Intelligence Artificielle et intelligence collective : des nouveaux eldorados pour rendre les textes patrimoniaux plus accessibles ?

Alix Chagué

Intelligence Artificielle et intelligence collective : des nouveaux eldorados pour rendre les textes patrimoniaux plus accessibles ?

Alix Chagué

Université de Montréal, Montréal & ALMAnaCH, Inria, Paris

Article initialement publié sur Museonum le 24 mai 2022, accessible à https://medium.com/museonum/intelligence-artificielle-et-intelligence-collective-des-nouveaux-eldorados-pour-rendre-les-c8c4e214d4e6

Le 4 avril 2022, paraissait dans *The Guardian* un article intitulé « 'Mind-blowing': Ai-Da becomes first robot to paint like an artist » présentant le projet Ai-Da. C'est un robot, d'apparence humaine, qui peint des images décrites comme « ultra-réalistes », à renfort d'intelligence artificielle et d'un pinceau tenu par un bras mécanique. Pour son créateur Aidan Meller, qui est aussi propriétaire de deux galeries d'art en Grande-Bretagne, il n'est pas question d'interroger le statut des œuvres ainsi créées, mais plutôt d'utiliser Ai-Da pour explorer le rapport de notre société à l'idée qu'un robot doué d'une intelligence artificielle puisse être considéré comme une artiste.

<lien vers AI-Da, *The Intersection of Art and AI*. TEDxOxford, 2020. https://www.youtube.com/watch?v=XaZJG7jiRak.>

Les GLAMs et l'IA

Ai-Da n'est qu'un exemple parmi d'autres de l'utilisation de l'apprentissage automatique (*machine learning*) dans le monde des GLAMs (*Galleries, Libraries, Archives and Museums*) : en dehors du domaine de la création artistique, ces nouvelles technologies peuvent répondre à de très nombreux enjeux. À l'inverse, du caractère par moment déplaisamment dystopique de la conférence TEDx « donnée par Ai-Da » en 2020, les professionnel·les des GLAMs voient dans l'emploi de l'IA un moyen d'optimiser certains processus et d'innover. C'est le cas de Brendan Ciecko, fondateur de Cuseum, qui listait en 2017 plusieurs exemples d'usages qui concernent les domaines économique (anticipation de l'affluence et vente de tickets), de la communication (utilisation du *sentiment analysis* simultanément sur les cartels d'un musée et les publications de son public sur les réseaux sociaux), ou encore de la documentation des collections ainsi que leur consultation en ligne par le public (rapprochement d'œuvres selon des critères formels ou chromatiques automatiquement extraits des images) (Ciecko, 2017).

L'un des domaines d'application de l'IA les plus pertinents pour la documentation des collections de musées est sans nul doute celui de la vision par ordinateur (*computer vision*). Il s'agit de permettre aux machines de recevoir un input visuel fixe ou animé (image ou vidéo) et de l'interpréter de manière à en extraire des informations telle que le thème représenté ou encore la présence d'objets ou de motifs et leur position dans l'espace. Il s'agit, en fait, tout simplement de reproduire le mécanisme de vision d'un organisme vivant. Appliquée aux musées, cette technologie permettrait d'assister au catalogage des collections

en proposant un étiquetage automatique des numérisations. C'est une stratégie explorée à partir de 2018 par l'Auckland Museum, en Nouvelle Zélande, mais qui a rapidement montré ses limites lorsqu'elle a été appliquée à des fonds de photographies anciennes et liées aux Maori car les modèles employés avaient été entraînés à partir d'exemples contemporains et majoritairement occidentaux (Moriarty, 2019). L'expérimentation de l'Auckland Museum confirme que la connaissance du contexte culturel associé à une image n'est pas encore suffisamment maîtrisée par les outils disponibles, en particulier pour les collections non-occidentales, au point de pouvoir systématiser son application pour le catalogage.

La transcription automatique est l'un des sous-domaines de la vision par ordinateur les plus prometteurs pour les GLAMs. L'OCR (pour *Optical Character Recognition*) désigne la transcription automatique quand elle est appliquée aux textes imprimés. Elle est considérée comme un problème résolu (Cao & Natarajan, 2014) à l'exception des documents imprimés anciens (Reul et al., 2018) ou de la presse (Nguyen et al., 2019). L'HTR (pour *Handwritten Text Recognition*) correspond au même processus mais appliqué aux textes manuscrits. C'est un domaine encore très expérimental, quoique de plus en plus accessible au grand public. Cette technologie est particulièrement importante pour les bibliothèques et les centres d'archives, qui possèdent des millions de kilomètres linéaires de documents rédigés à la main. Le fait de posséder des textes qu'il est possible d'interroger avec des outils numériques, quel que soit le mode d'acquisition de ces données textuelles, impacte autant les professionnel·les de la documentation que les usager·es. En 2021, Victoria A. Van Hyning et Mason A. Jones présentaient deux cas d'usage très fréquents : 1) l'existence d'un texte compatible avec un moteur de recherche facilite la découvrabilité des documents, et 2) elle permet à des publics d'accéder à des textes qu'ils n'auraient autrement pas su déchiffrer (Van Hyning & Jones, 2021).

L'automatisation de l'acquisition des textes manuscrits reste globalement perçue comme un domaine encore pionnier pour les GLAMs. Dans sa communication de 2022, le MoMu parle même de « déverrouiller » (*unlock*) les sources historiques, tant la maîtrise de l'HTR peut constituer un réel point de bascule entre deux modes de consultations des collections documentaires (MoMu Antwerp, 2022). En évoquant le projet « From Quill to Byte », la Uppsala Universitet insistait d'ailleurs sur l'idée qu'il y aurait « une fortune à se faire » (Svensson, 2015) par ce biais.

"If someone today had an algorithm to carry out large-scale digital searches of things like the collection of manuscripts in the Vatican Library, it would be worth a fortune." *(Anders Brun, cité par* Svensson, 2015*)*

Le texte par quatre chemins

La transcription automatique est loin d'être le seul moyen pour une institution d'extraire les données textuelles de ses fonds patrimoniaux. Tout d'abord, il est possible pour une institution d'avoir recours à un prestataire extérieur, qui se charge de réaliser la transcription « à la main ». C'est certainement le cas de l'entreprise AEL Data Service, basée à Chennai en Inde, qui a produit des transcriptions d'une partie du corpus du Sir Hans Sloan's Miscellanies Catalogue du British Museum (Humbel & Nyhan, 2019), sur lequel nous reviendrons plus tard. Certaines entreprises peuvent aussi baser leurs services aux

institutions culturelles sur le recours à l'HTR. On peut mentionner ici l'entreprise Teklia dont le modèle économique repose sur la mise en place et l'exécution de chaînes de traitement complètes, allant de la numérisation à l'indexation, en passant par la transcription automatique. C'est aussi le cas de la Adam Matthew Digital Ltd. qui intègre la transcription automatique de manuscrits, d'imprimés et d'audio dans Quartex, sa solution de gestion d'actifs numériques (GAN ou DAM pour *Digital Asset Management*) et de création d'expositions numériques.

D'autre part, dans le cadre de positionnements stratégiques visant à renforcer l'interaction avec les publics des institutions, on observe que plusieurs musées utilisent le *Crowdsourcing* pour acquérir ces textes. Le *Crowdsourcing* désigne le fait, pour une institution ou un organisme, de solliciter des contributions de diverses formes de la part d'individus généralement issus du grand public, sur la base du volontariat. Des plateformes de *Crowdsourcing* comme Zooniverse, l'une des plus connues dans le domaine patrimonial, permettent de mettre en relation des institutions porteuses d'un projet avec des contributeur·rices, anonymes ou non, chargé·es de réaliser des tâches définies par l'institution. Outre la transcription d'éléments textuels plus ou moins longs présents sur des images, ces tâches peuvent consister à étiqueter des images ou des parties d'images ou encore à classer des vidéos. Lorsqu'il est bien constitué, doté de ressources pour animer la communauté et qu'il rencontre un public, un projet de *Crowdsourcing* peut être spectaculairement efficace. Le portail de la Library of Congress, By the People, en est un exemple. En avril 2022, il recense 31 campagnes de *Crowdsourcing* pour un total de plus de 470 000 pages transcrites sur les plus de 694 000 proposées (The Library of Congress, 2018).

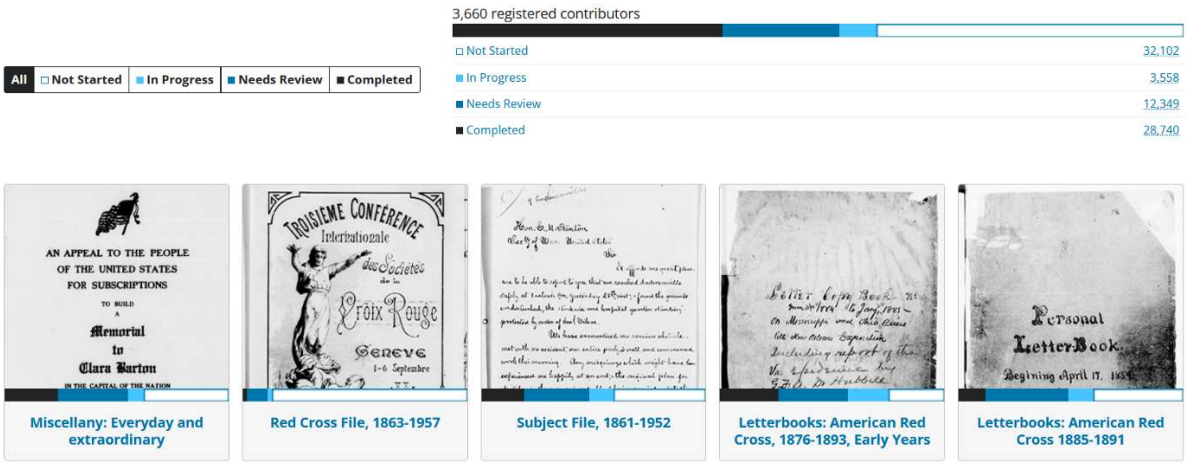
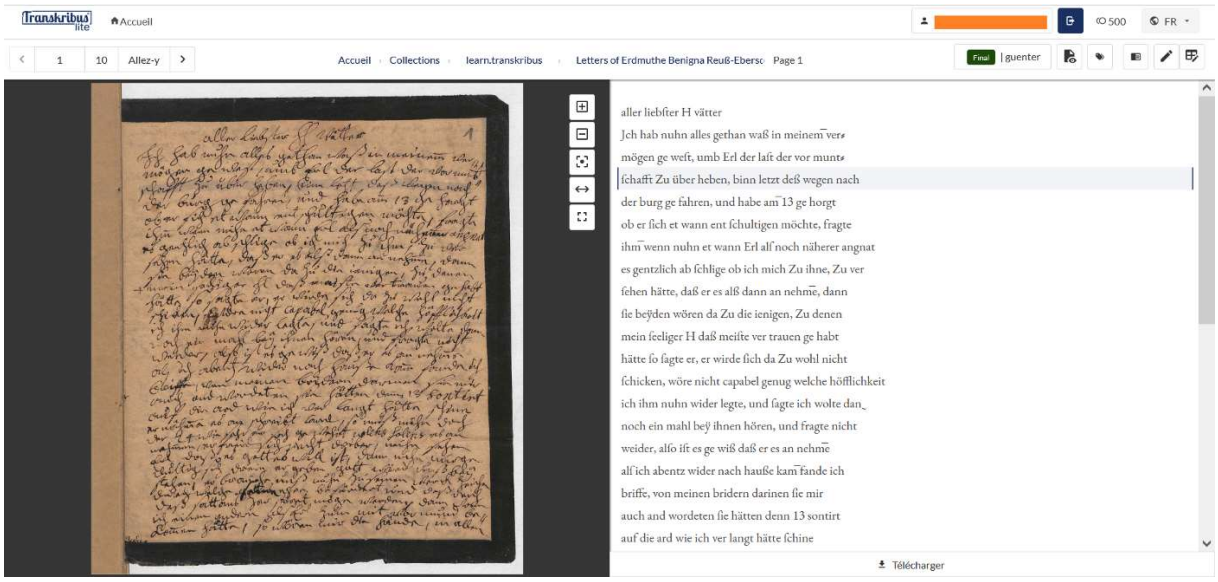


Tableau de bord de la campagne « *C*lara Barton » dans By The People, où l'on peut voir l'état d'avancement de la transcription et les différentes étapes associées à chaque document (à transcrire, en cours, à réviser, etc).

Le *Crowdsourcing* repose généralement sur l'idée que la contribution est de petite ampleur et gratuite. C'est l'accumulation des contributions qui fait la force de ces projets. Il est néanmoins possible de monétiser l'exécution de ces tâches, basculant dans une quatrième voie, à mi-chemin entre *Crowdsourcing* et prestations externalisée. Adam Moriarty la désigne par l'expression de « *gig economy* ». Dans cette configuration, des tâches relativement semblables à celles qui sont proposées sur une plateforme de *Crowdsourcing* sont rémunérées à une poignée de centimes chacune. Elle est cependant peu utilisée, car

elle pose des graves problèmes éthiques liés notamment à la précarité des “emplois” générés (notamment à l’impossibilité de contrôler les conditions de travail) et ne constitue pas une garantie de qualité (Moriarty, 2019).

En dehors des bibliothèques et des archives, le recours des institutions patrimoniales à l’HTR est relativement discret. Le logiciel Transkribus, développé par READ COOP depuis près de 10 ans, a joué un rôle très important dans l’appropriation de l’HTR par les institutions culturelles en général. Sa particularité est de permettre aux utilisateur·rices de gérer de manière autonome leur campagnes de transcription : c’est à l’utilisateur·rice d’appliquer les différentes étapes de traitement permettant d’obtenir un texte aussi parfait que possible à partir d’une numérisation. Il existe désormais d’autres logiciels semblables, mais les quelques musées qui font publiquement état d’une utilisation de l’HTR ont presque tous utilisé Transkribus : c’est le cas du MoMu à Anvers, que nous avons déjà évoqué, pour des documents manuscrits des XVIIe et XVIIIe siècles (MoMu Antwerp, 2022), du Sedgwick Museum of Earth Sciences à Cambridge pour les carnets du début du XIXe siècle de la collection « Sedgwick 200 », du Brunel’s SS Great Britain, et du British Museum pour le Sir Hans Sloane’s Miscellanies Catalogue, qui a utilisé les données de références produites par son prestataire indien pour entraîner un modèle dans Transkribus (Humbel & Nyhan, 2019). En avril 2022, on compte également une poignée de musées parmi les près de 80 partenaires de READ COOP : le Germanisches National Museum, le Museum für Musikinstrumente de l’Université de Leipzig, le Munch Museum, et le Museum d’Histoire Naturelle du Royaume-Uni à travers son département « Library and Archive ».



L’interface de transcription de Transkribus Lite permet d’interagir simultanément avec l’image et la transcription.

Si l’on considère que l’HTR est encore une technologie en cours d’élaboration, c’est principalement pour deux raisons. Premièrement, il est presque toujours nécessaire de créer d’abord, à la main, des transcriptions de références pour entraîner des modèles « taillés sur mesure » pour chaque variation d’écriture. Deuxièmement, il est quasiment impossible d’acquérir des transcriptions parfaites à 100% : il reste inévitablement des fautes dans un texte manuscrit transcrit automatiquement. À cela s’ajoutent des difficultés posées par les abréviations et les mises en page particulières.

Dans ces conditions, HTR et *Crowdsourcing* sont en fait souvent perçus comme des modalités complémentaires d'acquisition du texte. En plus de fournir des transcriptions de référence, les campagnes de *Crowdsourcing* peuvent servir à nettoyer le résultat de l'HTR, comme en rendent compte le Brunel's SS Great Britain un an après avoir commencé à utiliser Transkribus, et Karl-Magnus Johansson et Mats Jönsson en 2020 pour la transcription de documents conservés aux Archives nationales de Suède (Johansson & Jönsson, 2020). Inversement, l'HTR peut contribuer à faciliter la tâche des participant·es au *Crowdsourcing* en proposant de générer automatiquement des suggestions de transcription pour les mots difficiles. C'est l'une des applications envisagées dès 2018 par les porteur·ses du projet Transribe Bentham (Causer et al., 2018)

"We could never have anticipated that the work of volunteer transcribers would be used as 'ground truth' data for training HTR models, or that we would envisage and test a transcription interface in which volunteers could ask an HTR engine for suggestions for words which they were struggling to decipher." *(*Causer et al., 2018*)*

Infrastructures et culture(s) professionnelle(s)

Comme le rappellent Van Hyning et Jones, il est fréquent que les institutions patrimoniales rencontrent des difficultés pour intégrer, au sein de leur logiciels et bases de données, les transcriptions générées grâce à des campagnes de *Crowdsourcing*. Van Hyning et Jones pointent du doigt un manque de préparation de la part des institutions sur ces questions alors que de véritables enjeux éthiques se posent. En effet, l'une des contre-parties du *Crowdsourcing* est la promesse d'une mise à disposition libre des transcriptions obtenues grâce à l'aide des participant·es.

"it's not right to invite people to help researchers and increase accessibility… if the resulting data aren't made widely available as well as understandable — meaning that the methods of collection are well described" *(*Van Hyning & Jones, 2021*)*

Fournir une infrastructure de consultation voire de téléchargement adaptée, documenter le processus de création des transcriptions et les lier aux données déjà enregistrées dans les bases prend tant de temps que l'une des solutions « rapides » adoptées par les institutions consiste à déposer les données sur des entrepôts plus ou moins adéquats comme Github (Van Hyning & Jones, 2021). On peut tirer le même constat pour l'HTR car l'arrivée de données textuelles en masse pose un véritable défi technique et suppose de repenser le parcours des utilisateur·rices. En outre, en s'appuyant sur les travaux de Chern Li Liew (Liew, 2016), Van Hyning et Jones signalent qu'il demeure dans les GLAMs des réticences à mettre en ligne des données générées par l'HTR ou par le *Crowdsourcing* : il y aurait une certaine « anxiété » liée à une qualité des transcriptions pressentie comme insuffisante.

Dans de nombreux cas, les campagnes d'HTR et/ou de *Crowdsourcing* relèvent encore de la preuve de concept permettant à des institutions de tester les chaînes de traitement existantes et de se confronter aux limites des outils professionnels en place. Il resterait encore à trancher entre deux paradigmes : continuer à progresser lentement vers la mise en

ligne des collections en ne diffusant qu'au compte-goutte des données vérifiées et de qualité, quitte à limiter la capacité des utilisateur·rices à explorer les collections de manuscrits; ou bien débloquer massivement les données, mais au risque de générer beaucoup de bruit, voir de graves erreurs de transcription. C'est le choix qu'a fait la Bibliothèque nationale de France pour les documents imprimés diffusés sur le portail Gallica, quitte à proposer ultérieurement des campagnes de soutien à la correction des certains corpus mal transcrits (Michez, 2021).

Vers une mise en ligne massive et libre ?

Finalement, on voit qu'extraire les transcriptions de documents manuscrits et patrimoniaux demeure un traitement coûteux même si l'intérêt est indéniable pour les GLAMs. Externalisé chez un prestataire, le processus représente une dépense financière importante ; soumis aux aléas de l'IA, il suppose généralement de faire le sacrifice de la perfection ; atomisé en *gigs*, son coût éthique peut peser lourd dans la balance ; et enfin proposé à l'examen de l'intelligence collective, il mobilise des ressources humaines conséquentes.

Traiter des unités logiques au sein d'une collection dans le cadre de prestations de numérisation reste une pratique courante et « sûre ». Mais le succès rencontré par plusieurs campagnes de transcription participative et les avancées technologiques nombreuses dans le domaine de la transcription automatique pourraient bien rapidement changer la donne. Et avec un peu de chance (et de militantisme !), elles entraîneront au passage les institutions patrimoniales vers la mise en ligne massive et libre de précieux contenus textuels.

Bibliographie

- Cao, H., & Natarajan, P. (2014). Machine-Printed Character Recognition. In D. Doermann & K. Tombre (Éds.), *Handbook of Document Image Processing and Recognition* (p. 331-358). Springer. https://doi.org/10.1007/978-0-85729-859-1_44
- Causer, T., Grint, K., Sichani, A.-M., & Terras, M. (2018). 'Making such bargain' : Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*, *33*(3), 467-487. https://doi.org/10.1093/llc/fqx064
- Ciecko, B. (2017). *Examining the Impact of Artificial Intelligence in Museums — MW17 : Museums and the Web 2017*. Museums and the Web 2017, Cleveland, Ohio. https://mw17.mwconf.org/paper/exploring-artificial-intelligence-in-museums/
- Humbel, M., & Nyhan, J. (2019, juillet 1). *The Application of HTR to Early-modern Museum Collections : A Case Study of Sir Hans Sloane's Miscellanies Catalogue*. DH2019, Utrecht, Netherlands. https://www.researchgate.net/publication/334458820_The_Application_of_HTR_to_Early-modern_Museum_Collections_a_Case_Study_of_Sir_Hans_Sloane%27s_Miscellanies_Catalogue
- Johansson, K.-M., & Jönsson, M. (2020, octobre 22). *The Detective Section : On Machine Learning and Local Knowledge processes in A Recently Initiated Handwritten Text Recognition and Citizen Science Project*. Museum Big Data, [Online].

https://www.gu.se/en/news/handwritten-text-recognition-and-citizen-science-methods-presented-at-museum-big-data-2020

- Liew, C. L. (2016). Social Metadata and Public-Contributed Contents in Memory Institutions : “Crowd Voice” Versus “Authenticated Heritage”? *Preservation, Digital Technology & Culture*, *45*(3), 122-133. https://doi.org/10.1515/pdtc-2016-0017
- Michez, G. (2021, septembre 22). Aidez-nous à donner une bonne correction à Gallica ! [Blog]. *Le blog de Gallica*. https://gallica.bnf.fr/blog/22092021/aidez-nous-donner-une-bonne-correction-gallica?mode=desktop
- MoMu Antwerp. (2022). How MoMu digitises 17th & 18th century copybooks and ledgers. *MoMu*. https://www.momu.be/en/magazine/melijnproject
- Moriarty, A. (2019). A Crisis of Capacity : How can Museums use Machine Learning, the Gig Economy and the Power of the Crowd to Tackle Our Backlogs. *MW19*. https://mw19.mwconf.org/paper/a-crisis-of-capacity-how-can-museums-use-machine-learning-the-gig-economy-and-the-power-of-the-crowd-to-tackle-our-backlogs/
- Nguyen, T.-T.-H., Jatowt, A., Coustaty, M., Nguyen, N.-V., & Doucet, A. (2019). Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 29-38. https://doi.org/10.1109/JCDL.2019.00015
- Reul, C., Springmann, U., Wick, C., & Puppe, F. (2018). Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. *arXiv:1802.10038 [cs]*. http://arxiv.org/abs/1802.10038
- Svensson, J. (2015, avril 28). *A Google for handwriting — Uppsala University, Sweden* [Blog]. Uppsala Universitet News; Uppsala University, Sweden. https://www.uu.se/en/news/article/?id=4574&typ=artikel
- The Library of Congress. (2018). *By the People Campaigns*. By the People. https://crowd.loc.gov/campaigns-topics/
- Van Hyning, V. A., & Jones, M. A. (2021). Data’s Destinations : Three Case Studies in Crowdsourced Transcription Data Management and Dissemination. *Startwords*, *2*. https://doi.org/10.5281/zenodo.5750691