Katabase: À la recherche des manuscrits vendus

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

[1]Université de Genève – Suisse
[2]Université de Neuchâtel – Suisse
[3]École nationale des chartes – École nationale des Chartes, Ecole Nationale des Chartes – France
[4]Histoire, Archéologie et Littératures des mondes chrétiens et musulmans médiévaux – École Normale Supérieure - Lyon – France

Résumé

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

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 passees. La documentation est donc disparate et fragmentaire, concernant une ressource de premier ordre pour les collectionneurs, mais aussi pour les philologues en quete de sources, les historiens du livre ou les adeptes de la Rezeptionsgeschichte qui peuvent s'interesser aux prix ou aux noms des collectionneurs.

Enjeu

Si les principaux problèmes poses par la numerisation de catalogues comme la Revue des autographes sont connus, tout comme les enjeux de la detection d'un manuscrit revenant plusieurs fois sur le marché, parfois sous une forme fragmentaire [8], il nous a paru important d'ameliorer notre algorithme de classification. Ce dernier doit en effet etre implementable dans une application web disponible en ligne, tout en etant capable de traiter de grandes quantites de donnees avec un maximum de precision.

L'enjeu est donc la conception d'un algorithme de classification assez precis pour reconnaitre un meme document, mais assez souple pour s'accommoder de variations plus ou moins importantes.

Strategie

Afin d'accelerer le traitement de l'information et d'alleger le poids des fichiers mis en ligne, l'encodage XML-TEI, qui n'est qu'un format pivot, est abandonne au profit du JSON.

*Intervenant

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 ramenes à un nombre decimal ("une page et demie" → 1.5, "un quart de page" →0.25. . .)

- 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 signee) a le code 7, tandis que P.a.s. (Pièce autographe signee) a le code 2.

Les informations en JSON sont alors transformees pour faire une base de donnees orientee graphe, afin de faciliter la reconciliation des donnees.

Reconciliation

La transformation des donnees en graphe permet de simplifier le mecanisme de reconciliation : si chaque nœud represente un document vendu, il suffit d'ajouter une ar- rete entre deux nœuds une fois atteint un certain degre de similarite.

Nous parlons de similarite et non d'identite stricte, car il n'est pas souhaitable de rechercher cette dernière : deux entrees differentes peuvent en effet renvoyer à un meme document pour des raisons internes (deux fragments d'un meme manuscrit) comme externes (une faute d'OCR). Il faut donc contourner ce problème via un algorithme de classification apte à gerer ces discrepances.

A partir de la liste des documents vendus, chaque entree est comparee avec les autres. Cette comparaison se fait sur la bases des informations clefs standardisees dans le fichier JSON : pour chacune de ces informations, un système de bonus/malus est applique. Si le score obtenu est superieur à 0.6, alors les entrees sont considerees comme renvoyant à un meme manuscrit.

La valeur de ces bonus/malus a ete trouvee de manière experimentale, sur la base de tests unitaires evaluant l'efficacite de l'algorithme. Ces valeurs sont susceptibles d'evoluer avec l'ajout de nouveaux manuscrits.

Applications

Une application en ligne [https://katabase.herokuapp.com] s'appuie sur les donnees en JSON pour l'affichage des catalogues, qui sont disponibles à la lecture, et sur l'algorithme de classification afin de proposer un double mode de lecture des resultats pour une requete dans la base : la liste des ventes et la liste des manuscrits vendus.

Les donnees disponibles proviennent pour l'instant presque essentiellement de catalogues de vente à prix marques, publies dans le dernier tiers du XIXe siècle à Paris par Gabriel Charavay (le detail precis des catalogues numerises est disponible dans l'application).

En faisant tourner l'algorithme sur ces donnees preliminaires, nous pouvons dejà offrir quelques premiers resultats. Nous avons pu definir un ratio de retour sur le marche des manuscrits : pour 44 333 manuscrits vendus, 3 364 sont ont ete vendus au moins deux fois, soit environ 7,5%. A première vue, ces retours sur le marche sont marques par une nette tendance baissière, notamment pour les manuscrits les plus chers, peu importe l'epoque de l'auteur – la faible variation du franc à cette periode et le de court laps de temps etudie permet par ailleurs une comparaison des prix malgre l'evolution du cours de la monnaie.

Recherches futures

Du point de vue philologique, la base de donnees ainsi que les capacites de classement developpees pour l'application devraient permettre de retrouver plus aisement les sources des futures editions, mais aussi de garantir l'authenticite des documents. Ces donnees devraient aussi etre exploitables dans le cadre d'une approche distante du corpus afin d'etudier, par exemple, la construction du canon via la valeur marchande des auteurs.

Donnees et application

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

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

Bibliographie

https://fr.artprice.com.

American Book-Prices Current, New York, 1894/95-. ABPC tend avec le temps à repertorier de plus en plus de ventes europeennes.

Book-Auction Records, London, 1902–1997 et Book Prices Current, London, 1887-1952.

Jahrbuch der Auktionspreise fur Bucher, Handschriften und Autographen, Hamburg, 1950-. Au debut Jahrbuch der Auktionspreise fur Bucher und Autographen.

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

Autograph Prices Current, London, 1914-1922

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.

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