Institut national de la recherche scientifique

WIKIMEDIA CANADA

SYNTHÈSE DE LA JOURNÉE D'ÉTUDE COMMUNS DU NUMÉRIQUE ET IA

1er novembre 2024
WikiConvention francophone
Québec

WikiConvention francophone Québec 2024

Direction et rédaction de la synthèse
Nathalie Casemajor

Codirection de l'événement
Louis Germain

Comité de lecture de la synthèse
Remy Gerbet, Lucie Gianola, Thomas Mboa, Jean-Philippe Moreux, Stéphane Nepton, Thérèse Ottawa, Viriya Thach, Liam Wyatt

Révision et mise en forme
Sophie Valade

Édité par : INRS et Wikimédia Canada, Montréal

2025

L'édition de la WikiConvention francophone à Québec est rendue possible grâce au soutien de la Fondation Wikimédia et du gouvernement du Québec, en vertu des programmes de soutien en matière de francophonie canadienne.

Québec

INTRODUCTION

Ce document propose une synthèse des échanges qui ont eu lieu lors de la journée d'étude *Communs du Numérique et IA*. Tenue le 1er novembre 2024 à Québec, lors de la WikiConvention francophone, cette journée était organisée par Wikimédia Canada et l'INRS. Elle a rassemblé une quarantaine de personnes issues du mouvement Wikimédia, du milieu des bibliothèques et archives, du milieu gouvernemental (ministères de la Culture en France et au Québec), du monde universitaire et du milieu de l'IA.

Nous tenons à remercier les personnes suivantes, invitées à intervenir lors de l'événement, pour avoir partagé leur expertise et enrichi le contenu de cette synthèse :

- Jenny Ebermann (directrice exécutive de Wikimédia Suisse)
- Georges Fodouop (responsable informatique et formateur Wikimédia, co-fondateur du Wikimedians of Cameroon User Group)
- Rémy Gerbet (directeur exécutif de Wikimédia France)
- Lucie Gianola (chargée de mission pour les technologies, la recherche et l'innovation au ministère de la Culture de France)
- Maryana Iskander (PDG de la Fondation Wikimédia)
- Thomas Mboa (chercheur en résidence au Centre d'expertise international de Montréal en intelligence artificielle)
- Jean-Philippe Moreux (chef de mission IA à la Bibliothèque nationale de France)
- Stéphane Nepton (coordonnateur de projet au Printemps numérique, ambassadeur numérique autochtone)
- Christophe Prévost (Conseiller en intelligence artificielle, réglementation et télécommunications, ministère de la Culture et des Communications du Québec)
- Ayla Rigouts Terryn (professeure à l'Université de Montréal)
- Anastasia Stasenko (co-fondatrice de la start-up Pleias)
- Viriya Thach (responsable de la gouvernance des données à Bibliothèque et Archives nationales du Québec)
- Daniel van Strien (*machine learning librarian* à Hugging Face)
- Liam Wyatt (gestionnaire de programme à Wikimedia Enterprise)

Merci également à Thérèse Ottawa, Sabrina MacGregor, Benoît Rochon, Sophie Valade et Stéphane Couture pour leur aide précieuse dans l'organisation de l'événement et l'animation des ateliers.

Nous exprimons également notre reconnaissance au Consulat général de France à Québec, à l'Institut national de la recherche scientifique (INRS) et au Centre de recherche interuniversitaire sur les humanités numériques (CRIHN) pour leur soutien financier à cette journée d'étude.

Objectif

Cette synthèse vise à ouvrir une discussion sur la place des communs numériques, et en particulier **les communs de la connaissance, comme ressource d'entraînement pour les systèmes d'intelligence artificielle (SIA)[1] dans les territoires francophones**. Les grands modèles de langage (LLM, tels que ceux à la base de services comme ChatGPT) sont majoritairement entraînés sur des corpus anglo-américains, ce qui cause une sous-représentation des contenus et schémas culturels des autres territoires. L'entraînement d'IA sur des corpus dans d'autres langues, dont le français, mais aussi les langues des territoires francophones (langues autochtones, langues régionales), revêt donc une importance cruciale pour l'accès au savoir dans une pluralité de contextes culturels et linguistiques.

Les communs de la connaissance désignent des ressources informationnelles, culturelles ou artistiques produites, partagées et gouvernées collectivement par une communauté autogérée, selon des règles établies par les personnes participantes. Ils se caractérisent par leur accessibilité, leur non-rivalité (leur usage par une personne n'empêche pas celui par une autre) et leur enrichissement continu grâce à la collaboration ouverte. Dans l'environnement numérique, un exemple emblématique est constitué par les projets Wikimédia. Ils incluent l'encyclopédie Wikipédia (en 340 langues), la base de données liées multilingue Wikidata (115 millions d'éléments), la base multimédia Wikimédia Commons (112 millions de fichiers) et le projet Lingua Libre (près d'1 million d'enregistrements en 170 langues). Leurs contenus sont publiés sous une licence libre autorisant la réutilisation, y compris commerciale. Ces communs constituent des données ouvertes précieuses pour l'entraînement des modèles d'IA, incluant dans des langues à ressources limitées.

De même, **les fonds patrimoniaux publics** (bibliothèques et archives nationales, fonds audiovisuels) sont convoités pour entraîner des IA francophones. Ces fonds et collections revêtent une valeur inestimable en tant que documents historiques et ressources clés pour l'étude et la transmission du patrimoine culturel. Leur statut juridique est cependant complexe : les œuvres les plus anciennes appartiennent au domaine public, ce qui signifie qu'elles ne sont plus protégées par des droits d'auteur et peuvent (théoriquement) être librement utilisées. Certaines de ces

[1] Les systèmes d'intelligence artificielle (SIA) sont des systèmes technologiques qui utilisent des algorithmes avancés, tels que l'apprentissage automatique, le traitement du langage naturel et la vision par ordinateur, pour accomplir des tâches de classification, de raisonnement, de génération de contenu ou de prise de décision. Ces systèmes reposent sur une modélisation des relations entre des données d'entrée et des données de sortie. L'entraînement de ces modèles nécessite de grands ensembles de données (texte, des images, des vidéos ou des sons, dont une grande partie sont collectées sur le Web) ainsi que des données annotées pour l'entraînement supervisé. Dans cette synthèse, c'est en particulier les cas de l'entraînement de grands modèles de langage et des services d'IA générative conversationnelle qui ont servi de cadre de réflexion.

ressources, accompagnées d'ensembles de métadonnées descriptives, sont rendues accessibles sous forme de données ouvertes, à la fois sur les portails internes de ces institutions et via des plateformes tierces. En revanche, les œuvres plus récentes restent protégées par le droit d'auteur, limitant leur exploitation sans autorisation ou licence appropriée. Par ailleurs, ces fonds sont souvent soumis à un empilement de couches de droits (droits sur les reproductions, droits voisins, droits moraux), rendant leur mise à disposition pour l'entraînement des modèles d'IA particulièrement difficile à évaluer.

Il existe des différences significatives entre les ressources informationnelles gérées par le mouvement Wikimédia et par les GLAM. Ces deux cas de figure se distinguent tant en termes de production que de gestion et de financement. Les ressources de Wikimédia sont produites par des bénévoles, gérées selon des principes de gouvernance collective des communs[2] et financées essentiellement par des dons. En revanche, les fonds patrimoniaux des bibliothèques et archives nationales constituent un héritage national, valorisé par des professionnels, administré par des entités publiques sous la tutelle de l'État et largement financé par des subventions publiques.

Cependant, les fonds patrimoniaux publics et les projets wikimédiens partagent certaines missions et valeurs fondamentales : la préservation du patrimoine collectif, l'accessibilité des ressources, leur potentiel de réutilisation (notamment pour les œuvres dans le domaine public) et leur contribution au bien commun. Dans les deux cas se pose la question des types de relations à construire avec les acteurs de l'industrie de l'IA sur le plan économique, juridique et éthique.

Or le dialogue entre les projets Wikimédia et les établissements publics patrimoniaux concernant ces enjeux reste très limité à ce jour. Il est capital d'approfondir ces échanges à un moment où se cristallisent des relations et des modèles qui auront un impact sur les années à venir. Cette synergie est particulièrement importante dans le contexte de la montée de la désinformation et du désengagement des plateformes numériques commerciales envers la vérification des faits.

[2] Présence d'une communauté active autour d'une ressource disponible (matérielle ou immatérielle), organisée par une structure de gouvernance collective et régie par des règles claires d'accès et d'usage.

Contenu de la synthèse

La première partie propose un résumé détaillé des interventions réalisées lors des trois panels, abordant les relations entre **Wikimédia et les grands modèles de langage** (LLM), le rôle des **bibliothèques francophones** dans la constitution de corpus d'entraînement pour les SIA, ainsi que la **diversité linguistique et culturelle** dans les SIA de la francophonie.

La seconde partie propose une synthèse des discussions issues des ateliers, organisée en cinq dimensions principales pour structurer les résultats et les pistes d'action identifiées :

1. **La production des jeux de données**
2. **La pluralisation linguistique et culturelle** des données et des SIA
3. **La mise en accès des jeux de données**
4. **Les outils juridiques** applicables aux jeux de données
5. **Les modèles économiques**

Les enjeux et les pistes d'action identifiés dans chacune de ces dimensions ont été regroupés selon les principes suivants :

Équité	Nécessité de garantir d'un accès juste et inclusif aux bénéfices de l'exploitation des ressources de données en tenant compte des disparités économiques et sociales, ainsi que de la diversité culturelle et linguistique, tant chez les producteurs et que chez les utilisateurs de ces ressources.
Souveraineté	Capacité des organisations et des collectifs concernés par la constitution et l'exploitation des jeux de données à prendre des décisions éclairées et à agir de manière autonome pour assurer la gouvernance de ces ressources.
Découvrabilité	Capacité à rendre visibles les producteurs de données ainsi que leur offre de jeux de données, afin de reconnaître leur rôle et leur contribution, tout en mettant en valeur les spécificités culturelles et linguistiques dans les contenus ou résultats générés par les SIA.
Utilisabilité	Garantir que les jeux de données soient pertinents, de haute qualité, structurés, publiés dans des formats diversifiés et interopérables, juridiquement clairs et soutenus par une infrastructure robuste et extensible, afin de répondre efficacement aux besoins variés des utilisateurs tout en maximisant leur accessibilité et leurs usages.
Soutenabilité	Assurer un usage responsable et durable des ressources de données en équilibrant les dimensions économiques, sociales et environnementales, grâce à des mécanismes de régulation et à la négociation de contreparties permettant d'assurer leur préservation et leur développement.

Certains des enjeux et pistes d'action identifiés dans cette synthèse sont propres au mouvement Wikimédia, tandis que d'autres concernent davantage le contexte des

institutions GLAM (Galeries, Bibliothèques, Archives et Musées). Les structures, les ressources et les modes de gouvernance de ces deux types d'acteurs présentent des différences significatives, qui peuvent engendrer des écarts dans leurs priorités et leurs modes d'action. Il est crucial de reconnaître ces distinctions et leurs implications, sans pour autant négliger les opportunités de complémentarité et de synergie. **En mutualisant les réflexions et les pistes d'action, il devient possible de concevoir des stratégies d'intervention croisées et de renforcer les dynamiques collectives.**

La conclusion trace des pistes d'action prioritaires et des questions de recherche à creuser. Une section Ressources propose une liste de références bibliographiques pour creuser différents aspects du problème.

SYNTHÈSE DES PANELS

1- RELATIONS ENTRE WIKIMÉDIA ET LES GRANDS MODÈLES DE LANGAGE (LLM)

Animé par Jenny Ebermann (directrice exécutive de Wikimédia Suisse), ce panel visait à explorer les relations entre les données de Wikimédia et les grands modèles de langage (LLM) avec les objectifs suivants :

- **comprendre l'utilisation des données de Wikimédia** dans le contexte des grands modèles de langage (LLM), de l'IA générative et des systèmes RAG (Retrieval-Augmented Generation) ;
- **explorer les relations possibles entre Wikimédia et un écosystème d'innovation ouverte** réunissant des acteurs publics, privés et associatifs ;
- **identifier les valeurs, modèles économiques et contreparties nécessaires** pour soutenir la vitalité des communs de la connaissance et soutenir les communautés qui les produisent.

Les échanges entre les intervenants et avec la salle ont permis d'identifier plusieurs pistes d'action pour renforcer les relations entre Wikimédia et les acteurs de l'IA. Anastasia Stasenko a plaidé pour une démocratisation des outils d'entraînement, notamment en fournissant des données adaptées à des entreprises de taille moyenne en dehors des grands groupes. Liam Wyatt a insisté sur l'importance de structurer les données de manière à abaisser le seuil d'entrée pour les petits acteurs tout en conservant une traçabilité efficace. Rémy Gerbet a souligné la nécessité de coordonner les efforts au sein du mouvement Wikimédia et de se mobiliser pour influencer les législations européennes, en particulier dans le cadre des nouvelles régulations sur le droit d'auteur et l'IA.

Enfin, des préoccupations ont été soulevées sur la transparence des usages des données issues du mouvement Wikimédia, sur la reconnaissance du travail des contributeurs et contributrices bénévoles, ainsi que sur les biais géographiques dans les processus décisionnels. Ces discussions ont mis en lumière les tensions entre la préservation des valeurs des communs et la réponse aux pressions des grands acteurs commerciaux, tout en offrant des perspectives pour construire des solutions durables et inclusives.

Pleias : le rôle des données ouvertes dans l'écosystème de l'IA

La première intervenante, Anastasia Stasenko, est cofondatrice de **pleias, une start-up spécialisée dans les modèles de langage basés sur des données**

ouvertes. Son collègue Pierre-Carl Langlais, autre cofondateur de l'organisme, est par ailleurs un contributeur actif et administrateur au sein de Wikipédia. Pleias soutient que les données ouvertes constituent une opportunité stratégique pour favoriser le développement des systèmes d'intelligence artificielle. Sa vision s'appuie sur l'auditabilité des données d'entraînement, un principe souvent négligé depuis l'émergence de systèmes fermés tels que GPT-3.

Selon A. Stasenko, **valoriser la qualité et la factualité des données ouvertes permet de développer des modèles fiables et adaptés à des usages sensibles**. Ces données de haute qualité, notamment celles issues de Wikimédia, jouent un rôle essentiel dans la capacité des modèles à reproduire un raisonnement crédible et argumenté[3]. Cependant, la disponibilité des données ouvertes reste un défi ; les bases de données telles Common Crawl[4], massivement utilisées par les industries de l'IA**,** en contiennent relativement peu. Toutefois, il existe des ressources de grande valeur qui pourraient être davantage valorisées sous la forme de données ouvertes, tels des dépôts de PDF.

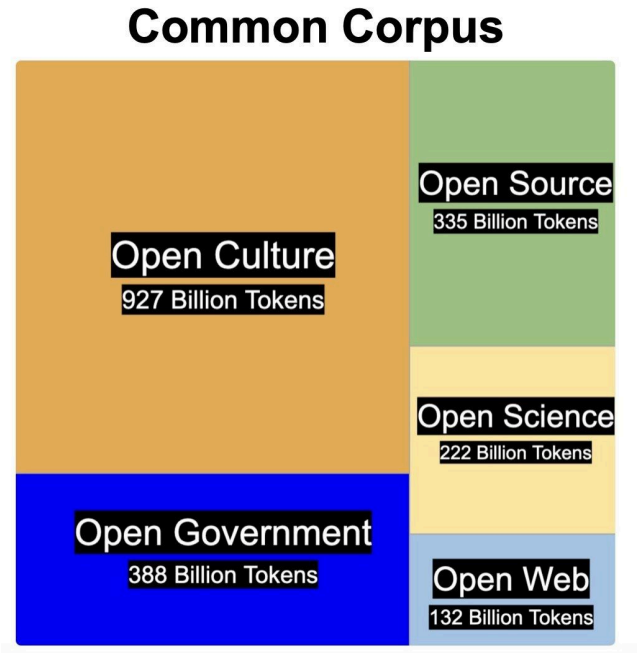


Figure 1 : Composition du Common Corpus. Source : pleias

Dans son Common Corpus[5] (publié en 2024), Pleias a rassemblé deux trillions de tokens ouverts, constituant à ce jour le plus grand corpus multilingue entièrement

[3] Les pages de discussion des articles sont en particulier des lieux d'argumentation.
[4] L'archive web de Common Crawl se compose de pétaoctets de données collectées depuis 2008. Des versions filtrées de Common Crawl sont utilisées pour entraîner de nombreux grands modèles de langage.
[5] https://huggingface.co/blog/Pclanglais/common-corpus

fondé sur des données ouvertes. Selon A. Stasenko, ils sont d'une qualité nettement supérieure à celle des données couramment utilisées dans des corpus comme Common Crawl. Pleias a également développé sa première suite de LLM (3b, 1b, 350m) sur un corpus ouvert qui intégrait des corpus de données substantiels provenant de Wikipédia. Ces derniers ont été fournis par Wikimedia Enterprise sous des formats structurés (HTML et JSON). Elles ont été intégrées dans un sous-corpus plus large appelé Open Web, composé de contenus sous licences permissives. Ce type de données est aussi utile pour aligner les modèles et améliorer leur performance. Wikidata constitue par ailleurs une ressource clé pour cibler des données factuelles et **alimenter les systèmes de Retrieval-Augmented Generation (RAG).**

A. Stasenko a mis en avant la nécessité de renforcer les liens entre Wikimédia et les acteurs de l'écosystème de l'IA. Pour y parvenir, il est essentiel selon elle d'intégrer activement des données ouvertes de haute qualité – notamment issues du projet Wikidata – et de favoriser une collaboration étroite entre les différents acteurs.

Wikimédia France : accompagner les projets au service des communautés Wikimédia

Rémy Gerbet, directeur exécutif de l'association Wikimédia France (WMFr), a présenté le rôle stratégique que WMFr peut jouer dans le développement de modèles de langage open source et les bénéfices potentiels pour l'ensemble du mouvement Wikimédia. Dans un contexte de sollicitations croissantes de la part des pouvoirs publics sur ces questions, **WMFr se positionne comme un acteur clé de médiation et d'accompagnement, en proposant des réponses aux défis politiques posés par l'articulation entre innovation technologique et gouvernance communautaire du mouvement.**

La collaboration de WMFr avec Pleias a constitué un tournant majeur de sa réflexion sur la manière dont **les outils d'IA peuvent être rendus utiles aux communautés de Wikimédia.** Ce partenariat a souligné la capacité de WMFr à jouer un rôle de facilitateur entre les acteurs de l'IA et les communautés internes du mouvement. Dans cette optique, WMFr s'investit activement dans plusieurs initiatives visant à développer des outils liés aux IA génératives.

Le projet **PROMPT (Predictive Research On Misinformation & Propagation Trajectories)**[6], financé par l'Observatoire Européen des Récits (ENO - Union européenne), vise à détecter les formes de désinformation, notamment dans

[6] https://disinfo-prompt.eu/

PROMPT
Narrative Intelligence for Information Integrity

Wikipédia. Ce projet pilote, mené par Opsci[7], vise à développer un modèle de langage ouvert (LLM) de manière à identifier des récits malveillants. Trois études de cas sont ciblées : la guerre en Ukraine, les droits LGBTQI+ et les élections européennes de 2024. PROMPT se concentre sur la reconnaissance approfondie des motifs formels dans les médias, les réseaux sociaux et Wikipédia afin de lutter contre la propagation de fausses informations.

Le projet AuditLLM, porté par OpSci et Wikimédia France dans le cadre du programme France 2030, vise à analyser les biais présents dans les corpus Wikimédia avant leur utilisation pour l'entraînement de modèles de langage. En démarrage début 2025, ce projet, soutenu par le Campus Cyber, ambitionne de développer une gamme innovante de grands modèles de langage dédiés à l'audit des corpus et des modèles.

Pour finir, R. Gerbet a mis en avant l'importance de coordonner les efforts au sein du mouvement Wikimédia pour répondre aux **défis législatifs, face à la complexité croissante des lois européennes sur le droit d'auteur et l'intelligence artificielle**. L'association accorde une attention particulière à l'attribution rigoureuse des contenus, en conformité avec les principes des licences libres (CC-BY).

Wikimedia Enterprise : un modèle commercial pour les grands usagers commerciaux

Liam Wyatt est gestionnaire de programme à Wikimedia Enterprise[8], un projet lancé par la Fondation Wikimédia en 2021. **Ce nouveau service, à vocation commerciale, s'adresse aux grands utilisateurs commerciaux des contenus de Wikimédia.** Il a pour objectif de générer des revenus propres afin de soutenir le mouvement Wikimédia, en monétisant des services d'accès aux données fiables, continus et à haut volume via des API avec des fonctions développées sur mesure pour répondre aux besoins de formats et de débit des différents clients. Son intervention a abordé le rôle de Wikimedia Enterprise et ses API dans l'établissement de relations contractuelles avec les acteurs de l'IA, ainsi que les termes à négocier pour aligner ces collaborations avec les valeurs de Wikimédia.

L. Wyatt a insisté sur le fait que **les projets de Wikimédia sont conçus "par et pour les humains",** mettant en avant le principe essentiel de maintenir "l'humain dans la

[7] https://www.opsci.ai/fr
[8] https://enterprise.wikimedia.com/

boucle" de l'IA. Il a insisté sur la nécessité de renforcer les pratiques en accord avec les valeurs fondamentales de Wikimédia, en prenant pour exemple la collaboration de Wikimedia Enterprise et pleias, une start-up qui privilégie l'attribution rigoureuse des contenus et le respect des principes du libre.

En ce qui concerne le modèle de Wikimedia Enterprise, il a expliqué que son objectif est de répondre aux besoins des grands utilisateurs, tels que les GAFAM et autres acteurs commerciaux majeurs, tout en amplifiant la portée et l'impact des contenus Wikimédia. Il est en effet de plus en plus **difficile pour les projets Wikimédia de soutenir l'énorme volume d'utilisation généré par ces entreprises** avec une infrastructure financée principalement par de petits dons. Pour répondre à ce défi, la Fondation Wikimédia a adopté une stratégie visant à diversifier ses sources de revenus, notamment en monétisant l'accès aux réseaux à haut débit indispensables à ces utilisateurs, tout en préservant l'accès libre et égalitaire au contenu. "**Nous ne vendons pas un meilleur Wikipédia, mais le tuyau par lequel il est transmis – l'eau reste la même"**, a-t-il précisé.

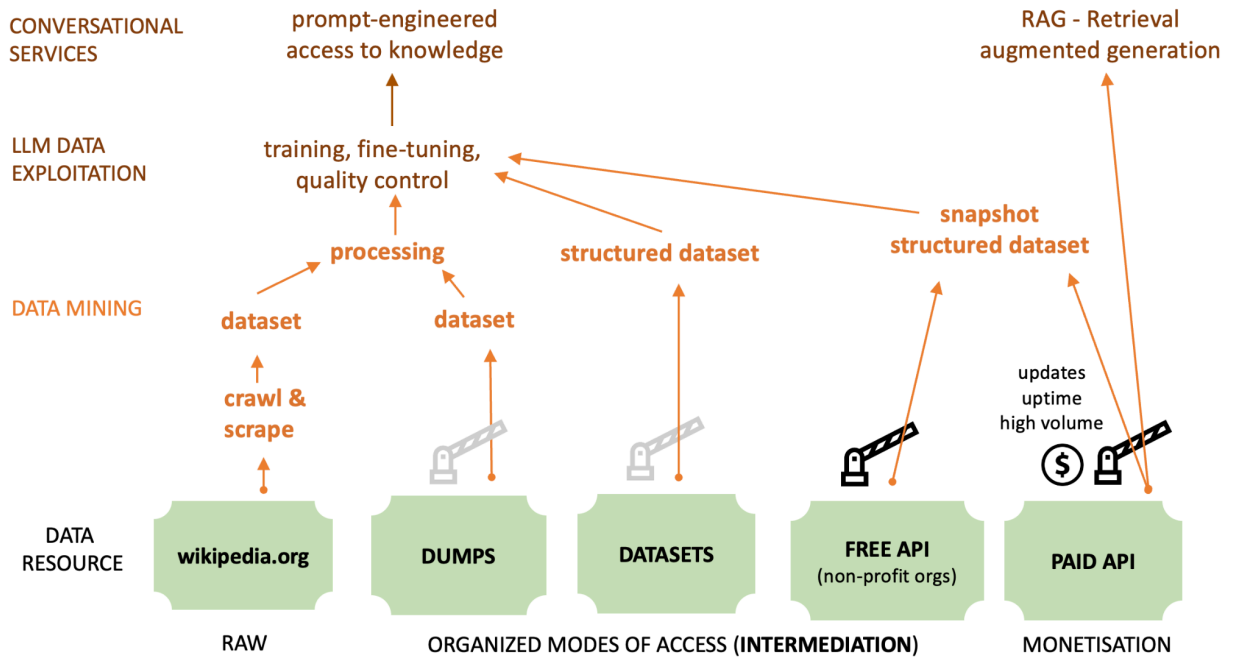
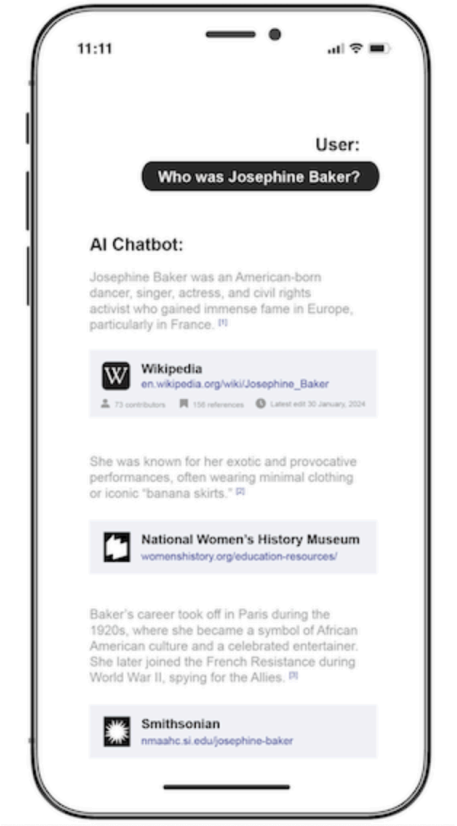


Figure 2 : Schéma des différents modes d'accès aux données de Wikipédia pour des usages d'apprentissage automatique, de services d'IA conversationnelle et de RAG. Source : Nathalie Casemajor, CC-BY-SA.

Fin 2024, **Wikimedia Enterprise a publié sur la plateforme Hugging Face une version bêta d'un jeu de données** issu des versions anglaise et française de Wikipédia. Ce jeu de données structuré est lisible par les machines et dérivé d'une nouvelle fonctionnalité appelée "Structured Contents" développée sur l'API Snapshot de Wikimedia Enterprise. **La publication du jeu de données inclut des informations détaillées sur l'attribution des données,** ainsi que des exemples illustrés d'interface utilisateur d'un service d'IA conversationnelle attribuant à Wikipédia des résultats d'un prompt.



Wikipedia Attribution						
Application	**Contributors.** Celebrate the millions of people who volunteer their time to keep Wikipedia current, trustworthy and reliable.	**Brand.** The Wikipedia brand is associated with trustworthy, reliable information.	**Access.** Offering one click access to Wikipedia.org allows your end-users a chance to give back to Wikipedia through donations and new edits.	**Sources.** Highlight Wikipedia's verifiability by including sources and references to drive trust.	**Last update.** Highlight that Wikipedia is up to date updated every 3.3 seconds	*Disclaimer
Outputs using Wikipedia in-line	XX contributors in plain text	Wikipedia Favicon	Hyperlink to Wikipedia article/s	XXX References	Last updated MM/YYYY	N/A
Outputs using Wikipedia non-specifically	XX contributors in plain text	Wikipedia Favicon	Hyperlink to Wikipedia article/s	XXX References	Last updated MM/YYYY	Sources

Figure 3 : Exemples d'interface utilisateur pour l'attribution de résultats à Wikipédia dans un service d'IA conversationnelle. Source : Wikimedia Enterprise.

En fin d'intervention, L. Wyatt a détaillé **trois axes de réflexion actuels pour la Fondation Wikimédia :** les usages commerciaux des données Wikimédia, l'exploitation de l'IA par les lecteurs pour diffuser des contenus issus de Wikipédia (ex. des courtes vidéos sur TikTok ou Instagram), et l'utilisation de l'IA par les contributeurs au sein des projets Wikimédia.

2- BIBLIOTHÈQUES FRANCOPHONES ET CORPUS D'ENTRAÎNEMENT DES IA

Ce panel avait pour objectif d'explorer le rôle des bibliothèques francophones dans la constitution et l'utilisation de corpus d'entraînement pour les modèles d'IA. Animé par Nathalie Casemajor (INRS), il réunissait Viriya Thach (BAnQ), Jean-Philippe Moreux (BnF) et Daniel van Strien (Hugging Face). Ce panel s'est tenu peu après le *XIX^e Sommet de la Francophonie*, organisé à Villers-Cotterêts et Paris les 4 et 5 octobre 2024. À cette occasion, les gouvernements français et québécois ont signé **la *Déclaration de Villers-Cotterêts*[9], reconnaissant la contribution des institutions patrimoniales à l'essor de l'entraînement en français des systèmes d'intelligence artificielle**.

Les présentations ont porté sur les moyens par lesquels les institutions patrimoniales peuvent répondre aux demandes croissantes d'utilisation de leurs ressources pour l'entraînement des modèles d'IA. Deux principaux angles ont été abordés :

- **les enjeux liés à la création et à l'accessibilité des jeux de données** provenant des milieux documentaires pour l'entraînement des IA (préparation des données, plateformes d'hébergement, licences) ;
- **la gouvernance des relations entre institutions patrimoniales et acteurs de l'IA** dans l'utilisation des ressources publiques (négociation des usages, bénéfices réciproques, attribution et traçabilité des sources).

Lors des échanges avec la salle, plusieurs enjeux ont été mis en lumière par les intervenants, notamment la **difficulté de déterminer où publier les jeux de données en fonction des publics cibles, des besoins des utilisateurs et des contraintes d'hébergement.** Les options incluent les sites internes aux institutions patrimoniales, les portails étatiques[10] ou des plateformes comme Hugging Face, fréquentées par des spécialistes en informatique. Référencer les jeux de données sur plusieurs plateformes a été identifié comme un moyen de maximiser leur visibilité.

---

9 https://www.francophonie.org/sites/default/files/2024-10/Declaration_XIXeSommet_SOM_XIX_05102024_vf.pdf

10 Voir également le Cultural Heritage Cloud (Union Européenne), une initiative de l'Union européenne visant à créer une infrastructure numérique qui connectera les institutions et les professionnels du patrimoine culturel à travers l'UE. https://research-and-innovation.ec.europa.eu/research-area/social-sciences-and-humanities/cultural-heritage-and-cultural-and-creative-industries-ccis/cultural-heritage-cloud_en?prefLang=fr

Par ailleurs, les intervenantes et intervenants ont indiqué que **les biais et lacunes des collections patrimoniales sont mieux documentés lorsque l'accès aux données est effectué en collaboration avec les institutions détentrices.** Cette approche collaborative ajoute une valeur significative en termes de contextualisation et de médiation des contenus, tout en valorisant le rôle essentiel des archivistes et bibliothécaires.

BAnQ : une banque de données reflétant le contexte québécois

Viriya Thach est responsable de la gouvernance des données à Bibliothèque et Archives nationales du Québec (BAnQ). BAnQ est un acteur clé de la conservation, de la diffusion et de la valorisation du patrimoine documentaire québécois. Elle assume en outre un rôle-conseil en gestion de l'information auprès des ministères et des organismes publics. Ces missions l'amènent à jouer un rôle clé dans la gouvernance des données culturelles au Québec. BAnQ est également détentrice et productrice de données, avec à ce jour, **trois pétaoctets de données numériques (équivalent à 1500 milliards de pages imprimées), incluant des millions d'éléments des collections patrimoniales sur le Web.**

Le rapport *Prêt pour l'IA,* publié par le Conseil de l'innovation du Québec en 2024, a souligné la nécessité de mieux adapter les jeux de données aux différents contextes nationaux et locaux, de manière à refléter les réalités sociales, culturelles, politiques et historiques du Québec :

> ***le Québec doit produire et valoriser des données de haute qualité pour entraîner des systèmes d'IA qui tiennent compte de ses spécificités, de sa réalité et de ses besoins.*** *En particulier, les systèmes d'IA générative que les Québécois utilisent ne répondent pas toujours de manière optimale à leurs besoins et attentes. En effet, cela est en partie dû au fait que les modèles statistiques sont construits à partir d'une trop faible quantité de données en langue française et de données québécoises. Pour favoriser le développement et le déploiement hautement responsable de SIA qui sont véritablement centrés sur les besoins des Québécois,* ***le Conseil recommande au gouvernement du Québec d'imiter la Suède en soutenant le développement d'une banque de données culturelles nationales de haute qualité en français et en langues autochtones*** *(RP-9). (p. xiii)*

Le projet suédois mentionné dans le rapport est **The Nordic Pile[11]. Ce jeu de données de haute qualité de 1,2 To a été produit et mis en accès par AI Sweden pour soutenir le développement des LLM dans les langues nordiques,** où les

[11] https://github.com/AI-Nordics/the-nordic-pile

corpus textuels sont limités (Öhman et al., 2023). Il inclut des textes dans les principales langues germaniques septentrionales (danois, islandais, norvégien et suédois), ainsi que des données en anglais de haute qualité.

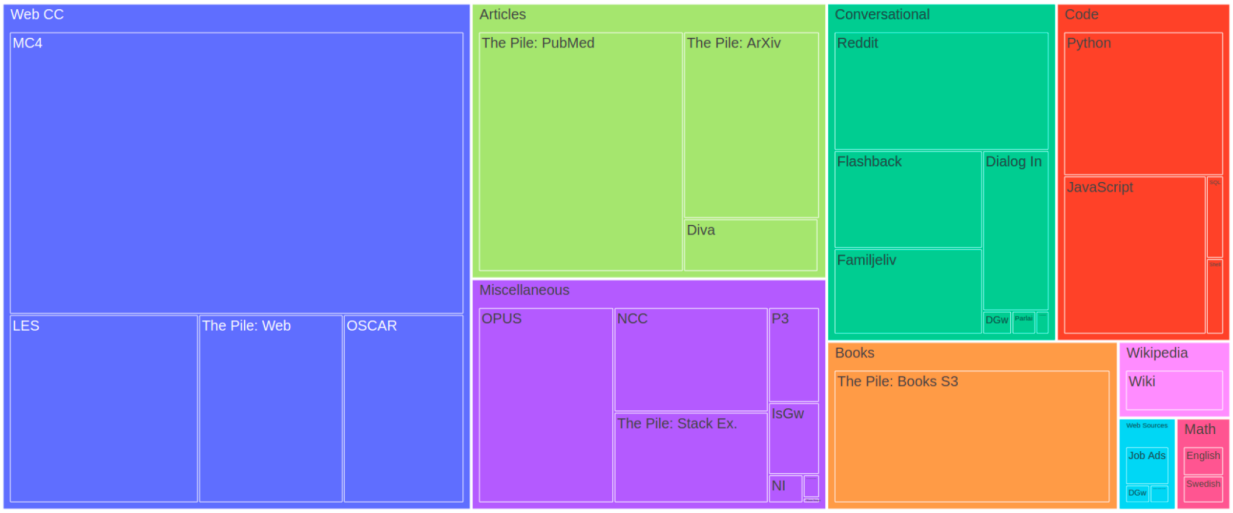


Figure 4 : Carte proportionnelle (treemap) visualisant le jeu de données final de The Nordic Pile. Source : Öhman et al., 2023.

En 2024, BAnQ s'est vu confier par le ministère de la Culture et des Communications du Québec un mandat pour étudier la faisabilité d'un pôle francophone regroupant les données publiques culturelles sous la forme d'une **Banque de données gouvernementales et culturelles québécoises en français et en langues autochtones.** Ce projet de mutualisation vise à réduire les biais sociolinguistiques dans les systèmes d'IA, à renforcer la découvrabilité des contenus culturels francophones, ainsi qu'à préserver et la promouvoir de l'identité culturelle du Québec.

En parallèle, BAnQ adhère fermement à la volonté du milieu culturel d'investir des efforts pour la normalisation des données culturelles de haute qualité, utiles notamment pour accroître l'efficacité de l'entraînement des systèmes d'intelligence artificielle. De plus, le fait d'établir des normes communes relatives aux métadonnées et aux descriptions de contenu, cela faciliterait l'interopérabilité entre différentes plateformes et services, améliorant ainsi la découvrabilité des contenus culturels. À ce titre, le rapport du Comité-conseil sur la découvrabilité des contenus culturels paru en 2024 a émis une recommandation sur le besoin urgent de normalisation des données culturelles (notamment les données descriptives des œuvres) en travaillant à des normes communes par secteur, tout en veillant à l'interopérabilité avec les normes d'autres juridictions.

Enfin, BAnQ s'engage résolument envers le principe d'approvisionnement responsable des données, en s'inspirant de **certifications comme Fairly Trained**[12], qui valorisent la transparence, l'équité et la protection de la vie privée. La collaboration internationale, en particulier avec la France, est également une priorité pour construire une stratégie conjointe en matière de métadonnées de qualité et d'accès élargi, en réponse aux pressions des grands acteurs commerciaux du numérique.

BnF : corpus du patrimoine et IA générative – repenser les collaborations

Jean-Philippe Moreux, chef de mission IA à la Bibliothèque nationale de France (BnF), a décrit l'évolution des usages de l'IA à la BnF et les chantiers actuels.

Le réservoir de données de la BnF est unique par sa profondeur historique et la diversité de ses contenus culturels. Il contient plusieurs centaines de milliards de mots issus des imprimés océrisés, 11 millions de documents numérisés dans la bibliothèque numérique de la BnF Gallica, 2 pétaoctets de données audiovisuelles et multimédia, 2 pétaoctets de données du dépôt légal du Web, ainsi que 18 millions de notices bibliographiques. Dans les années à venir, les collections numériques issues du dépôt légal numérique instauré en 2021 continueront d'alimenter ce réservoir.

Durant les années 2000 et 2010, les usages de l'IA à la BnF visaient surtout à répondre aux besoins internes de l'institution. Cette approche a évolué vers **un modèle de production et de partage des données**. Les premières réflexions sur l'élaboration de la Feuille de route IA de la BnF ont démarré en 2019, pour aboutir à une publication en 2021. Cette Feuille de route identifie cinq axes prioritaires, dont le lancement d'un programme pluriannuel de projets IA. Les projets menés depuis concernent entre autres l'extraction d'informations à partir des collections numériques[13], la transcription automatisée du texte (OCR et HTR) et l'aide à l'indexation et à la valorisation des collections.

[12] https://www.fairlytrained.org/

[13] Par exemple, le projet DALGOCOL se concentre sur la fouille de données et l'élaboration d'algorithmes pour prédire l'état des collections, en croisant diverses informations de conservation-restauration.

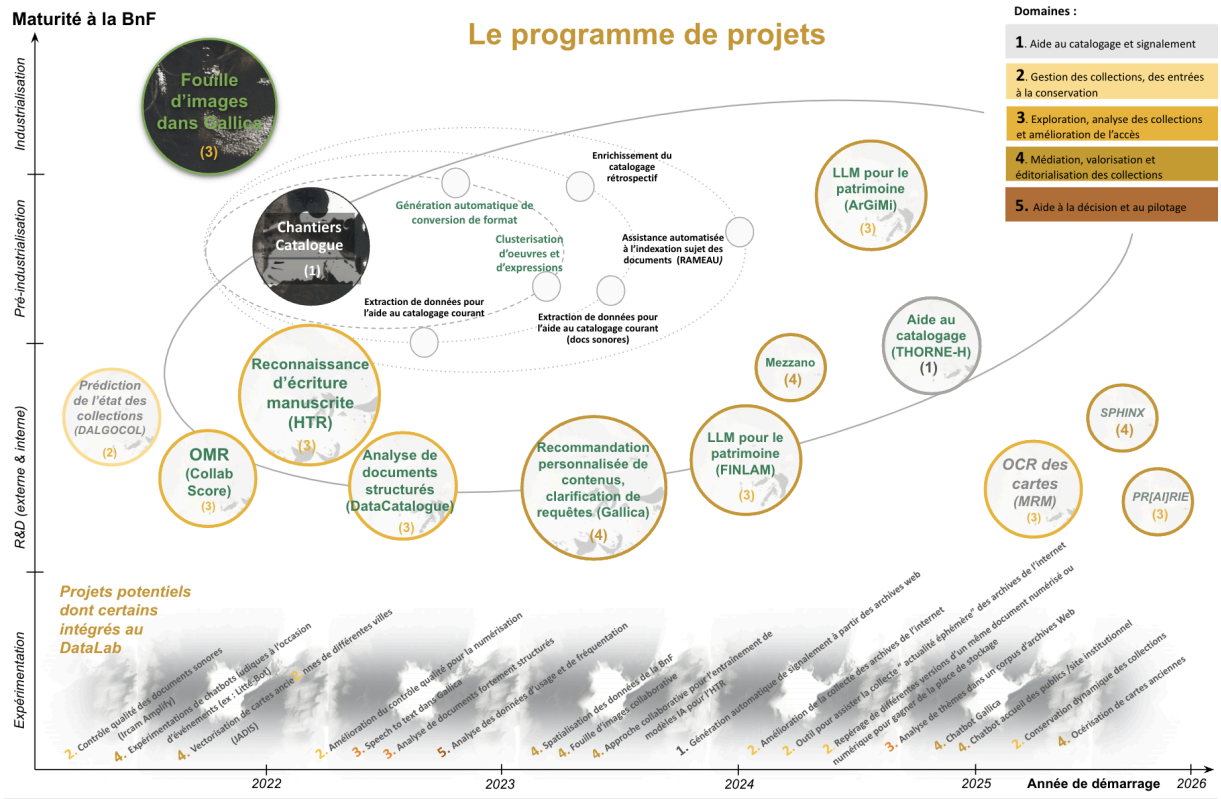


Figure 5 : Projets du programme pluriannuel dans la Feuille de route sur l'IA 2021-2026 de la BnF. Source : BnF.

Le BnF Datalab[14] est un volet important de l'action de la BnF dans le domaine. Mis en place 2021, il s'agit d'un service opéré en partenariat avec la très grande infrastructure de recherche Huma-Num[15] et dédié aux chercheurs souhaitant exploiter les collections numériques de la BnF. Ses activités témoignent de l'essor de l'usage de l'IA dans la communauté des humanités numériques.

L'année 2022 a constitué un tournant, avec la croissance des LLM et des services d'IA conversationnelle. **Les données de la BnF ont suscité un intérêt croissant de la part du secteur industriel de l'IA.** Ces acteurs industriels ont déployé des robots pour aspirer divers contenus sur le portail Gallica. Certains d'entre eux se sont manifestés pour engager un dialogue au sujet de la fourniture de données textuelles, mais aussi sonores et visuelles.

L'un des défis auquel la BnF est confrontée est donc l'impact des usages intensifs de ses données sur ses infrastructures. De ce point de vue, Jean-Philippe Moreux a mis en lumière **les limites du paradigme "Collection as data"**[16]**.** En 2023, la BnF a

[14] https://www.bnf.fr/fr/bnf-datalab
[15] https://www.huma-num.fr/
[16] https://collectionsasdata.github.io/

constaté que l'utilisation de son API Gallica ouverte et les pratiques de scraping dépassaient les capacités de l'infrastructure, impactant la qualité de service de ses applications web. Pour remédier à cette situation, il a par exemple été nécessaire de limiter le débit autorisé des API et d'intégrer un gestionnaire d'API afin de réguler le flux des usages.

D'autres institutions patrimoniales ont refusé de voir leurs données se faire aspirer par des robots. C'est le cas de la bibliothèque nationale des Pays-Bas qui a publié en janvier 2024 la déclaration suivante :

> *La KB (Koninklijke Bibliotheek) ne souhaite pas que des entreprises commerciales utilisent ses ressources numériques sans autorisation pour entraîner des intelligences artificielles. Cela contrevient aux principes d'IA établis par la KB. En conséquence, la KB a pris des mesures pour restreindre cet usage.*[17]

Ce nouveau contexte a conduit à une mise à jour de la Feuille de route sur l'IA en 2023[18]. Les récentes sollicitations visant à utiliser les collections patrimoniales pour l'entraînement de modèles d'IA nécessitent de **repenser les modalités de collaboration avec les acteurs qui souhaitent exploiter les données de la BnF**. Un exemple emblématique est le projet "Communs numériques et IA générative" (2024), financé dans le cadre de l'appel à projets France 2030. Ce projet a pour objectif d'exploiter les données patrimoniales de Gallica, et plus spécifiquement celles du domaine public, pour entraîner des modèles en langue française, contribuant ainsi à la création de communs numériques francophones dans le domaine des grands modèles de langage (LLM).

De tels projets doivent s'inscrire dans le cadre réglementaire actuel, mais aussi contribuer à organiser la mise à disposition de données patrimoniales libres de droit dans le cadre de partenariats équitables. Dans cette perspective, la BnF contribue à la réflexion engagée par le Conseil supérieur de la propriété littéraire et artistique (CSPLA) relative à l'élaboration d'un modèle de rémunération pour les auteurs et éditeurs dont les données seraient utilisées par les entreprises de l'IA.

Hugging Face : un partenaire pour la mise en accès des jeux de données

Daniel van Strien est bibliothécaire spécialisé en apprentissage automatique chez Hugging Face, une entreprise dédiée à la démocratisation de l'apprentissage automatique et au développement d'outils open-source. Il travaille à l'amélioration

[17] Traduit de l'anglais. KB (2024) "KB restricts access to collections for training commercial AI", https://www.kb.nl/en/news/kb-restricts-access-to-collections-for-training-commercial-ai
[18] https://www.bnf.fr/fr/feuille-de-route-ia#bnf-feuille-de-route

du Hugging Face Hub, un dépôt central pour le partage de modèles et de jeux de données en apprentissage automatique. Ce Hub héberge **plus de 900 000 modèles, 200 000 jeux de données et 300 000 applications de démonstration**. Auparavant, D. van Strien a été conservateur numérique à la British Library dans le cadre du projet Living with Machines[19], une collaboration entre la British Library et l'Alan Turing Institute, axée sur l'application de la science des données et des méthodes d'apprentissage automatique aux collections historiques numérisées.

Hugging Face s'intéresse depuis plusieurs années à l'utilisation de corpus patrimoniaux pour l'apprentissage automatique. Un défi souligné par D. van Strien pour les institutions patrimoniales est la gestion de l'ouverture par défaut des collections. À ce jour, quelques bibliothèques utilisent Hugging Face pour documenter et partager les collections patrimoniales :

- **La Bibliothèque nationale de Norvège** a publié en 2020 le "Norwegian Colossal Corpus"[20], un jeu de données textuelles de 45 Go, accessible sur Hugging Face sous certaines conditions. Ce corpus, ainsi que plusieurs autres, est aussi disponible sur un site interne à l'institution via son AI-lab[21]. Le site de ce lab met également en accès des modèles entraînés par la bibliothèque sur ses propres collections numériques. La bibliothèque a également développé le projet "The Norwegian Language Bank"[22], une infrastructure nationale pour les technologies du langage, mettant à disposition un catalogue de ressources accessibles en ligne.
- **La British Library** met en accès sur le site de Hugging Face une série de jeux de données et de modèles[23]. Plusieurs de ces jeux de données sont issus d'un partenariat de numérisation de livres anciens mené avec Microsoft. Parmi les modèles publiés par la bibliothèque, l'un est conçu pour prédire si un livre de sa collection numérisée est une fiction ou une non-fiction, en se basant sur son titre.

D'autres jeux de données intégrant des ressources patrimoniales sont publiés par des tiers sur Hugging Face. C'est le cas du **projet PD12M[24], lancé fin 2024. Il s'agit d'un vaste jeu de données d'images ouvertes intégrant des légendes générées par intelligence artificielle.** Composé de 12,4 millions de paires image-légende, il s'agit du plus grand ensemble de données image-texte du domaine public à ce jour. Les images ont été collectées à partir de *dumps* et d'API de bibliothèques, de

[19] https://livingwithmachines.ac.uk/
[20] https://aclanthology.org/2022.lrec-1.410.pdf
[21] https://ai.nb.no/
[22] https://www.nb.no/sprakbanken/en/
[23] https://huggingface.co/TheBritishLibrary
[24] https://huggingface.co/datasets/Spawning/PD12M

musées, d'institutions patrimoniales, ainsi que de Wikimedia Commons. Toutefois, les métadonnées descriptives de chacune de ces images ne sont pas incluses dans le jeu de données.

D. van Strien souligne par ailleurs que Hugging Face s'efforce de soutenir la diversité linguistique et culturelle en collaborant avec des institutions patrimoniales pour valoriser les données locales. Ces efforts incluent un projet d'évaluation de modèles de langage en arabe ainsi que le **lancement de vastes jeux de données multilingues.** D. van Strien a également insisté sur la nécessité de responsabiliser les utilisateurs dans la création de leurs propres SIA, tout en offrant des outils pour faciliter l'interaction avec les communautés locales.

3- DIVERSITÉ LINGUISTIQUE ET CULTURELLE DANS LES SIA

Animé par Ayla Rigouts Terryn (professeure à l'Université de Montréal), ce panel a exploré les **enjeux et initiatives liés à la diversité culturelle et linguistique dans les SIA francophones.** Il a mis en lumière les enjeux spécifiques des langues sous-dotées dans la constitution de corpus d'entraînement de SIA. Les interventions de Lucie Gianola (Ministère de la Culture de France), Thomas Mboa (CEIMIA) et Stéphane Nepton (Printemps numérique) ont porté sur les initiatives inspirantes en la matière, mais aussi sur les freins techniques et politiques spécifiques aux langues sous-dotées, ainsi que sur les questions de gouvernance et d'équité dans le développement de SIA.

Les discussions ont souligné l'importance de mutualiser les efforts et de créer des dynamiques collectives, notamment en élaborant des **jeux de données regroupant plusieurs langues sous-dotées**. Cependant, la prédominance de l'écrit dans le développement des LLM entre en contradiction avec l'oralité caractéristique de nombreuses langues, en particulier dans les contextes autochtones et africains. Il est donc essentiel de **considérer l'oralité comme un vecteur clé pour la vitalité des langues peu documentées**.

DGLFLF : Technologies de la langue pour la diversité linguistique

Lucie Gianola, chargée de mission pour les technologies, la recherche et l'innovation au ministère de la Culture, au sein de la Délégation générale à la langue française et aux langues de France (DGLFLF). Elle a souligné l'importance de **replacer les technologies des systèmes d'intelligence artificielle (SIA) dans un cadre plus large englobant les technologies des langues**, et plus particulièrement le traitement automatique des langues (TAL). Dans son intervention, elle a présenté

diverses initiatives de création de technologies adaptées en faveur des langues régionales et minoritaires en France et dans le contexte européen.

L'Union européenne est un espace de grande diversité linguistique, avec 3 alphabets, 24 langues officielles et de nombreuses langues régionales. L'article 22 de la *Charte des droits fondamentaux de l'Union européenne* consacre ce principe en énonçant que "l'Union respecte la diversité culturelle, religieuse et linguistique"[25]. En France, on dénombre plus de 75 langues régionales, auxquelles s'ajoutent des langues non territoriales telles que le yiddish et la langue des signes française[26].

Dès 2015, la DGLFLF a **lancé une initiative d'accompagnement technologique pour les langues régionales de France**[27]. Ce projet avait favorisé le partage d'expertise et la diffusion d'exemples de pratiques inspirantes, dont celles menées par le milieu associatif breton et l'Office public de la langue bretonne. Ces échanges ont par la suite contribué au développement d'outils tels que des synthèses vocales et des traducteurs pour les langues régionales.

Une nouvelle initiative phare est en cours de développement à la DGLFLF. Il s'agit d'**un futur centre de référence dédié aux technologies de la langue, baptisé LANGU:IA**, qui sera installé à la Cité internationale de la langue française à Villers-Cotterêts. Ses activités incluront la recherche, l'accompagnement des projets industriels et institutionnels, la valorisation de données, la création et le partage de corpus ainsi que la médiation scientifique. En 2024, la préfiguration des activités de LANGU:IA, en partenariat avec l'Institut national de recherche en sciences et technologies du numérique (INRIA) et la Direction interministérielle du numérique (DINUM), a abouti à la création du dispositif Compar:IA[28], qui permet de tester deux IA anonymes en parallèle afin de comparer leurs réponses.

Les activités de LANGU:IA seront en étroite interaction avec **le projet européen Alliance pour les technologies des langues (ALT-EDIC)**, dont le pilotage a été confié à la France. Ce projet vise à soutenir les initiatives industrielles en faveur des technologies européennes des langues. L'offre de services d'ALT-EDIC inclut la mise à disposition de données et le développement de modèles de langue ouverts et souverains dans toutes les langues de l'Union européenne. L'ALT-EDIC soutiendra ainsi un écosystème d'IA respectueux du droit d'auteur et des droits voisins, et

[25] https://fra.europa.eu/fr/eu-charter/article/22-diversite-culturelle-religieuse-et-linguistique
[26] https://www.culture.gouv.fr/Thematiques/langue-francaise-et-langues-de-france/Agir-pour-les-langues/Promouvoir-les-langues-de-France
[27] https://www.culture.gouv.fr/content/download/136599/1465800
[28] https://www.comparia.beta.gouv.fr/

relèvera les défis des grands modèles de langue comme les biais culturels et linguistiques, la frugalité, l'explicabilité et la réutilisabilité des modèles.

Thomas Mboa : enjeux dans le contexte africain

Thomas Mboa est chercheur en résidence au Centre d'expertise international de Montréal en intelligence artificielle (CEIMIA). Son expertise porte sur l'inclusion et **la représentation de l'Afrique dans l'écosystème international de l'intelligence artificielle**. Dans sa présentation, il aborde la nature duale des technologies numériques, dont l'intelligence artificielle, perçue à la fois comme un poison et un remède au sens du pharmakon[29]. Il insiste sur la nécessité d'en maîtriser les dynamiques afin de **développer des solutions adaptées, répondant en priorité aux besoins locaux.** Pour ce faire, il souligne d'entrée de jeu l'importance de ne pas se focaliser uniquement sur les bénéfices de l'IA pour la société tout en minimisant ses effets négatifs, pourtant tout aussi significatifs. D'où le concept de technocolonialité qu'il introduit.

La technocolonialité se manifeste lorsque l'utilisation de la technologie induit un mode de pensée colonial visant, consciemment ou non, à exercer le pouvoir, le contrôle et la domination. Ce phénomène contribue souvent à la reproduction des schémas d'oppression hérités de la colonisation. Bien que la technocolonialité soit présente aux quatre coins du monde, y compris en Occident, **elle peut être exacerbée dans certains contextes, notamment en Afrique, en raison de son lourd passé colonial**. La technocolonialité se décline en quatre dimensions principales : le discours *techno-utopique*, la *colonialité des savoirs*, le *transfert des technologies* et les *pratiques néocapitalistes* (Mboa, 2020).

Dimensions de la technocolonialité	Description
Discours techno-utopique	Le discours techno-utopique sur l'IA promet une révolution aux retombées économiques considérables et affirme que les outils de l'IA pourront résoudre une multitude de problèmes. Ce discours techno-utopique masque souvent les inconvénients, les échecs et les dangers associés à ces technologies. Or, cela peut avoir des conséquences aliénantes pour les aspirations, car les promesses qui sous-tendent le discours techno-utopique sont généralement inspirées par des réalités étrangères.

[29] Dans son essai sur la *Pharmacie de Platon,* et dans un contexte plus récent, Derrida donne une interprétation moderne et philosophique de ce rituel ; il souligne l'ambiguïté du terme pharmakon qui peut signifier à la fois médicament et poison.

Colonialité des savoirs	Ce discours contribue à imposer l'histoire globale de l'Occident aux peuples non-occidentaux, reléguant les histoires locales et régionales au second plan. Dans le cas de l'IA, la colonialité du savoir est maintenue par le biais d'ensembles de données d'entraînement qui, dans la plupart des cas, sont centrés sur l'Occident. Les outils de reconnaissance vocale en sont un exemple : ils sont disponibles pour l'anglais, mais ne sont pas adaptés à de nombreuses autres langues.
Transfert des technologies	Cet aspect de la colonialité se manifeste dans le transfert de technologie du Nord vers le Sud. Cela se traduit notamment par l'adoption de formes de technologie structurellement similaires à celles de l'Occident, sans aucun effort de contextualisation.
Pratiques néocapitalistes	Derrière ce transfert de technologie se cachent des pratiques néo-capitalistes ancrées dans l'extractivisme et le colonialisme des données de certains géants occidentaux de la technologie qui exploitent souvent les données des pays en développement sans compensation adéquate.

Sous le prisme de la technocolonialité, **des préoccupations émergent quant au développement des systèmes d'IA, à la collecte de données et à la création de normes, qui se déroulent majoritairement dans le Nord global**, où se concentrent la plupart des institutions de recherche en IA et des grandes entreprises technologiques. Cette situation suscite l'inquiétude de nombreux acteurs dans les nations en développement, qui remettent en question l'imposition des « valeurs » occidentales dans les systèmes d'IA. Ils s'interrogent notamment sur la pertinence d'appliquer les conceptions occidentales de l'« équité » et de l'égalité de manière universelle, ainsi que sur la nécessité d'adapter ces principes aux réalités locales. Ils explorent également la possibilité d'intégrer des perspectives alternatives propres à ces pays afin de concevoir des systèmes d'IA éthiques, mieux ancrés dans leurs contextes socioculturels, tout en étant acceptés à l'échelle mondiale (USAID, 2023).

Les pays africains sont également confrontés à divers défis liés aux pratiques d'extractivisme numérique. Celles-ci se traduisent notamment par l'incitation des individus à partager leurs données sans contrepartie, par leur rôle de testeurs bêta sans implication dans le développement initial des outils d'IA, et par les effets pervers du *digital labor*, tels que l'exploitation des compétences pour des salaires dérisoires (Neema, 2022).

T. Mboa a identifié plusieurs enjeux d'IA propres à l'Afrique, qu'il a classés en quatre grandes catégories : techniques, éthiques et sociaux, économiques et politiques.

Catégorie	Sous-catégorie	Description
Enjeux techniques	**Fracture numérique**	L'accent croissant mis sur l'accès numérique aux collections et les fonctionnalités basées sur l'IA pourrait élargir le fossé entre les communautés ayant des niveaux d'accès technologique variés, conduisant à un accès inégal aux ressources culturelles.
	Dépendance envers les fournisseurs de technologie	La dépendance envers les fournisseurs technologiques externes pour les solutions d'IA.
	Manque de données	Les initiatives en IA dépendent de données diversifiées et de haute qualité, mais l'Afrique manque cruellement de données appropriées. Cette pénurie entrave la capacité des systèmes d'IA à fournir des réponses précises et pertinentes pour les populations africaines diverses.
Enjeux éthiques et sociaux	**Faible littératie en IA**	Des compétences théoriques et pratiques sont nécessaires pour le développement, la mise en œuvre et l'utilisation des applications d'IA dans divers secteurs à travers l'Afrique.
	Attitudes des utilisateurs	Les Africains sont souvent sceptiques quant à l'adoption et à l'utilisation de nouvelles technologies, en raison de l'influence de la culture et des normes sociales.
Enjeux économiques	**Manque de financement**	Les projets d'IA, en particulier ceux de grande envergure, nécessitent des investissements conséquents. L'accès limité aux financements et au capital-risque dans de nombreux pays africains rend difficile le développement de solutions basées sur l'IA pour les start-ups et les innovateurs.
Enjeux politiques	**Absence de politiques nationales sur l'IA**	Il existe un manque général de politiques pertinentes qui pourrait prioriser la conception et la mise en œuvre de l'IA, tout en abordant ses impacts potentiels sur la société.

Étant donné que l'IA se nourrit de données, qui véhiculent inévitablement des biais, le secteur de l'industrie culturelle, notamment celui des institutions GLAM (Galeries, Bibliothèques, Archives et Musées), n'échappe pas aux défis qu'elle pose en Afrique et ailleurs dans le monde. Toutefois, en tenant compte des avantages qu'offrent les applications de l'IA dans ce domaine, T. Mboa ne prône pas une attitude réfractaire à cette technologie, mais plutôt une adoption responsable par les organisations et les professionnelles et professionnelles. Pour ce faire, il propose **une démarche basée sur l'appropriation décolonisée de l'IA par les Africains** (Mboa, 2023).

Stéphane Nepton : la revitalisation des langues autochtones

Stéphane Nepton est coordonnateur de projet en médiation numérique pour les Premières Nations au sein de l'organisme Printemps numérique. Innu de Mashteuiatsh, il a présenté des pistes de réflexion sur **les usages des SIA pour valoriser les cultures autochtones dans les communautés des Premières Nations.** À travers le projet Uhu Lab, il a expérimenté diverses technologies comme la photogrammétrie et la réalité augmentée pour favoriser la transmission intergénérationnelle des savoirs autochtones.

S. Nepton a mis de l'avant le potentiel des **technologies de la langue pour intéresser les jeunes à parler la langue de leur communauté** et à l'utiliser de manière quotidienne. Cependant, il a souligné les défis majeurs liés à la création de modèles de langage dans des langues comme l'innu, marqués par **un manque de données disponibles pour l'entraînement.** Par ailleurs, la langue innue se décline en plusieurs dialectes, posant un défi supplémentaire de standardisation de la langue.

Un exemple inspirant mentionné dans cette présentation est le projet FLAIR[30], dirigé par le chercheur cheyenne Michael Running Wolf et développé au MILA. À partir d'outils de reconnaissance vocale, ce projet vise à réduire considérablement les besoins en données pour l'entraînement de modèles en langues autochtones, de façon à permettre la création rapide de modèles personnalisés pour les langues en danger. Enfin, S. Nepton a insisté sur **les principes d'autodétermination et de gouvernance locale** qui doivent guider le développement de modèles de langue des Premières Nations.

[30] https://mila.quebec/fr/ia-pour-lhumanite/projets-appliques/initiative-flair

SYNTHÈSE DES ATELIERS

1- PRODUCTION DES JEUX DE DONNÉES

La production de jeux de données est un enjeu stratégique pour le développement des systèmes d'intelligence artificielle (SIA). Dans le contexte actuel où une part croissante des contenus diffusés sur le Web provient de sources non vérifiées et sont eux-mêmes massivement générés par des SIA, **les données de qualité, vérifiables et gouvernées par des règles transparentes sont une ressource rare.** Les données issues de Wikimédia et des institutions GLAM (Galeries, Bibliothèques, Archives et Musées) jouissent d'une réputation de qualité, appuyée sur une expertise de curation de l'information, contrairement à d'autres sources de données (telles que Reddit) couramment utilisées par les grands modèles de langage (LLM). La valeur de leurs données est donc prisée par les acteurs de l'IA pour développer des SIA compétitifs.

Principes	Enjeux	Pistes d'actions
Utilisabilité	**Identifier et prioriser les données les plus pertinentes à produire sous la forme de jeux de données d'entraînement**	> Organiser des concours et événements locaux pour identifier les données les plus pertinentes stimuler la création de jeux de données
	Améliorer la qualité des données produites	> Renforcer la gouvernance des projets Wikimédia en consolidant les critères de sélection et d'évaluation des données, notamment au sein de Wikimedia Commons
	Améliorer la normalisation des données et leur interopérabilité	> Élaborer des normes communes et des standards de données, notamment à travers des fiches de métadonnées et des standards définis par consortiums thématiques, afin de se positionner en leader dans son domaine
		> Identifier les normes émergentes > Mettre en œuvre les normes telles ISO/IEC 42001 destinée aux organismes qui fournissent ou utilisent des systèmes d'IA
		> Identifier les meilleurs formats et standards pour constituer les jeux de données > Promouvoir l'encodage des données dans le format UTF-8

		> Renforcer et diversifier les formations axées sur la normalisation et la standardisation des jeux de données
Découvrabilité	**Désintermédiation - Les services d'IA conversationnelle, tels que ChatGPT, exploitent des ensembles de données sans attribuer leur provenance, ce qui nuit à la reconnaissance et à la réputation des producteurs de données. Cette situation constitue une menace pour leurs modèles de production des données à plusieurs niveaux :**	> Mesurer et faire valoir le volume d'usage des données par les industries de l'IA > Faire connaître l'offre en mettant en valeur les avantages : des données de qualité ; Montrer la valeur du travail de la communauté pour la production de user-generated content : mettre en valeur le nombre d'éditions récentes sur les articles (pratique de plus en plus répandue par exemple dans reddit tiktok google etc)
	- effets négatifs sur la motivation des professionnels et des contributeurs et contributrices bénévoles	> Imposer l'attribution des sources dans la présentation des résultats par les services d'IA
	- baisse de la fréquentation des sites web des producteurs de données	> Imposer un hyperlien vers le site de référence dans l'attribution des sources
	- diminution du renouvellement des contributeurs et contributrices bénévoles engagés dans la production des contenus	> Idem
Soutenabilité	**Envisager comment les outils d'IA générative peuvent soutenir le travail de production de contenus de qualité**	> Définir des critères d'utilisation de ces outils (pour Wikimédia, spécifier les critères d'admissibilité des contenus générés par des SIA) > Créer une boîte à outils pour accompagner l'utilisation de ces nouveaux outils
	Considérer la consommation énergétique des SIA	> Privilégier les approches frugales

	L'engouement pour l'IA, les grands modèles de langage et les services actuels peut occulter les autres usages structurants à long terme	> Diversifier les cas d'usage des données dans le domaine du traitement automatique des langues (TAL)

2- PLURALISATION LINGUISTIQUE ET CULTURELLE

La pluralisation désigne un processus de coexpression des différences (culturelles, linguistiques, sociales et politiques) au sein d'un même espace de données. **Ces capacités d'expression sont inégalement distribuées**, certaines étant freinées tandis que d'autres sont amplifiées par des inégalités structurelles. Ces freins concernent particulièrement les langues sous-dotées ainsi que les variantes géographiques et culturelles non-dominantes au sein d'une même langue. La pluralisation linguistique et culturelle des jeux de données contribue à adapter les SIA au contexte linguistique et culturel des utilisateurs et utilisatrices (schémas culturels, expressions locales, accents) ainsi qu'à étendre leurs coordonnées de raisonnement.

Principes	Enjeux	Pistes d'action
Équité	**Infrastructure de production de données :** défis d'accès aux télécommunications dans certains territoires des Premières Nations et des Suds (coupures d'électricité, faible vitesse et instabilité d'Internet, ou absence de réseau dans les territoires).	> Renforcer les infrastructures de données avec la participation des communautés locales. > Développer la formation d'experts locaux.
	Multilinguisme : risque que les langues et cultures moins dotées soient marginalisées, insuffisamment ou incorrectement représentées dans les outils d'intelligence artificielle, compromettant leur visibilité, leur intégrité et leur vitalité.	> Développer la qualité et quantité des données concernant les langues et cultures sous-dotées. > Produire des données de manière structurée et bien documentée. > Favoriser les projets menés par et pour les groupes concernés.
	Variations linguistiques intra-langue : les usages du français varient selon les territoires, posant la question de leur prise en compte dans les outils d'IA.	> Produire à des jeux de données bien documentés qui reflètent les usages linguistiques des différents territoires francophones et des Premières Nations.

	Les LLM maîtrisent relativement bien certaines langues, mais **le contenu qu'elles véhiculent reste souvent déconnecté des référents culturels spécifiques** des locuteurs, surtout dans les langues et cultures sous-dotées.	> Développer des modèles sensibles au contexte culturel de l'utilisateur.
	Méconnaissance des enjeux culturels liés à l'IA chez de nombreux acteurs, notamment dans les communautés autochtones et les territoires des Suds, limitant leur capacité à prendre des décisions éclairées en matière de gouvernance des données.	> Sensibiliser l'ensemble des acteurs, en particulier aux enjeux spécifiques pour les langues et cultures moins dotées.
Souveraineté	**Appropriation de certaines ressources ethnolinguistiques** exploitées sans le consentement des groupes concernés.	> Favoriser le développement de jeux de données par et pour les groupes concernés. > Pour les données liées aux Premières Nations, se référer aux lignes directrices de la CSSSPNQL[31].
	Identifier **les risques des outils de traduction automatique**, notamment leur tendance à imposer des structures linguistiques et conceptuelles exogènes, ce qui peut entraîner une rupture avec les modes de pensée et les cadres épistémiques locaux.	> Encourager les groupes concernés à mobiliser leurs propres connaissances et cadres culturels pour traduire des termes, en évitant de s'appuyer uniquement sur des solutions automatisées, sans réflexion critique.

[31] Commission de la santé et des services sociaux des Premières Nations du Québec et du Labrador.

	Identifier les risques liés à la divulgation ouverte de connaissances considérées comme rituelles, sacrées ou stratégiques par les Premières Nations.	> Concertation avec les groupes concernés. > Se référer aux lignes directrices de la CSSSPNQL.
	Les restrictions d'accès et d'usage des jeux de données peuvent nuire à la découvrabilité des contenus culturels dans les services d'IA.	> Évaluer les avantages et inconvénients de ces restrictions.

3- MISE EN ACCÈS DES JEUX DE DONNÉES

La mise en accès concerne le choix des lieux et des formes de publication des jeux de données de manière à les rendre disponibles pour le développement de SIA. Les sites web de Wikimédia et des GLAM font l'objet d'un moissonnage intensif par des robots d'indexation qui collectent leurs contenus sans consultation préalable. Certaines de ces données se retrouvent publiées par des tiers sur des plateformes externes, telles que Hugging Face, sans consultation préalable. Ces pratiques conduisent à **un phénomène de désintermédiation**, où les détenteurs de données ne sont plus en capacité de négocier la diffusion et l'usage de leurs ressources.

Principes	Enjeux	Pistes d'actions
Équité	Adapter la mise en accès des jeux de données **en fonction des profils d'usage.**	> Identifier, différencier et documenter les profils d'usagers (secteur, mission) et les types d'usage des données.
Utilisabilité	**Capacité limitée des infrastructures informatiques** à gérer un accès simultané à haut volume pour certains utilisateurs industriels, entraînant des contraintes de performance et de disponibilité du service.	> Mettre en place un service d'API spécifique pour les utilisateurs de haut volume de données.
	Préférences et besoins variés des utilisateurs concernant les formats de données dans un contexte technologique en constante évolution.	> Publier les données dans des formats multiples.
Découvrabilité	**Publication de jeux de données par des tierces parties (non autorisées ou autorisées)** et perte de maîtrise sur la qualité et l'intégrité des données diffusées et sur leurs métadonnées (incomplètes ou erronées).	> Réintermédiation : prendre l'initiative de publier ses propres jeux de données sur des plateformes clés afin d'assurer la qualité des données et des métadonnées, tout en se positionnant comme un interlocuteur de référence et de renforcer la visibilité et l'influence au dans le domaine.
	Identifier les meilleurs points de diffusion des jeux de données pour promouvoir les ressources tout en maîtrisant leur diffusion.	> Évaluer les avantages et inconvénients de publier les jeux données sur des plateformes internes ou des plateformes tierces (telles Hugging Face ou Huma-Num). > Identifier les plateformes tierces les plus pertinentes où diffuser les jeux de données ; analyser leurs fonctionnalités en termes de traçabilité des usages.

		> Créer une boîte à outil technologique et mutualiser les efforts.
Soutenabilité	**Non-respect des conditions d'utilisation** des jeux de données.	> Mettre en place des outils de détection pour tracer les usages des jeux de données par les producteurs de modèles (action de régulation internationale).

4- OUTILS JURIDIQUES

La mise en accès et l'usage des jeux de données soulèvent des enjeux complexes liés à la protection des droits d'auteur, des droits voisins et des données personnelles. L'adaptation des cadres législatifs est amorcée (AI Act dans l'Union européenne, Loi sur l'intelligence artificielle et les données — LIAD — en examen au Canada) mais leur mise en œuvre pratique soulève des défis. Comment, par exemple, appliquer le mécanisme d'*opt out* (opposition éventuelle à toute fouille de données) et l'obligation à fournir un résumé des données utilisées pour entraîner un modèle, tel que prévu par le droit européen[32] ? Parmi les chantiers de réflexion actuels, certains portent sur les nouvelles formes de licences, les exceptions pour la fouille de données et l'évolution du dépôt légal. Ces dynamiques reflètent un équilibre fragile entre la gouvernance collective des communs numériques et les initiatives des acteurs publics et privés visant à encadrer l'accès et l'utilisation des données.

Principes	Enjeux	Pistes d'actions
Souveraineté	**Méconnaissance des enjeux législatifs spécifiques à l'IA.** Littératie numérique des décideurs politiques insuffisante.	> Favoriser le dialogue entre les acteurs des communs, les acteurs publics, les acteurs industriels et les gouvernements. > Mettre en place des actions de sensibilisation pour les décideurs politiques.
	Complexité et évolution rapide du cadre juridique entourant l'IA	> Faire de la veille législative sur le sujet.
	Droits émergents en lien avec les SIA et nouvelles applications des lois existantes (dont les droits de la personnalité).	> Mener une réflexion approfondie sur l'exploitation des données incluant les attributs personnels des wikimédien.ne.s, tels que la personnalité sociale exprimée à travers les échanges sur les pages de discussion et les voix enregistrées via Lingua Libre, qui peuvent être synthétisées par des systèmes d'intelligence artificielle.
	Fenêtre d'opportunité restreinte pour adapter et améliorer la législation.	> Travailler en collaboration à l'échelle internationale à travers des groupes de concertation. > Appuyer les actions de think tanks et de coalitions autour des communs du numérique. > Mener des consultations publiques.

[32] À ce sujet, voir le *Rapport de mission relative à la mise en œuvre du règlement européen établissant des règles harmonisées sur l'intelligence artificielle* (CSPLA, 11 décembre 2024).

Utilisabilité	**Ambiguïté** de l'application des outils juridiques existants dans le cadre de l'entraînement des SIA.	> Faire de la veille législative sur le sujet. > Mener des consultations publiques.
Soutenabilité	**Pression croissante exercée par les groupes et individus détenteurs de droits sur les données** pour encadrer et réguler leur utilisation par les SIA.	> Consultations publiques

5- MODÈLES ÉCONOMIQUES

La demande croissante par les chercheurs et les industries de l'IA pour des jeux de données issus de Wikimédia et des GLAM pose la question de leurs formes de valorisation économique. Le risque réside dans une exploitation commerciale des ressources publiques et communes sans qu'une contrepartie équitable ne soit assurée aux communautés ou institutions ayant contribué à leur création. Dans le contexte de la gouvernance des communs et du service public, l'objectif de la valorisation économique n'est pas la maximisation des profits, mais plutôt **la contribution des utilisateurs qui génèrent de forts profits commerciaux à partir de l'exploitation de ces données pour soutenir et développer un service accessible à tous et toutes**.

Principes	Enjeux	Pistes d'actions
Équité	Comment **faire payer l'accès à une ressource libre** pour les grands acteurs commerciaux tout en conservant une version gratuite pour les autres usages ?	> Établir que chaque utilisateur des données contribue financièrement de manière proportionnelle à ses usages et aux bénéfices qu'il en retire, tout en prenant en considération leurs disparités de moyens.
		> Développer des licences spécifiques d'usage, établissant des contributions financières adaptées au profil des réutilisateurs (paiement proportionnel aux revenus générés par l'utilisation des données) et introduisant de nouvelles typologies de licences équitables.
Souveraineté	**À qui appartiennent les communs ?** Comment et par qui peuvent-ils être exploités dans le contexte des SIA ?	> S'appuyer sur l'histoire et les évolutions contemporaines de la pensée des communs et du savoir libre pour mieux comprendre les enjeux liés à la propriété, à la possession, ainsi qu'à l'ouverture et à l'accessibilité des données.
	Est-ce que la commercialisation des données est **alignée avec les valeurs du mouvement de libre partage du savoir** et des connaissances ?	> S'appuyer sur l'histoire des choix du mouvement Wikimédia (licences Creative Commons avec autorisation d'exploitation commerciale) pour mener la réflexion à ce sujet.

	Comment adopter une posture stratégique pour **devenir un interlocuteur clé des industries de l'IA ?**	> Interrompre le flux de données dans le cas d'usage abusif (ex. téléchargement massif perturbant le service normal) pour forcer les acteurs industriels à agir à visage découvert ; > Négocier des contrats commerciaux avec les grands acteurs industriels pour établir des obligations et des conditions d'utilisation (terms of service) garantissant un accès fluide et stable pour les usages intensifs des données ; > Interrompre de manière aléatoire le transfert de données si les conditions d'utilisation ne sont pas respectées (impact sur la fiabilité des services commerciaux des industries de l'IA).
Utilisabilité	Comment **établir la valeur commerciale** d'une ressource libre et gratuite ?	> Traduire la valeur des communs en des termes que les acteurs économiques peuvent comprendre. > Montrer la valeur ajoutée des jeux de données en intégrant des indicateurs de qualité fiables, des mises à jour régulières, une structuration rigoureuse, une disponibilité dans des formats diversifiés adaptés aux besoins des utilisateurs, des conditions d'utilisation claires, des services personnalisés répondant à des besoins spécifiques, ainsi qu'une infrastructure solide assurant un accès fluide et continu, y compris pour des usages intensifs.
Découvrabilité	Les services d'IA conversationnelle, tels que ChatGPT, exploitent des ensembles de données sans attribuer leur provenance, ce qui nuit à la fréquentation des sites web des producteurs de données et peut mener à une **baisse des dons.**	> Compenser la baisse des dons avec la signature de contrats commerciaux d'exploitation des données.

Soutenabilité	Comment faire en sorte que **les géants commerciaux soutiennent financièrement** l'écosystème des communs de manière durable ?	> Imposer aux utilisateurs intensifs de contribuer financièrement pour l'accès aux données, tout en préservant la gratuité des usages alignés avec les valeurs et les missions des producteurs de données. > Monétiser non pas les données elles-mêmes, mais l'infrastructure qui permet un usage intensif (haut volume, haut débit, accès en tout temps aux dernières mise à jour des données). Ce modèle repose sur la vente de l'accès au "tuyau" de transmission, où ce qui est commercialisé est la capacité technique liée à leur transfert et leur traitement, notamment via des API dédiées.
	Comment financer les projets **sans dépendre des sources de revenu commercial ?**	> Diversifier les sources de financement pour assurer une pérennité financière tout en limitant la part des revenus commerciaux dans les recettes globales (à 30% dans le cas de la Fondation Wikimédia) afin de préserver l'indépendance et les valeurs fondamentales des producteurs de données.
	Concentration des outils et services d'IA entre les mains de quelques grands acteurs commerciaux	> Encourager le développement de solutions libres et open source.

CONCLUSION

Les échanges et réflexions issus de cette journée d'étude ont permis de mettre en lumière des enjeux stratégiques cruciaux pour les communs de la connaissance face aux défis posés par le développement des SIA. Ces défis nécessitent des actions concertées et structurées autour de cinq axes prioritaires :

1. **Réintermédiation stratégique** - Afin de se positionner comme intermédiaire clé, il est essentiel de promouvoir une visibilité proactive des collectifs et organismes chargés de la production et de l'intendance des jeux de données en développant des services d'API et en étant actif sur des plateformes adaptées et fréquentées par les usagers des données. Une attention particulière doit être accordée à l'attribution des sources, à la traçabilité des données d'entraînement et à la négociation d'ententes ou de contrats commerciaux de manière à clarifier les conditions d'utilisation et à assurer une rétribution équitable pour l'usage de ces données.

2. **Documentation des usages** - Il est nécessaire de recenser et d'analyser les principaux cas d'utilisation des données issues de Wikimédia et des fonds patrimoniaux, ainsi que les profils des principaux usagers. Une telle démarche permettrait de concevoir des stratégies adaptées pour la production et la diffusion des jeux de données. Ces exemples d'application mettraient en lumière la contribution essentielle de ces données à l'entraînement des modèles d'intelligence artificielle et aux systèmes de réponse automatique générative (RAG).

3. **Évaluation des plateformes existantes pour la mise en accès des jeux de données** - Une analyse comparée des infrastructures actuelles dédiées à la mise en accès des données permettrait d'évaluer leur pertinence respective en matière de découvrabilité, de qualité de la documentation des jeux de données, de traçabilité des usages, de modes de gouvernance et de coûts d'investissement.

4. **Production d'un référentiel de pratiques** - La création d'un guide partagé des pratiques exemplaires contribuerait à harmoniser les approches et à renforcer les capacités des différents acteurs à produire et diffuser des données dans un objectif de service public et de bien commun.

5. **Plaidoyer et coalition entre les acteurs du libre et les GLAM** - Un partage de ressources et d'expertise entre ces acteurs permettrait de mieux défendre les principes de pluralisation culturelle et linguistique et d'équité dans les relations avec les acteurs industriels de l'intelligence artificielle.

Ces axes d’intervention constituent des leviers essentiels pour répondre aux enjeux de souveraineté, de découvrabilité et de soutenabilité des communs numériques. La poursuite de ces efforts requiert un engagement collectif et une vision commune, ancrée dans la valorisation des données comme ressource stratégique au service du bien commun.

RÉFÉRENCES

Adams Stan (2024) “AI for the people: How machines can help humans improve Wikipedia”, blogue de la Fondation Wikimédia.

Baack Stefan, et al. (2025) "Towards Best Practices for Open Datasets for LLM Training" arXiv preprint ; arXiv:2501.08365.

Beaudoin Louise, Duhaime Clément, Guèvremont Véronique et Patrick Taillon (2024) *La souveraineté culturelle du Québec à l’ère du numérique: rapport du comité-conseil sur la découvrabilité des contenus culturels*, Québec : Gouvernement du Québec.

Bensamoun Alexandra (2024) *Rapport de mission relative à la mise en œuvre du règlement européen établissant des règles harmonisées sur l’intelligence artificielle*. Paris : CSPLA / Ministère de la culture (France).

CEIMIA - Centre d’expertise international de Montréal en intelligence artificielle (2024). W*hite Paper: Responsibly Scaling AI in Africa*. CEIMIA’s Strategic Framework for Increasing Impact. https://doi.org/10.5281/zenodo.13743603

Commission de la santé et des services sociaux des Premières Nations du Québec et du Labrador (2024) *Éthique du numérique et de l’intelligence artificielle : répertoire de références*, Wendake : CSSSPNQL.

Conseil de l’innovation du Québec (2024) *Prêt pour l’IA. Répondre au défi du développement et du déploiement responsables de l’IA au Québec*. Québec: Ministère de l’Économie, de l’Innovation et de l’Énergie.

Ebberman Jenny (2024) “L’IA remodèle le savoir : risques et enjeux pour une démocratie éclairée”’, Wikimedia Suisse.

Frenzel Janna, McKelvey Fenwick et Bart Simon (2022) *Imagining an AI Commons Report*. Montréal : Machine Agencies, Milieux Institute for Arts, Culture and Technology, Concordia University.

Gerbet Rémy (2024) “Qu’est-ce qu’un commun numérique ?”, Wikiconvention francophone, Québec.

Grégoire Marie et Latifa Moftakir (2024) « Pour une IA au service du bien commun qui reflète la diversité linguistique et culturelle mondiale », *Le Soleil*.

GPAI (2023) *Designing Trustworthy Data Institutions: Scanning the Local Data Ecosystem in Climate-Induced Migration in Lake Chad Basin - Pilot Study in Cameroon*, Rapport de recherche, Global Partnership on AI.

Mboa Nkoudou Thomas Hervé (2023) “We need a decolonized appropriation of AI in Africa”, *Nature Human Behaviour*, 7(11).

Mboa Nkoudou Thomas Hervé (2020). *Les makerspaces en Afrique francophone, entre développement local durable et technocolonialité : Trois études de cas au Burkina Faso, au Cameroun et au Sénégal* [Doctorat]. Université Laval.

Ministère de la Culture - Délégation générale à la langue française et aux langues de France (2015), Les technologies pour les langues régionales de France, Actes du colloque. Paris : Ministère de la culture.

Ministère de la Culture (2024), *Politiques culturelles : la stratégie numérique du ministère de la Culture*, Paris : Ministère de la culture.

Neema Iyer, Garnett Achieng, Uri Ludger, & Favour Borokini. (2022). *Automated imperialism, expansionist dreams: Exploring Digital Extractivism in Africa*. Standford PACS. https://www.are.na/block/12479796

Öhman, Joey, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. "The Nordic Pile: A 1.2 Tb Nordic Dataset for Language Modeling." *arXiv preprint arXiv:2303.17183* (2023).

OpenLLM France (2024) *Manifeste*, github.com.

Organisation Internationale de la Francophonie (2024) *Déclaration de Villers-Cotterêts*, XIX[e] Sommet de la francophonie, Villers-Cotterêts et Paris.

Stasenko Anastasia et Pierre-Carl Langlais (2024) "Announcing Common Corpus", *Builders Mozilla*.

Tarkowski, Alek (2023) "How Wikipedia can shape the future of AI", Open Future Foundation.

Tarpin, Mathieu (2025) "Wikimedia, les bibliothèques, et l'entraînement des IA", *Bulletin des bibliothèques de France*.

United States International Development (2023) *Artificial Intelligence (AI), Ethics Guide*, Washington: USAID.

Wikimedia Enterprise (2024) "Wikipedia Hugging Face Dataset using Structured Contents Snapshot", blogue de Wikimédia Enterprise.