps de texte délimité par le modèle précédent, le modèle intitulé « *Dictionary Body Segmentation* » détermine les limites de chaque entrée lexicale : elles sont délimitées par la paire de balises <entry>.

Le troisième niveau d'étiquetage du texte a pour objectif de délimiter les différentes

68. M. Khemakhem, L. Foppiano et L. Romary. *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*. *electronic lexicography*, eLex 2017, Septembre 2017, Leyde, Pays-Bas. hal-01508868v2

69. E. Martienne, V. Claveau et P. Gros, « Application des Champs Conditionnels Aléatoires à l'étiquetage de flux télévisuel », *RFIA - Reconnaissance des Formes et Intelligence Artificielle*, janvier 2012, Lyon, p. 5. hal-00656547

2.2. Présentation détaillée et évaluation des outils retenus

composantes d'une entrée lexicale :

- la forme de base du mot à définir (<form>) ;
- son étymologie (<etym>) ;
- son ou ses multiples sens (<sense>) ;
- les autres entrées du dictionnaire desquelles il peut être rapproché (<re>).

entry	num	27	num	form	Balzac (Honoré de), le grand romancier réaliste	form	.	sense	L. a. s. à
lb	lb	Schlésinger, 1837, 1 p. in-8.	7	sense		entry			
entry	num	28	num	form	Bancel, célèbre député de la Drôme et de Paris	form	.	sense	Sentence
lb	lb	philosophique aut. sig. ; Gand, 2 mars 1860, 1/2 p. in-4., avec							
lb	lb	son portrait photographié. 3	sense		entry				
entry	num	29	num	form	Banderali (David), célèbre chanteur et compositeur	form	.	sense	Pièce
lb	lb	sig., 1830, 1 p. in-4. 2 »							
lb	lb	Reçu de ses appointements comme professeur au Conservatoire de							
lb	lb	Paris	sense		entry				
entry	num	30	num	form	Banville (Théodore de), célèbre poète contemporain	form	.	sense	Reçu
lb	lb	sig. de son indemnité littéraire, 1850, 1 p. in-f. 2	sense		entry				
entry	num	31	num	form	Barante (le baron de), historien des ducs de Bourgogne, de				
lb	lb	l'Acad. fr	form	.	sense	L. a. s., 1850, 1 p. in-8. 2	sense		entry

Figure 11 – Données d'entraînement où sont annotées manuellement les différentes composantes des entrées (extrait de la *Revue des Autographes* n° 50)

Les évolutions de la structure de GROBID-dictionaries ont notamment été déterminées par les retours des utilisateurs du logiciel. Au troisième niveau de segmentation, il est désormais possible de spécifier la présence d'un numéro à l'aide de l'élément <num>, ce qui se révèle très utile dans le cas de nos listes de documents en vente.

Le même principe de segmentation et d'étiquetage au sein des composantes déjà identifiées par les modèles précédents trouve à s'appliquer à l'intérieur des blocs balisés <form> et <sense>. Cependant, la structure des niveaux inférieurs a largement évolué depuis la production du schéma ci-dessous et sa présentation détaillée ne serait ici pas pertinente, puisque les entrées de catalogue que nous encodons sont de nature plus simple que les entrées de dictionnaire détaillées et complexes. Nous présenterons et justifierons ici le mode d'encodage finalement retenu pour les catalogues d'autographes.

Un premier modèle de structuration des entrées de catalogues avait été proposé à la conférence annuelle de la TEI en 2018. Suite à des évolutions du logiciel au cours de l'année

2.2. Présentation détaillée et évaluation des outils retenus

passée et dans une optique de simplification, nous avons proposé un modèle légèrement différent.

Modèle proposé en 2018 ⁷⁰	Modèle proposé en 2019 ⁷¹
<pre> <entry> <num>54</num> <form type="lemma"> <surname>Lassalle</surname> <addName>(A.-Ch.-L. de)</addName>, <desc>le plus brillant général de cavalerie des guerres de la République et de l'Empire, né à Metz, tué à la bataille de Wagram</desc> </form> <sense> .- <def> <bibl>L. a. s. au général Dugua; 1 p. in-f.</bibl> <num type="price">10 </num> </def> <note>Superbe lettre sur la campagne d'Egypte...</note> </sense> </entry> </pre>	<pre> <entry> <num>54</num> <form type="lemma"> <name>Lassalle (A.-Ch.-L. de</name> <pc>),</pc> <desc>le plus brillant général de cavalerie des guerres de la République et de l'Empire, né à Metz, tué à la bataille de Wagram</desc> </form> <pc> .-</pc> <sense> <subSense> L. a. s. au général Dugua; 1 p. in-f. 10 </subSense> <pc> » </pc> <note>Superbe lettre sur la campagne d'Egypte...</note> </sense> </entry> </pre>

Figure 12 - Modèles d'encodage successifs d'un autographe
proposé à la vente dans un catalogue à prix marqués

Pour des raisons d'efficacité mais également pour produire un modèle d'annotation susceptible de s'accorder avec les différents types de catalogues d'autographes traitables *via* GROBID-dictionaries, nous avons retenu des découpages binaires parmi les possibilités ménagées par les modèles du niveau <sense> et du niveau <form>.

La séquence textuelle présente dans l'élément <form> correspond à la description de l'auteur et elle est séparée entre les différentes composantes du patronyme (réunies dans un élément <name>) et les informations complémentaires intéressant cette personne (réunies dans un élément <desc>). Pour la séquence <sense>, la description codifiée du document en vente (présentation, éventuelles mentions de destinataire, date et lieu de rédaction, format, éventuelle mention d'un prix dans le cas d'un catalogue à prix marqués) est délimitée dans un élément <subSense>. Si des informations complémentaires sont procurées dans un deuxième paragraphe, elles sont encodées comme <note>.

70. M. Khemakhem, L. Romary, S. Gabay, H. Bohbot *et al.*, « Automatically Encoding Encyclopedic-like Resources in TEI », The annual TEI Conference and Members Meeting, Septembre 2018, Tokyo, Japon. hal-01819505

71. L. Rondeau du Noyer, S. Gabay, M. Khemakhem et L. Romary, « Scaling up Automatic Structuring of Manuscript Sales Catalogues », The annual TEI Conference and Members Meeting, Septembre 2019, Graz, Autriche.

2.2. Présentation détaillée et évaluation des outils retenus

form	name	Bonaparte (Laetitia	name	,	desc	mère de Napoléon Ier	desc	form
form	name	Borel (Petrus	name	,	desc	célèbre littérateur romantique, dit le Lycan-		
lb	lb	throe	desc	form				
form	name	Bouet-Willaumez (Ed. comte	name	,	desc	amiral, commandant de la		
lb	lb	flotte de la Baltique pendant la guerre avec la Prusse	desc	form				
form	name	Bourbaki (Ch.	name	,	desc	brave général en chef de l'armée de l'Est pen-		
lb	lb	dant la dernière guerre, né à Pau le 22 avril 1816	desc	form				
form	name	Brillat-Savarin (J.-A.	name	,	desc	constituant, le spirituel auteur de la		
lb	lb	Physiologie du goût, né à Belley (Ain)	desc	form				
form	name	Buchez (P.-J.-B.	name	,	desc	chef d'une école socialiste, président de		
lb	lb	l'Assemblée constituante de 1848, né dans les Ardennes	desc	form				
form	name	Capoul (J.	name	,	desc	célèbre chanteur de l'Opéra	desc	form

Figure 13 – Données d'entraînement où sont annotées manuellement les différentes composantes des descriptions d'auteurs (extrait de la *Revue des Autographes* n° 35)

Le modèle d'encodage pourrait être raffiné ou complexifié. Il pourrait être notamment intéressant d'encoder, comme il était prévu dans le modèle présenté en 2018, les prix marqués. L'usage de l'élément <measure> serait en réalité plus adapté que l'élément <num> car il permettrait plus aisément de mentionner des fourchettes de prix. Cependant, au vu des performances de GROBID-dictionaries exposées dans les deux sections suivantes, il paraît complexe d'automatiser cette reconnaissance. L'utilisation d'expressions régulières serait plus indiquée.

Une autre piste d'amélioration du modèle pourrait être de procéder à la reconnaissance et la désambiguïsation des entités nommées présentes dans les catalogues, c'est-à-dire des noms propres renvoyant à des villes, des pays, des personnes, etc. Une solution technique envisageable est GROBID-NERD, autre sous-projet de GROBID déjà mentionné. L'utilisation de ce logiciel nécessite cependant un matériel informatique très performant et il n'est pas certain que le seul recours à la base de données préconstituée par le développeur du logiciel suffise à identifier la majorité des entités nommées présentes dans les catalogues. Un travail de relecture important et de constitution d'une gazette complémentaire⁷² serait nécessaire.

72. C. Riondet et L. Foppiano, *History Fishing When engineering meets History. Text as a Resource. Text Mining in Historical Science #dhiha7*, Institut Historique Allemand, Paris, 2017. hal-01830713

2.2. Présentation détaillée et évaluation des outils retenus

Dans les limites temporelles de mon stage, j'ai choisi de me concentrer sur la production de données d'entraînement permettant d'encoder le maximum de catalogues de vente d'autographes, mais le modèle de structuration et d'encodage mis au point ne s'oppose nullement à des enrichissements dans le futur.

2.2.2.2. Production de données d'entraînement et protocole d'évaluation

2.2.2.2.1. Présentation du principe de l'apprentissage supervisé et des modes d'évaluation

Pour utiliser GROBID-dictionaries, il n'est pas nécessaire de recourir à une machine virtuelle car le logiciel est conteneurisé : il suppose seulement l'installation de son prédécesseur GROBID. GROBID-dictionaries repose sur l'apprentissage supervisé : à partir des données d'entraînement fournies, il crée un modèle permettant la structuration automatique de documents similaires.

La première étape d'utilisation de GROBID-dictionaries consiste donc en la création et l'annotation de données d'entraînement et d'évaluation à partir de fichiers PDF. À partir de chaque PDF d'origine, cinq fichiers différents sont générés. Les deux premiers sont spécifiques du fichier de base :

- un fichier `.rawtxt` qui contient le texte extrait du PDF. Il n'est pas utilisé pour l'entraînement du logiciel ;
- un fichier `.tei.xml` qui est destiné à devenir l'un des fichiers « étalons » (*gold standard files*) sur lequel le nouveau modèle va se fonder. C'est le type de fichier le plus important pour le développement d'un nouveau modèle GROBID-dictionaries. Il est donc nécessaire de l'annoter manuellement et avec le plus grand soin.

Les trois autres fichiers générés sont identiques quelles que soient les propriétés des fichiers PDF d'origine :

- un fichier `.template` qui contient une matrice de *features* (écrite en CRF++) : c'est sa combinaison avec les fichiers `.tei.xml` qui permet la production du modèle ;
- un fichier `.css` qui permet d'utiliser un éditeur XML (par exemple Oxygen) en mode « auteur ». Cela facilite le balisage des données d'entraînement, ce qui est une condition favorable à la production de données d'entraînement de qualité ;
- un fichier `.rng` qui décrit pour chaque niveau d'encodage les éléments de balisage autorisés. Il a la même fonction que le fichier précédent : rendre le balisage des documents source plus précis et plus rapide.

Après l'encodage à tous les niveaux de ces données générées, vient l'étape d'entraînement du logiciel à partir de ces données spécifiques. Pour évaluer la qualité du

2.2. Présentation détaillée et évaluation des outils retenus

modèle, une partie des données annotées ne servent pas à l'entraînement mais sont utilisées comme données d'évaluation. Il est à noter que lorsqu'un modèle est déjà en place, des données d'entraînement ou d'évaluation pré-annotées peuvent être générées. Il suffit de les corriger puis de les placer dans le fichier approprié et d'entraîner à nouveau le modèle pour augmenter le nombre de données d'entraînement disponibles ou soumettre le modèle à une évaluation plus stricte.

```
===== Token-level results =====
label          accuracy    precision    recall      f1
<note>         98.13      97.39      98.77      98.08
<pc>           99.97      98.77      98.77      98.77
<subSense>     98.1       98.73      97.35      98.04
all fields     98.73      98.07      98.07      98.07 (micro average)
               98.73      98.3       98.29      98.29 (macro average)

===== Field-level results =====
label          accuracy    precision    recall      f1
<note>         96.71      94.19      94.19      94.19
<pc>           99.34      98.77      98.77      98.77
<subSense>     95.72      95.16      94.4       94.78
all fields     97.26      95.88      95.55      95.71 (micro average)
               97.26      96.04      95.78      95.91 (macro average)

===== Instance-level results =====
Total expected instances: 128
Correct instances:        120
Instance-level recall:    93.75

Evaluation for org.grobid.core.GrobidModels$1@5d3af31c model is realized in 417 ms
```

Figure 14 – Capture d'écran du terminal après entraînement de GROBID-dictionaries au niveau « *Sense* »

Le modèle construit grâce aux données d'entraînement est appliqué aux données d'évaluation et le résultat automatique est comparé à l'annotation manuelle des fichiers. Pour chaque élément de structure, la conformité du résultat automatique à l'annotation manuelle est exprimée en pourcentage. Un résultat de 100% correspond à une stricte équivalence. Il faut que le résultat soit supérieur à 90%, voire à 95%, pour que les résultats de l'apprentissage supervisé soient considérés comme satisfaisants.

Les résultats de l'évaluation fournie par le logiciel dans le terminal sont composés de deux tableaux, comme le montre la figure 14. Le premier évalue l'étiquetage au niveau de chaque *token* : dans le cas présent, le résultat est considéré comme correct si un mot donné se trouve entre deux balises correspondant effectivement à sa localisation au sein de l'entrée. Le

2.2. Présentation détaillée et évaluation des outils retenus

deuxième intitulé *field-level results* évalue que l'étiquetage est correct pour l'intégralité des séquences, c'est-à-dire que les balises XML-TEI sont bien placées. Ces deux tableaux possèdent les mêmes intitulés de colonnes : ils renvoient à des variables statistiques. L'exactitude (*accuracy*) correspond au pourcentage de mots ou de séquences correctement étiquetés. Pour chaque étiquette, la précision (*precision*) est le rapport entre le total de mots ou de séquences correctement étiquetés et le total de mots ou de séquences étiquetés avec ce *label*. Le rappel (*recall*) correspond, pour chaque étiquette, au rapport entre le total de mots ou de séquences bien étiquetés et l'ensemble des mots ou des séquences qui auraient dû être étiquetés avec ce *label*. Le *f1 score* (ou, en français, la F-mesure) est fonction de la précision et du rappel et suit la formule ci-dessous :

$$F = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

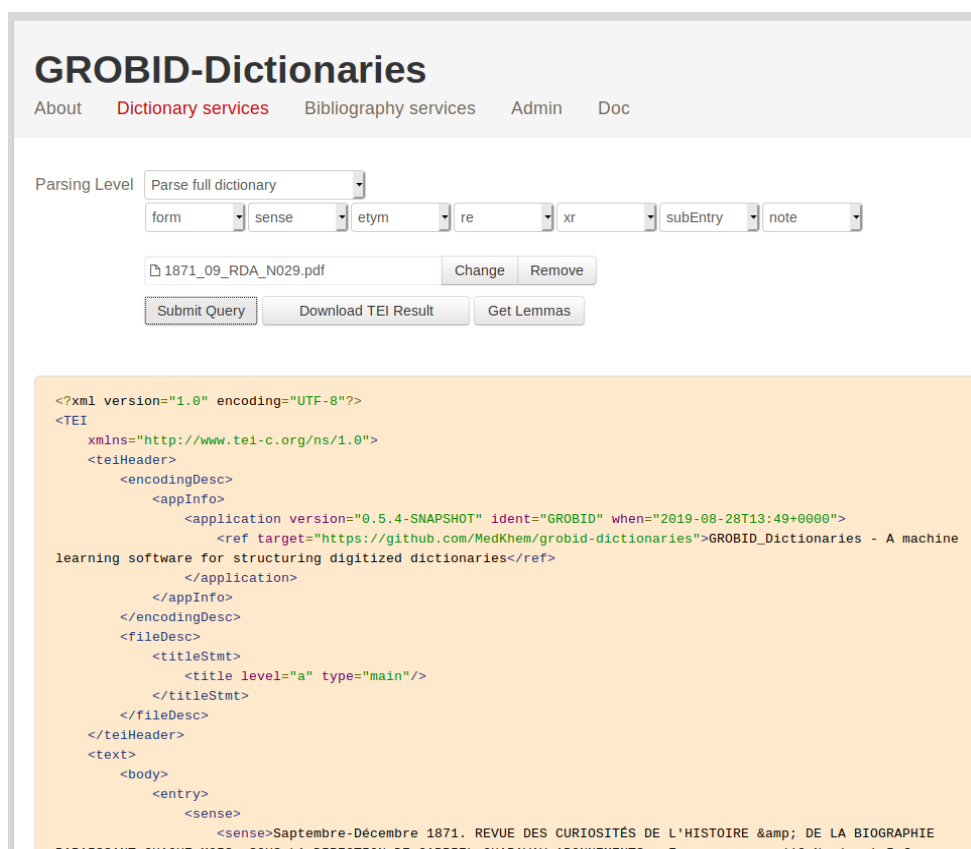


Figure 15 – Capture d'écran de l'interface graphique de GROBID-dictionaries après sélection et traitement d'un catalogue au format PDF

2.2. Présentation détaillée et évaluation des outils retenus

Dans le domaine de l'identification automatique, le *f1 score* est retenu comme indicateur car il permet de tenir compte à la fois de la précision et du rappel, tout en mesurant mieux que l'exactitude (*accuracy*) la proportion de faux négatifs (c'est-à-dire des mots ou des séquences qui n'ont pas reçu la bonne étiquette et ne sont donc pas comptés comme appartenant à leur catégorie véritable). Si les résultats de l'évaluation sont bons, il est ensuite possible de lancer l'interface graphique de GROBID-dictionaries. L'utilisateur n'a plus à utiliser le terminal et les lignes de commande. Il lui suffit de sélectionner chaque document PDF qu'il souhaite structurer automatiquement et d'appuyer sur un bouton pour récupérer un fichier XML-TEI correspondant.

2.2.2.2.2. Premiers résultats décevants et importance du *feature engineering*

En dépit des résultats encourageants de GROBID-dictionaries présentés lors de la conférence TEI de 2018 et du soin apporté à la préparation des données d'entraînement et d'évaluation, aussi bien sur le plan de la reconnaissance optique de caractères que de l'encodage, les premiers résultats obtenus se sont révélés très insuffisants à partir du niveau « *Lexical entry* ». Du fait de l'organisation en cascade du logiciel, cela empêchait son utilisation au-delà des deux premiers niveaux de structuration.

```
<entry>
  <num>12</num>
  <form type="lemma">Bertoni (Ferdinando), célèbre compositeur vénitien, maître de
chapelle de Saint</form>
  <sense>-Marc</sense>
  <form type="lemma">,</form>
  <sense>éminent élève</sense>
  <form type="lemma">du Père Martini</form>
  <pc>.-</pc>
  <sense>L. a. s. au Père J</sense>
  <pc>.-</pc>
  <sense>B. Martini (son</sense>
  <form type="lemma">maître</form>
  <sense>), à Bologne; Venise, 25 mai 1776, 1 p. in-4. Belle pièce. Rare. 18</sense>
</entry>
<pc>></pc>
```

Figure 16 - Un exemple d'entrée mal structurée

2.2. Présentation détaillée et évaluation des outils retenus

De manière surprenante, la forte imprécision constatée au niveau « *Lexical Entry* » demeurait en dépit de l’augmentation des données d’entraînement.

Suite à ces résultats décevants, deux pistes d’amélioration complémentaires ont été envisagées : l’amélioration de la qualité des données de départ et le *feature engineering*. En premier lieu, l’utilisateur de GROBID-dictionaries doit s’assurer de la qualité des données traitées par le logiciel. Il s’agit d’arriver à une reconnaissance optique de caractères la plus satisfaisante possible mais également de tenir scrupuleusement compte des conventions d’encodage présentées dans la documentation de GROBID-dictionaries⁷³. Dans la section 2.1.1., nous avons exposé comment les différents modèles développés sur Transkribus garantissent cette qualité. Il est en outre à rappeler que le logiciel avait déjà démontré ses capacités à obtenir de bonnes évaluations même lorsque le fichier PDF d’origine était bruité⁷⁴.

La persistance de mauvais scores en dépit de l’augmentation substantielle des données d’entraînement paraissait écarter l’hypothèse que c’était le manque de données ou leur mauvaise qualité qui empêchait d’obtenir de bons résultats. Suite à la communication de ces résultats, Mohamed Khemakhem, concepteur de GROBID-dictionaries et doctorant en informatique dans l’équipe ALMA^{na}CH (*Automatic Language Modelling and Analysis & Computational Humanities*) de l’Institut national de recherche en informatique et en automatique (Inria), a exploré l’autre voie d’amélioration des résultats, à savoir le *feature engineering*.

Dans le domaine de l’apprentissage supervisé, ce terme renvoie à la mise en place, au sein de la *pipeline* du logiciel, d’un traitement des données « brutes » garantissant le meilleur résultat possible à la sortie⁷⁵. Un *feature* se définit comme une représentation numérique d’un aspect de la donnée brute. C’est en quelque sorte une donnée préparée pour l’algorithme. Les *features* servent donc d’intermédiaires entre des données brutes et les modèles utilisés : ils en sont en fait le véritable *input*.

Le *feature engineering* consiste donc à déterminer quels sont le ou les *features* les plus pertinents à extraire des données brutes disponibles pour arriver aux meilleurs résultats. Dans les faits, le *feature engineering*, de même que les opérations de « nettoyage de données » (dont il n’est pas toujours distingué dans la littérature technique et la presse spécialisée) occupe une large partie de l’activité des programmeurs et des *data scientists*.

Le *feature engineering* est contextuel, c’est-à-dire qu’il dépend à la fois des données à

73. La documentation de GROBID-dictionaries est disponible à l’adresse suivante : <https://grobid-dictionaries.readthedocs.io/en/latest/>.

74. Dans l’article déjà cité « Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories », les résultats d’évaluation à tous les niveaux dépassaient les 90% malgré la qualité médiocre des numérisations d’annuaires.

75. A. Casari et A. Zheng, *Feature Engineering for Machine Learning*, Sebastopol, O’Reilly Media, 2018.

2.2. Présentation détaillée et évaluation des outils retenus

traiter, du modèle d'apprentissage supervisé utilisé et de l'objectif final. Dans le cas étudié ici, les données de départ sont un ensemble de catalogues de vente d'autographes sous forme de PDF exportés depuis Transkribus avec une couche de texte. Il s'agit en appliquant un modèle probabiliste reposant sur les champs aléatoires conditionnels de structurer automatiquement et en XML-TEI le texte reconnu par Transkribus. La démarche du *feature engineering* consiste alors à déterminer quels sont les aspects du fichier d'entrée sur lesquels le modèle d'apprentissage va s'appuyer en particulier. Dans le cas qui nous intéresse, une attention certaine a été portée aux séquences de ponctuation signalées par les balises <pc>.

Le *feature engineering* s'est traduit par le remplacement d'un modèle unigramme par un modèle bigramme, appliqué aux mêmes données d'entraînement et d'évaluation. Un modèle unigramme associe une étiquette (*label*) à un *token* fourni en fonction des seules caractéristiques de cet élément. Un modèle bigramme prend lui en compte l'étiquette appliquée au *token* précédent pour améliorer l'exactitude de l'étiquetage. Ce changement de modèle a été rendu possible par l'envoi d'un nouveau *template* (c'est-à-dire d'une nouvelle matrice de *features*), ce qui nous a permis d'obtenir des résultats très supérieurs et plus que satisfaisants, comme le montre le tableau *infra* :

	<form>	<num>	<pc>	<sense>	<all fields>
sans <i>feature engineering</i>	36.23	99.39	95.12	38.71	67.36
avec <i>feature engineering</i>	99.35	99.35	100	98.7	99.35

Tableau 1 - Résultats f1 level-field obtenus au niveau « *Lexical entry* »
(10 pages de données d'entraînement, 5 pages d'évaluation)

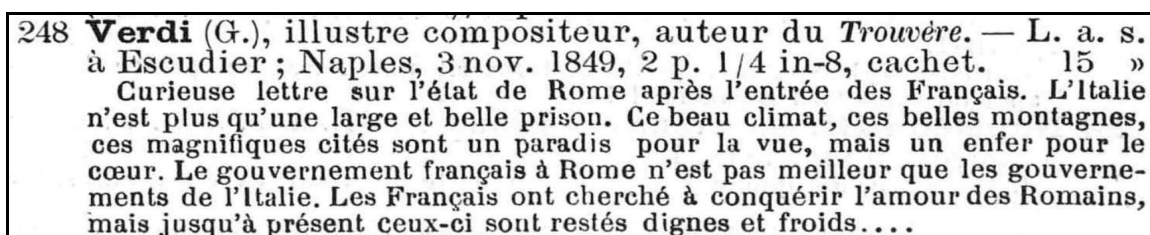
Cette première étape de mon stage m'a conduite à constater que, dans le cas fréquent où le chercheur en humanités numériques n'est pas le développeur du logiciel utilisé, il est fondamental qu'il puisse rapidement établir un dialogue avec le programmeur pour dépasser des résultats décevants et progresser dans sa recherche. Il ne faut pas non plus négliger les moyens humains nécessaires à la production de données d'entraînement et d'évaluation en quantité et qualité suffisantes.

2.2. Présentation détaillée et évaluation des outils retenus

2.2.2.2.3. Modèle spécifique ou modèle général ?

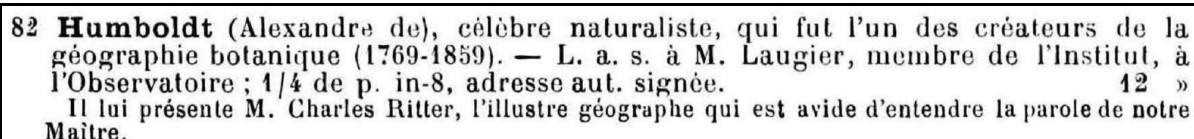
Après l'obtention de résultats satisfaisants pour une série précise de catalogues d'autographes à prix marqués, l'étape suivante consistait à évaluer si ce premier modèle entraîné par GROBID-dictionaries pouvait être appliqué à des catalogues ne présentant pas exactement la même mise en page, si chacun des types de présentation nécessitait la production de données d'entraînement propres ou si la meilleure solution était la production d'un modèle reposant sur des données d'entraînement diversifiées.

Pour répondre à cette question, nous avons constitué un corpus composé de quatre types de catalogues différents.



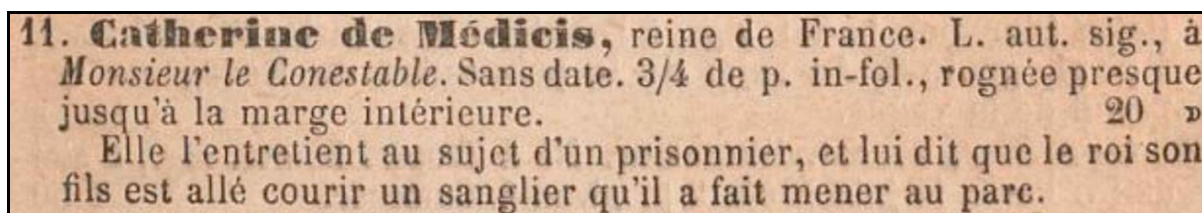
248 **Verdi** (G.), illustre compositeur, auteur du *Trovère*. — L. a. s. à Escudier ; Naples, 3 nov. 1849, 2 p. 1/4 in-8, cachet. 15 »
Curieuse lettre sur l'état de Rome après l'entrée des Français. L'Italie n'est plus qu'une large et belle prison. Ce beau climat, ces belles montagnes, ces magnifiques cités sont un paradis pour la vue, mais un enfer pour le cœur. Le gouvernement français à Rome n'est pas meilleur que les gouvernements de l'Italie. Les Français ont cherché à conquérir l'amour des Romains, mais jusqu'à présent ceux-ci sont restés dignes et froids....

Figure 17 - Exemple d'entrée de type 1 : *Revue des autographes, des curiosités de l'histoire et de la biographie*, Gabriel Charavay. (Première série N°42, Décembre 1874)



82 **Humboldt** (Alexandre de), célèbre naturaliste, qui fut l'un des créateurs de la géographie botanique (1769-1859). — L. a. s. à M. Laugier, membre de l'Institut, à l'Observatoire ; 1/4 de p. in-8, adresse aut. signée. 12 »
Il lui présente M. Charles Ritter, l'illustre géographe qui est avide d'entendre la parole de notre Maître.

Figure 18 - Exemple d'entrée de type 2 : *Revue des autographes, des curiosités de l'histoire et de la biographie*, Gabrielle Charavay (Seconde série N°56, 1934)



11. **Catherine de Médicis**, reine de France. L. aut. sig., à Monsieur le Conestable. Sans date. 3/4 de p. in-fol., rognée presque jusqu'à la marge intérieure. 20 »
Elle l'entretient au sujet d'un prisonnier, et lui dit que le roi son fils est allé courir un sanglier qu'il a fait mener au parc.

Figure 19 - Exemple d'entrée de type 3 : *Catalogue de lettres autographes et manuscrits*, Auguste Laverdet (N°1, avril 1856.)

2.2. Présentation détaillée et évaluation des outils retenus

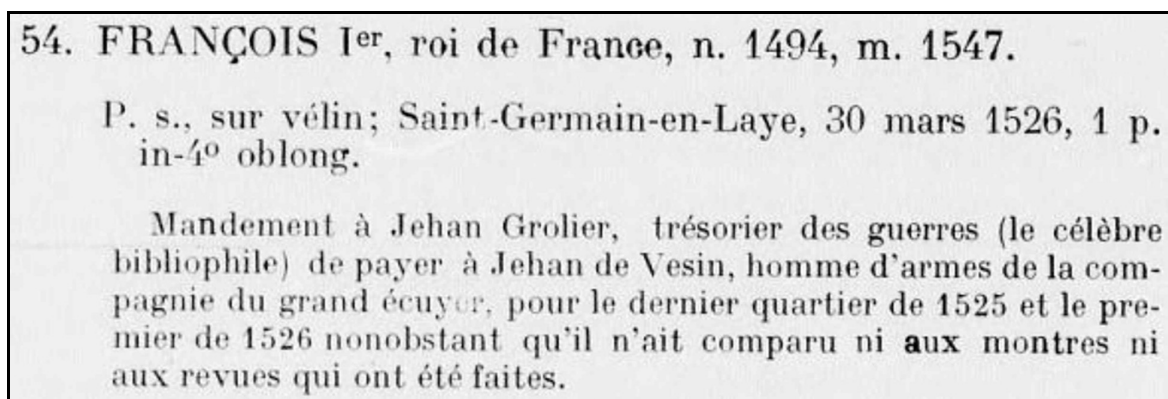


Figure 20 - Exemple d'entrée de type 4 : *Catalogue d'une intéressante collection de lettres autographes...*, Etienne Charavay (14 décembre 1908)

Le type 1 correspond à la première série de la *Revue des autographes, des curiosités de l'histoire et de la biographie*.

Le type 2 reprend la même structuration des entrées mais celles-ci sont listées sur deux colonnes par page.

L'originalité du type 3 de catalogue, plus ancien que la *Revue des autographes, des curiosités de l'histoire et de la biographie*, est de marquer de manière moins univoque et moins nette le passage de la description de l'auteur à celle du document. Un simple point les sépare, alors que dans les deux cas précédents, elles sont distinguées par l'enchaînement d'un point, d'un espace et d'un tiret cadratin.

Le type 4 a été choisi pour tester l'adaptabilité du logiciel aux catalogues de vente aux enchères. L'absence de prix mentionné dans les entrées rend plus difficile la détection automatique du passage entre la description standardisée du document à vendre et une éventuelle note complémentaire. GROBID-dictionaries ne peut pas s'appuyer sur le changement de paragraphe car la reconnaissance optique de caractères pratiquée par Transkribus ne permet pas le repérage et l'encodage des retours à la ligne.

Pour évaluer l'efficacité d'un modèle général et l'effet du cumul des quatre types de données sélectionnées, nous avons annoté quinze pages pour chaque type de catalogue : dix pour les données d'entraînement et cinq pour l'évaluation. Les pages choisies avec soin présentaient des entrées de longueur variée afin de ne pas introduire de biais dans les quatre modèles spécifiques. Nous avons ensuite comparé les résultats obtenus à la suite de l'entraînement de chacun de ces quatre modèles avec les évaluations recueillies par le modèle général reposant sur toutes les données d'entraînement produites.

2.2. Présentation détaillée et évaluation des outils retenus

2.2.2.3. Bilan : des résultats inégaux en fonction du type de catalogue

Les résultats obtenus démontrent encore une fois que les modèles bigrammes fournissent systématiquement de meilleurs résultats que les modèles unigrammes. Ils permettent de réaffirmer l'importance du *feature engineering* qui autorise des gains très sensibles, quel que soit le type de catalogue.

L'évaluation menée ne permet cependant pas d'établir définitivement s'il est plus efficace de recourir à des modèles spécifiques ou à un modèle général. Comme le montre les deux tableaux ci-dessous, l'utilisation des données cumulées permet d'obtenir de meilleurs résultats aux niveaux « *Form* » et « *Sense* » mais tend à diminuer le résultat du niveau « *Lexical entry* ».

Pour la suite de mon stage et en plus du modèle général⁷⁶, j'ai donc constitué trois ensembles de données. L'un correspond aux seules données issues de différentes séries de la *Revue des autographes, des curiosités de l'histoire et de la biographie*. Pour encoder le plus efficacement les catalogues s'apparentant aux types 3 et 4, il existe désormais des modèles qui utilisent des données spécifiques au niveau « *Lexical Entry* » et des données cumulées pour les niveaux inférieurs. La composition de chaque jeu de données et la marche à suivre sont détaillées sur une des pages Github du projet e-ditiones⁷⁷.

	Modèles GROBID		
Catalogues	Lexical Entry	Form	Sense
Type 1	72.49	93.72	73.62
Type 2	57.07	84.42	60.5
Type 3	60.07	74.71	48.07
Type 4	60.58	92.03	40.91
Tous types confondus	63.99	86.42	54.81

Tableau 2 - All fields F1-score des **modèles unigrammes**

76. L. Rondeau du Noyer, S. Gabay, M. Khemakhem et L. Romary. *General training and evaluation data for encoding Manuscript Sales Catalogues with GROBID-dictionaries*. Le jeu de données est disponible à l'adresse suivante : https://github.com/lairaines/grobid_TEL_2019.

77. L. Rondeau du Noyer, S. Gabay, M. Khemakhem et L. Romary, 2019, *Specific training and evaluation data for encoding Manuscript Sales Catalogues with GROBID-dictionaries*. Le jeu de données est disponible à l'adresse suivante : <https://github.com/e-ditiones/GROBID>.

2.2. Présentation détaillée et évaluation des outils retenus

	Modèles GROBID		
Catalogues	Lexical Entry	Form	Sense
Type 1	98.05	100	98.71
Type 2	99.16	95.02	90.78
Type 3	96.01	92.89	88.1
Type 4	92.78	96.83	86.64
Tous types confondus	95.43	97.44	92.77

Tableau 3 - All fields F1-score des **modèles bigrammes**

2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement

2.3.1. Un traitement intermédiaire pour rendre les fichiers générés compatibles avec le schéma général de la TEI et faciliter leur correction manuelle

Les données d'entraînement et d'évaluation produites au cours de mon stage, de même que le corpus de catalogues annotés qui est présenté en complément de ce mémoire, ont été générés à l'aide de la version 5.4.0 de GROBID-dictionaries. Or, les résultats téléchargés à partir de l'interface graphique du logiciel ne sont pas validés par le schéma général développé par le consortium de la TEI. Un tel état de fait réduit considérablement l'intérêt de l'utilisation du langage XML-TEI et n'est donc pas satisfaisant⁷⁸.

Afin d'y remédier, une feuille de style XSL permet de munir le fichier sorti de GROBID-dictionaries d'un élément <teiHeader> respectant les recommandations (*guidelines*) générales édictées par la TEI. En outre, c'est grâce à cette feuille de style qu'il est possible de remplacer les balises lexicographiques utilisées par GROBID-dictionaries par l'encodage plus léger et plus exact sémantiquement que nous avons déjà présenté dans la section 2.2.2.1.

Cette première transformation effectuée, le document XML-TEI est associé à un

78. L. Burnard « What is TEI Conformance, and Why Should You Care ? », *Journal of the Text Encoding Initiative*, n°12, janvier 2019. DOI : 10.4000/jtei.1777

2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement

premier schéma Relax NG intitulé `schema_grobid_output.rng`. Celui-ci vise à faciliter la correction manuelle des entrées de catalogue structurées automatiquement par GROBID-dictionaries. Pour ce faire, il fait en partie appel au langage Schematron. Complémentaire au langage Relax NG, Schematron considère comme valides toutes les parties du document XML qui ne sont pas régies par les règles qu'il formule.

Les règles rédigées en langage Schematron visent à vérifier deux aspects de l'encodage automatique des catalogues de vente d'autographes par GROBID-dictionaries. Le premier concerne l'enchaînement de la séquence des composantes de chaque élément `<item>`. Il s'agit de vérifier la présence, au minimum, d'un numéro de lot, d'un nom d'auteur et d'une description standardisée. La deuxième règle vérifie que les lots sont numérotés de manière continue. Ces deux règles ne sont pas retenues dans le schéma final du projet e-ditiones. Il est en effet possible que des catalogues moins structurés ou adoptant un système plus souple et discontinu de numérotation des lots soient à l'avenir intégrés au projet.

À l'usage, ce mode de correction s'est révélé satisfaisant et efficace. Une dizaine de pages recensant des lots d'autographes se traite et se corrige en une vingtaine de minutes.

2.3.2. Deuxième vague de transformations XSLT et schéma de validation final : penser l'insertion des catalogues de vente dans l'architecture globale du projet e-ditiones

Après la correction manuelle de chaque fichier structuré automatiquement grâce à GROBID-dictionaries, une deuxième feuille de style XSL lui est appliquée afin d'attribuer un identifiant unique à chaque entrée de catalogue. Cette génération automatique d'un attribut `xml:id` prépare l'éventuelle mise en relation d'une description d'un document autographe en vente avec l'édition électronique du texte de ce document. Un nouveau schéma Relax NG intitulé `odd_editiones.rng`, moins restrictif que le précédent, lui est ensuite associé. Ce schéma a été construit afin de permettre d'encoder l'ensemble des documents susceptibles d'être produits au format XML-TEI dans le cadre du projet e-ditiones.

2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement

```
<item n="183" xml:id="CAT_000069_e183">
  <num>183</num>
  <name type="author">Sandeau (Jules),</name>
  <trait>
    <p>romancier et auteur dramatique, de l'Acad. fr</p>
  </trait>
  <desc>L. a. s., 3/4 de p. in-4. Belle lettre. 5</desc>
</item>

<item n="184" xml:id="CAT_000069_e184">
  <num>184</num>
  <name type="author">Sardou (Victorien),</name>
  <trait>
    <p>le célèbre auteur dramatique, de l'Acad. fr</p>
  </trait>
  <desc>L. a. s. à M. Bridault, 1 p. in-8. 4</desc>
</item>

<item n="185" xml:id="CAT_000069_e185">
  <num>185</num>
  <name type="author">Saxe-Cobourg-Gotha (Victoria de),</name>
  <trait>
    <p>duchesse de Kent, mère de la reine Victoria</p>
  </trait>
  <desc>L. a. s., 1846, 3 p. in-18. Papier à son chiffre. Rare. 20</desc>
</item>
```

Figure 21 – Exemples d'entrées annotées et numérotées (extraites du n° 82 de la *Revue des autographes, des curiosités de l'histoire et de la biographie*)

La constitution d'une base de données recensant les documents autographes et manuscrits passés en vente sur le marché parisien au XIX^e siècle n'est en effet qu'un des axes du projet e-ditiones. Celui-ci vise également à recenser tous les manuscrits français du XVII^e siècle qui ont circulé sur le marché, à en proposer des descriptions approfondies grâce au module <msDesc> développé par la communauté TEI⁷⁹, et à plus long terme des reproductions et des éditions scientifiques électroniques.

79. Des exemples de descriptions de manuscrits sont en ligne sur la plateforme Github du projet e-ditiones : il s'agit de lettres de Madame de Sévigné. L'ensemble est disponible à l'adresse suivante : <https://github.com/e-ditiones/MS>.

2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement

```
<item n="279" xml:id="CAT_000007_e279" corresp="#MS_000008" rend="facs">
  <num>279</num>
  <name type="author">SEVIGNE (Marie de Rabutin-Chantal Marquise de),</name>
  <trait>
    <p>l'illustre épistolière
      (1626-1696). Lettre autographe à du Plessis. Aux Rochers, 20 août (1690).</p>
  </trait>
  <desc>4 pp. pet. in-4°. très rare.</desc>
  <note>Belle et précieuse lettre publiée dans l'édition de la Pléiade (Tome III, p. 765). Elle
    est adressée à M. du Plessis, précepteur du jeune marquis de Grignan, puis du petit
    marquis de Vins. «J'ay envie de comancer ma lettre come vous me conancés la vostre
    et vous dire que je vous écrirais trop souvent sy je le faisais toutes les fois que je
    pense à vous. Vous ne sauriez croire mon cher Monsieur combien je suis touchée des
    sujets de chagrin qui ont noircy... la gayeté et la vivacité de vostre belle jeunesse, c'est
    un meurtre que d'avoir chassé tout cela de chez vous...» Elle lui donne des conseils
    pour l'aider à remonter son moral défaillant... «Il faut profiter de tout pour l'éternité,
    j'ay fait icy des lectures admirables qui m'ont donné une telle foi, que sy mon cœur
    estait aussi touché que mon esprit est convaincu, je serais une sainte. Je suis toujours
    persuadée que quand vous aurés remis votre petit poussin (le jeune comte de Vins)
    sous l'aile de son brave père, vous rentrerez dans le giron de cette tribu de GRIGNAN
    où vous estes fort aymé...» Elle ne veut plus parler des drgns que le fils de Pomponne
    commandait à Fleurus: «Ce sont des démons, ils ont le diable au corps, mais
    je suis en furie contre le Mercure galant qui loue tous ceux qui ont esté à cette batille...
    et ne dit pas un seul mot du chevalier de Pomponne... - Adieu Monsieur... mon fils vous
    ayme toujours à la folie».</note>
</item>
```

Figure 22 – Notice du catalogue de la vente Robert Schuman reliée manuellement par Simon Gabay à la description d'un manuscrit aujourd'hui conservé à la Fondation Bodmer (Suisse)

Deux ensembles de fiches au format XML-TEI sont en cours de constitution par Simon Gabay. L'un recense un certain nombre d'auteurs et de destinataires récurrents ainsi qu'une sélection d'informations les concernant⁸⁰ ; l'autre est consacré aux acteurs de la vente des autographes et manuscrits⁸¹. À titre d'exemple, nous reproduisons ci-dessous un extrait de la fiche consacrée à Gabriel Charavay.

80. Au mois d'août 2019, les fiches recensent des personnalités appartenant au cercle épistolaire de Madame de Sévigné. Elles sont consultables à l'adresse <https://github.com/e-ditiones/PE1>.

81. Ce deuxième index est disponible à l'adresse suivante : <https://github.com/e-ditiones/PE2>.

2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement

```
<text>
  <body>
    <listPerson>
      <person>
        <persName>Gabriel <name type="label">Charavay</name></persName>
        <birth>
          <placeName>Lyon</placeName>
          <date when="1818-08-07">7 août 1818</date>
        </birth>
        <death>
          <placeName>Paris</placeName>
          <date when="1879-05-22">22 mai 1879</date>
        </death>
        <occupation>Marchand d'autographes</occupation>
        <sex value="m"/>
        <nationality key="fra"/>
        <note type="bio">
          Il est le frère de <persName>Jacques Charavay</persName> et le père d'<persName>Eugène Charavay</persName>
          <lb/>
          Il édite la <title>Revue des autographes</title>.
        </note>
        <listBibl>
          <listBibl type="link">
            <bibl>
              <ptr type="wikipedia" target="https://fr.wikipedia.org/wiki/Gabriel_Charavay"/>
            </bibl>
          </listBibl>
        </listBibl>
      </person>
    </listPerson>
  </body>
</text>
```

Figure 23 - Capture d'écran de la [fiche descriptive de Gabriel Charavay](#)

La documentation du projet e-ditiones rédigée sous forme d'ODD lors de mon stage et annexée à ce mémoire ne vise donc pas uniquement à spécifier la structure des catalogues de vente. Elle doit aussi permettre de documenter et de contrôler l'encodage de descriptions et éditions électroniques de manuscrits ainsi que la constitution d'index de noms.

Troisième partie

La constitution et la valorisation

d'une base de données test :

la première série de la *Revue des Autographes*,

des curiosités de l'histoire et de la biographie

3.1. Le corpus test : une vitrine de l'activité commerciale des Charavay

Après la production de données d'entraînement spécifiques pour rendre GROBID-dictionaries applicable aux catalogues de vente d'autographes et la rédaction d'une documentation et d'un schéma englobant tous les aspects du projet e-ditiones, la dernière étape de mon stage a consisté en la construction d'une application permettant de consulter et d'interroger les fichiers numériques obtenus à partir des numérisations de catalogues de vente.

Les numéros 19 à 95 de la première série de la *Revue des autographes, des curiosités de l'histoire et de la biographie* ont été choisis pour peupler la première version de notre base de données car il n'existe actuellement pas de possibilité de les consulter en ligne. Le travail de génération des fichiers numériques correspondants a été conduit à partir des numérisations des microfilms de la *Revue des Autographes, des curiosités de l'histoire et de la biographie* conservés à la Bibliothèque nationale de France.

Pourquoi se concentrer sur la *Revue des Autographes, des curiosités de l'histoire et de la biographie* ? Cette revue paraît pendant tout le dernier tiers du XIX^e siècle, période considérée comme le premier point culminant du « culte des autographes⁸² ». Cette popularité est attestée par le développement d'une « science des autographes⁸³ » comme par la multiplication des éditions de correspondances d'auteurs.

Lorsqu'il crée la *Revue des Autographes, des curiosités de l'histoire et de la biographie* en janvier 1866, Gabriel Charavay est déjà reconnu dans le milieu des amateurs d'autographes. En 1865, Alphonse de Lescure écrit de *L'Amateur d'autographes*, première entreprise éditoriale du cadet des Charavay, qu'il est « destiné à devenir le *Moniteur* de cette branche de la curiosité⁸⁴. ». Avec les premiers numéros de la *Revue des Autographes, des curiosités de l'histoire et de la biographie*, Gabriel Charavay poursuit la même ambition et annonce que sa nouvelle publication mensuelle est « destinée à devenir un traité complet des autographes, tout en accordant une large place aux recherches d'érudition⁸⁵. ».

En décembre 1868, paraît le numéro 19 de la *Revue des autographes, des curiosités de l'histoire et de la biographie*, après huit mois d'interruption de la publication. Le manque d'abonnés contraint Gabriel Charavay à transformer le contenu de son périodique et à diminuer le prix de l'abonnement de moitié. En préambule de ce numéro, il annonce : « La

82. P.-J. Dufief, « Correspondances et histoire littéraire (1850-1900) », art.cit, p. 127.

83. Cette expression est volontiers reprise par Étienne Charavay qui l'a choisie comme titre à l'avant-propos qu'il donne en 1887 à la seconde édition du catalogue Alfred Bovet.

84. A. de Lescure, *Les autographes et le goût des autographes...*, op. cit., p. 52

85. *Revue des Autographes, des curiosités de l'histoire et de la biographie*, n°1, janvier 1866, p. 1.

3.1. Le corpus test : une vitrine de l'activité commerciale des Charavay

première partie contiendra des articles, et la seconde les autographes que nous mettons en vente à prix marqués⁸⁶. ». À partir de cette date, il apparaît donc clairement que la *Revue des autographes, des curiosités de l'histoire et de la biographie* devient avant tout un catalogue de vente à prix marqués. Certaines de ses livraisons sont constituées uniquement d'une liste d'autographes à vendre et il est rare que les articles occupent plus de la moitié des seize pages de la revue.

Du fait de la concision des notices et de la permanence de la mise en page du périodique, la chaîne de traitement mise en place au cours de mon stage s'est révélée efficace pour structurer automatiquement les annonces de vente de la *Revue des autographes, des curiosités de l'histoire et de la biographie*. GROBID-dictionaries est particulièrement adapté à la valorisation de cette source sérielle dont la Bibliothèque nationale de France possède une collection presque complète.

Comment exploiter les informations contenues dans la *Revue des autographes, des curiosités de l'histoire et de la biographie* ? Il est important de garder à l'esprit que la revue est avant tout une vitrine. Le libraire n'y présente qu'une sélection raisonnée de son stock. La mise en base de données de la revue ne vise pas à reconstituer un état exhaustif des autographes et documents passés par la librairie de Gabriel et Eugène Charavay. Elle ouvre cependant des pistes de réflexion et de recherche quant aux pratiques commerciales et publicitaires des marchands d'autographes. Certains documents vendus, autographes ou non, apparaissent sous forme de lots ou intégrés à des recueils composites. Les unités documentaires sont difficilement identifiables mais leur somme reste un bon indicateur des tendances du marché du patrimoine écrit. La recherche par nom propre peut permettre de cerner dans le temps le rôle que tiennent les autographes d'un auteur donné dans la stratégie commerciale générale des Charavay, ainsi que leur apparition, leur valorisation ou leur disparition au sein de la *Revue des autographes, des curiosités de l'histoire et de la biographie*.

3.2. L'application de consultation : mise en relation de la base de données des ventes avec les autres fichiers du projet e-ditiones

Pour rendre disponible à la consultation et à l'interrogation l'ensemble des passages de la *Revue des Autographes, des curiosités de l'histoire et de la biographie* relevant du catalogue à prix marqués, j'ai travaillé à la construction d'une base de données XML native. Le recours à ce type de base de données a été privilégié car il ne nécessitait aucune

86. *Revue des Autographes, des curiosités de l'histoire et de la biographie*, n°19, décembre 1868, p. 1.

3.2. L'application de consultation : mise en relation de la base de données des ventes avec les autres fichiers du projet e-ditiones

transformation supplémentaire du corpus de catalogues de vente d'autographes constitué précédemment grâce à GROBID-dictionaries.

Contrairement à une base de données relationnelle dont le préalable est la conception d'un modèle de données fixe et immuable, une base de données XML ne prédétermine pas la forme des informations qui y seront versées au cours du temps. Il est ainsi possible de la peupler avec des documents de forme diversifiée. Dans le cadre du projet e-ditiones, cette souplesse est un atout et ménage la possibilité que le schéma `odd_editiones.rng` soit modifié ou que de nouveaux types de documents soient ajoutés sans qu'aucune des deux évolutions ne se traduise par la refonte de la structure de la base de données.

Parmi les divers systèmes de gestion de bases de données XML, j'ai choisi eXist-db, car il comprend, en plus des fonctions basiques de gestion, un environnement de développement intégré (EDI, en anglais *IDE*) pour le développement d'applications Web.

Le système eXist-db a également été sélectionné car il permet de configurer et de constituer facilement, grâce au module Lucene, des index prenant en compte tout le contenu des documents et autorisant la recherche plein texte. Cette fonction est fondamentale pour que la recherche parmi les catalogues soit aisée et efficace : elle permet aux utilisateurs de rechercher des mots-clés de toute nature et quelle que soit la place qu'ils occupent dans les entrées de catalogue.

Les résultats de la recherche plein texte permettent aussi bien d'accéder à la version électronique du catalogue disponible en ligne qu'aux références bibliographiques précises du catalogue dont sont extraits les résultats. Une attention particulière est apportée au référencement du catalogue dans lequel le mot recherché apparaît car ces informations bibliographiques manquent dans certains éléments des « fichiers Charavay ». Ce sont la possibilité de recherche plein texte et l'accès au catalogue complet qui apportent une plus-value à cette première base de données de catalogues de vente.

Au terme de mon stage, l'application e-ditiones permet de :

- consulter la liste complète des catalogues de vente faisant partie du corpus-test ;
- visualiser en ligne ces catalogues ;
- télécharger directement leur contenu au format XML-TEI ;
- explorer le contenu des catalogues en se servant de la recherche plein texte.

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

Le premier objectif de l'application conçue pendant mon stage est le repérage et le suivi, sur le marché privé, de documents autographes et manuscrits. Une campagne complémentaire de structuration automatique de catalogues numérisés de vente aux enchères et à prix marqués est maintenant nécessaire pour étoffer la base de données et offrir aux chercheurs et philologues la possibilité de requêter un plus large corpus de catalogues.

En l'état, il est possible d'avancer que la *Revue des autographes, des curiosités de l'histoire et de la biographie* est une source documentaire utile à divers domaines de recherche et à de nombreux chercheurs.

Les résultats de la recherche d'un nom propre dans la base de données ne se limitent pas au repérage et au comptage des lettres d'un auteur mentionnées dans la *Revue des Autographes, des curiosités de l'histoire et de la biographie*.

Si l'on recherche le terme « Sévigné », on constate qu'aucune de ses lettres n'a été proposée à prix marqués par la *Revue des Autographes* entre 1868 et 1885. Cependant, son patronyme apparaît à de multiples reprises dans la revue. Dans le numéro 90 publié en avril 1885 où 187 lettres sont proposées à la vente, sept descriptions comportent le terme « Madame de Sévigné ». L'épistolière n'en est ni l'auteur, ni la destinataire. Ces lettres sont présentées avec des mentions telles que « ami de Madame de Sévigné » pour en faire la promotion et attirer l'œil des amateurs. C'est le signe que les liens qui existent entre le nom d'un auteur et la valorisation économique de documents écrits ne se résument pas à la valeur marchande de ses propres lettres. Un autre exemple intéressant est le lot 144 du numéro 32 de la *Revue des Autographes*. Gabriel Charavay y propose pour la somme élevée de 150 francs une lettre de la petite-fille de Madame de Sévigné.

144
Simiane (Pauline de Grignan, marquise de), fille de Mme de Grignan et petite-fille de Mme de Sévigné, comme elles célèbre épistolaire
L. a. s.; 24 octobre, 3 p. in-4. Belle pièce. Rare. 150 francs

144	Simiane (Pauline de Grignan, marquise de), fille de M ^{me} de Grignan et petite-fille de M ^{me} de Sévigné, comme elles célèbre épistolaire.— L. a. s.; 24 octobre, 3 p. in-4. Belle pièce. Rare. 150. »
-----	---

Figure 24 - Identification grâce à l'application e-ditiones d'une lettre vendue avec la mention publicitaire « petite-fille de M^{me} de Sévigné » et entrée originale correspondante

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

Le relevé de l'ensemble des personnalités associées dans les catalogues des Charavay à Madame de Sévigné permet de percevoir comment les choix de présentation et de rédaction des marchands d'autographes constituent en galaxie autour de personnalités célèbres, à des fins commerciales, un ensemble d'individus et d'objets à vendre.

Pour montrer la diversité des documents présentés dans la *Revue des autographes, des curiosités de l'histoire et de la biographie*, une requête avec le mot-clé « Danton » a été effectuée. Elle retourne onze entrées (reproduites en annexe 3) commençant par son nom. Elles correspondent à dix pièces distinctes car la même lettre signée du 25 août 1792 a été présentée deux fois à la vente dans la *Revue des autographes, des curiosités de l'histoire et de la biographie* (numéro 22 de mai-juillet 1869 puis numéro 32 d'août 1872).

L'exploration de la base de données confirme qu'au XIX^e siècle des pièces qui sont en réalité des archives publiques signées par des personnages historiques étaient vendues en tant qu'autographes. Parmi les dix documents, six ont été signés dans l'exercice d'une fonction officielle et publique.

Dans chacune des descriptions accompagnant les pièces en vente, la qualification de « lettre autographe » n'est pas utilisée. La mention « L. s. » (pour signifier « lettre signée ») est utilisée car seule la signature des documents est de la main de Danton. C'est cette signature, ainsi que l'intérêt historique, qui fondent la valeur commerciale de ces pièces néanmoins vendues sous la rubrique « Autographes à prix marqués ».

Les autres résultats de la requête permettent de constater que les Charavay utilisent, comme dans le cas de Madame de Sévigné, la notoriété du « célèbre conventionnel » pour promouvoir les lettres de sa veuve, de Camille Desmoulins, de Marie-Jean Hérault de Séchelles, de François-Joseph Westermann et de Fabre d'Églantine.

Dans le cadre du projet e-ditiones, la réconciliation automatique des différentes notices décrivant le même document dans plusieurs numéros de la revue est une question à approfondir. L'exemple ci-dessous conduit à remarquer que la même pièce n'est pas présentée identiquement d'un numéro à l'autre. La qualification de l'auteur, la description du document, son prix et la note complémentaire diffèrent, ce qui complexifie la démarche informatique à effectuer.

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

n° 22 lot 4374	Danton (Georges), le célèbre conventionnel L. s., comme ministre de la justice, au directoire du départ, de Paris; 25 août 1792, 1 p. in-f. Belle pièce. 27 francs Les prisonniers de Bicêtre, condamnés a y être détenus jusqu'à leur majorité, instruits du décret qui la fixe à 21 ans, demandent s'ils doivent jouir du bénéfice de cette loi. « L'affirmative, dit-il, n'est pas douteuse. Les citoyens français doivent, sous tous les rapports, jouir du bénéfice de la loi qui avance la majorité, et cette loi est trop formelle pour notre point applicable aux détenus à Bicêtre. »
n° 32 lot 36	Danton (Georges), célèbre conventionnel, sacrifié à la jalousie de Robespierre L. s., au Directoire de Paris; 25 août 1792, 1 p. in-f. 30 francs Alors ministre de la justice, et consulté sur la question de savoir si les jeunes détenus de Bicêtre, condamnés a y être renfermés jusqu'à leur majorité, doivent jouir du bénéfice de la loi qui vient d'abaisser la majorité à 21 ans, il n'hésite pas à se prononcer pour l'affirmative. (Il s'agissait de la capacité de vote pour les élections à la Convention nationale.)

Tableau 4 - Extrait du relevé des entrées commençant dans la *Revue des autographes* par le nom de Danton. Les deux notices reproduites décrivent le même document.

Outre les philologues et les historiens, les spécialistes de l'histoire de l'art peuvent également extraire des informations intéressantes de la consultation de l'application e-ditiones. La recherche d'un nom d'artiste fait émerger des lettres dont la date, la description et les destinataires fournissent des indications biographiques. Elle retourne aussi des entrées correspondant à des dessins signés. La recherche plein texte « Corot » permet ainsi de repérer huit documents dans le corpus actuel : sept lettres (dont une de recommandation et deux sur les modalités de vente et les prix de ses tableaux) et un « dessin au crayon conté » dans le numéro 61 du périodique. La mise en vente de ce dessin au prix marqué de « 25 francs » est également une indication à retenir pour le spécialiste soucieux de reconstituer les variations et l'évolution de la cote de l'artiste.

33
Corot (J.-B.-C.), un des meilleurs peintres de notre époque
Paysage, dessin au crayon conté signé, in-4. 25 francs

33 Corot (J.-B.-C.), un des meilleurs peintres de notre époque.—
Paysage, dessin au crayon conté signé, in-4. 25 »

Figure 25 - Identification grâce à l'application e-ditiones
d'un dessin de Corot vendu par les Charavay et entrée originale correspondante

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

La mise en base de données de la *Revue des autographes, des curiosités de l'histoire et de la biographie* permet de mettre en lumière la diversité des documents vendus dans les catalogues d'autographes : lettres autographes évidemment, mais aussi documents historiques signés ou non, multiples pièces relatives à l'histoire d'un département ou d'une ville, dessins, etc. Cette variété se lit dans la structure même de certains numéros de la *Revue des autographes, des curiosités de l'histoire et de la biographie* qui possèdent des sous-sections thématiques : par exemple, une partie non négligeable de son numéro 86 est spécifiquement dédiée à la vente de dessins.

AUTOGRAPHES A PRIX MARQUÉS
MAISON Gabriel CHARAVAY
Dirigée par **Eugène CHARAVAY Flis**, expert en autographes,
8, QUAI DU LOUVRE, A PARIS.

AUTOGRAPHES CLASSÉS PAR DÉPARTEMENTS
PERSONNAGES QUI Y SONT NÉS OU S'Y RATTACHENT.

AIN.

- 1 ANGEVILLE (le comte d'), agronome, député de l'Ain. — L. a. s., 1844, 1 p. in-4. 2 »
- 2 ANGEVILLE (la comtesse d'), célèbre par son ascension du Mont-Blanc, auteur de plusieurs ouvrages sur ses voyages. — L. a. s. au baron Volland, 3 p. in-8. Cachet. Charmante épître. 3 »
- 3 ARBELLE (le baron André d'), littérateur, journaliste et préfet, né à Montluel, tué en 1825. — L. a. s., 1815, 1 p. in-8. 2 »
- 4 BAILLOD (le baron J.-P.), brave général de la République et de l'Empire, député, né à Songieu. — L. a. s., 1837, 2 p. in-8. 2 »
- 5 BOULLÉE (A.-Aug.), biographe de d'Aguesseau et de Portalis, historien des États-Généraux, né à Bourg. — L. a. s., 1838, 5 p. in-8. 2 »

Il propose une série d'articles pour l'*Encyclopédie du XIX^e siècle*, et en donne la liste.

Figure 26 – Extrait de la *Revue des Autographes* n° 58
proposant un classement des autographes par département

Comme la requête par nom propre ou l'analyse typologique des documents vendus par les Charavay, porter attention aux descriptions successives d'un même auteur dans le temps peut se révéler intéressant relativement à l'histoire de la célébrité et à l'étude de la réception des œuvres littéraires. Les qualificatifs apposés aux noms propres ont, en plus de leur aspect commercial et publicitaire, valeur d'indices des hiérarchies culturelles et artistiques du temps.

Relever grâce à une recherche plein texte les différentes caractérisations de Gérard de

3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives

Nerval permet de constater l'évolution positive de son statut dans le temps. Décrit en mars 1875 comme un « célèbre littérateur », il est qualifié en mai 1882 de « célèbre écrivain fantaisiste » et en novembre 1885 de « célèbre écrivain, traducteur de *Faust*, né en 1808, mort par suicide en 1855 ». Ce relevé rapide mériterait d'être mis en perspective par le recours aux études de la réception nervalienne. Il gagnerait à être généralisé à d'autres auteurs présents dans la base de données. Les catalogues de vente d'autographes à prix marqués sont donc susceptibles de retenir, entre autres, l'attention des praticiens de l'histoire culturelle et littéraire.

Conclusion

Un des principaux objectifs de mon stage à l'Institut de littérature française de l'UniNE était d'évaluer quel était l'apport des technologies numériques à l'exploration et à l'exploitation des catalogues de vente de lettres autographes, sources historiques riches mais aujourd'hui encore dispersées et relativement sous-exploitées.

Au cours de ce mémoire, j'ai cherché à démontrer et illustrer que cette contribution était multiforme. La chaîne de traitement que j'ai contribué à mettre en place permet à la fois de rassembler, de transcrire dans un format standard et pérenne et de rendre accessible et interrogeable de façon standardisée et simple un corpus de catalogues de vente d'autographes.

Plus encore, le traitement numérique des catalogues d'autographes contribue au développement d'un nouveau rapport à ces sources. L'application proposée en complément de ce mémoire permet de ne pas se contenter d'explorer les catalogues à partir des noms placés au début des entrées, alors que c'est la seule perspective ménagée par un ensemble de fiches papiers. Elle permet au contraire d'alterner rapidement entre la consultation des listes d'autographes dans leur intégralité et l'interrogation de tout un corpus par mots-clés, ce qui ouvre de nouvelles perspectives de recherche. Contrairement à la constitution d'un fichier papier, la mise en base de données de catalogues ne pose pas de risque de perte d'information ou de séparation entre une entrée donnée et son contexte documentaire d'origine. Au contraire, l'adoption du langage XML-TEI établit un lien étroit entre le contenu du catalogue et ses métadonnées.

À l'issue de mon stage, il apparaît que la structuration grâce à des outils informatiques et automatiques renouvelle la consultation et la compréhension des catalogues de vente d'autographes, ce qui permet de formuler deux constats.

Sur le plan général, le travail mené s'inscrit dans une tendance forte et actuelle des humanités numériques. Alors que les technologies de numérisation et de reconnaissance de texte sont arrivées à un point de développement certain, le nouvel enjeu et la nouvelle frontière sont de pallier le manque de structuration des sources historiques numérisées, qui reste aujourd'hui encore, selon certains, un frein à l'exploitation des « *big data of the past*⁸⁷ ».

Plus spécifiquement, le traitement numérique des catalogues de lettres autographes semble le moyen le plus efficace pour répondre aux appels au dépouillement systématique de ces sources qu'avaient formulés certains chercheurs des générations passées. Le fait que ce dépouillement systématique soit désormais autorisé par les technologies numériques ne doit

87. F. Kaplan et I. di Lenardo, « Big Data of the Past. », *Front. Digit. Humanit*, 4:12, 2017. DOI : 10.3389/fdigh.2017.00012

Conclusion

cependant pas dissimuler qu'il suppose encore la mise en œuvre au préalable de moyens humains importants et nécessite de pouvoir établir un dialogue et des échanges réguliers avec les programmeurs qui développent les solutions de structuration automatique des données.

Parmi tous les outils décrits dans ce mémoire, il convient de remarquer que GROBID-dictionaries s'est imposé comme une solution de structuration automatique très satisfaisante, à condition d'avoir les moyens de produire un certain nombre de données d'entraînement et de conduire ou de faire conduire un travail de *feature engineering*.

Mon travail de stage a permis de confirmer une caractéristique de GROBID-dictionaries : son adaptabilité à tout document possédant une présentation régulière et systématique. Ainsi, toute forme de catalogue, pour peu que l'information y soit organisée de manière standardisée, peut être traitée à l'aide de la chaîne de traitement décrite dans ce mémoire.

Il ne faut cependant pas dissimuler que, dans sa version actuelle, GROBID-dictionaries ne se prête pas à la production d'un modèle général unique permettant d'encoder plusieurs types de catalogues aussi efficacement que certains des jeux de données spécifiquement développés pour chaque mise en page. L'efficacité du logiciel reste en outre dépendante de la présence de séparateurs typographiques fréquents et variés dans le texte à structurer. Plus les catalogues traités sont structurés explicitement, plus la chaîne de traitement développée au cours de mon stage est efficace. Ainsi, les résultats actuels restent meilleurs pour les catalogues à prix marqués que pour les catalogues de vente aux enchères, sources pourtant indispensables à la constitution d'un catalogue des manuscrits du XVII^e siècle français, objectif final et principal du projet e-ditiones.

L'application que j'ai développée permet actuellement de consulter et d'interroger des catalogues de vente du XIX^e siècle. De nouvelles étapes sont à prévoir pour qu'elle puisse s'enrichir d'un volet permettant la consultation de descriptions et d'éditions électroniques des manuscrits vendus comme autographes sur le marché parisien aux XIX^e et XX^e siècles et la circulation entre les notices de manuscrits et les entrées des catalogues de vente.

À l'avenir, il est aussi possible que le modèle d'encodage des entrées de catalogues proposé dans ce mémoire soit raffiné pour distinguer au sein des entrées les entités nommées, les prix de vente ou les estimations, etc. Cependant, l'amélioration de la granularité passerait par l'utilisation d'autres logiciels que GROBID-dictionaries, car la distinction de ces informations spécifiques n'est pas matérialisée dans la structure même des catalogues. Le choix d'utiliser à la fois le format XML-TEI pour encoder les catalogues et de les réunir dans une base de données XML native laisse ouvertes de telles évolutions, de même qu'il facilite la réutilisation des données encodées et leur éventuelle intégration dans d'autres projets.

Grâce au premier corpus test de catalogues d'autographes proposé, il ressort que ces

Conclusion

publications commerciales proposent à la vente une diversité de documents non négligeable. Il démontre que les catalogues de vente produits par le marché privé sont une source pertinente pour différents domaines de la recherche, en particulier pour les chercheurs qui s'attellent à étudier dans le détail quelles étaient au XIX^e siècle les modalités de la valorisation marchande et non-marchande du patrimoine écrit, un sujet qui reste encore d'actualité. Il établit l'intérêt d'étudier le marché des autographes du point de vue des libraires et des experts, acteurs jusqu'ici restés dans l'ombre des amateurs et des collectionneurs.

L'objet de mon stage n'était pas de traiter dans leur intégralité les catalogues et revues-catalogues produits par les deux librairies Charavay. L'étude minutieuse de leur contenu éditorial, articulée à l'enrichissement de la base de données des documents vantés dans leurs publications ouvre la perspective d'une étude complète de cette dynastie de libraires et experts dont les continuateurs sont aujourd'hui encore, sur le marché privé parisien, des acteurs importants du commerce de l'écrit.

Annexes

Annexe A – Extraits de la documentation du projet e-ditiones portant sur l'encodage des catalogues

Why encoding manuscript sales catalogs in XML-TEI?

The aim of the project is to define an efficient way of processing and structuring scanned catalogs in order to build a searchable database of manuscripts for sale on the private market during the 19th and the early 20th century. It is also meant to link some of the registered sales with other sources : manuscripts, indices, etc.

On the private market, manuscripts and autographs can be sold either at a fixed price or during an auction sale. The process described below can be apply to fixed prices catalogs and auction sale catalogs. In both cases, the catalog is often not only composed of a list of entries presenting documents for sale. Fixed prices catalog are often inserted in reviews that also feature articles, obituaries, advertising, etc. The first pages of auction sales catalogs provide information about the material organization of the date (location, people involved, etc.).

The aim of the project is not to provide scholarly editions of manuscript sale catalogs and periodicals dedicated to autographs, thus only list of entries are encoded in the final documents.

Processing manuscripts sale catalogs, from PDF to XML-TEI

From the scanned catalog to the final XML-TEI document, several transformations are applied on the electronic file:

Preparing PDF

Using the tool `cpdf` (for “Coherent PDF Command Line Tools”, more information available [here](#)), a new PDF is created, containing only the pages that record lots for sale. This extraction aimed at improving the quality of output in the subsequent steps. Indeed, OCR and HTR models do not fully perform on very ornate pages, such as catalogs cover. Furthermore, GROBID-dictionaries is a software meant to deal only with encyclopedic-like data and not with articles or advertising.

After downloading and building `cpdf` on one’s computer (the steps to follow are [here](#)) and going to the relevant folder, the command line to extract some pages from a PDF is: `cpdf PDF.pdf X-Y -o selection.pdf` where X is the first page with encyclopedic-like entries and Y the last one

OCRising and HTRising catalogs

This trimmed PDF is imported in Transkribus and processed in two steps. An OCR model is firstly applied to recognize the layout of the document. A HTR model specifically trained on manuscript catalogs such as the *Revue des autographes* is then applied to enhance the quality and the accuracy of the output. A file that contains the PDF and the text is finally exported from Transkribus.

Automatically structuring catalog entries

This exported PDF is processed by GROBID-dictionaries, a software designed to automatically structure lexicographical resources. Because catalog entries, especially fixed-prices publications, have a structure similar to dictionary entries.

GROBID-dictionaries is a machine-learning software. Models need to be trained before being used. Specific training data were specifically developed for automatically encoding manuscripts sale catalogs. There can be downloaded from [this repository](#).

In function of documents that are processed, restraining at certain levels training data can provide better results.

At dictionary body segmentation level, the best results are obtained when using training data that most closely resemble the processed document.

- If the document is composed of fixed-price entries with a punctuation between the lot number and the author of the letter, the more relevant data to use are type RDA 1 and type RDA2.
- If the document is composed of fixed-price entries without a punctuation between the lot number and the author of the letter, the more relevant data to use are type LAV.
- If the document is a auction sale catalog, the more relevant data to use are type AUC.

At lexical entry, form and sense levels:

- If the document has exactly the same layout as the *Revue des autographes*, the results are better if you only use training data from type RDA1 and type RDA2.
- In all other cases, the results are better if you use all training data available

From GROBID output to a valid XML-TEI valid document

Currently (in May 2019), the encoded output by GROBID-dictionaries does not comply with the TEI P5 guidelines. Hence, it is necessary to transform the output downloaded from GROBID-dictionaries thanks to the XSLT stylesheet `grobid_transformation.xslt` before trying to validate it with the schema.

Assessment and correction of the final XML-TEI output

The next step is to associate the document with the schema `schema_grobid_output.rng`. If using Oxygen XML editor, it is important to tick the box "Add additional associations for embedded schematron rules". These schematron rules ensure that each entry is well structured (with at least a number, an author name and an autograph description) and that the entries numbers follow each other. It enables to spot and correct inaccuracies in the final document structure.

If the encoded catalog does not include lot numbers or does not feature a continuous numbering, it is important to create the numerotation by hand.

- If the catalog is structured by vacations, the number to associate with the item is of the following form : `CATXXXXX_dYeZZZ`, with `XXXXXX` the number of the catalogue, `Y` the number of the day of the vacation and `ZZZ` the number of the entry in a given vacation.
- If the catalog does not have any specific structure nor numbering, the number to associate with the item is of the following form : `CATXXXXX_pYeZZZ`, with `XXXXXX` the number of the catalogue, `Y` the number of the page and `ZZZ` the number of the entry in a given vacation.

After the global structure of the document has been checked and corrected, the XML-TEI document is transformed thanks to a second XSLT stylesheet named `catalog_final_transformation.xsl`. Every item is especially given a single `xml:id` and can therefore be linked with a relevant manuscript description. Finally, the document is associated with the global schema of e-ditiones named `odd_editiones.rng`

The structure of an encoded catalog

The TEI header

When a file is processed by GROBID-dictionaries, a header is automatically generated. However, it needs to be completed and restructured to meet the requirements of the TEI P5 Guidelines. As it is not easy to train an OCR or a HTR on first pages, the information available will not be automatically processed but dispatched in the `SourceDesc` element of the TEI Header.

The `<fileDesc>`

It contains four elements:

- the `<titleStmt>` with child `<title>` (with the same name as the file name composed of `CAT`, an underscore and six digits. The distinctive number corresponds with the order in which the documents are encoded) and `<respStmt>`. This last element contains a `<persName>` (with the name of the person responsible for the

creation of the electronic file) and an element `<date>` included in an element `<resp>`

- the `<extent>` to be completed when the document is ready to be uploaded.
- the `<publicationStmt>`, identical from a document to another

```
<publicationStmt>
  <publisher>Université de Neuchâtel</publisher>
  <availability status="restricted">
    <licence target="https://creativecommons.org/licenses/by/2.0">Attribution 2.0
    Generic (CC BY 2.0)</licence>
  </availability>
</publicationStmt>
```

The `<sourceDesc>`

The first child element of the `<sourceDesc>` is the `<bibl>`. This element contains a short bibliographical description of the print catalog.

- The `<title>` is the name of the catalog. If the catalog is a periodical, the full title is reproduced in this field. If the catalog is an auction sale catalog, in keeping with the cataloging choices made by the libraries that keep such records, we chose to indicate in this field only the first part of the title – the one which is printed bigger than the rest - and the date of the sale. The cuts in the title are replaced by the sign [...]. The rest of the relevant information present in the full title is indicated in subsequent fields of the `<sourceDesc>`.
- The `<num>` is used if the catalog is a periodical publication.
- The `<editor>` corresponds to the person responsible for the establishment of its content. When the catalog is inserted in a review, the editor is the publication director, that is to say the person responsible for the accuracy of the content.
- The `<publisher>` corresponds to the person in charge of its distribution. There can be several publishers for one catalog, especially when it is an auction sale advertised internationally.
- The `<pubPlace>` corresponds to the address of the publisher.
- The `<date>` corresponds to the date of the publication of the catalog (which can differ with the date of the advertised sale).

The second child element of the `<sourceDesc>` is the `<listEvent>`. This section aims at recording the events related to the sale described in a catalog. They are only relevant in case of an auction sale and they can be of four different types :

- the auction sale itself. In an `<event>` with the attribute `@type` of value `auction` people and organizations that took part in the sale, its location (in `<address>` and `<addrLine>` and its date are listed). When the name of an organization is mentioned, it is included in an element `<orgName>`. When the name of a person is mentioned, it is included in an element `<persName>` and supplemented with an attribute `@type` that indicates his role in the sale. Only the following values are possible :
 - “auctioneer” (person leading the sale)
 - “collector” (person whose collection is sold)
 - “expert” (person guaranteeing the authenticity of the items for sale)
 - “sales_assistant” (if the assistant is not granted the title of expert)
 - “seller” (if the person at the origin of the sale is not the person who collected the documents, e. g. their heir).
- the publication of a record of the sale and prices reached by the manuscripts (e.g. in the press). If such a record exists, it is recorded in an `<event>` type=“record” and its reference is reproduced in an element.
- the existence of corresponding auctioneer minutes. If such a record exists, it is recorded in an `<event>` with the attribute `@type` minutes” and its reference is reproduced in a `<p>`.
- the existence of an identification of the collector if the sale was anonymous. If such a record exists, it is recorded in an `<event>` with the attribute `@type` “identification” and the sources are listed in a `<listBibl>`.

The third section of the `<sourceDesc>` is the `<listWit>`, This element records and describes the different instances of the catalog that were consulted. Sales catalogs are generally hard to find and copies may vary given the fact that some of them are annotated. Each element witness has an attribute `xml:id`. If an online version of one witness is available, the link is indicated by a child element `<ptr>` and an attribute `@target`.

This fourth section of the `<sourceDesc>` is the element `<listBibl>`. It lists the bibliographical sources that give addition information on the sale.

The `<encodingDesc>`

In the `<encodingDesc>`, four different elements are needed.

- The first one is the element `<samplingDecl>`(with a child element `<p>`). It states the choice to process only the parts of catalogs that are list of entries: “This electronic version of the catalog only reproduces the entries that correspond to documents for

sale. All text preceding or succeeding the list of documents for sale is not reproduced below.”. This precision is added automatically thanks to the XSLT stylesheet.

- The element `<application>` concerning Transkribus is also automatically created. The version used – if not 1.6 – and the date have to be added by hand.

```
<application version="1.6"
  ident="Transkribus" when="2019-05-15">
  <label>Transkribus</label>
  <ptr target="https://transkribus.eu/Transkribus/" />
</application>
```

- The element `<application>` concerning GROBID-dictionaries and its child elements have to be transformed to be TEI-P5 conforming.
 - The attribute `version` has to be trimmed down, by hand, to a series of numbers.
 - The attribute `date` has to be trimmed down, by hand, to a date whose format is YYYY-MM-DD.
 - The element `<ref>` has to be transformed in a element `<label>` (XSLT transformation).
 - The attribute `@target` has to be associated to a new element `<ptr>` (XSLT transformation).

The element `<listPref>` is used to declare prefixes to simplify the declaration of indexed names described in the folders PE1 or PE2.

The body

Each entry in the GROBID-dictionaries output is transformed in an element `<item>`. Because 19th century manuscripts catalogs do not always provide the reader with extensive information about documents for sale, using the element `<item>` rather than the element `<object>` seems sufficient. Each element `<item>` can contain up to six child elements:

- the mandatory element `<num>` for the lot number.
- the mandatory element `<name>` for the name mentioned at the beginning of the entry. Most of the times, it is the author of the letter that is mentioned. However, when it comes to really famous characters, the letters sent to them will also be listed under their name (e.g. Louis XIV to register a letter sent to the king), hence the two possible values of the `@type` : "author" and "recipient".

Annexes

- the non-mandatory element `<trait>` (with a child element `<p>`) for the description of the author provided by the catalog.
- The mandatory element `<desc>` for the description of the document(s) for sale. In a fixed-price catalogs, this sequence ends with the price of the autograph(s). Each element `<desc>` can have an attribute `@ref` in order to point to another XML-TEI document. The aim is to link a letter or a series of letters to already known manuscripts, described thanks to an element `<msDescription>`.
- The non-mandatory element `<note>` if the description of the document is followed by a quotation or details about the circumstances in which the letter was sent.
- The non-mandatory element `<add>` for the auction sales catalogs where the price reached during the auctions is recorded. The element `<add>` has two attributes: “`@place`”, to specify where the note is written in relation to the note, and “`@hand`” to refer to the `xml:id` of the catalog instance where the prices are written.

Some catalogs are ordered by themes, hence divided by subheads. Using different lists is the best way to encode them. This operation must be done by hand. Each subhead is then encoded as the head of one list.

Scaling up Automatic Structuring of Manuscript Sales Catalogues

Lucie Rondeau du Noyer

{surname.name@chartes.psl.eu}

Ecole des Chartes, Paris

Simon Gabay

{surname.name@unine.ch}

Université de Neuchâtel

Mohamed Khemakhem

{name.surname@inria.fr}

Inria, team ALMAAnCH, Paris

Centre Marc Bloch, Berlin

Université Paris Diderot, Paris

Laurent Romary

{name.surname@inria.fr}

Inria, team ALMAAnCH

Keywords: Machine learning, manuscript sales catalogues, 19th c. France

Manuscript Sales Catalogues (MSC) are highly important for authenticating documents and studying the reception of authors. Their regular publication throughout Europe since the beginning of the 19th c. has consequently raised the interest around scaling up the means for automatically structuring their contents.

Following successful first encoding tests with *GROBID-Dictionaries* [1,2] on a single MSC collection [3], we aim in this paper to present the results of more advanced tests of the system's capacity to handle a larger corpus with MSC of different dealers, and therefore multiple layouts.

Corpus

Four different types of catalogues published between the middle of the 19th c. and the beginning of the 20th c. have been tested.

248 **Verdi** (G.), illustre compositeur, auteur du *Trouvère*. — L. a. s. à Escudier ; Naples, 3 nov. 1849, 2 p. 1/4 in-8, cachet. 15 »
Curieuse lettre sur l'état de Rome après l'entrée des Français. L'Italie n'est plus qu'une large et belle prison. Ce beau climat, ces belles montagnes, ces magnifiques cités sont un paradis pour la vue, mais un enfer pour le cœur. Le gouvernement français à Rome n'est pas meilleur que les gouvernements de l'Italie. Les Français ont cherché à conquérir l'amour des Romains, mais jusqu'à présent ceux-ci sont restés dignes et froids....

Figure 1 - Type 1: *Revue des autographes*, Gabriel Charavay. (Première série N°42, Decembre 1874)

82 **Humboldt** (Alexandre de), célèbre naturaliste, qui fut l'un des créateurs de la géographie botanique (1769-1859). — L. a. s. à M. Laugier, membre de l'Institut, à l'Observatoire ; 1/4 de p. in-8, adresse aut. signée. 12 »
Il lui présente M. Charles Ritter, l'illustre géographe qui est avide d'entendre la parole de notre Maître.

Figure 2 - Type 2: *Revue des autographes, des curiosités de l'histoire et de la biographie*, Gabrielle Charavay (Seconde série N°56, 1934)

11. **Catherine de Médicis**, reine de France. L. aut. sig., à Monsieur le Conestable. Sans date. 3/4 de p. in-fol., rognée presque jusqu'à la marge intérieure. 20 »
Elle l'entretient au sujet d'un prisonnier, et lui dit que le roi son fils est allé courir un sanglier qu'il a fait mener au parc.

Figure 3 - Type 3: *Catalogue de lettres autographes et manuscrits*, Auguste Laverdet (N°1, April 1856.)

54. **FRANÇOIS I^{er}**, roi de France, n. 1494, m. 1547.
P. s., sur vélin; Saint-Germain-en-Laye, 30 mars 1526, 1 p. in-4° oblong.
Mandement à Jehan Grolier, trésorier des guerres (le célèbre bibliophile) de payer à Jehan de Vesin, homme d'armes de la compagnie du grand écuyer, pour le dernier quartier de 1525 et le premier de 1526 nonobstant qu'il n'ait comparu ni aux montres ni aux revues qui ont été faites.

Figure 4 - Type 4: *Catalogue d'une intéressante collection de lettres autographes...*, Etienne Charavay (December, 14th 1908)

Experiment

To parse the presented MSC corpus we followed the same encoding presented in earlier experiments [3]. We tried then to focus our experiments on two aspects: feature engineering and cumulative samples training.

For the former we tested tuning the GROBID models at three levels of segmentation following two variations: unigram and bigram features. The difference between the two

categories is that a label predicted by a trained model is based only on the features of the input token - case of unigram - where a bigram feature template takes also the label of the previous token into consideration.

For the second experimenting aspect we tested the performance of the system on models trained separately for each layout and a general model with all the data.

To that end we used 10 annotated pages for training and 5 others for evaluation, chosen to be representative of each series of catalogues. [4]

	GROBID Models		
MSC	Lexical Entry	Form	Sense
<i>Type 1</i>	72.49	93.72	73.62
<i>Type 2</i>	57.07	84.42	60.5
<i>Type 3</i>	60.07	74.71	48.07
<i>Type 4</i>	60.58	92.03	40.91
<i>All types mixed</i>	63.99	86.42	54.81

Table 1: All fields F1-score of **Unigram Feature Templates**

	GROBID Models		
MSC	Lexical Entry	Form	Sense
<i>Type 1</i>	98.05	100	98.71
<i>Type 2</i>	99.16	95.02	90.78
<i>Type 3</i>	96.01	92.89	88.1
<i>Type 4</i>	92.78	96.83	86.64
<i>All types mixed</i>	95.43	97.44	92.77

Table 2: All fields F1-score of **Bigram Feature Templates**

Conclusion

Two important conclusions can be drawn from this test. First, bigram feature templates are more efficient than unigram templates. Second, a general model potentially increases scores for certain levels (*form* and *sense*) but not all of them (*lexical entry*), which raises the question of the pertinence of a hybrid model, choosing the best solution for each level.

References

- Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, et al.. Automatically Encoding Encyclopedic-like Resources in TEI. *The annual TEI Conference and Members Meeting*, Sep 2018, Tokyo, Japan.
- Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *electronic lexicography*, eLex 2017, Sep 2017, Leiden, Netherlands.
- Mohamed Khemakhem, Axel Herold, Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX workshop at LREC 2018*, May 2018, Miyazaki, Japan. 2018.
- Lucie Rondeau du Noyer, Simon Gabay, Mohamed Khemakhem, Laurent Romary. *Training and evaluation data for encoding Manuscript Sales Catalogues with GROBID-dictionaries*, Paris: École nationale des chartes (PSL)/Neuchâtel: université de Neuchâtel, 2019, https://github.com/lairaines/grobid_TEI_2019.

Biographies

Lucie Rondeau du Noyer is a “History and New Technologies” masters’ student at the Ecole Nationale des Chartes and a graduate student at the Ecole Normale supérieure (Paris).

Simon Gabay is post-doc at the University of Neuchâtel (Switzerland), where he teaches DH and carries research on 17th c. French literature and modern manuscripts. He is currently working on a database of sold manuscripts in 19th c. France.

Mohamed Khemakhem is a PhD candidate at Inria, team ALMAAnaCH (Paris), Paris 7 University and Centre Marc Bloch (Berlin). His research is focused on parsing lexical and encyclopedic legacy resources using standard-based machine learning models.

Laurent Romary is senior researcher at Inria, team ALMAAnaCH and works on data modelling and standards in humanities computing.

Annexe C - Entrées correspondant à la requête « Danton » après requête sur e-ditiones

n° 22 lot 4374	Danton (Georges), le célèbre conventionnel L. s., comme ministre de la justice, au directoire du départ, de Paris; 25 août 1792, 1 p. in-f. Belle pièce. 27 francs Les prisonniers de Bicêtre, condamnés a y être détenus jusqu'à leur majorité, instruits du décret qui la fixe à 21 ans, demandent s'ils doivent jouir du bénéfice de cette loi. « L'affirmative, dit-il, n'est pas douteuse. Les citoyens français doivent, sous tous les rapports, jouir du bénéfice de la loi qui avance la majorité, et cette loi est trop formelle pour notre point applicable aux détenus à Bicêtre.”
n 24 lot 76	Danton (Georges), le célèbre conventionnel Belle pièce sig. de lui, Ch. Delacroix, Camus et Gossuin, comme commissaires de la Convention à l'armée de Belgique; Liège, 11 janv. 1793, 1/2 p. in-4. 25 francs Les pièces signées de Danton sont très-rares, surtout comme commissaire, en Belgique
n 26 lot 97	Danton (Georges), conventionnel célèbre L. s. à l'Agent du Trésor public; 16 août 1792, 1/2 p. in-f. Rare. 15 francs
n 32 lot 36	Danton (Georges), célèbre conventionnel, sacrifié à la jalousie de Robespierre L. s., au Directoire de Paris; 25 août 1792, 1 p. in-f. 30 francs Alors ministre de la justice, et consulté sur la question de savoir si les jeunes détenus de Bicêtre, condamnés a y être renfermés jusqu'à leur majorité, doivent jouir du bénéfice de la loi qui vient d'abaisser la majorité à 21 ans, il n'hésite pas à se prononcer pour l'affirmative. (Il s'agissait de la capacité de vote pour les élections à la Convention nationale.)
n 36 lot 39	Danton (Georges), le célèbre conventionnel L. sig., comme ministre de la justice, à M. Le Roux, chef du bureau du sceau, 30 août 1792, 1 p. in-f. 30 francs Belle et curieuse lettre relative au renouvellement du personnel administratif. « Les circonstances, dit-il, m'obligeant de mettre à la tête de mes bureaux des personnes de confiance, et qui, m'étant connues depuis très-longtemps, me sont également indiquées par l'opinion publique, j'ai dû disposer en leur faveur d'une grande partie des places de mon administration. » Celle que M. Le Roux remplit depuis 23 ans étant de ce nombre, il l'invite à faire valoir ses droits à la retraite
n 40 lot 29	Danton (Georges), célèbre conventionnel Pièce sig. comme ministre de la justice, 20 août 1792, 4 p. in-f. 25 francs Décret de l'Assemblée nationale par lequel elle envoie des commissaires à l'armée du maréchal Luckner : curieuse pièce historique sur les premiers représentants envoyés aux armées dans les départements.

Annexes

n 42 lot 54	Danton (G.), célèbre conventionnel Pièce 2 fois signée, sig. aussi des autres membres du Conseil exécutif, 4 sept. 1792, doublée. 20 francs Sauf-conduit pour le citoyen Bruault sous-directeur des hôpitaux ambulants au camp sous Paris, se rendant aux environs de la ville. (C'était le moment des massacres dans les prisons. Les portes de Paris étaient fermées, et il fallait un ordre officiel de ce genre pour sortir.) Outre la double signature de Danton, la pièce porte celles de Clavière, de Monge, et de Grouvelle, secrétaire du Conseil exécutif
n 51 lot 32	Danton (Georges), le célèbre conventionnel L. sig. de Danton, Barère, Lindet, Cambon et autres, comme membres du Comité de salut public, à la Municipalité de Charleville; Paris, 26 av. 1793, 1 p. 1/2 in-f. 40 francs Ordre à la Municipalité de Charleville de faire transporter 2,000 fusils à l'arsenal de Péronne. (Danton n'ayant été que très-peu de temps membre du Comité de salut public, ses signatures en cette qualité sont très-rares.)
n 56 lot 30	Danton, célèbre conventionnel. Pièce sig. ; 30 août 1792, 1 p. 1/4 in-f. 25 francs Relative à l'arrestation de l'économe des missionnaires du Mont- Valérien (près Paris), et à la conservation du mobilier de cette maison
n 57 lot 27	Danton (Georges), le célèbre conventionnel Pièce sig. comme président du district des Cordeliers; 5 nov. 1789, 1 p. in-f., cachet. 50 francs Intéressante pièce historique. Le district, informé que le général Lafayette a demandé le rappel des gardes du corps (éloignés après les journées des 5 et 6 octobre), considérant que la garde de la personne sacrée du roi ne peut être mieux confiée qu'à la nation elle-même, proteste contre la formation de tout corps particulier qui tendrait à priver l'ensemble des citoyens de la garde d'un prince restaurateur de la liberté française. -Il a ainsi signé : D'Anton, président
n 59 lot 60	Danton (Georges), le célèbre conventionnel Pièce sig., an IV, 1 p. 1/2 in-f., coupure dans une marge atteignant le texte. 20 francs

Bibliographie

1. Histoire du livre et de l'écrit : généralités

- BARBIER (Frédéric), DUBOIS (Thierry) et SORDET (Yann), *De l'argile au nuage, une archéologie des catalogues : II^e millénaire av. J. C. - XXI^e siècle*, Paris / Genève, Bibliothèque Mazarine / Bibliothèque de Genève / Éditions des Cendres, 2015.
- DELSAERDT (Pierre) et SORDET (Yann) (éd.), *Lectures princières & commerce du livre. La bibliothèque de Charles III de Croÿ et sa mise en vente (1614)*, 2 vol., Enghien / Paris, Fondation d'Arenberg / Éditions des Cendres, 2017.
- DOCQUIER (Gilles), « Le document autographe, une « non-réalité » pour l'historien ? », *Le Moyen Âge*, Tome CXVIII, n° 2, août 2012, p. 387-410.
- FAURE (Chantal) (dir.), *Catalogues de libraires et d'éditeurs, 1811-1924*, Paris, Bibliothèque nationale de France, 2003.
- FRAISSE (Luc) (dir.), *Le manuscrit littéraire. Son statut, son histoire, du Moyen Âge à nos jours*, Paris, Klincksieck / ADIREL, 1998.
- FERRAND (Nathalie), « L'Ancien et le Nouveau Régime des manuscrits de travail », *Genesis. Manuscrits – Recherche – Invention*, n° 34, avril 2012, p. 7-17.
- SOREL (Patricia) et LEBLANC (Frédérique), *Histoire de la librairie française*, Paris, Éditions du Cercle de la librairie, 2008.
- TURCAN-VERKERK (Anne-Marie) et BERTRAND (Paul), « BIBLISSIMA : Bibliotheca bibliothecarum novissima, an observatory for the written cultural heritage of the Middle Age and the Renaissance », dans SAOU-DUFRENE (Bernadette) et BARBIER (Benjamin) (éd.), *Heritage and Digital Humanities. How should training practices evolve?*, Berlin, Lit, 2014, p. 129-139.

2. Le marché des autographes depuis le XIX^e siècle

a. Sources imprimées

- *Inventaire des autographes et des documents historiques composant la Collection de M. Benjamin Fillon...*, 5 volumes, Paris / Londres, E. Charavay / F. Naylor, 1877-1883.
- *The collection of autograph letters and historical documents formed by Alfred*

Bibliographie

Morrison. Second series, 1882-1893, 3 vol, Londres, 1893.

- CHARAVAY (Étienne) et CALMETTES (Fernand), *Lettres autographes composant la collection de M. Alfred Bovet*, deuxième édition corrigée et augmentée, Paris, Charavay frères, 1887 [1884-1885].
- FONTAINE (Pierre-Jules), *Manuel de l'amateur d'autographes*, Paris, Paul Morta, 1836.
- LALANNE (Ludovic) et BORDIER (Henri), *Dictionnaire de pièces autographes volées aux bibliothèques publiques de la France, précédé d'observations sur le commerce des autographes*, Paris, Panckoucke, 1851.
- LALANNE (Ludovic) et BORDIER (Henri), *Dictionnaire de pièces autographes volées aux bibliothèques publiques de la France précédé d'observations sur le Commerce des autographes ... 3e et 4e livraisons*, Paris, Panckoucke, 1853.
- LESCURE (Adolphe) (de), *Les Autographes et le Goût des autographes*, Paris, J. Gay, 1865.
- CHARAVAY (Étienne), *La science des autographes, essai critique*, Paris, Charavay frères, 1887.
- MECKLENBURG (Günther), *Vom Autographensammeln. Versuch einer Darstellung seines Wesens und seiner Geschichte im deutschen Sprachgebiet*, Marbourg, J.A. Stargardt, 1963.

b. Perspectives artistiques, littéraires et anthropologiques

- DUFIEF (Pierre-Jean), « Correspondances et histoire littéraire (1850-1900) » dans FRAISSE (Luc) (dir.), *L'histoire littéraire à l'aube du XXI^e siècle : Controverses et consensus*, Paris, Presses Universitaires de France, 2005.
- FRAISSE (Luc), *Les fondements de l'histoire littéraire : de Saint-René Taillandier à Lanson*, Paris, Honoré Champion, 2002.
- GUICHARD (Charlotte) (éd.), *De l'authenticité : une histoire des valeurs de l'art (XVI^e-XX^e siècle)*, Paris, Publications de la Sorbonne, 2014.
- PETY (Dominique), *Poétique de la collection au XIX^e siècle : Du document de l'historien au bibelot de l'esthète*, Nanterre, Presses universitaires de Paris Nanterre, 2012.
- POTIN (Yann), « Le prix de l'écrit », *Genèses*, n° 105/4, novembre 2016, p. 3-7.
- PREISS (Nathalie), éd., *Le XIX^e siècle à l'épreuve de la collection*, Reims, ÉPURE, 2018.

c. Perspectives socio-économiques

- BOLTANSKI (Luc) et ESQUERRE (Arnaud), « La « collection », une forme neuve du capitalisme la mise en valeur économique du passé et ses effets », *Les Temps Modernes* n°679, octobre 2014, p. 5-72.
- BOLTANSKI (Luc) et ESQUERRE (Arnaud), *Enrichissement : une critique de la marchandise*. Paris, Gallimard, 2017.
- EVEN (Pascal), « Les archives : un marché? », *Pouvoirs*, n° 153/ 2, avril 2015, p. 95-107.
- MENDOZA (Ileana Miranda), « L'économie du patrimoine écrit : le marché des autographes », thèse de doctorat, Université Paris 1 Panthéon-Sorbonne, 2010.
- MENDOZA (Ileana Miranda), GARDES (François), GREFFE (Xavier) et PRADIER (Pierre-Charles), « *Are Autographs Integrating the Global Art Market? The Case of Hedonic Prices for French Autographs (1960-2005)* », *Documents de travail du Centre d'Économie de la Sorbonne*, n° 2014.53. halshs-01025095
- MARCILLOUX (Patrice), *Les ego-archives : traces documentaires et recherche de soi*, Rennes, Presses universitaires de Rennes, 2013.

d. Vendeurs et collectionneurs d'autographes depuis le XIX^e siècle

- BODIN (Thierry) et NEEFS (Jacques), « Les autographes. Entretien », *Genesis (Manuscripts-Recherche-Invention)*, n° 7, 1995, p. 177-184.
- BODIN (Thierry), « Les grandes collections de manuscrits littéraires » dans *Les ventes des livres et leurs catalogues, XVII^e-XX^e siècle*, Paris, Publications de l'École nationale des chartes, 2000.
- CAPPON (Lester J.), « *Walter R. Benjamin and the Autograph Trade at the Turn of the Century* », *Proceedings of the Massachusetts Historical Society*, n° 78, 1966, p. 20-37.
- CHARREIRE (Magali), « Un marchand d'histoire au XIX^e siècle ». *Genèses*, n° 105/4, novembre 2016, p. 36-56.
- FAIVRE d'ARCIER (Catherine), *Lovenjoul (1836-1907) : une vie, une collection*. Paris, Kimé, 2007.
- FAIVRE d'ARCIER (Catherine), « Lovenjoul et ses catalogues au cœur d'un service de commissions » dans *Le Livre entre le commerce et l'histoire des idées : Les catalogues de libraires (XV^e-XIX^e siècle)*, Paris, Publications de l'École nationale des

Bibliographie

chartes, 2011.

- LAUER (Joshua), « *Traces of the Real: Autographomania and the Cult of the Signers in Nineteenth-Century America* », *Text and Performance Quarterly* 2007 (27). DOI : 10.1080/10462930701251207
- LE BAIL (Marine), « L'amour des livres la plume à la main : écrivains bibliophiles du XIX^{ème} siècle », thèse de doctorat, Université Toulouse-Jean Jaurès, 2016.
- MACCIONI RUJU (Alessandra) et MOSTERT (Marco), *The Life and Times of Guglielmo Libri (1802-1869): Scientist, Patriot, Scholar, Journalist and Thief, A Nineteenth-Century Story*, Hilversum, Verloren Publishers, 1995.
- MUNBY (Alan Noel Latimer), *The Cult of the Autograph Letter in England*, Londres, Athlone Press, 1962.
- NICOLAS (Alain) (dir.), *Les autographes*, Paris, Maisonneuve & Larose, 1988.
- NORTIER (Michel), « Le sort des archives dispersées de la Chambre des Comptes de Paris », *Bibliothèque de l'Ecole des chartes*, n° 123/2, 1965, p. 460-537.

e. La famille Charavay

- TOURNEUX (Maurice), *Étienne Charavay : sa vie et ses travaux*, Paris, E. Charavay, 1900.
- RIBÉMONT (Thomas), « Les historiens chartistes au cœur de l'affaire Dreyfus », *Raisons politiques*, n° 18-2, 2005, p 97-116.
- <http://maitron-en-ligne.univ-paris1.fr/spip.php?article28465>, notice CHARAVAY frères [CHARAVAY Gabriel et CHARAVAY Jean] par Notice revue, corrigée et complétée par Jacques Grandjonc, version mise en ligne le 20 février 2009, dernière modification le 17 décembre 2018.

3. Le traitement numérique des catalogues de ventes

a. Généralités sur l'apprentissage supervisé

- MARTIENNE (Emmanuelle), CLAVEAU (Vincent) et GROS (Patrick), « Application des Champs Conditionnels Aléatoires à l'étiquetage de flux télévisuel », *RFIA - Reconnaissance des Formes et Intelligence Artificielle*, janvier 2012, Lyon. hal-00656547
- CASARI (Alice) et ZHENG (Amanda), *Feature Engineering for Machine Learning*,

Bibliographie

Sebastopol, O'Reily Media, 2018.

b. Structuration automatique des documents

- ARES OLIVEIRA (Sofia), SEGUIN (Benoit) et KAPLAN (Frederic), « *dhSegment: A generic deep-learning approach for document segmentation* », *CoRR*, avril 2018 (article révisé en août 2019).
- CUADRA (Ruth) et MICHELS (Suzanne), « *Publishing German Sales, A Look under the Hood of the Getty Provenance Index* », *IRIS / Behind the Scenes at the Getty*, 2013. Disponible à l'adresse :<https://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index>
- LOPEZ (Patrice) et ROMARY (Laurent), « *GROBID - Information Extraction from Scientific Publications* », janvier 2015. [hal-01673305](#)
- RIONDET (Charles) et FOPPIANO (Luca), « History Fishing When engineering meets History ». *Text as a Resource. Text Mining in Historical Science #dhiha7*, Paris, France : Institut Historique Allemand (Paris), 2017. [hal-01830713](#)

c. GROBID-dictionaries

- KHEMAKHEM (Mohamed), FOPPIANO (Luca) et ROMARY (Laurent), *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. electronic lexicography*, eLex 2017, Septembre 2017, Leyde, Pays-Bas. [hal-01508868v2](#)
- KHEMAKHEM (Mohamed), HEROLD (Axel) et ROMARY (Laurent), *Enhancing Usability for Automatically Structuring Digitised Dictionaries*, *Globalex 2018*, Globalex workshop at LREC 2018, Mai 2018, Miyazaki, Japon. 2018. [hal-01708137v2](#)
- BOHBOT (Hervé), FRONTINI (Francesca), LUXARDO (Giancarlo), KHEMAKHEM (Mohamed) et ROMARY (Laurent), *Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré*, *Globalex 2018*, Globalex workshop at LREC 2018, Mai 2018, Miyazaki, Japon. [hal-01728328](#)
- KHEMAKHEM (Mohamed), BRANDO (Carmen), ROMARY (Laurent), MÉLANIE-BECQUET (Frédérique) et PINOL (Jean-Luc), « Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. », *JADH2018, Leveraging Open Data*, Septembre 2018, Tokyo, Japon. [hal-01814189](#)

Bibliographie

- KHEMAKHEM (Mohamed), ROMARY (Laurent), GABAY (Simon), BOHBOT (Hervé), FRONTINI (Francesca) et al., *Automatically Encoding Encyclopedic-like Resources in TEI*, The annual TEI Conference and Members Meeting, Septembre 2018, Tokyo, Japon. [hal-01819505](#)
- LINDEMANN (David), KHEMAKHEM (Mohamed) et ROMARY (Laurent), *Retro-digitizing and Automatically Structuring a Large Bibliography Collection*, European Association for Digital Humanities (EADH) Conference, Décembre 2018, Galway, Irlande. [hal-01941534](#)
- RONDEAU du NOYER (Lucie), GABAY (Simon), KHEMAKHEM (Mohamed) et ROMARY (Laurent), « *Scaling up Automatic Structuring of Manuscript Sales Catalogues* », *The annual TEI Conference and Members Meeting*, Septembre 2019, Graz, Autriche.

d. Encoder en XML-TEI

- BISSON (Marie), GOLOUBKOFF (Anne) et KUHRY (Emmanuelle), *Éditer un inventaire XML-TEI P5*, Université de Caen / Pôle Document Numérique, « Les Manuels du Pôle », mis en ligne le 25 février 2019.
- BURNARD (Lou,) *What is the Text Encoding Initiative ?*, Marseille, OpenEdition Press, 2014.
- NELSON (Brent), « *Curating Object-Oriented Collections Using the TEI* », *Journal of the Text Encoding Initiative*, n° 9, septembre 2016 - décembre 2017. DOI : [10.4000/jtei.1680](#).
- STADLER (Peter), ILLETSCHKO (Marcel) et SEIFERT (Sabine), « *Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>* », *Journal of the Text Encoding Initiative*, n° 9, septembre 2016 - décembre 2017. DOI : [10.4000/jtei.1680](#).

e. Fédérer

- BURROWS (Toby), HYVÖNEN (Eero), RANSOM (Lynn) et WIJSMAN (Hanno), « *Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts* », *Manuscript studies* 2019, Vol. 3 : Iss. 1 , Article 13. Disponible à l'adresse : https://repository.upenn.edu/mss_sims/vol3/iss1/13

Table des figures

Figure 1	Extrait d'une liste de lettres autographes à prix marqués	15
Figure 2	Liste de lettres autographes non assorties de leur description	16
Figure 3	Extrait d'un catalogue de vente aux enchères rédigé par Noël Charavay	20
Figure 4	Représentation schématique de la chaîne de traitement	26
Figure 5	Extrait du <teiHeader> décrivant le catalogue de la vente préparant la dispersion de la collection d'autographes Schuman en 1965	28
Figure 6	Extrait du <teiHeader> décrivant les modalités de la vente préparant la dispersion de la collection d'autographes Schuman en 1965	29
Figure 7	Capture d'écran du logiciel Transkribus : l'analyse de la mise en page est la première étape de traitement des catalogues de vente	32
Figure 8	Capture d'écran du logiciel Transkribus : après application du modèle HTR, le logiciel affiche la transcription réalisée en regard du catalogue numérisé	32
Figure 9	Modèle de départ du logiciel GROBID-dictionaries	34
Figure 10	Données d'entraînement où sont annotées manuellement les différentes entrées	35
Figure 11	Données d'entraînement où sont annotées manuellement les différentes composantes des entrées	36
Figure 12	Modèles d'encodage successifs d'un autographe proposé à la vente dans un catalogue à prix marqués	37
Figure 13	Données d'entraînement où sont annotées manuellement les différentes composantes des descriptions d'auteurs	38
Figure 14	Capture d'écran du terminal après entraînement de GROBID-dictionaries au niveau « <i>Sense</i> »	40
Figure 15	Capture d'écran de l'interface graphique de GROBID-dictionaries après sélection et traitement d'un catalogue au format PDF	41
Figure 16	Un exemple d'entrée mal structurée	42
Figure 17	Exemple d'entrée de type 1	45
Figure 18	Exemple d'entrée de type 2	45
Figure 19	Exemple d'entrée de type 3	45
Figure 20	Exemple d'entrée de type 4	46
Figure 21	Exemples d'entrées annotées et numérotées	50
Figure 22	Notice d'un catalogue de vente reliée manuellement à la description d'un manuscrit	51

Table des figures

Figure 23	Capture d'écran de la fiche descriptive de Gabriel Charavay	52
Figure 24	Identification grâce à l'application e-ditiones d'une lettre vendue avec la mention publicitaire « petite-fille de M ^{me} de Sévigné »	58
Figure 25	Identification grâce à l'application e-ditiones d'un dessin de Corot vendu par les Charavay	60
Figure 26	Extrait de la <i>Revue des Autographes</i> n° 58 proposant un classement des autographes par département	61

Table des tableaux

Tableau 1	Résultats f1 level-field obtenus au niveau « <i>Lexical entry</i> » (10 pages de données d'entraînement, 5 pages d'évaluation)	44
Tableau 2	All fields F1-score des modèles unigrammes	47
Tableau 3	All fields F1-score des modèles bigrammes	48
Tableau 4	Extrait du relevé des entrées commençant dans la <i>Revue des autographes</i> par le nom de Danton	60

Table des matières

Résumé.....	iii
Remerciements.....	vii

Introduction.....	1
-------------------	---

Première partie

Le marché des lettres autographes : historiographie et typologie des sources.....	5
1.1. Les catalogues de vente, une source majeure pour l’histoire du livre et l’histoire de l’art.....	7
1.1.1. Catalogues et histoire du livre : multiplication des éditions électroniques et des bases de données.....	8
1.1.2. Catalogues et histoire de l’art : chantiers de numérisation et étude des circulations.....	9
1.2. Le commerce des lettres autographes, un secteur du marché de l’art encore peu étudié.....	11
1.2.1. Le poids historique de la parole de l’expert vendeur.....	11
1.2.2. « Le prix de l’écrit » : l’émergence d’un questionnement interdisciplinaire.....	12
1.3. Les catalogues de vente de lettres autographes : une typologie.....	14
1.3.1. Catalogues d’enchères et catalogues à prix marqués : le reflet d’un marché dual	14
1.3.2. Les revues-catalogues : des supports commerciaux hybrides	16
1.3.3. L’activité éditoriale de la famille Charavay : un exemple de diversification des catalogues.....	17

Deuxième partie

Du catalogue à la base de données : méthodologie et évaluation d’une chaîne de traitement.....	23
2.1. Présentation synthétique de la chaîne de traitement des catalogues numérisés.....	25
2.1.1. Un traitement qui se concentre sur les entrées de catalogues.....	27
2.1.2. Principes et choix de l’encodage manuel.....	27
2.2. Présentation détaillée et évaluation des outils retenus.....	31
2.2.1 La reconnaissance automatique de caractères : Transkribus	31
2.2.2 La structuration automatique des données : GROBID-dictionaries	33

Table des matières

2.2.2.1 L'intérêt et les conditions d'une utilisation « métaphorique » du logiciel	33
2.2.2.2. Production de données d'entraînement et protocole d'évaluation	39
2.2.2.2.1. Présentation du principe de l'apprentissage supervisé et des modes d'évaluation.....	39
2.2.2.2.2. Premiers résultats décevants et importance du <i>feature engineering</i>	42
2.2.2.2.3. Modèle spécifique ou modèle général ?.....	45
2.2.2.3. Bilan : des résultats inégaux en fonction du type de catalogue.....	47
2.3. Chaîne de transformation XSLT et contrôle de la qualité des fichiers générés automatiquement	48
2.3.1. Un traitement intermédiaire pour rendre les fichiers générés compatibles avec le schéma général de la TEI et faciliter leur correction manuelle	48
2.3.2. Deuxième vague de transformations XSLT et schéma de validation final : penser l'insertion des catalogues de vente dans l'architecture globale du projet e-ditiones.....	49
 Troisième partie	
La constitution et la valorisation d'une base de données test : la première série de la Revue des Autographes, des curiosités de l'histoire et de la biographie	53
3.1. Le corpus test : une vitrine de l'activité commerciale des Charavay.....	55
3.2. L'application de consultation : mise en relation de la base de données des ventes avec les autres fichiers du projet e-ditiones	56
3.3. Une application au service de champs disciplinaires variés : exemples et premières perspectives.....	58
 Conclusion.....	 63
 Annexes.....	 67
Annexe A – Extraits de la documentation du projet e-ditiones portant sur l'encodage des catalogues.....	69
Annexe B - Abstract soumis et accepté à la conférence annuelle de la TEI (2019).....	77
Annexe C - Entrées correspondant à la requête Danton après requête sur e-ditiones.....	81
 Bibliographie.....	 83
 Table des figures.....	 89
Table des tableaux.....	91
Table des matières.....	93