Image et texte : les livres d'heures manuscrits vus par l'intelligence artificielle

Dominique Stutzmann*[1] and Louis Chevalier*[2]

[1]Institut de recherche et d'histoire des textes (IRHT) – CNRS : UPR841 – 40 avenue d'Iéna 75116 Paris, France

[2]Institut de Recherche et d'Histoire des Textes (IRHT) – CNRS : UPR841 – France

Résumé

Le projet ANR *HORAE - Hours Recognition, Analysis, Editions* étudie les livres d'heures médiévaux, livres de prière à l'usage des laïcs et le " best-seller " du Moyen Âge. Produits en grand nombre à partir du XIIIe siècle dans toute l'Europe occidentale et plus nombreux que les manuscrits bibliques, ils documentent le sentiment religieux, les pratiques sociales, la circulation des textes et les connexions liturgiques entre des régions différentes, mais aussi les processus d'industrialisation de la production livresque car ils sont largement standardisés (Stutzmann 2019).
L'étude des livres d'heures est complexe. En effet, ces livres sont nombreux et volumineux, compilés à partir de centaines de pièces élémentaires. Ils ont découragé les études systématiques. Or il importe de pouvoir examiner un très grand nombre de ces documents pour saisir la relation qui lie leurs différentes parties, le degré de fréquence ou de rareté des chants, des lectures, oraisons et prières, et comprendre la circulation de ces textes et l'individualisation d'un produit de masse.

Le programme de recherche réunit trois partenaires complémentaires, spécialisés en humanités, en TALN et en IA (particulièrement apprentissage machine) : l'Institut de Recherche et d'Histoire des textes (CNRS), le Laboratoire des Sciences du numérique de Nantes, et Teklia. Les défis technologiques, maintenant relevés, portent sur analyse d'image, tels que la lecture des écritures médiévales par ordinateur ou l'analyse de la mise en page (Boillet et al. 2019 ; Boros et al. 2019), l'identification de textes par " alignement " ou *text reuse identification* (Hazem et al. 2019) et établissement automatique de tables de matières hiérarchiques ou " segmentation " (Hazem et al. 2020 ; Daille et al. 2019). L'analyse historique affronte particulièrement les problèmes de masse de données, de granularité et de complexité des réseaux (Stutzmann et al. 2019).

S'inscrivant dans l'axe 2 de l'appel (effet des méthodes computationnelles en SHS), cette communication présentera les premiers résultats à grande échelle de l'analyse automatique. Pour la démonstration, en utilisant un sous-corpus de 50 manuscrits (11800 images, 248932 lignes de texte, 674 miniatures), nous traiterons d'abord de la méthodologie historique. Les questions et typologies usuelles de la codicologie quantitative, d'une part, et les méthodes plus récentes d'analyse des réseaux textuels sont renouvelées (Ornato 1997 ; É. Cottereau-Gabillet 2015 ; E. Cottereau-Gabillet 2016 ; Julien 2016 ; Riva 2019). L'élaboration des critères d'exploitation et d'interrogation historique peut prendre appui sur des données autrefois soit

*Intervenant

inaccessibles, soit formalisées à un autre niveau de granularité : la comparaison du luxe des manuscrits peut tenir compte du nombre de miniatures, initiales, éléments de décoration autant que de leur superficie réelle ; la comparaison des contenus textuels peut se fonder sur les structures et les pièces unitaires dans des collections de citations ; l'on peut aussi corréler ces domaines généralement séparés.

La communication discutera aussi la mise en œuvre concrète de l'apprentissage automatique et de l'intelligence artificielle. Parfois inutile face aux méthodes dites 'traditionnelles' (temps de calcul, création de données directement utiles aux historiens), elle permet de croiser des critères sans corrélation préalablement connue. Elle pose aussi des problèmes heuristiques. Les *mid-level features* définis dans ce projet (sur ce concept, voir (Hassner et al. 2013)) aboutissent à une production massive de données largement brutes et " à plat ", posant des difficultés d'interprétation (discrétisation d'un continuum de données et clustering, poids excessif des critères d'analyse sur les savoir créés).

L'axe 3 (transmission des patrimoines culturels) sera concerné secondairement, car la recherche se fonde sur les numérisations disponibles via le protocole IIIF, la base de données Heurist produite est hébergée par Huma-Num et permet des valorisations vers le grand public : repérage automatique des miniatures, transcription et traduction de documents autrement hermétiques.

Boillet, Mélodie, Marie-Laurence Bonhomme, Dominique Stutzmann et Christopher Kermorvant. 2019. " HORAE: an annotated dataset of books of hours ". Dans *the 5th International Workshop on Historical Document Imaging and Processing*, 7-12. Sydney : ACM Press. https://doi.org/10.1145/3352631.3352633.

Boros, Emanuela, Alexis Toumi, Erwan Rouchet, Bastien Abadie, Dominique Stutzmann et Christopher Kermorvant. 2019. " Automatic Page Classification in a Large Collection of Manuscripts Based on the International Image Interoperability Framework ". Dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 756-762. Sydney : IEEE. https://doi.org/10.1109/ICDAR.2019.00126.

Cottereau-Gabillet, Émilie. 2015. " Manuscrits de luxe et distinction sociale à la fin du Moyen Âge ". Dans *Marquer la prééminence sociale*, édité par Jean-Philippe Genet et E. Igor Mineo, 283-301. Histoire ancienne et médiévale. Paris : Éditions de la Sorbonne. http://books.openedition.org/psorbonne/3352.

Cottereau-Gabillet, Emilie. 2016. " Revealing Some Structures and Rules of Book Production (France, Fourteenth and Fifteenth Centuries) ". Dans *Ruling the Script in the Middle Ages. Formal Aspects of Written Communication (Books, Charters, and Inscriptions)*, édité par Sébastien Barret, Dominique Stutzmann et Georg Vogeler, 129-163. Utrecht Studies in Medieval Literacy 35. Turnhout : Brepols.

Daille, Béatrice, Amir Hazem, Christopher Kermorvant, Martin Maarand, Marie-Laurence Bonhomme, Dominique Stutzmann, Jacob Currie et Christine Jacquin. 2019. " Handwritten text recognition and text segmentation adapted to manuscript books of hours ". *Traitement Automatique des Langues*, TAL et humanités numériques, 60 (3). ATALA : 13-36.

Hassner, Tal, Malte Rehbein, Peter A. Stokes et Lior Wolf. 2013. " Computation and Palaeography: Potentials and Limits ". *Dagstuhl Manifestos* 2 : 14-35. http://dx.doi.org/doi:10.4230/DagMan.2.1.1

Hazem, Amir, Beatrice Daille, Christopher Kermorvant, Dominique Stutzmann, Marie-Laurence Bonhomme, Martin Maarand et Mélodie Boillet. 2020. " Books of Hours: the First Liturgical Corpus for Text Segmentation ". Dans *Proceedings of The 12th Language Resources and Evaluation Conference*, 776–784. Marseille, France : European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.97.

Hazem, Amir, Béatrice Daille, Dominique Stutzmann, Jacob Currie et Christine Jacquin.

2019. " Towards Automatic Variant Analysis of Ancient Devotional Texts ". Dans *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 240–249. Firenze : Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4730.

Julien, Octave. 2016. " Délier, lire et relier ". *Hypotheses* 19 (1). Éditions de la Sorbonne : 211-224.

Ornato, Ezio. 1997. *La face cachée du livre médiéval: l'histoire du livre vue par Ezio Ornato, ses amis et ses collègues.* I libri di Viella 10. Roma : Viella.

Riva, Gustavo Fernandez. 2019. " Network Analysis of Medieval Manuscript Transmission ". *Journal of Historical Network Research* 3 (novembre) : 30-49. https://doi.org/10.25517/jhnr.v3i1.61.

Stutzmann, Dominique. 2019. " Résistance au changement? Les écritures des livres d'heures dans l'espace français (1200-1600) ". Dans *" Change " in medieval and Renaissance scripts and manuscripts. Proceedings of the 19th Colloquium of the Comité international de paléographie latine (Berlin, 16-18 September, 2015)*, édité par Eef Overgaauw et Martin J. Schubert, 97-116. Bibliologia 50. Turnhout : Brepols.

Stutzmann, Dominique, Jacob Currie, Béatrice Daille, Amir Hazem et Christopher Kermorvant. 2019. " Integrated DH. Rationale of the HORAE Research Project ". Dans . Utrecht. https://dev.clariah.nl/files/dh2019/boa/0192.html.