Maison des sciences humaines et sociales en Bretagne

# Offre de services autour de l'IA à la MSHB - volet données de la recherche

L'essor récent des intelligences artificielles génératives, telles que les chatbots ou les systèmes de production automatique de textes et d'images, a contribué à remettre « l'IA » au centre des débats sur les méthodes scientifiques et l'évolution du numérique dans la recherche, y compris dans les disciplines en sciences humaines et sociales (SHS). Cependant, l'IA ne se réduit pas à ses formes génératives : elle englobe un ensemble de modèles et de systèmes variés – apprentissage automatique (*machine learning*), apprentissage profond (*deep learning*), moteurs de recherche intelligents, systèmes de recommandation – qui participent depuis longtemps aux pratiques de recherche, souvent de manière invisible.

C. Muller et F. Clavert qualifient ces usages de « pratiques numériques discrètes » et rappellent que :

*« Plus discrète encore est la façon dont l'intelligence artificielle (IA) intègre désormais les pratiques de recherche via l'apprentissage machine (machine learning) et l'apprentissage profond (deep learning), sans que leurs utilisateurs et utilisatrices en aient conscience. Utiliser son smartphone dans un centre d'archives, comme l'on utiliserait un scanner, ou transformer une photographie de document en texte sont tout autant de gestes qui intègrent déjà dans le travail des doses d'intelligence artificielle. »*

(Frédéric Clavert et Caroline Muller, « L'histoire au temps des algorithmes : Une réflexion prospective sur l'introduction de l'intelligence artificielle en histoire au 21e siècle », *20 & 21. Revue d'histoire* 162, no 2 (2024): 13-26, https://doi.org/10.3917/vin.162.0013.)

Ces pratiques se manifestent dans de nombreux outils : moteurs de recherche, logiciels de traitement d'images et de textes, technologies de reconnaissance de l'écriture (ATR), bases de données intelligentes, outils de traduction automatique, ou de transcription de fichiers audio, et bien d'autres encore. L'IA participe donc activement à la transformation des méthodes et des outils de recherche en SHS, en influençant la collecte, le traitement et l'analyse des données.

**Dans ce contexte, la MSHB propose une offre de services autour de l'intelligence artificielle structurée en deux niveaux complémentaires :**

**Niveau 1. Sensibilisation aux enjeux et panorama des usages de l'IA en SHS**

La MSHB propose, dans le cadre des Datalabs un service d'accompagnement aux bonnes pratiques de gestion des données de la recherche ; à ce titre, les animateurs et les animatrices peuvent prendre en charge les premières questions relatives à l'IA et orienter les chercheurs et les chercheuses vers des ressources adaptées, notamment concernant les enjeux de biais, de fiabilité et de limites des IA génératives.

Par ailleurs, à Rennes, un atelier est proposé sous la forme d'un panorama des usages de l'IA en SHS. Celui-ci vise à clarifier ce qui est entendu par le terme « intelligence artificielle », en présentant les différents types de systèmes et de modèles regroupés sous cette appellation, leurs apports potentiels ainsi que leurs limites. L'objectif est de permettre aux chercheurs, aux chercheuses et aux personnels d'appui à la recherche d'évaluer la pertinence du recours à des outils d'IA dans le cadre de leurs travaux.

D'autres ressources sur les questions autour des usages de l'IA :

- Café IA – démarche mise en place par le Conseil national du numérique. Source d'informations et de ressources pédagogiques : https://cafeia.org/
- RAGaRenn – expérimentation menée par l'Université de rennes autour de l'IA générative sécurisée, déployée sur le datacenter régional Eskemm Data : https://ragarenn.eskemm-numerique.fr/index.html

**Niveau 2. Usage de l'IA pour des tâches spécifiques de traitement des données en SHS**

Un second niveau d'accompagnement concerne l'usage de méthodes et d'outils mobilisant l'IA pour des tâches spécifiques liées au traitement et à l'analyse de données en SHS, en lien avec les spécialités des deux plateformes technologiques de la MSHB : la plateforme Humanités numériques et la plateforme PUD-B.

Les ingénieurs et ingénieures des plateformes peuvent accompagner les équipes de recherche dans le traitement et l'analyse de leurs corpus, afin d'identifier les types de méthodes et d'outils les plus adaptés. À titre d'exemple, cet accompagnement peut mobiliser :

1. **Des logiciels permettant l'entraînement de modèles de reconnaissance automatique de l'écriture imprimé (OCR) et manuscrite (HTR)** :

- prise en main du logiciel eScriptorium, interface reposant sur le moteur Kraken, qui utilise l'entraînement de modèles (*machine learning*) pour la reconnaissance automatique de l'écriture (imprimé ou manuscrite) ; par exemple, pour des manuscrits à transcrire, un modèle personnalisé peut être entraîné afin d'automatiser la transcription de grands volumes de documents et d'identifier les différentes zones d'un document (segmentation, lignes, blocs de texte). Support de formation à la prise en main : https://zenodo.org/records/18374870

2. **l'usage de bibliothèques Python relevant du traitement automatique des langues (TAL), pour l'analyse du discours et d'autres tâches telles que la reconnaissance d'entités nommées, la lemmatisation ou l'analyse des sentiments, par exemple :**

- NLTK : bibliothèque qui permet des analyses textuelles rapides et basiques, comme la tokenisation, la recherche de mots-clés, les expressions régulières... Elle peut servir, par exemple, pour compter la fréquence de mots ou extraire des noms communs ou des verbes au sein d'un grand corpus.
- spaCy : bibliothèque qui utilise des modèles statistiques et probabilistes pour prédire ou classer des éléments de texte. Elle est adaptée à des tâches un peu plus avancées, comme la classification de documents ou reconnaissance d'entités nommées.
- Transformers (Hugging Face) : modèles d'apprentissage profond (*deep learning*), plus lents mais avec de très hautes capacités, capables d'apprendre des relations complexes dans le langage, par exemple : génération de texte, résumé automatique, classification fine, analyse de sentiments, question-réponse sur corpus.

3. **Des méthodes de traitement et d'analyse statistique des données en SHS, adaptées à des bases de données de taille intermédiaire à importante :**

- pour l'analyse de grands corpus (grande quantité d'individus et de variables dans une enquête, par exemple), des méthodes d'apprentissage automatique (lasso, *ridge, random forest, gradient boosting*), permettant d'optimiser la robustesse et la performance prédictive des modèles statistiques (régressions linéaires, régressions logistiques)
- pour des volumes très importants, des approches d'apprentissage profond (réseaux de neurones)

4. **Des informations sur les outils et services mis à disposition par l'IR*Huma-Num :**

- Aide à l'usage des outils embarqués dans le gestionnaire de fichiers **Sharedocs**
  Sharedocs propose des outils prêts à l'emploi pour un usage de type watchFolder (dossier de traitement automatique), parmi lesquels :
  - Whisper : modèle d'IA permettant la **transcription automatique d'audio et de vidéo**. Whisper est intégré à Sharedocs via le dossier « SpeechToText ». Voici un petit guide qui explique pas à pas comment utiliser Whisper via Sharedocs : https://zenodo.org/records/18407638

Pour plus d'informations, consulter la documentation officielle d'Huma-Num : https://documentation.huma-num.fr/sharedocs-traitement/#sharedocs-outils-de-traitement

**Mise à disposition d'infrastructure pour l'usage de l'IA:**

Des informations sur la mise à disposition par Huma-Num de ressources du Centre de calcul de l'IN2P3 - Huma-Num permet d'accéder aux services suivants du CC-IN2P3 :

- Plateforme de calcul : pour l'exécution de tâches de calcul intensives (*jobs*), nécessitant une puissance de calcul supérieure à celle d'un ordinateur personnel (mise à disposition de CPU/GPU)
- Plateforme Jupyter Notebooks : pour écrire/exécuter/partager du code, l'analyse de données et l'utilisation de modèles d'IA avec des grands modèles de langage (LLM)
- Possibilité d'héberger des applications web basées sur l'IA.

Documentation sur l'infrastructure pour l'usage de l'IA : https://documentation.huma-num.fr/infrastructure-IA/

***Ressources supplémentaires***

- Notebook Jupyter du HN Lab sur la mise en œuvre d'une solution de Retrieval-Augmented Generation (RAG) : https://gitlab.huma-num.fr/hnlab-public/notebook-jupyter-rag

  Ce projet, développé par le HN Lab, propose une solution technique pour l'exploration de ses documents (PDF, Word, etc.). Il permet de créer un agent conversationnel capable de répondre à des questions en s'appuyant sur les documents indexés dans le système RAG. Plus qu'un moteur de recherche, cette approche offre la possibilité d'interroger l'ensemble des documents en langage naturel, facilitant ainsi la navigation et l'analyse du corpus.

- Article de Stéphane Pouyllau sur les RAG en SHS : https://hnlab.huma-num.fr/blog/2024/03/17/RAG/#fn:0
  - Plusieurs ressources à retrouver sur le blog du HN Lab : https://hnlab.huma-num.fr/blog/

**Consortiums d'Huma-Num :**

Vous pouvez retrouver des ressources sur l'usage des technologies d'IA adaptées à différents types de corpus en SHS au sein des **consortiums d'Huma-Num**, par exemple :

- Pour le traitement de corpus iconographiques via le consortium **PictorIA** (retours d'expérience, tutoriels, etc.) : https://pictoria.hypotheses.org/
  - Notamment, pour les corpus iconographiques, une instance Huma-Num de la plateforme Arkindex a été mise en place : cette plateforme est consacrée au traitement et à l'analyse des images à l'aide d'outils d'intelligence artificielle (segmentation, annotation, reconnaissance de formes...)
  - Utilisation du logiciel Arkindex, tutoriel : https://pictoria.hypotheses.org/4362

- Pour le traitement de données textuelles, via le consortium **Ariane** (informations sur l'usage de LLM, algorithmes d'analyse textuelle...) https://consortiumariane.gitpages.huma-num.fr/axe2/

**Projets labellisés**

Par ailleurs, la MSHB soutient actuellement le projet PROMPT - Enseigner et apprendre avec l'intelligence artificielle générative. Ce projet « conduit par une équipe interdisciplinaire composée d'enseignants-chercheurs et d'ingénieurs pédagogiques, examine l'usage de l'intelligence artificielle générative (IAG) par les étudiants et les enseignants dans le contexte des formations universitaires. Il vise à élaborer des recommandations pour une intégration raisonnée de l'IAG dans les pratiques pédagogiques. Le projet porte une attention particulière à la rédaction des prompts et l'analyse critique des réponses produites par les outils d'IAG. »

Pour plus d'informations, contactez les porteurs du projet : Nicolas Thély (PTAC - UR 7472, Université Rennes 2) et Catherine Archieri (CREAD - UR 3875, Université de Bretagne Occidentale).

14 avril 2026