(Ne pas) mettre l'humain au service de la machine ? Vérité terrain, legacy data, données ouvertes et modèles entraînés

Laurence Bobis*[1], Joanna Fronska*[2], Simon Gabay*[3], Arsène Georges*[1], Thierry Kouamé*[4], Martin Morard*[2], and Claudia Rabel*[2]

[1]Bibliothèque interuniversitaire de la Sorbonne – UNIVERSITE PARIS 1 PANTHEON-SORBONNE, Université Paris 1, Panthéon-Sorbonne, Université Paris 1 - Panthéon-Sorbonne – France
[2]Institut de recherche et d'histoire des textes – Centre National de la Recherche Scientifique – France
[3]Université de Genève – Suisse
[4]Université Bourgogne Franche-Comté [COMUE] – Université Bourgogne Franche-Comté [COMUE] – France

Résumé

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. Il n'engage pas leur responsabilité scientifique.

Argumentaire
Ce qu'on appelle IA ou intelligence artificielle est aujourd'hui largement synonyme de " apprentissage machine " (*machine learning*). Dans les cas le plus favorable, le savoir est déjà formalisé ou disponible de façon semi-structurée et il est possible " d'apprendre " ou plutôt de faire apprendre à la machine à reproduire le comportement ou l'expertise humaine. Souvent néanmoins une phrase d'entraînement supplémentaire est nécessaire et peut prendre la forme d'annotations massive à l'usage de certaines tâches. Un exemple en sont les " captcha " qui ne servent pas, ou pas que, à prouver que nous ne sommes pas des robots, mais surtout apprendre à reconnaître des feux rouges, des ponts, des passages piétons à l'usage des futures voitures autonomes. Tant que la main-d'œuvre est volontaire et massive (les usager:e:s de tel ou tel moteur de recherche ou fournisseur de messagerie électronique), le tout est facilement intégré.

Dans le cas des communautés concernées par Biblissima+, les tâches qui pourraient être confiées à une machine portent sur des corpus à la fois restreints en nombre et complexes, et dont la complexité est précisément l'objet des études en SHS. Les équipes concernées peuvent vivre le besoin d'annotation pour entraîner une intelligence artificielle comme un asservissement des " humains " au service de la " machine " dans un processus qui n'est pas rentable pour les questions de recherche considérée.

En réponse aux problèmes posés par la nécessité de constituer des corpus d'entraînement

*Intervenant

pour permettre aux ordinateurs d'effectuer les tâches attendues, des solutions émergent. D'une part, le mouvement des " données ouvertes " (*open data*), qu'il s'agisse de bases de données textuelles ou d'images ou d'éditions électroniques, permet de penser la disponibilité d'une expertise formalisée et partagée au-delà du changement et du renouvellement des outils. D'autre part, des communautés commencent à se structurer autour de " modèles pré-entraînés " et publiés et réutilisables, soit librement (publications sur Zenodo par exemple), soit dans un cadre plus restreint (modèles publics de Transkribus).

Contribuant à ces évolutions, plusieurs projets de recherche cherchent à formaliser et exploiter des données qui ont été produites dans un autre cadre et, souvent, avec un autre objectif de recherche. Dans le cas d'éditions de textes médiévaux, certaines ont pu servir à entraîner des modèles d'HTR spécifiques (projet Himanis), d'autres sont converties pour de nouvelles exploitations.

Cette table ronde permettra de considérer les choix faits pour la base iconographique Initiale (http://initiale.irht.cnrs.fr/), pour l'étude des manuscrits brûlés de Chartres (https://www.manuscrits-de-chartres.fr/) et pour l'édition électronique des manuscrits glosés de la Bible (https://gloss-e.irht.cnrs.fr/) avec les porteur:se:s de ces projets de recherche. Elle permettra aussi d'interroger les expériences et résultats de projets de mise à disposition de données et de modèles, ainsi que l'expérience en cours soutenue par Biblissima+ pour la transformation du *Chartularium Universitatis Parisiensis* et de ses suites pour alimenter la base prosopographique ORESM (*Œuvres et Référentiels des Étudiants, Suppôts et Maîtres de l'Université de Paris, des écoles et collèges parisiens*, 1200-1600).

Intervenant:e:s

Laurence Bobis est conservatrice générale, directrice de la Bibliothèque Interuniversitaire de la Sorbonne, porteuse avec Thierry Kouamé du projet *ECRU- Editions critiques relatives à l'Université de Paris.*

Joanna Fronska est historienne de l'art, ingénieure de recherche à l'Institut de Recherche et d'Histoire des Textes, responsable de la base de données *Initiale : Catalogue des manuscrits enluminés* et collaboratrice du projet *Renaissance virtuelle des manuscrits sinistrés de la bibliothèque de Chartres.* Ses centres d'intérêt et ses publications concernent la production, la circulation et l'usage des manuscrits de droit au Moyen Âge, l'iconographie politique et juridique et l'histoire des collections de manuscrits médiévaux.

Simon Gabay est maître-assistant à l'université de Genève auprès de la chaire de Béatrice Joyeux-Prunel et y dirige les projets *E-ditiones: corpus et outils pour l'étude du français classique* (https://github.com/e-ditiones) et *Katabase: base de données des manuscrits en circulation sur le marché privé* (https://github.com/katabase). Outre la philologie française moderne, ses principaux domaines de recherche sont le traitement automatique des langues et la reconnaissance optique de caractères.

Arsène Georges est ingénieur d'études chargé de projets numériques, impliqué dans les projet ORESM et ECRU (conception du modèle de données, traitement et publication des données textuelles).

Thierry Kouamé est professeur à l'Université de Bourgogne Franche-Comté et spécialiste de l'histoire de l'université. Il porte avec Laurent Bobis le projet *ECRU- Editions critiques relatives à l'Université de Paris.*

Martin Morard est chercheur à l'IRHT, spécialiste de l'histoire de l'exégèse de la Bible latine et directeur du projet *gloss-e - Glossae latinae omnes Scripturae -electronicae* (http://gloss-e.irht.cnrs.fr), dans lequel il publie l'édition électronique des gloses et chaînes latines de la Bible, dont la *Catena aurea* de Thomas d'Aquin.
Claudia Rabel est historienne de l'art, spécialiste de l'iconographie et des manuscrits enluminés du Moyen Âge et responsable de la section des Manuscrits enluminés de l'IRHT

et du projet projet dédié à la numérisation et à l'étude des manuscrits brûlés de Chartres. Elle est aussi responsable du séminaire de recherche Les Ymagiers, consacré à l'iconographie médiévale.