Katabase: À la recherche des manuscrits vendus

Simon Gabay, Ljudmila Petkovic, Alexandre Bartz, Matthias Gille Levenson, Lucie Rondeau Du Noyer

Katabase: À la recherche des manuscrits vendus

Simon Gabay[1], Ljudmila Petkovic[2], Alexandre Bartz[3], Matthias Gille Levenson[4], and Lucie Rondeau du Noyer[5]

[1]Université de Genève (Suisse)
[2]Université de Neuchâtel (Suisse)
[3]Ecole nationale des Chartes (France) et université de Neuchâtel (Suisse)
[4]Ecole normale supérieure de Lyon (France)

Résumé

Databases exist for paintings, sculptures, drawings. . . circulating on the private market, but not for manuscripts. Such a tool being useful for scholars cataloguing sources or studying the reception of authors, we have designed an online application able not only to display sold items, but to reconcile several entries of a similar item sold multiple times. The algorithm developed for this reconciliation should open new horizons to those interested in the manuscript market, like economic historians.

Les marchés de l'art, des livres ou des manuscrits sont tous relativement anciens, mais ne bénéficient cependant pas des mêmes outils pour la recherche. Des bases de données comme *ArtPrice* [1] existent pour les beaux-arts (peinture, sculpture. . .) et permettent de recenser les ventes. Des outils équivalents existent pour les livres anciens aux États-Unis [2], au Royaume-Uni [3], en Allemagne [4] ou en France [5].

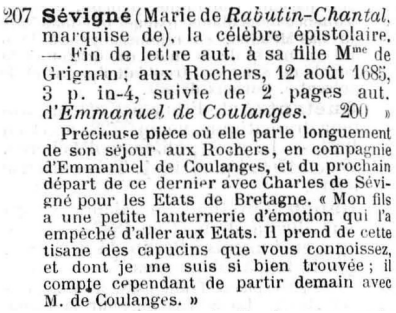
SÉVIGNÉ (Madame de). Lettres de sa Famille et de ses Amis. 10 vols., 1818. With Lettres inédites de Mme. de Sévigné, 1 vol., 1814; and Memoirs de M. de Coulanges suivis de lettres inédites de Madame de Sévigné, 1 vol., 1820. Paris, 1814-20. Together 12 vols., 8vo.
Lev. mor., unc., by Niédrée (with numerous extra illustrations, comprising portraits, plates, orig. drawings, and autograph letters, inserted), Meacham, A., Mar. 19, '17. (1176) $100.00.

Illustration 1 – *American Book Prices Current*, 1917, p. 35.

1. `https://fr.artprice.com`.
2. *American Book-Prices Current*, New York, 1894/95-. *ABPC* tend avec le temps à répertorier de plus en plus de ventes européennes.
3. *Book-Auction Records*, London, 1902–1997 et *Book Prices Current*, London, 1887-1952.
4. *Jahrbuch der Auktionspreise für Bücher, Handschriften und Autographen*, Hamburg, 1950-. Au début *Jahrbuch der Auktionspreise für Bücher und Autographen*.
5. *L'Argus mensuel du livre ancien et moderne*, Promodis, Paris, 1981-, devenu *L'Argus du livre de collection & de l'autographe*.

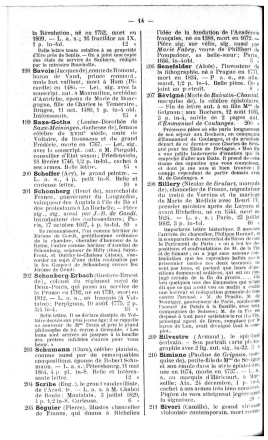
Si certains index pour les ventes de livres anciens recensent bien les autographes, tous le ne font pas [6], et les publications apparues tardivement ne reviennent pas sur les ventes passées. La documentation est donc disparate et fragmentaire, concernant une ressource de premier ordre pour les collectionneurs, mais aussi pour les philologues en quête de sources, les historiens du livre ou les adeptes de la *Rezeptionsgeschichte* qui peuvent s'intéresser aux prix ou aux noms des collectionneurs (cf. illustration 1).

1 Enjeu

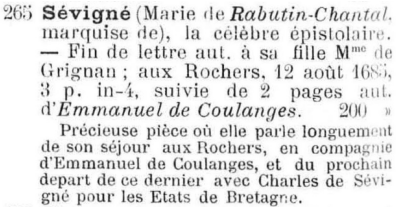

207 **Sévigné** (Marie de *Rabutin-Chantal*, marquise de), la célèbre épistolaire, — Fin de lettre aut. à sa fille Mme de Grignan ; aux Rochers, 12 août 1685, 3 p. in-4, suivie de 2 pages aut. d'*Emmanuel de Coulanges*. 200 »

Précieuse pièce où elle parle longuement de son séjour aux Rochers, en compagnie d'Emmanuel de Coulanges, et du prochain départ de ce dernier avec Charles de Sévigné pour les Etats de Bretagne. « Mon fils a une petite lanternerie d'émotion qui l'a empêché d'aller aux Etats. Il prend de cette tisane des capucins que vous connoissez, et dont je me suis si bien trouvée ; il compte cependant de partir demain avec M. de Coulanges. »

(a) Mai 1894, lot n°207

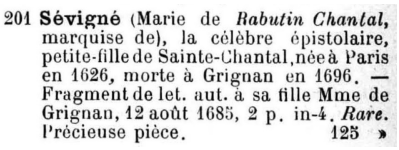


(b) Mai 1894


265 **Sévigné** (Marie de *Rabutin-Chantal*, marquise de), la célèbre épistolaire. — Fin de lettre aut. à sa fille Mme de Grignan ; aux Rochers, 12 août 1685, 3 p. in-4, suivie de 2 pages aut. d'*Emmanuel de Coulanges*. 200 »

Précieuse pièce où elle parle longuement de son séjour aux Rochers, en compagnie d'Emmanuel de Coulanges, et du prochain depart de ce dernier avec Charles de Sévigné pour les Etats de Bretagne.

(c) Juillet 1897, lot n°265


201 **Sévigné** (Marie de *Rabutin Chantal*, marquise de), la célèbre épistolaire, petite-fille de Sainte-Chantal, née à Paris en 1626, morte à Grignan en 1696. — Fragment de let. aut. à sa fille Mme de Grignan, 12 août 1685, 2 p. in-4. *Rare.* Précieuse pièce. 125 »

(d) Avril 1902, lot n°201

Illustration 2 – *Revue des autographes*

Si les principaux problèmes posés par la numérisation de catalogues comme la *Revue des autographes* (*e.g.* illustration. 2b) sont connus [7], tout comme les enjeux de

6. Ainsi l'apparition en Angleterre d'une publication éphémère : *Autograph Prices Current*, London, 1914-1922.

7. Simon Gabay, Lucie Rondeau Du Noyer et Mohamed Khemakhem, « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile : sfide e prospettive per l'Informatica Umanistica*, Milan, Italy, 2020 (Quaderni di Umanistica Digitale), p. 113-118, URL : `https://hal.archives-ouvertes.fr/hal-02388407`.

la détection d'un manuscrit revenant plusieurs fois sur le marché (cf. illustration. 2c et 2a), parfois sous une forme fragmentaire (cf. illustration. 2c et 2a vs illustration. 2d) [8], il nous a paru important d'améliorer notre algorithme de classification. Ce dernier doit en effet être implémentable dans une application web disponible en ligne, tout en étant capable de traiter de grandes quantités de données avec un maximum de précision.

RDA, May 1894 (N°166)	RDA, July 1897 (N°200)	RDA, April 1902 (N°257)
Sévigné	Sévigné	Sévigné
Fin de lettre aut.	Fin de lettre aut.	Fragment de let. aut.
12 août 1685	12 août 1685	12 août 1685
3 p.	3 p.	2 p.
in-4	in-4	in-4
200 francs	200 francs	125 francs

Tableau 1 – Informations clefs sur trois manuscrits.

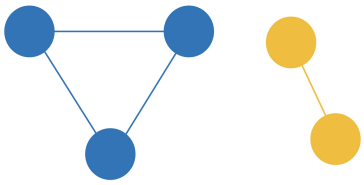
L'enjeu est donc la conception d'un algorithme de classification assez précis pour reconnaître un même document, mais assez souple pour s'accommoder de variations plus ou moins importantes (cf. tableau. 1).

2 Stratégie

Afin d'accélérer le traitement de l'information et d'alléger le poids des fichiers mis en ligne, l'encodage XML-TEI, qui n'est qu'un format pivot, est abandonné au profit du JSON (cf. code 3a).


```
"CAT_000055_e55_d1": {
    "price": 100.0,
    "author": "Scudéry",
    "date": "1686",
    "number_of_pages": 3.0,
    "format": "#document_format_4",
    "term": "#document_type_7",
    "sell_date": "1881-03",
    "desc": "L. a. s. à Huet, (1686), 3 p. in-4."}
```


(a) Données au format JSON



(b) Données sous forme de graphe

Illustration 3 – *Transformation des données*

Chaque fois que c'est possible une *string* est convertie en *integer* ou en *float* :

— Pour la longueur (`number_of_pages`) les documents incomplets sont ramenés à un nombre décimal («une page et demie» → `1.5`, «un quart de page» → `0.25`...)

8. S. Gabay, L. Rondeau Du Noyer, Matthias Gille Levenson, Ljudmila Petkovic et Alexandre Bartz, « Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? », dans *Digital Humanities DH2020*, Ottawa, Canada, 2020 (DH2020 Book of Abstracts), URL : `https://hal.archives-ouvertes.fr/hal-02898929` (visité le 23/11/2020).

— Pour le format (`format`) le nombre de pliage est le chiffre retenu («in-4°» → `4`, «in-folio» → `1`...)
— Pour la date (`date`) on utilise le format ISO `YYYY-MM-DD` («3 mai 1645» → `1645-05-03`, «septembre 1736» → `1736-09`...).
— Le type de document (`term`) est converti en chiffre : ainsi *L.a.s.* (*Lettre autographe signée*) a le code 7, tandis que *P.a.s.* (*Pièce autographe signée*) a le code 2.

Les informations en JSON sont alors transformées pour faire une base de données orientée graphe (cf. illustration 3b), afin de faciliter la réconciliation des données.

3 Réconciliation

La transformation des données en graphe permet de simplifier le mécanisme de réconciliation : si chaque nœud représente un document vendu, il suffit d'ajouter une arrête entre deux nœuds une fois atteint un certain degré de similarité.

Nous parlons de similarité et non d'identité stricte, car il n'est pas souhaitable de rechercher cette dernière : deux entrées différentes peuvent en effet renvoyer à un même document pour des raisons internes (deux fragments d'un même manuscrit) comme externes (une faute d'OCR). Il faut donc contourner ce problème via un algorithme de classification apte à gérer ces discrépances.

À partir de la liste des documents vendus, chaque entrée est comparée avec les autres. Cette comparaison se fait sur la base des informations clefs standardisées dans le fichier JSON : pour chacune de ces informations, un système de bonus/malus est appliqué (cf. tableau 2). Si le score obtenu est supérieur à 0.6, alors les entrées sont considérées comme renvoyant à un même manuscrit.

La valeur de ces bonus/malus a été trouvée de manière expérimentale, sur la base de tests unitaires évaluant l'efficacité de l'algorithme. Ces valeurs sont susceptibles d'évoluer avec l'ajout de nouveaux manuscrits.

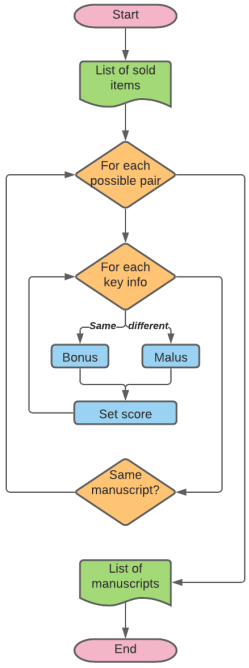


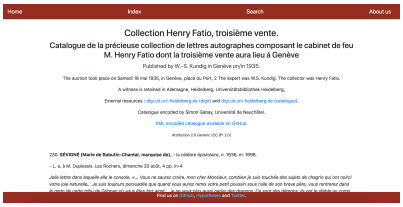
Illustration 4 – Flowchart de l'algorithme de réconciliation

ID	Bonus	Malus
Type de document	+ 0.2	- 0.1
Date d'écriture	+ 0.5	- 0.5
Longueur	+ 0.1	- 0.1
Format	+ 0.1	- 0.3
Prix	+ 0.1	- 0.1

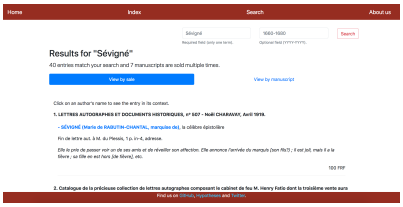
Tableau 2 – Bonus et malus utilisés

4 Applications

Une application en ligne s'appuie sur les données en JSON pour l'affichage des catalogues (cf. illustration 5a), qui sont disponibles à la lecture, et sur l'algorithme de classification afin de proposer un double mode de présentation des résultats pour une requête dans la base : la liste des ventes et la liste des manuscrits vendus (cf. illustration 5b).



(a) Lecture du catalogue de vente



(b) Affichage des manuscrits vendus

Illustration 5 – Application *Katabase*

Les données disponibles proviennent pour l'instant presque essentiellement de catalogues de vente à prix marqués, publiés dans le dernier tiers du XIX^{e} siècle à Paris par Gabriel Charavay [9].

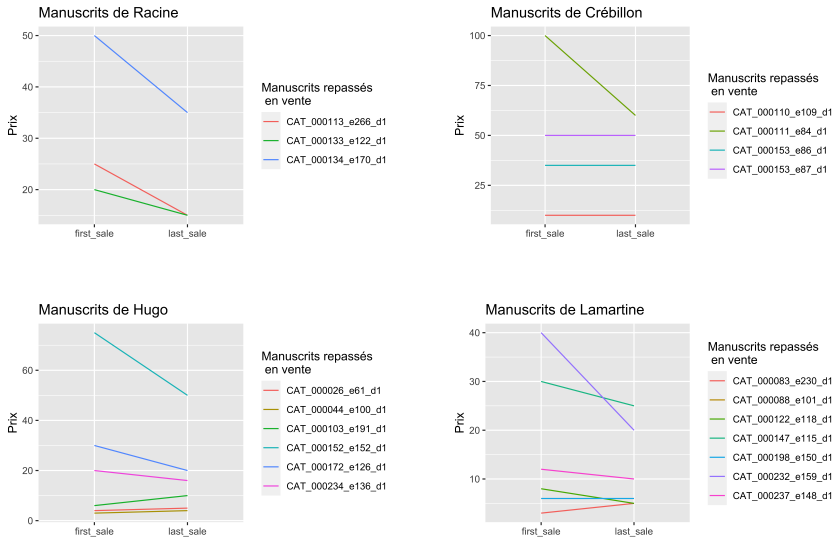


Illustration 6 – Évolution du prix en franc entre le premier et le dernier passage en vente pour les manuscrits de Racine, Crébillon, Lamartine et Hugo.

9. Le détail précis des catalogues numérisés est disponible dans l'application.

En faisant tourner l'algorithme sur ces données préliminaires, nous pouvons déjà offrir quelques premiers résultats. Nous avons pu définir un ratio de retour sur le marché des manuscrits : pour 44 333 manuscrits vendus, 3 567 sont ont été vendus au moins deux fois, soit environ 7,5%. À première vue, ces retours sur le marché sont marqués par une nette tendance baissière, notamment pour les manuscrits les plus chers, peu importe l'époque de l'auteur (cf. illustration 6) – la faible variation du franc à cette période et le de court laps de temps étudié permet par ailleurs une comparaison des prix malgré l'évolution du cours de la monnaie.

5 Recherches futures

Du point de vue philologique, la base de données ainsi que les capacités de classement développées pour l'application devraient permettre de retrouver plus aisément les sources des futures éditions, mais aussi de garantir l'authenticité des documents. Ces données devraient aussi être exploitables dans le cadre d'une approche distante du corpus afin d'étudier, par exemple, la construction du canon *via* la valeur marchande des auteurs.

Données et application

L'application web est disponible à l'adresse suivante : `https://katabase.herokuapp.com`.

Toutes les données utilisées pour ce projet sont disponibles en ligne à l'adresse suivante : `https://github.com/katabase`.

Remerciements

Merci à Jean-Baptiste Camps pour ses conseils (plus ou moins) avisés sur R (et tant d'autres choses).

Références

American Book-Prices Current, New York, 1894/95-.

Autograph Prices Current, London, 1914-1922.

Book Prices Current, London, 1887-1952.

Book-Auction Records, London, 1902–1997.

GABAY (Simon), RONDEAU DU NOYER (Lucie) et KHEMAKHEM (Mohamed), « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile : sfide e prospettive per l'Informatica Umanistica*, Milan, Italy, 2020 (Quaderni di Umanistica Digitale), p. 113-118, URL : https://hal.archives-ouvertes.fr/hal-02388407.

GABAY (Simon), RONDEAU DU NOYER (Lucie), GILLE LEVENSON (Matthias), PETKOVIC (Ljudmila) et BARTZ (Alexandre), « Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? », dans *Digital Humanities DH2020*, Ottawa, Canada, 2020 (DH2020 Book of Abstracts), URL : https://hal.archives-ouvertes.fr/hal-02898929 (visité le 23/11/2020).

Jahrbuch der Auktionspreise für Bücher, Handschriften und Autographen, Hamburg, 1950-.

L'Argus mensuel du livre ancien et moderne, Promodis, Paris, 1981-.