Rendre visible la face cachée de l'iceberg. Explorer les documents manuscrits à l'aide de l'intelligence artificielle

Emmanuelle Perrin, Philippe Chassignet

Rendre visible la face cachée de l'iceberg

PAR EMMANUELLE PERRIN ET PHILIPPE CHASSIGNET · 27 MARS 2020

Explorer les documents manuscrits à l'aide de l'intelligence artificielle

La 3e conférence des utilisateurs de Transkribus (*Transkribus User Conference 2020*) a réuni 160 participants venus de 26 pays différents les 6 et 7 février 2020 à l'université d'Innsbruck (Autriche). Nous utilisons cette plateforme de transcription automatique de textes manuscrits et imprimés dans le cadre du projet Bulliot, Bibracte et moi, lauréat de l'appel à projets Services numériques innovants 2019, financé par le ministère de la Culture. Porté par le laboratoire Archéorient, ce projet a pour objectif la transcription collaborative des archives manuscrites du XIXe siècle du site archéologique de Bibracte.

Développé entre 2016 et 2019 à l'université d'Innsbruck dans le cadre du projet européen READ (*Recognition and Enrichment of Archival Documents*), Transkribus est fondé sur l'intelligence artificielle. Un moteur de reconnaissance d'écriture manuscrite (*Handwritten Text Recognition* ou *HTR*), qui peut aussi être utilisé pour les imprimés anciens sur lesquels la reconnaissance optique de caractères (*optical character recognition* ou OCR) classique donne de piètres résultats, est configuré à partir de données d'apprentissage d'au moins 20 000 mots. La technologie est fondée sur la segmentation ou détection des lignes qui permet d'associer l'image au texte et de restituer le contexte dans lequel les lettres sont employées (Fig. 1).

Fig.1 : La reconnaissance de caractères en contexte (cliché E. Perrin)

Transkribus compte aujourd'hui 32 000 utilisateurs inscrits, soit 200 à 400 utilisateurs connectés par jour pour 1 000 utilisateurs uniques par semaine. Transkribus c'est aussi 4 000 moteurs de reconnaissance d'écriture manuscrite configurés, 8,7 millions d'images téléchargées sur la plateforme pour former, selon l'équipe de l'université d'Innsbruck, la plus grande collection au monde de documents historiques. Pour « révolutionner » l'accès aux documents manuscrits et les transcrire à grande échelle, l'idée fondatrice de Transkribus est d'impliquer tous les acteurs du traitement des documents historiques : archivistes, bibliothécaires, chercheurs, informaticiens et grand public. La plateforme procure les services nécessaires à chacun de ces groupes (Fig. 2).

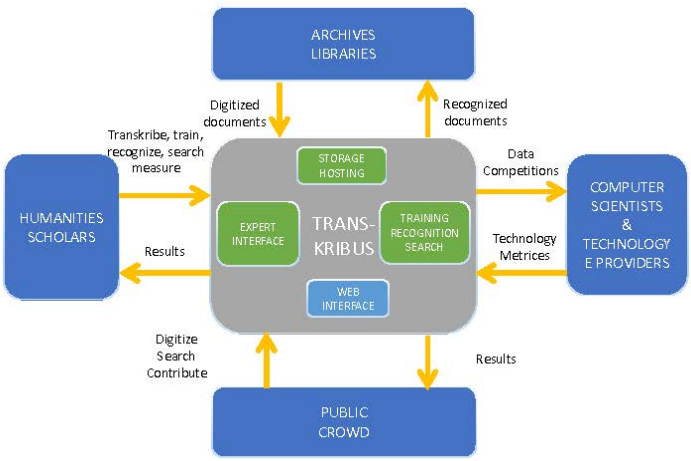

Fig.2 : Le traitement de documents historiques avec Transkribus (Schéma G. Mülberger, source : http://savoirs.ens.fr/uploads/videos/diffusion/2019_10_17_Digit_Hum_2019_Gunter_Muhlberger.mp4)

De nouvelles fonctionnalités

Les méthodes d'extraction des données à des fins de recherches universitaires font l'objet d'une constante optimisation, comme l'illustre le partenariat de Naver Labs Europe (centre de recherche en intelligence artificielle) avec les départements d'histoire et des sciences de l'informatique de l'université de Cambridge, pour l'amélioration de l'analyse des tableaux complexes (Fig. 3).

Fig. 3. : Un atelier sur l'analyse des tableaux complexes (cliché P. Chassignet).

Des ateliers étaient consacrés à la présentation des nouvelles fonctionnalités de la plateforme :

- L'outil **P2PaLA** (*Page to Page Layout Analysis*) permet la détection automatique de la structure du document. Pour l'équipe de Transkribus, c'est la deuxième grande étape de l'analyse des documents après la reconnaissance de l'écriture. Cette fonction s'applique sur des documents imprimés dont la structure est répétitive comme les livres (numéro de page, titre, *marginalia*, note de bas de page) et les journaux. Il faut que ces caractéristiques soient discernables par leur localisation géométrique sur la page ou par leur apparence visuelle (police de caractère). Le projet sur les cadastres autrichiens utilise cette fonctionnalité pour reconnaître automatiquement la colonne indiquant les propriétaires et en extraire les noms propres.
- L'outil **Text2Image** permet d'intégrer dans la plateforme des transcriptions existantes à partir de fichiers texte sans mise en forme. Le procédé est fondé sur la correspondance entre les lignes du fichier texte importé et celles qui ont été détectées sur l'image.
- **Les modèles publics.** Une autre fonction permet de partager les modèles HTR créés par les utilisateurs. Les services d'archives ont notamment travaillé à la configuration de modèles à grande échelle, capable de traiter des mains différentes. Le programme IJsberg des archives nationales des Pays-Bas met à disposition différents modèles allant du XVII^{e} au XIX^{e} siècle à partir des archives de la *Dutch East India Company* et d'actes notariés.

Une communauté dynamique

Les multiples projets présentés ont donné une image très dynamique et inventive de la communauté internationale réunie lors de la conférence. Certains portent sur l'édition de manuscrits (The Ambraser Heldenbuch, Documents of Early England Data Set), d'autres sur la publication d'archives ou de bibliothèques numériques. Le Trinity College de

Dublin par exemple, avec Beyond 2022,vise la reconstitution virtuelle du Public Record Office of Ireland, détruit par un incendie en 1922. Outre la modélisation du bâtiment, une vaste campagne de récupération, de numérisation et de traitement des documents est lancée (Fig. 4). Le projet Hakki Tarik Us Collection, sur les archives ottomanes d'Istanbul, tente de résoudre la coupure dans l'accès aux archives, entraînée par l'adoption, en 1925 en Turquie, de l'alphabet latin. Transkribus est ingénieusement détourné en un outil de translittération des caractères arabes en caractères latins sur un corpus de 1 500 périodiques (400 000 pages). Le projet Alpenwort porte sur le journal autrichien du Club alpin (*Zeitschrift des Österreichischen Alpenvereins*) et fournit un corpus en caractère gothique de près de 50 000 pages, lisibles par les machines. Ce journal est mis à disposition pour des études linguistiques et lié à une base de données sur les noms de personnes et les toponymes.

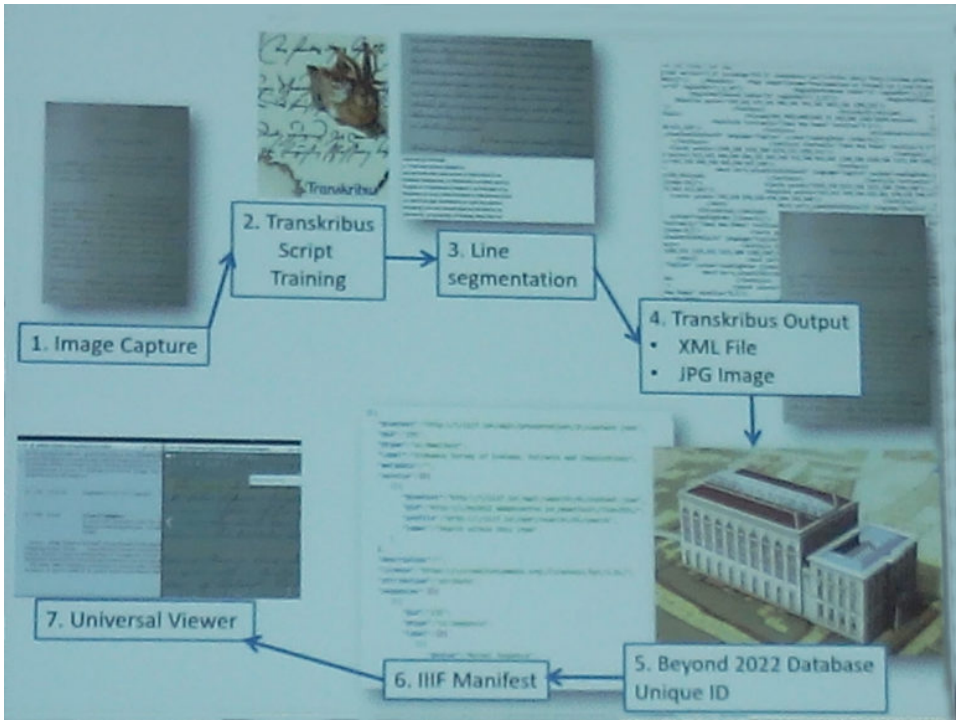

Fig. 4. : Le projet Beyond 2022 (cliché P. Chassignet)

À ces projets de publication s'articulent le plus souvent l'extraction de données spécifiques, comme les noms de personnes ou de lieux, pour naviguer dans un corpus, ou l'analyse structurelle des textes. On peut citer, par exemple, l'indexation, avec la technologie du Key Word Spotting fondée sur le traitement de l'image, des actes des tribunaux conservés aux archives nationales de Finlande. L'exploitation des 80 km d'archives historiques de la Banco di Napoli (*Il Cartastorie*) est également orientée vers l'extraction des 17 millions de noms de personnes consignés depuis le XVI^{e} siècle. Le projet européen NewsEye réunit la Bibliothèque nationale autrichienne, la Bibliothèque nationale de France et la Bibliothèque nationale de Finlande, autour de l'analyse de la presse historique. Il associe la reconnaissance de la structure des articles à celle des entités nommées et travaille à la détection des événements.

Humanités numériques et *crowdsourcing*

Humanités numériques et *crowdsourcing* (sciences participatives)apparaissent intimement liés dans l'esprit des porteurs de projets de transcriptions massives. Certains projets envisagent de recourir au *crowdsourcing* à plus ou moins court terme (Cadastres autrichiens, *Il Cartastorie*). Selon Andrea Zappulli (*Banco di Napoli Foundation*), il est fondamental de disposer au préalable d'une interface dédiée. Le public peut contribuer sous différentes formes : préparation des documents (segmentation ou *layout analysis*), transcription des documents avec ou sans HTR et vérification des résultats. Pour Günther Mülberger, coordinateur du projet Transkribus, les transcripteurs doivent être les relecteurs et il n'envisage pas de vérification finale par des chercheurs. Les volontaires se forment tout au long du projet

et développent donc une réelle expertise. Il lui paraît également important de rendre le processus de relecture transparent.

Le public visé peut être circonscrit aux étudiants et aux passionnés d'histoire locale, pour s'élargir au grand public avec l'opération lancée par les archives de la ville d'Amsterdam sur les archives des notaires (1578-1915) : *Crowd leert computer lezen* (la foule apprend à lire à l'ordinateur). Le projet bénéficie d'un cadre ludique et stimulant. Les actes à transcrire sont classés selon leur niveau de difficultés ou par les notaires. Les participants disposent d'une page personnelle et de statistiques. Ils gagnent des points par actes transcrits, qui leur permettent de participer à des événements et à des expositions aux Archives nationales. Il y a aussi des jeux comme le « Rembrandt Bingo », gagnant si l'on transcrit un acte signé par le peintre. L'auto-organisation de la communauté des transcripteurs par des « super participants » se montre efficace et le projet développe une audience internationale (Fig. 5). Avec le Key Word Spotting Demonstrator, qui permet de naviguer dans les actes notariés, Pauline van den Heuvel, des Archives de la ville d'Amsterdam, voit se réaliser le rêve des humanités numériques de combiner reconnaissance d'écriture manuscrite, indexation, reconnaissance des entités nommées et thesaurus.

Fig. 5. : L'audience internationale du projet Crowd leert computer lezen (source : https://velehanden.nl/account/gebruikers/map/project_id/270)

L'avenir de Transkribus

Au terme du financement du projet Read, le modèle économique de Transkribus évolue avec la création d'une coopérative européenne, READ-COOP SCE, qui va proposer des services payants pour traiter de grandes quantités de documents ou adapter le logiciel à des besoins spécifiques. La coopérative tire actuellement l'essentiel de ses revenus de projets à grande échelle conclus avec les Archives nationales de Finlande et des Pays-Bas. Les services gratuits seront limités en performance et en nombre de pages. Les tarifs devraient être mis au point et publiés d'ici l'été 2020.

L'avenir de Transkribus passe aussi par la consolidation de la communauté avec la construction d'un réseau de formateurs et une réflexion sur les moyens de communication mis à la disposition des utilisateurs.

Nous avons apprécié, en plus des aspects techniques du développement du logiciel et des projets présentés, l'engouement international et communicatif des participants autour de Transkribus. Nous espérons pouvoir candidater à

la prochaine conférence pour présenter le projet Bulliot, Bibracte et moi car la France, comme l'archéologie, ne sont guère représentées au sein de cette communauté internationale en devenir.

Bibliographie

Bonhomme M.-L. 2018. *Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture : Rapport exploratoire*, Inria. [En ligne] https://hal.archives-ouvertes.fr/hal-02386180.

Het Utrechts Archief 2020. *AI helps humans to search through historical handwritten documents.* [En ligne] https://youtu.be/gS3GvvelN4o.

Massot M.-L., Sforzini A., Ventresque V. 2019. Transcribing Foucault's handwriting with Transkribus, *Journal of Data Mining and Digital Humanities*. [En ligne] https://hal.archives-ouvertes.fr/hal-01913435

Mülberger G. 2019. Transkribus. Transcribe. Collaborate. Share, communication à l'atelier Digit-Hum 2019, Paris, 17 octobre 2019.
[En ligne] http://savoirs.ens.fr/uploads/videos/diffusion/2019_10_17_Digit_Hum_2019_Gunter_Muhlberger.mp4.

— *et al*. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, *Journal of Documentation* 75/ 5, 954-976. [En ligne] https://doi.org/10.1108/JD-07-2018-0114

Transkribus User Conference 2020. [En ligne] https://read.transkribus.eu/2020/02/13/transkribus-user-conference-2020-we-keep-transcribing-as-read-coop/

Les auteurs

Emmanuelle Perrin, en contrat pour le projet ANR données ouvertes HisArc-RDF au sein du laboratoire Archéorient, formée aux humanités numériques, participe à la coordination scientifique du projet *Bulliot, Bibracte et moi*.

Philippe Chassignet, archéologue et ingénieur d'étude contractuel au CNRS rattaché au laboratoire Archéorient, fait partie de l'équipe de coordination scientifique du projet *Bulliot, Bibracte et moi*.

Pour citer ce billet : Emmanuelle Perrin et Philippe Chassignet. Rendre visible la face cachée de l'iceberg, *ArchéOrient - Le Blog*, 27 mars 2020, **[En ligne] https://archeorient.hypotheses.org/14807**