ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Sarah Marcq
licenciée ès histoire et histoire de l'art

DES USAGES ARCHIVISTIQUES POUR L'INTELLIGENCE ARTIFICIELLE

LE CAS DE L'AUTOMATISATION DE L'INVENTAIRE DES ARCHIVES À LA CHAMBRE DES DÉPUTÉS DU GRAND-DUCHÉ DE LUXEMBOURG

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2024

Résumé

Résumé en français : Ce mémoire est une prise de recul après un stage de quatre mois sur un projet d'automatisation en contexte archivistique luxembourgeois. Il examine la pertinence de l'intelligence artificielle (IA) comme moyen d'automatisation dans le domaine et comme solution aux défis rencontrés par les producteurs d'archives publiques au Luxembourg. Le contexte public et archivistique luxembourgeois présente des conditions favorables au lancement de projets d'IA. Les contributions potentielles de systèmes basés sur du *machine learning* sont multiples pour les services d'archives. Elles ne se limitent pas à l'automatisation de tâches métier. Les apports entre le domaine des archives et de l'intelligence artificielle peuvent être connexes. Les ambitions sont élevées mais le déploiement d'outils basés sur ces technologies reste complexe. Des problématiques d'ordre éthique sont à prendre en compte et de nombreux prérequis techniques sont à penser en amont.

Résumé en anglais : This thesis is a reflection following a four-month internship on an automation project in the Luxembourgish archival context. It examines the relevance of artificial intelligence (AI) as a means of automation in the field and as a solution to the challenges faced by public archive producers in Luxembourg. The public and archival context in Luxembourg provides favorable conditions for launching AI projects. The potential contributions of systems based on machine learning are numerous for archival services. They are not limited to the automation of business tasks. The contributions between the fields of archives and artificial intelligence can be interconnected. The ambitions are high, but the deployment of tools based on these technologies remains complex. Ethical issues must be considered, and numerous technical prerequisites need to be addressed in advance.

Mots-clés : intelligence artificielle ; machine learning ; archives numériques ; archivistique ; parlement.

Informations bibliographiques : Sarah Marcq, *Des usages archivistiques pour l'intelligence artificielle. Le cas de l'automatisation de l'inventaire des archives à la Chambre des Députés du Grand-Duché de Luxembourg*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Florian Cafiero, École nationale des chartes, 2024.

Remerciements

Je tiens tout d'abord à exprimer ma gratitude envers la Cellule Archives de la Chambre des Députés pour m'avoir offert l'opportunité de réaliser ce stage. Je remercie également toutes les personnes qui m'ont accueillie chaleureusement au cours de cette expérience. J'adresse en particulier mes remerciements à Amandine Gorse pour son encadrement attentif tout au long de mon stage ainsi que ses conseils sur la rédaction de ce mémoire. Merci à Jonathan Baud pour son aide dans la réalisation des schémas qui figurent dans ce mémoire, ainsi que pour nos discussions enrichissantes sur l'usage de l'intelligence artificielle dans les parlements. Je remercie François-Marie Giraud pour ses recommandations bibliographiques avisées et ses réponses à mes nombreuses questions. Un grand merci également à Michel Cottin et Camille Forget des Archives nationales, ainsi qu'à Christine Mayr de la Chambre des Députés, pour leur contribution lors de la relecture de ce mémoire, notamment pour leurs commentaires pertinents et recommandations sur le deuxième chapitre. Je remercie Florian Cafiero, mon directeur de mémoire, pour ses conseils et son accompagnement tout au long de ce travail.

Je tiens à exprimer ma reconnaissance envers ma famille pour son soutien et sa relecture attentive, et plus particulièrement à mon frère Hugo pour ses précieux éclairages concernant les systèmes d'information. Je remercie également mes amis pour leur appui. Un merci particulier à Manon pour nos après-midi de travail et nos pauses café prolongées qui m'ont été d'un grand réconfort, ainsi qu'à mes camarades de promotion, devenus des amis, pour leur soutien et nos discussions enrichissantes autour de nos stages et mémoires.

Bibliographie

Archives

Banat-Berger (Françoise), « La prise en charge des archives electroniques en France dans le secteur public », *Archives*, 40–1 (2008), p. 27-69, URL : https://www.archivistes.qc.ca/revuearchives/vol40_1/40_1_banat-berger.htm (visité le 30/08/2024).

— « « Un métier à part entière, l'archiviste un généraliste de l'information » : qu'en est-il en 2012 dans le nouvel environnement numérique ? », *La Gazette des archives*, 226–2 (2012), p. 117-126, DOI : 10.3406/gazar.2012.4901.

Baron (Jason R.) et Payne (Nathaniel), « Dark Archives and Edemocracy : Strategies for Overcoming Access Barriers to the Public Record Archives of the Future », dans *2017 Conference for E-Democracy and Open Government (CeDEM)*, Krems, Austria, 2017, p. 3-11, DOI : 10.1109/CeDEM.2017.27.

Béchard (Lorène), Fuentes Hashimoto (Lourdes) et Vasseur (Édouard), *Les archives électroniques*, 2e éd., enrichie et mise à jour, Paris, 2020 (Les petits guides des archives).

Comoy (Patrick), « Archives LGBTQI+ en France : de la « déplacardisation » à l'autonomie », *La Gazette des archives*, 255–3 (2019), p. 141-152, DOI : 10.3406/gazar.2019.5836.

Cristià (Elisenda), *Information management function at the Amsterdam City Archives*, en, janv. 2020, URL : http://arxivers.com/ladada/information-management-function-at-the-amsterdam-city-archives/ (visité le 14/08/2024).

Forastier (Sylvie), « Archiviste : un métier protéiforme ? », *La Gazette des archives*, 240–4 (2015), p. 305-311, DOI : 10.3406/gazar.2015.5310.

Grailles (Bénédicte) et Ducol (Laurent), « Les enjeux de la normalisation dans les services d'archives », *La Gazette des archives*, 228–4 (2012), p. 9-22, DOI : 10.3406/gazar.2012.4980.

Guyon (Céline), « Une archivistique sous influence », *Revue d'histoire culturelle. XVIIIe-XXIe siècles*–5 (oct. 2022), DOI : 10.4000/rhc.3466.

— « Théorie, technique et pratique archivistique en environnement numérique », dans *Colloque International sur le Document Electronique : Document et archivage : pra-*

tiques formelles et informelles dans les organisations, Grenoble, France, 2023, URL : https://hal.science/hal-04371177 (visité le 04/08/2024).

LAPPIN (James), *The science of recordkeeping systems - a realist perspective*, en, thèse, Loughborough University, 2024.

MOTTE (Alice), « La normalisation de la description archivistique : enjeux et actualités », *La Gazette des archives*, 238–2 (2015), p. 121-128, DOI : 10.3406/gazar.2015.5262.

NOUGARET (Christine), « Vers une normalisation internationale de la description des archives. La norme ISAD(G) du Conseil international des archives », *La Gazette des archives*, 169–1 (1995), p. 274-292, DOI : 10.3406/gazar.1995.3353.

PITTI (Daniel), « Encoded Archival Description : The Development of an Encoding Standard for Archival Finding Aids », dir. Jackie Dooley, *The American Archivist*, 60–3 (juill. 1997), p. 268-283, DOI : 10.17723/aarc.60.3.f5102tt644q123lx.

POUPEAU (Gautier), « Le « lac de données », une infrastructure technique pour déployer la gouvernance des données à l'Ina », dans *Les nouveaux paradigmes de l'archive*, dir. Claire Scopsi, Clothilde Roullier, Martine Sin Blima-Barru et Édouard Vasseur, Pierrefitte-sur-Seine, 2024 (Actes), URL : https://books.openedition.org/pan/7253 (visité le 22/07/2024).

RAJOTTE (David), « La réflexion archivistique à l'ère du document numérique : un bilan historique », *Archives*, 42–2 (2010), p. 69-105.

Records in Contexts (11) : versie 1.0 gelanceerd !, nl, 19 mars 2024, URL : https://www.amsterdam.nl/stadsarchief/organisatie/blog-bronnen-bytes/records-contexts-(11)-versie-1-0/ (visité le 14/08/2024).

« Scarcity or Abundance ? Preserving the Past in a Digital Era », *The American Historical Review* (, juin 2003), DOI : 10.1086/ahr/108.3.735.

WIERINGA (Maranke), « The Fragility of Digital Media Content : On Preservation and Loss : Sketching the Pilgrimage of Future Scholars to Recover Our Digital Vellum », *Junctions : Graduate Journal of the Humanities*, 2–2 (sept. 2017), p. 27, DOI : 10.33391/jgjh.33.

Intelligence artificielle - généralités

AGHION (Philippe), ANTONIN (Céline) et BUNEL (Simon), « Intelligence artificielle, croissance et emploi : le rôle des politiques », *Economie et Statistique / Economics and Statistics*–510-511-512 (2019), p. 153, DOI : 10.24187/ecostat.2019.510t.1994.

BURDEN (John), CLARKE (Sam) et WHITTLESTONE (Jess), « 9. From Turing's Speculations to an Academic Discipline : A History of AI Existential Safety », dans *The Era of Global Risk*, dir. Sj Beard, Martin Rees, Catherine Richards et Clarissa Rios Rojas, 1re éd., Cambridge, UK, 2023, p. 201-236, DOI : 10.11647/OBP.0336.09.

Crawford (Kate), *Contre-atlas de l'intelligence artificielle : les coûts politiques, sociaux et environnementaux de l'IA*, trad. par Laurent Bury, 18 cm. Bibliogr. et webliogr. p. 311-362. Index., Paris Veules-les-roses, 2023 (Z a).

Duca (Mihail), *Artificial intelligence. Myth and reality ? | Annals of Philosophy, Social &amp ; Human Disciplines | EBSCOhost*, fr, janv. 2023, URL : https://openurl.ebsco.com/contentitem/gcd:173354185?sid=ebsco:plink:crawler&id=ebsco:gcd:173354185 (visité le 13/08/2024).

Elish (M.C.), *Don't Call AI Magic*, en, janv. 2018, URL : https://medium.com/datasociety-points/dont-call-ai-magic-142da16db408 (visité le 13/08/2024).

Guillory (Thomas), Tilmant (Cyprien), Trécourt (Alexis) et Gaillot-Durand (Lucie), « Impacts environnementaux du numérique et de l'intelligence artificielle, à l'heure de la pathologie digitale », *Annales de Pathologie* (, juin 2024), DOI : 10.1016/j.annpat.2024.05.006.

Piorkowski (David), Park (Soya), Wang (April Yi), Wang (Dakuo), Muller (Michael) et Portnoy (Felix), « How AI Developers Overcome Communication Challenges in a Multidisciplinary Team : A Case Study », *Proc. ACM Hum.-Comput. Interact.* 5–CSCW1 (avr. 2021), 131 :1-131 :25, DOI : 10.1145/3449205.

Poussing (Nicolas), « Résultats de la consultation publique relative aux opportunités et aux défis de l'Intelligence Artificielle (IA). » Dans 2021, URL : https://liser.elsevierpure.com/fr/publications/r%C3%A9sultats-de-la-consultation-publique-relative-aux-opportunit%C3%A9s-e (visité le 03/08/2024).

Smuha (Nathalie A.), « From a 'race to AI' to a 'race to AI regulation' : regulatory competition for artificial intelligence », *Law, Innovation and Technology*, 13–1 (janv. 2021), p. 57-84, DOI : 10.1080/17579961.2021.1898300.

Thibout (Charles), « L'intelligence artificielle, une géopolitique des fantasmes », *Etudes digitales*, Religiosité technologique 5–1 (août 2019), DOI : 10.15122/isbn.978-2-406-09290-2.p.0105.

Vaswani (Ashish), Shazeer (Noam), Parmar (Niki), Uszkoreit (Jakob), Jones (Llion), Gomez (Aidan N.), Kaiser (Lukasz) et Polosukhin (Illia), *Attention Is All You Need*, arXiv :1706.03762 [cs], août 2023, URL : http://arxiv.org/abs/1706.03762 (visité le 16/08/2024).

LLM (Large language models)

Bai (Yuntao), Kadavath (Saurav), Kundu (Sandipan), Askell (Amanda), Kernion (Jackson), Jones (Andy), Chen (Anna), Goldie (Anna), Mirhoseini (Azalia), McKinnon (Cameron), *et al.*, *Constitutional AI : Harmlessness from AI Feedback*, arXiv :2212.08073 [cs], déc. 2022, DOI : 10.48550/arXiv.2212.08073.

Bansal (Parikshit) et Sharma (Amit), *Large Language Models as Annotators : Enhancing Generalization of NLP Models at Minimal Cost*, juin 2023, url : https://arxiv.org/abs/2306.15766 (visité le 10/08/2024).

Blog ML Explained, *Large Language Model (LLM) Evaluation Metrics – BLEU and ROUGE*, en, juill. 2023, url : https://mlexplained.blog/2023/07/08/large-language-model-llm-evaluation-metrics-bleu-and-rouge/ (visité le 30/08/2024).

Huggingface, *DistilBERT*, url : https://huggingface.co/docs/transformers/model_doc/distilbert (visité le 12/08/2024).

Lewis (Patrick), Perez (Ethan), Piktus (Aleksandra), Petroni (Fabio), Karpukhin (Vladimir), Goyal (Naman), Küttler (Heinrich), Lewis (Mike), Yih (Wen-tau), Rocktäschel (Tim), *et al.*, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, en, mai 2020, url : https://arxiv.org/abs/2005.11401v4 (visité le 30/08/2024).

Meta, *Introducing Meta Llama 3 : The most capable openly available LLM to date*, 18 avr. 2024, url : https://ai.meta.com/blog/meta-llama-3/ (visité le 02/08/2024).

Priyanshu (Aman), Maurya (Yash) et Hong (Zuofei), *AI Governance and Accountability : An Analysis of Anthropic's Claude*, arXiv :2407.01557 [cs], mai 2024, doi : 10.48550/arXiv.2407.01557.

Staab (Robin), Vero (Mark), Balunović (Mislav) et Vechev (Martin), *Large Language Models are Advanced Anonymizers*, en, févr. 2024, url : https://arxiv.org/abs/2402.13846 (visité le 31/08/2024).

Torene (Spencer), *Do LLMs Reason ?*, en, 9 oct. 2023, url : https://medium.com/@spencertorene/do-llms-reason-d33fa885872f (visité le 10/08/2024).

Wei (Jason), Bosma (Maarten), Zhao (Vincent Y.), Guu (Kelvin), Yu (Adams Wei), Lester (Brian), Du (Nan), Dai (Andrew M.) et Le (Quoc V.), *Finetuned Language Models Are Zero-Shot Learners*, arXiv :2109.01652 [cs], févr. 2022, doi : 10.48550/arXiv.2109.01652.

Yang (Tianyu), Zhu (Xiaodan) et Gurevych (Iryna), *Robust Utility-Preserving Text Anonymization Based on Large Language Models*, arXiv :2407.11770 [cs], juill. 2024, doi : 10.48550/arXiv.2407.11770.

Intelligence artificielle dans le secteur public

Bertolucci (Marius), « L'intelligence artificielle dans le secteur public : revue de la littérature et programme de recherche », *Gestion et management public*, Pub. anticipées–5 (2024), p. 118-139, doi : 10.3917/gmp.pr1.0008.

Commission européenne, *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Com-*

mittee, and the Committee of the Regions. Artificial Intelligence for Europe, en, 2018, URL : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN (visité le 23/07/2024).

IPU, « Expert perspectives on AI in parliament », *Innovation tracker*, Issue 16 (2023), URL : https://www.ipu.org/innovation-tracker/story/expert-perspectives-ai-in-parliament (visité le 28/07/2024).

Le Conseil d'État, *Intelligence artificielle et action publique : construire la confiance, servir la performance*, fr, août 2022, URL : https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance (visité le 02/08/2024).

Le gouvernement du Grand-Duché de Luxembourg, *Intelligence artificielle : une vision stratégique pour le Luxembourg*, fr, text, mai 2019, URL : http://gouvernement.lu/fr/publications/rapport-etude-analyse/minist-digitalisation/artificial-intelligence/artificial-intelligence/intelligence-artificielle.html (visité le 30/08/2024).

Meneceur (Yannick), « Les trois grands défis posés par la gouvernance de l'intelligence artificielle et de la transformation numérique », *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, 23 (déc. 2021), DOI : 10.4000/ethiquepublique.6323.

Mergel (Ines), Dickinson (Helen), Stenvall (Jari) et Gasco (Mila), « Implementing AI in the public sector », *Public Management Review* (, juill. 2023), p. 1-14, DOI : 10.1080/14719037.2023.2231950.

Ministère de la Digitalisation, *L'initiative AI4Gov*, fr, text, mars 2021, URL : http://mindigital.gouvernement.lu/fr/dossiers/2021/AI4Gov.html (visité le 24/07/2024).

Portail public du Grand-Duché de Luxembourg, *Meluxina, le superordinateur du Luxembourg*, fr, mars 2023, URL : http://luxembourg.public.lu/fr/investir/innovation/meluxina-superordinateur.html (visité le 30/08/2024).

Intelligence artificielle dans les archives, bibliothèques et la recherche en SHS

Alquier (Eleonore), « L'intelligence artificielle à l'INA : de l'expérimentation à l'industrialisation », dir. Association des archivistes français, *Archivistes !*–147 (2024).

Andresen (Herbjørn), « A discussion frame for explaining records that are based on algorithmic output », *Records Management Journal*, 30–2 (nov. 2019), p. 129-141, DOI : 10.1108/RMJ-04-2019-0019.

BNL, *Eluxemburgensia.lu s'est doté d'un nouveau chatbot*, fr, oct. 2023, URL : http://bnl.public.lu/fr/a-la-une/actualites/communiques/2023/chatbot-eluxemburgensia.html (visité le 10/07/2024).

Bunn (Jenny), « Working in contexts for which transparency is important : A recordkeeping view of explainable artificial intelligence (XAI) », *Records Management Journal*, 30–2 (1er janv. 2020), p. 143-153, DOI : 10.1108/RMJ-08-2019-0038.

Chan (Peter), Phillips (Mark E.), Cebra (Jessica) et Jacobs (James), *Leveraging ChatGPT for Efficient Metadata Creation of Government Reports*, en, [Note : Cet article n'a pas encore été publié au moment de la rédaction de ce mémoire], 2024.

Chow (Eric H. C.), Kao (T. J.) et Li (Xiaoli), *An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations*, arXiv :2403.16424 [cs], juill. 2024, DOI : 10.48550/arXiv.2403.16424.

Clavaud (Florence), Romary (Laurent), Charbonnier (Pauline), Terriel (Lucas), Piraino (Gaetano) et Verdese (Vincent), « NER4Archives (named entity recognition for archives) : Conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD. » Dans *Atelier Culture-INRIA*, Pierrefitte sur Seine, France, 2022, URL : https://hal.science/hal-03625734 (visité le 01/07/2024).

Cottin (Michel), Forget (Camille) et Gaudier (Richard), « Traitement des vracs bureautiques et IA : un premier pas dans la porte », dir. Association des archivistes français, *Archivistes !*–147 (2024).

Durukan (Gürcan), Nar (Meryem Tuğba), Özcan (Abdullah), Çakil (Lütfü) et Kara (Hüseyin), « Multimodal Classification Algorithm for Turkish Document Archiving : Improving Digital Document Storage by Unifying Image and Text-Based Classifiers », dans *Innovative Methods in Computer Science and Computational Applications in the Era of Industry 5.0*, dir. D. Jude Hemanth, Utku Kose, Bogdan Patrut et Mevlut Ersoy, Cham, 2024, p. 1-12, DOI : 10.1007/978-3-031-56322-5_1.

Gayer (A.), Ershova (D.) et Arlazarov (V.), « Fast and Accurate Deep Learning Model for Stamps Detection for Embedded Devices », *Pattern Recognition and Image Analysis*, 32–4 (déc. 2022), p. 772-779, DOI : 10.1134/S1054661822040046.

Gesnouin (Joseph), Tannier (Yannis), Da Silva (Christophe Gomes), Tapory (Hatim), Brier (Camille), Simon (Hugo), Rozenberg (Raphael), Woehrel (Hermann), Yakaabi (Mehdi El), Binder (Thomas), *et al.*, *LLaMandement : Large Language Models for Summarization of French Legislative Proposals*, arXiv :2401.16182 [cs], janv. 2024, DOI : 10.48550/arXiv.2401.16182.

Girard-Chanudet (Camille), « Le travail de l'Intelligence Artificielle : concevoir et entraîner un outil de pseudonymisation automatique à la Cour de Cassation », *RESET. Recherches en sciences sociales sur Internet*–12 (mars 2023), DOI : 10.4000/reset.4731.

Grailles (Bénédicte), *Pêle-mél. Plate-forme d'exploration, de livraison et d'évaluation des méls. Rapport d'évaluation des usages*, 31 mars 2023.

Hamdi (Ahmed), Linhares Pontes (Elvys), Boros (Emanuela), Nguyen (Thi Tuyet Hai), Hackl (Günter), Moreno (Jose G.) et Doucet (Antoine), « A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers », dans *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2021 (SIGIR '21), p. 2328-2334, doi : 10.1145/3404835.3463255.

Jaillant (Lise) et Aske (Katherine), « Are Users of Digital Archives Ready for the AI Era ? Obstacles to the Application of Computational Research Methods and New Opportunities », *J. Comput. Cult. Herit.* 16–4 (janv. 2024), 87 :1-87 :16, doi : 10.1145/3631125.

Jaillant (Lise) et Rees (Arran), « Applying AI to digital archives : trust, collaboration and shared professional ethics », *Digital Scholarship in the Humanities*, 38–2 (mai 2023), p. 571-585, doi : 10.1093/llc/fqac073.

Jost (Clémence), *Comment la BNL a développé son chatbot basé sur ChatGPT*, fr, 16 févr. 2024, url : https://www.archimag.com/bibliotheque-edition/2024/02/16/comment-bnl-developpe-son-chatbot-base-sur-chatgpt (visité le 10/07/2024).

Kimaid (Luís), *Artificial Intelligence-Driven Archibot : Transforming Access to European Union Parliament Archives*, en, url : https://library.bussola-tech.co/p/artificial-intelligence-archibot-eu-parliament (visité le 05/08/2024).

Lee (Christopher A.), « Computer-Assisted Appraisal and Selection of Archival Materials », dans *2018 IEEE International Conference on Big Data (Big Data)*, 2018, p. 2721-2724, doi : 10.1109/BigData.2018.8622267.

Meeks (Elijah) et Weingart (Scott B.), « The Digital Humanities Contribution to Topic Modeling », *Journal of Digital Humanities*–1 (2012), url : https://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/ (visité le 17/07/2024).

Poibeau (Thierry), « Le traitement automatique des langues pour les sciences sociales », *Réseaux*, 188–6 (2014), p. 25-51, doi : 10.3917/res.188.0025.

Souza (Renato Rocha), Coelho (Flavio Codeco), Shah (Rohan) et Connelly (Matthew), *Using Artificial Intelligence to Identify State Secrets*, 1er nov. 2016, doi : 10.48550/arXiv.1611.00356, arXiv : 1611.00356[cs].

Tarekegn (Adane Nega), *Large Language Model Enhanced Clustering for News Event Detection*, arXiv :2406.10552 [cs], juill. 2024, doi : 10.48550/arXiv.2406.10552.

Van Hooland (Seth) et Coeckelbergs (Mathias), « Unsupervised machine learning for archival collections : Possibilities and limits of topic modeling and word embedding », *Revista catalana d'arxivística*, 41 (2018), p. 73, url : https://arxivers.com/wp-content/uploads/2018/10/1.4_-Dossier_SVHooland_MCoeckelbergs.pdf (visité le 06/07/2024).

Velonis (Adrian), « Topic Modeling, Named-Entity Recognition, and Network Analysis of Literary Corpora » (, 2022), url : http://hdl.handle.net/10066/24507 (visité le 15/07/2024).

Wagh (Vedangi), Khandve (Snehal), Joshi (Isha), Wani (Apurva), Kale (Geetanjali) et Joshi (Raviraj), « Comparative Study of Long Document Classification », dans *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, arXiv :2111.00702 [cs], 2021, p. 732-737, doi : 10.1109/TENCON54134.2021.9707465.

Zhang (Shitou), Hou (Jingrui), Peng (Siyuan), Li (Zuchao), Hu (Qibiao) et Wang (Ping), *ArcGPT : A Large Language Model Tailored for Real-world Archival Applications*, arXiv :2307.14852 [cs], juill. 2023, doi : 10.48550/arXiv.2307.14852.

— « Archives Meet GPT : A Pilot Study on Enhancing Archival Workflows with Large Language Models », *iConference 2024 Proceedings* (, mars 2024), url : https://hdl.handle.net/2142/122806 (visité le 30/08/2024).

Informatique - généralités

Bygstad (Bendik), Hanseth (Ole) et Le (Dan Truong), « From IT Silos to Integrated Solutions. A Study in E-Health Complexity », *ECIS 2015 Completed Research Papers* (, mai 2015), doi : 10.18151/7217283.

Depaz (Pierre), *The role of aesthetics in understanding source code*, These de doctorat, Paris 3, 2023, url : https://theses.fr/2023PA030084 (visité le 12/08/2024).

Pucheu (David), « Effacer l'interface : Une trajectoire du design de l'interaction hommemachine », *Interfaces numériques*, 5–2 (mai 2018), p. 257-276, doi : 10.25965/interfaces-numeriques.3044.

Gestion de projet

Beck (K.), Beedle (M.), Bennekum (V. A.), Cockburn (A.), Cunningham (W.), Fowler (M.), Grenning (J.), Highsmith (J.), Hunt (A.), Jeffries (R.), *et al.*, *Manifesto for Agile Software Development*, 2001, url : https://agilemanifesto.org/ (visité le 28/07/2024).

Cockburn (Alistair), *Writing effective use cases*, Boston, 2001 (The Crystal series for software development).

Garel (Gilles), « Pour une histoire de la gestion de projet », *Annales des mines*, Gérer et comprendre–74 (déc. 2003).

Grebe (Michael), Franke (Marc Roman) et Heinzl (Armin), « Artificial intelligence : how leading companies define use cases, scale-up utilization, and realize value », *Informatik Spektrum*, 46–4 (août 2023), p. 197-209, doi : 10.1007/s00287-023-01548-6.

Shein (Cyndi), Robinson (Hannah E.) et Gutierrez (Hana), « Agility in the Archives : Translating Agile Methods to Archival Project Management », *RBM : A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 19–2 (nov. 2018), p. 94, doi : 10.5860/rbm.19.2.94.

Westenberger (Jens), Schuler (Kajetan) et Schlegel (Dennis), « Failure of AI projects : understanding the critical factors », *Procedia Computer Science*, International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021 196 (janv. 2022), p. 69-76, doi : 10.1016/j.procs.2021.11.074.

Lois et réglementations

Chambre des Députés du Grand-Duché de Luxembourg, *Loi du 5 décembre 1958 ayant pour objet l'organisation de la Bibliothèque Nationale et des Archives de l'Etat*, 5 déc. 1958, url : https://www.stradalex.lu/fr/slu_src_publ_leg_mema/document/mema_1958A15511?access_token=a91bb2cd70dac12a291757375cb9de2f2520b197 (visité le 14/07/2024).

— *Loi du 28 décembre 1988 portant réorganisation des instituts culturels de l'Etat*, 28 déc. 1988, url : https://legilux.public.lu/eli/etat/leg/loi/1988/12/28/n1/jo (visité le 14/07/2024).

— *Loi du 25 juin 2004 portant réorganisation des instituts culturels de l'Etat*, 25 juin 2004, url : https://legilux.public.lu/eli/etat/leg/loi/2004/06/25/n7/jo (visité le 14/07/2024).

— *Loi du 25 juillet 2015 portant création du système de contrôle et de sanction automatisés et modification de la loi modifiée du 14 février 1955 concernant la réglementation de la circulation sur toutes les voies publiques*, 25 juill. 2015, url : https://legilux.public.lu/eli/etat/leg/loi/2015/07/25/n2/jo (visité le 14/07/2024).

— *Loi du 17 août 2018 relative à l'archivage*, 17 août 2018, url : https://legilux.public.lu/eli/etat/leg/loi/2018/08/17/a706/jo (visité le 14/07/2024).

— *La Chambre se dote d'une Charte sur l'intelligence artificielle | Chambre des députés du grand-duché de Luxembourg*, fr, 2024, url : https://www.chd.lu/fr/charteIA (visité le 01/08/2024).

ICA, *Code de déontologie de l'ICA*, fr, url : https://www.ica.org/fr/resource/code-de-deontologie-de-lica/ (visité le 12/08/2024).

Parlement européen, *Règlement (UE) 2016/679 du Parlement européen*, et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant

la directive 95/46/CE (règlement général sur la protection des données), Article 89, 4 mai 2016, URL : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A02016R0679-20160504 (visité le 13/07/2024).

PARLEMENT EUROPÉEN, *Loi sur l'IA de l'UE : première réglementation de l'intelligence artificielle*, fr, juin 2023, URL : https://www.europarl.europa.eu/topics/fr/article/20230601STO93804/loi-sur-l-ia-de-l-ue-premiere-reglementation-de-l-intelligence-artificielle (visité le 01/08/2024).

Introduction

L'intelligence artificielle attire actuellement un grand afflux d'investissements, à tel point que certains craignent l'émergence d'une bulle spéculative susceptible d'éclater. Nous n'en sommes néanmoins pas encore là : selon son bilan publié fin août 2024, les bénéfices de l'entreprise Nvidia, qui domine le marché des processeurs graphiques pour l'intelligence artificielle, auraient bondi de plus de 150% en un an. Les technologies d'IA promettent de stimuler la croissance en automatisant divers processus, mais il reste à voir si ces investissements porteront leurs fruits à long terme, en particulier dans des domaines spécifiques comme les archives.

C'est dans ce contexte que le projet *InventAIre* a vu le jour à la Chambre des Députés du Grand-Duché de Luxembourg, animé par la volonté d'automatiser le processus d'inventaire des fonds d'archives. Cet inventaire fournit une description de ces derniers et permet d'automatiser le calcul des délais de communicabilité. Il s'agit d'un modèle provenant des Archives nationales du Grand-Duché (ANLux), choisi par l'équipe dans une volonté d'harmonisation des outils de description au niveau national. Cet inventaire est un fichier *Excel* contenant 17 colonnes :

- Cote
- Localisation
- Identification de la série du plan de classement
- Code série du tableau de tri
- Titre
- Description
- Période de création : de
- Période de création : à
- Soumis au droit d'auteur
- Données à caractère personnel
- Acte d'état civil
- Acte notarié
- Atteinte aux relations extérieures, à la sécurité du Grand-Duché ou à l'ordre public
- Affaires portées devant les instances juridictionnelles, extrajudiciaires ou disciplinaires
- Prévention, recherche de faits punissables
- Données commerciales et industrielles
- Secret fiscal

Le projet InventAIre est un projet pilote dont l'objectif principal était le développement d'un prototype permettant d'automatiser la rédaction de cet inventaire. Il a vu le jour à la Chambre des Députés du Grand-Duché, organe législatif du Luxembourg, responsable de l'élaboration et de l'adoption des lois. Composée de soixante députés élus au suffrage universel pour une durée de cinq ans, elle joue un rôle central dans le processus législatif et la surveillance du gouvernement en vertu de la séparation des pouvoirs. Les

fonds les plus anciens conservés par l'administation datent d'après 1945. L'occupant a en effet transféré l'ensemble des fonds de la Chambre aux Archives de l'État en 1940. Avant ce transfert, les premiers documents conservés remontaient aux débuts de l'institution, fondée en 1848, moment où une nouvelle constitution fait du Luxembourg une monarchie constitutionnelle.

Le projet d'automatisation du remplissage de l'inventaire de ces archives s'est déroulé dans le cadre de notre stage de quatre mois à la Chambre. Il s'agissait d'évaluer la faisabilité de cette automatisation et de produire un prototype d'outil. Nous avons également produit une note méthodologique dont le but était d'expliquer les choix réalisés dans le cadre du projet, les difficultés rencontrées, et de proposer des recommandation en cas de suite du projet. Elle se trouve en annexe [1]. Ces réflexions ont alimenté la rédaction de ce mémoire, qui constitue une prise de recul problématisée sur le stage et les usages archivistiques potentiels de l'intelligence artificielle au Luxembourg. Par « usage », nous entendons les différentes manières dont les technologies peuvent être appliquées pour optimiser les processus archivistiques, qu'il s'agisse de gestion, conservation, communication, description ou recherche dans les archives. L'« intelligence artificielle » (IA), quant à elle, se réfère à des systèmes informatiques qui seraient capables de simuler des capacités cognitives humaines, comme l'apprentissage et la prise de décision, afin de traiter et analyser des données à grande échelle. Cependant, ce terme est parfois utilisé de manière floue pour désigner ce que l'on appelle plus précisément l'apprentissage machine ou *machine learning*. Dans ce mémoire, nous emploierons parfois les termes « intelligence artificielle » et « *machine learning* » de manière indistincte pour simplifier la compréhension, tout en étant consciente que le terme « intelligence artificielle » peut parfois manquer de précision.

Ces systèmes d'intelligence artificielle ne datent pas d'hier. Elle est théorisée par plusieurs penseurs tels qu'Alan Turing et les chercheurs Warren McCulloch et Walter Pitts, qui, dès les années 1940, posent les bases des réseaux de neurones. Le *Perceptron*, algorithme de classification développé par Frank Rosenblatt en 1958, marque un premier pas dans son développement technique, introduisant un premier modèle capable d'apprendre à partir de données. Cependant, les années 1970 et 1980 connaissent un premier « hiver » de l'IA, caractérisé par un manque de moyens techniques et un pessimisme croissant parmi les chercheurs face aux limites des technologies de l'époque. Ce déclin est suivi par une renaissance dans les années 1980 grâce aux systèmes experts, programmes informatiques conçus pour imiter le jugement et le comportement d'un expert humain dans des domaines spécifiques en utilisant des règles de décision et des bases de connaissances. Néanmoins, à partir de 1987, un second hiver de l'IA survient en raison des défis techniques persistants. Les années 1990 marquent un tournant avec des événements comme la défaite du cham-

1. N'étant pas public, le document a dû être retiré de la version diffusée du mémoire. S'adresser à la Chambre des Députés pour le consulter.

pion mondial d'échecs Garry Kasparov contre *Deep Blue* en 1997, illustrant la puissance croissante des systèmes d'IA. Une autre évolution arrive en 2008 avec l'émergence du deep learning, qui transforme radicalement les capacités des machines, leur permettant apprendre et de traiter des données complexes. En 2016, *AlphaGo* de Google DeepMind bat le champion du monde du jeu de go. Les systèmes IA commencent à être en capacité de maîtriser des jeux de stratégie complexes. En 2017, l'introduction de l'architecture *Transformer* marque une nouvelle ère, permettant le développement de grands modèles de langage qui seront capables de comprendre et de générer du texte de manière plus fluide et contextualisée. Enfin, fin 2022, le lancement de *ChatGPT* par *OpenAI* propulse les IA génératives [2] sur le devant de la scène, rendant l'intelligence artificielle plus accessible et interactive que jamais. Elle est de plus en plus présente dans le quotidien du grand public et il en a désormais davantage conscience.

Aujourd'hui, les projets se multiplient dans les secteurs publics et privés face à des possibilités d'automatisation qui semblent infinies. Dans le secteur public luxembourgeois, les projets d'intelligence artificielle commencent à émerger. Les administrations publiques s'intéressent à ce type de technologies mais peu de projets d'envergure ont pour l'instant abouti. Ils en sont souvent encore à la phase de pilote. Ce facteur a rendu la recherche de sources difficile pour ce mémoire. Nous nous sommes basée sur beaucoup de prépublications, d'articles de revues ou de communications lors de conférences. Il y a encore peu d'ouvrages généraux sur les enjeux et usages de l'IA dans le secteur public et encore moins dans le domaine archives.

Les administrations ont néanmoins compris les avantages de l'intelligence artificielle et ont beaucoup d'ambition. En ce qui concerne les services d'archives publics, les usages de systèmes basés sur le *machine learning* sont pour l'instant réduits. Les services doivent se concentrer sur les traitements les plus urgents, qui concernent le papier et sont difficilement automatisables. En effet, les législations sont récentes. Les services ont un arriéré important à traiter et manquent souvent de personnel. Au delà du défi de l'arriéré, les administrations doivent aussi faire face au défi du numérique : la production documentaire augmente et les pratiques ne sont pas encore complètement formalisées. L'intelligence artificielle, via l'automatisation de certains processus, paraît pouvoir fournir une réponse à certaines de ces problématiques. Cette situation soulève plusieurs enjeux : bien que l'automatisation offre de nombreuses possibilités et que les projets d'intelligence artificielle, très en vogue et ambitieux, soient en pleine expansion, les services d'archives publics sont-ils réellement prêts ? Il paraît important de déterminer les prérequis nécessaires à la mise en place de ce type de projets. Un travail de prise de recul sur les apports, aussi divers soient-ils, et les complexités de mise en place des systèmes IA s'impose.

2. Branche de l'intelligence artificielle dont les modèles créent de nouvelles données, telles que du texte, des images ou de la musique.

N.B. La majorité des définitions de ce mémoire a été rédigée à l'aide de ChatGPT.

L'intelligence artificielle est-elle une solution aux défis archivistiques rencontrés par les producteurs d'archives publiques luxembourgeois ?

Dans une première partie, nous verrons en quoi le contexte public luxembourgeois est propice au lancement de projets IA malgré leur complexité de mise en place. Les ambitions des pouvoirs publics et des parlements sont importantes, un dialogue et un cadre se mettent en place. Nous présenterons plus en détail les défis auxquels sont confrontés les producteurs d'archives publiques luxembourgeois afin d'explorer les nombreuses potentialités d'automatisation. En dépit de ce contexte public favorable et de ces larges possibilités, la mise en place de projets IA est complexe et nécessite des précautions. Nous verrons ensuite qu'une fois menés, les projets IA peuvent avoir des apports importants pour les services d'archives. Ces derniers ne se situent pas forcément là où on les imagine et peuvent s'avérer connexes entre archives et IA. Enfin, notre dernière partie sera consacrée aux précautions éthiques et prérequis techniques spécifiques à l'IA. Elle sera l'occasion de réfléchir sur des facteurs parfois ignorés à prendre en compte avant la mise en production d'outils basés sur du *machine learning*.

Première partie

Un contexte public luxembourgeois propice au lancement de projets IA dans les archives malgré la complexité de leur mise en place

Chapitre 1

Un contexte européen et luxembourgeois favorable au développement de projets IA dans le secteur public

1. Ambitions et bénéfices pour les administrations publiques

L'innovation dans le domaine de l'intelligence artificielle est poussée par l'Union européenne et l'État luxembourgeois. Les ambitions sont élevées et, en cas de succès, les bénéfices peuvent être nombreux pour le pays et ses administrations.

Les publicités pour les téléphones portables, ordinateurs et voitures intégrant de l'IA se multiplient. L'Intelligence artificielle envahit notre quotidien et est devenue un véritable argument marketing. Elle est entourée d'une mythologie la présentant comme la solution à de nombreux problèmes. L'IA générative a été introduite au grand public par l'arrivée ChatGPT en novembre 2022. C'est à ce moment que ce dernier a pour la première fois réellement pu s'approprier une technologie basée sur de l'apprentissage machine ou *machine learning*. Les possibilités semblent infinies lorsqu'on commence à utiliser les IA génératives les plus puissantes du marché, mais la plus grande partie des utilisateurs n'a pas d'idée précise de ce qu'il y a derrière. Le fonctionnement des IA génératives est obscur pour le Grand public et pour beaucoup d'administrations, dont le personnel n'est souvent pas encore formé. Les possibilités semblent donc infinies mais leurs usages sont difficiles à définir concrètement. Un article de blog de l'anthropologue Madeleine Clare Elish aborde l'idée de magie liée à l'intelligence artificielle [1]. Ce champs lexical de la magie souvent utilisé pour décrire ces technologies est révélateur de leur opacité, la magie étant quelque chose qui produit un résultat mais qui ne s'explique pas. L'intelligence arti-

1. M.C. Elish, *Don't Call AI Magic*, en, janv. 2018, URL : `https://medium.com/datasociety-points/dont-call-ai-magic-142da16db408` (visité le 13/08/2024).

ficielle est associée à un certain nombre de fantasmes, à une forme de mythologie [2]. Cette dernière est héritée de la science fiction, dans laquelle l'imaginaire des machines dont les capacités égaleraient celles des humains et prendraient le contrôle a été vivement exploité. Le terme-même d'Intelligence artificielle est un symptôme de ces fantasmes en évoquant l'idée d'une intelligence comparable à celle des humains pour les machines, nourrissant les mythes et peurs véhiculés par la science-fiction. Le terme d'apprentissage machine ou *machine learning* est souvent plus adapté mais reste peu utilisé car moins impactant. Cette mythologie qui entoure l'intelligence artificielle la rend ainsi source de grandes ambitions pour les acteurs publics ou privés. Elle est fréquemment exploitée dans les discours marketing qui contribuent à entretenir des attentes irréalistes et des perceptions erronées de ce que l'IA peut accomplir et de son fonctionnement [3]. Ces ambitions et cette obscurité se sont ressenties au lancement du projet *InventAIre*, dont les ambitions sont assez élevées et surtout peu cadrées au départ parce que le projet a mûri au sein d'une équipe et d'une administration qui n'avaient pas encore une grande maîtrise des technologies de *machine learning*. La première partie de notre stage a ainsi été consacrée à un travail de définition précis d'un périmètre. Créer un outil capable de remplir l'intégralité des colonnes de l'inventaire des Archives nationales était impossible en quatre mois. Les ambitions élevées et le fait que l'intelligence artificielle soit un sujet d'actualité peuvent être néanmoins bénéfiques en facilitant le financement de projets IA, dans un secteur public où ils sont souvent complexes à obtenir. Plusieurs bénéfices de l'intelligence artificielle dans les administrations publiques ont été identifiés dans un article récemment paru dans la revue *Gestion et management public* à partir de la littérature scientifique sur cette dernière [4]. Nous pouvons les regrouper en quatre catégories : efficacité et gestion des ressources, sécurité, transparence, qualité de service. Ce sont surtout les gains économiques, obtenus grâce à l'automatisation des tâches redondantes, qui sont les plus intéressants dans un service public cherchant à réaliser des économies. Ambitions et potentiels bénéfices sont ainsi élevés au sein des administrations.

À l'échelle des états, une « course à l'IA » est en marche depuis quelques années et les pousse à investir dans ces technologies. Un article de Charles Thibout, chercheur en sciences politiques, publié en 2018 fait remonter le début de cette course au tournant des années 2010 [5]. Les états de l'Union européenne ne veulent pas se sentir dépassés par

2. Mihail Duca, *Artificial intelligence. Myth and reality ? | Annals of Philosophy, Social &amp ; Human Disciplines | EBSCOhost*, fr, janv. 2023, URL : `https://openurl.ebsco.com/contentitem/gcd:173354185?sid=ebsco:plink:crawler&id=ebsco:gcd:173354185` (visité le 13/08/2024).

3. *Ibid.*

4. Marius Bertolucci, « L'intelligence artificielle dans le secteur public : revue de la littérature et programme de recherche », *Gestion et management public*, Pub. anticipées–5 (2024), p. 118-139, DOI : `10.3917/gmp.pr1.0008`.

5. Charles Thibout, « L'intelligence artificielle, une géopolitique des fantasmes », *Etudes digitales*, Religiosité technologique 5–1 (août 2019), DOI : `10.15122/isbn.978-2-406-09290-2.p.0105`.

ces nouvelles technologies. À l'échelle de l'Union, une stratégie en matière d'intelligence artificielle a été rédigée en 2018 par la Commission européenne [6]. Les bénéfices mis en avant pour les pays de l'UE sont d'ordre éthique : les outils produits dans ces pays ont beaucoup plus de chances d'être conformes aux règles européennes. Les enjeux sont toutefois principalement économiques : d'après plusieurs économistes l'intelligence artificielle pourrait permettre de stimuler la croissance [7]. Ces systèmes vecteurs de croissance ont donc une grande valeur. Un pays pourra générer beaucoup de bénéfices en exportant ces innovations. Les états, et plus largement le secteur public, souhaitent par ailleurs anticiper les évolutions numériques pour éviter d'être dépassés. S'engager activement dans le développement et l'innovation permet d'éviter de subir une transition numérique trop précipitée. Il est dans leur intérêt de ne pas dépendre du secteur privé. Dans le cas de l'Union européenne ne pouvant rivaliser avec les grandes puissances telles que la Chine et les États-Unis, développer des compétences, infrastructures et outils techniques contribue à limiter la dépendance envers ces dernières. De plus, pour les mêmes raisons, l'accent est mis sur un développement éthique et responsable de l'IA, des valeurs qui s'intègrent bien aux ambitions des institutions publiques [8].

Le Luxembourg n'échappe pas à cette course vectrice d'investissements. Une stratégie nationale pour l'IA a été publiée en mai 2019, un peu plus d'un an après la publication de la stratégie française [9]. Une partie de la stratégie du Grand-Duché est consacrée à « L'IA au service du secteur public ». Il y est expliqué que les systèmes IA peuvent améliorer la qualité des services. Des actions sont prévues pour pousser le développement de tels systèmes : évaluation des projets potentiels, échanges avec des autres états membres de l'Union européenne, promotion de la recherche et de l'innovation, développement de solutions pour l'administration et développement de bases de données publiques [10]. Des investissements ont été réalisés. Un superordinateur d'une valeur de 30,4 millions d'euros a par exemple été inauguré en 2021 dans le but de pouvoir traiter des grands volumes de données et entraîner des modèles de *machine learning* [11]. Le secteur public luxem-

6. Commission européenne, *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions. Artificial Intelligence for Europe*, en, 2018, URL : `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN` (visité le 23/07/2024).

7. Philippe Aghion, Céline Antonin et Simon Bunel, « Intelligence artificielle, croissance et emploi : le rôle des politiques », *Economie et Statistique / Economics and Statistics*–510-511-512 (2019), p. 153, DOI : `10.24187/ecostat.2019.510t.1994`.

8. Nathalie A. Smuha, « From a 'race to AI' to a 'race to AI regulation' : regulatory competition for artificial intelligence », *Law, Innovation and Technology*, 13–1 (janv. 2021), p. 57-84, DOI : `10.1080/17579961.2021.1898300`.

9. Le rapport de Cédric Villani intitulé « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne » est publié en mars 2018.

10. Le gouvernement du Grand-Duché de Luxembourg, *Intelligence artificielle : une vision stratégique pour le Luxembourg*, fr, text, mai 2019, URL : `http://gouvernement.lu/fr/publications/rapport-etude-analyse/minist-digitalisation/artificial-intelligence/artificial-intelligence/intelligence-artificielle.html` (visité le 30/08/2024).

11. Portail public du Grand-Duché de Luxembourg, *Meluxina, le superordinateur du Luxembourg*,

bourgeois présente certains avantages qui le rendent propice au développement de l'IA. Il apparaît comme un meilleur garant de l'éthique des algorithmes que le secteur privé. L'IA est définie comme « une technologie puissante, entièrement sous notre contrôle, et débordante de possibilités » [12] dans l'introduction de la stratégie nationale rédigée par Xavier Bettel, alors Premier ministre et ministre de la Digitalisation. L'expression « sous notre contrôle » souligne l'importance accordée à une gouvernance responsable et à une maîtrise des impacts de l'IA.

Ainsi, le contexte actuel d'émulation, la mythologie qui entoure l'intelligence artificielle et ses bénéfices pressentis encouragent le secteur public à investir dans ce type de technologies. Il subsiste malgré tout des inquiétudes qui poussent les états de l'Union européenne vers des questionnements d'ordre éthique et vers la mise en place de cadres régulateurs.

2. La mise en place d'un cadre propice au développement de l'IA dans les institutions publiques

Comme évoqué précédemment, l'Union européenne a pris le virage de la responsabilité dans la course à l'intelligence artificielle. La question des risques était présente dès les débuts de la théorisation d'une potentielle intelligence des machines par des chercheurs tels qu'Alan Turing et Irving John Good. Leurs craintes étaient centrées sur un potentiel dépassement de l'intelligence humaine par la machine [13]. L'imprévisibilité de la machine et le fait qu'elle ne soit pas dotée de sentiments inquiétait. Le développement des IA génératives a soulevé des préoccupations concernant la désinformation, la manipulation de l'opinion publique, le respect du droit d'auteur, du RGPD (Règlement général de protection des données) et les biais algorithmiques. Face à ces risques, un cadre est à mettre en place.

La question de la réglementation ne date pas d'hier. Elle est abordée dans plusieurs stratégies nationales, dont celle du Luxembourg.

> Compte tenu de l'importance stratégique et de la grande complexité de ce sujet, le Luxembourg tient à investir dans un cadre amélioré propice à l'IA. Cet objectif implique d'envisager une nouvelle réglementation, garantissant un marché des données fonctionnel, par exemple afin d'éliminer les obstacles au développement d'une IA fiable [14]

fr, mars 2023, URL : `http://luxembourg.public.lu/fr/investir/innovation/meluxina-superordinateur.html` (visité le 30/08/2024).

12. Le gouvernement du Grand-Duché de Luxembourg, *Intelligence artificielle…*

13. John Burden, Sam Clarke et Jess Whittlestone, « 9. From Turing's Speculations to an Academic Discipline : A History of AI Existential Safety », dans *The Era of Global Risk*, dir. Sj Beard, Martin Rees, Catherine Richards et Clarissa Rios Rojas, 1re éd., Cambridge, UK, 2023, p. 201-236, DOI : `10.11647/OBP.0336.09`.

14. Le gouvernement du Grand-Duché de Luxembourg, *Intelligence artificielle…*

Toutefois, il a fallu attendre que les systèmes se démocratisent, notamment via les IA génératives, ayant suscité de vives d'inquiétudes, pour que la question soit étudiée plus en profondeur. L'Union européenne a commencé le travail avec l'*AI Act*, publié au Journal officiel le 12 juillet 2024. Ce règlement de l'UE concernant l'intelligence artificielle est une première dans monde. Il part d'une approche basée sur les risques, classés en cinq catégories : inacceptable, haut risque, risque spécifique, risque associé aux IA d'usage général et risque systémique associé aux IA d'usage général. Le règlement interdit les systèmes à risque inacceptable. Une liste de ces derniers est fournie dans l'article 5. Ils incluent :

> La manipulation cognitivo-comportementale de personnes ou de groupes vulnérables spécifiques : par exemple, des jouets activés par la voix qui encouragent les comportements dangereux chez les enfants
>
> Un score social : classer les personnes en fonction de leur comportement, de leur statut socio-économique, de leurs caractéristiques personnelles
>
> Une catégorisation et une identification biométriques des personnes
>
> Des systèmes d'identification biométrique en temps réel et à distance, tels que la reconnaissance faciale [15]

Les systèmes à hauts risques doivent être évalués régulièrement et les systèmes risques limités ont des obligations de transparence à respecter et doivent respecter le droit d'auteur. Cette approche par risque a l'avantage d'être assez vague pour traiter des grandes menaces dont on ne peut pas encore forcément prédire la forme, mais peut néanmoins sembler assez floue. Le monde de l'IA est voué à évoluer. D'autres législations sont à prévoir même s'il s'agit d'une base importante d'établissement d'un cadre légal pour une IA plus responsable.

Des cadres se construisent également à l'échelle institutionnelle afin de guider les futurs projets et les futurs usages du personnel des administrations et de leur public. La question de la confiance des utilisateurs envers l'IA est récurrente dans les guides et documents d'études émanant des états. Une étude du Conseil d'État français datant de 2022 associait dans son titre la confiance à la performance : « Intelligence artificielle et action publique : construire la confiance, servir la performance [16] ». En effet, la confiance permet d'aller jusqu'au bout des projets et assure une utilisation des outils IA développés.

15. Parlement européen, *Loi sur l'IA de l'UE : première réglementation de l'intelligence artificielle*, fr, juin 2023, URL : `https://www.europarl.europa.eu/topics/fr/article/20230601STO93804/loi-sur-l-ia-de-l-ue-premiere-reglementation-de-l-intelligence-artificielle` (visité le 01/08/2024).

16. Le Conseil d'État, *Intelligence artificielle et action publique : construire la confiance, servir la performance*, fr, août 2022, URL : `https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance` (visité le 02/08/2024).

Dans le rapport d'une « Consultation publique relative aux opportunités et aux défis de l'Intelligence Artificielle » datant de 2021 menée par le *LISER (Luxembourg Institute of Socio-economic Research)*, 58 % des personnes interrogées avaient une confiance moyenne dans une IA mise en œuvre dans le secteur public contre 41 % dans le privé [17]. Le secteur public luxembourgeois a pour avantage d'être vu comme un secteur plus cadré, qui priorise davantage l'éthique que le secteur privé. Cela facilite la confiance de son public. Toutefois, les usagers ne sont pas les seules personnes concernées par l'IA. Le personnel des administrations est le premier acteur humain impliqué dans les processus. La mise en place d'un cadre devrait permettre une meilleure confiance de sa part en les outils IA qu'ils seront amenés à utiliser. Cette idée est évoquée à propos de l'usage de l'IA pour le traitement des archives numériques dans un article récent intitulé « Applying AI to digital archives : trust, collaboration and shared professional ethics » [18]. Les auteurs y expliquent que l'IA peut être un outil performant pour les archivistes, mais pour exploiter son potentiel, les professionnels doivent être d'accord sur ce qui est éthique et ce qui ne l'est pas. Ils proposent une collaboration des différents acteurs : producteurs d'archives, professionnels des archives et chercheurs, pour développer des codes de conduite. Pour guider l'usage de l'IA et assurer cette performance dans le secteur public, des chartes ont été rédigées ou sont en cours de rédaction. Cela a été le cas à la Chambre des Députés, où une charte IA a été publiée fin juillet 2024. Elle expose « 10 lignes directrices que la Chambre des Députés suivra pour ses futurs projets en lien avec l'intelligence artificielle, notamment en matière de transparence, d'éthique et de responsabilité [19] ». Elle a été rédigée par une équipe composée de personnel de différents services, dont le service informatique, la Cellule archives, la Cellule scientifique, ou encore le Service du compte-rendu. Les acteurs sont donc divers, regroupant des personnes des métiers traditionnels de l'administration parlementaire, du monde de l'informatique, de la recherche et les archivistes. Cette collaboration lui octroie une plus grande légitimité et visibilité au sein de l'administration. Elle est courte et facilement compréhensible. La charte est un premier cadre qui permettra le développement croissant de projets IA et facilitera leur mise en production. Il s'agit en quelque sorte d'une étape de fondation de la politique de conduite du changement sur le sujet. La charte de la Chambre a donné lieu à plusieurs articles dans les médias luxembourgeois. Elle peut aussi être une forme de vitrine pour les administrations publiques, les présentant comme modernes et à la pointe de l'innovation.

17. Nicolas Poussing, « Résultats de la consultation publique relative aux opportunités et aux défis de l'Intelligence Artificielle (IA). » Dans 2021, URL : `https://liser.elsevierpure.com/fr/publications/r%C3%A9sultats-de-la-consultation-publique-relative-aux-opportunit%C3%A9s-e` (visité le 03/08/2024).

18. Lise Jaillant et Arran Rees, « Applying AI to digital archives : trust, collaboration and shared professional ethics », *Digital Scholarship in the Humanities*, 38–2 (mai 2023), p. 571-585, DOI : `10.1093/llc/fqac073`.

19. Chambre des Députés du Grand-Duché de Luxembourg, *La Chambre se dote d'une Charte sur l'intelligence artificielle | Chambre des députés du grand-duché de Luxembourg*, fr, 2024, URL : `https://www.chd.lu/fr/charteIA` (visité le 01/08/2024).

Ces différents cadres devraient permettre la mise en place de projets IA de manière plus sûre et avec une confiance plus accrue du personnel et des utilisateurs des outils développés. Grâce à ces structures, les projets IA peuvent désormais être élaborés avec une approche plus rigoureuse et transparente. Il s'agit d'un grand avantage au sein des parlements, où s'écrivent et se votent normalement l'établissement de ces cadres, et où ces initiatives d'intelligence artificielle commencent à se déployer.

3. Le cas des parlements : vers les premières mises en production d'outils basés sur l'IA

Les projets pilotes IA se multiplient dans les parlements. Un grand nombre a été présenté lors d'un séminaire intitulé « Use of artificial intelligence for parliamentary research and documentation » organisé par l'ECPRD (*European Center for Parliamentary Research and Documentation*) à Rome en mai 2024. L'intelligence artificielle est un grand sujet de discussion dans les administrations parlementaires. Dans le *Bulletin de l'innovation* de l'Union interparlementaire (l'organisation internationale des parlements) d'octobre 2023, trois personnes interrogées, membres des parlements européens, brésiliens et grecs, insistaient sur l'importance du réseau inter-parlementaire pour soutenir les initiatives IA, coopérer et partager des expériences [20]. L'Union inter-parlementaire, l'ECPRD, et les pratiques d'échanges entre parlements permettent en effet de mutualiser les connaissances sur des technologies encore récentes dans ces institutions.

Différents projets ont été poussés suite à la montée de l'IA générative. Les grands modèles de langage génératifs sont pré-entraînés et nécessitent ainsi la mise à disposition de moins de données et moins de connaissances techniques. Parmi ces projets pilotes d'envergure en contexte législatif, nous pouvons citer le projet *LlaMandement*, mené par la Direction Générale des Finances Publiques en France, qui présente des similarités avec le projet *InventAIre* de la Chambre des Députés. Il s'agit de générer automatiquement des résumés neutres d'amendements législatifs. Pour cela, le grand modèle de langage *Llama 2* développé par l'entreprise Meta a été *fine-tuné* [21], c'est à dire qu'il a été ajusté et affiné sur un ensemble spécifique de données pertinentes pour le traitement des amendements. Cette phase de *fine-tuning* permet au modèle pré-entraîné d'adapter ses réponses et ses capacités de synthèse aux exigences particulières des résumés législatifs. Cette idée de générer des résumés fait écho au remplissage des colonnes « titre » et « description » de notre inventaire. L'étendue du travail réalisé par l'équipe montre que l'acte de résumer

20. IPU, « Expert perspectives on AI in parliament », *Innovation tracker*, Issue 16 (2023), URL : `https://www.ipu.org/innovation-tracker/story/expert-perspectives-ai-in-parliament` (visité le 28/07/2024).

21. Joseph Gesnouin, Yannis Tannier, Christophe Gomes Da Silva, Hatim Tapory, Camille Brier, Hugo Simon, Raphael Rozenberg, Hermann Woehrel, Mehdi El Yakaabi, Thomas Binder, *et al.*, *LLaMandement : Large Language Models for Summarization of French Legislative Proposals*, arXiv :2401.16182 [cs], janv. 2024, DOI : `10.48550/arXiv.2401.16182`.

n'est pas neutre dans un contexte législatif et que si l'on voulait obtenir les meilleurs descriptions et les meilleurs titres possibles, il faudrait idéalement être en mesure de *fine-tuner* un modèle, ce qui était pour nous impossible en quatre mois avec des données non étiquetées et le matériel dont nous disposions. Les données ayant servi au *fine-tuning* dans le cadre du projet *LlaMandement* ont été postés sur le web, elles contiennent un peu plus de 9 000 documents avec leur résumé [22]. Les apprentissages du projet *LlaMandement* ont ainsi pu guider notre approche au début du projet *InventAIre*, ce qui illustre à quel point la mutualisation des savoirs entre les institutions menant des projets IA en contexte législatif est importante.

Dans le domaine des archives, quelques projets ont été mis en production dans des parlements. Ce sont souvent des projets de reconnaissance automatique de caractères sur des documents : *OCR (optical character recognition)*, ou *HTR (Handwritten Text Recognition)*. Ce sont des technologies plus anciennes, donc davantage maîtrisées et l'impact est assez faible en cas d'erreur. En dehors des projets d'*OCR*, c'est le service des archives du parlement européen qui paraît le plus actif, autour de l'équipe de Ludovic Delépine. L'IA y est actuellement utilisée pour classifier automatiquement des documents, générer des résumés et faciliter la recherche dans les archives. Ils ont mis en production en avril 2024 *Archibot 3.0*, *chatbot* permettant de faire une recherche en langage naturel dans un corpus d'un peu plus de 450 000 documents [23]. Il fonctionne grâce à un grand modèle de langage, *Claude 3 Sonnet*, développé par la société Anthropic, et grâce au RAG (Retrieval Augmented Generation) [24], technologie permettant de sélectionner des documents correspondant à une requête dans une base de connaissance, qui sera expliquée plus en détail dans le chapitre 5. Le futur de l'IA au service les parlements et de leurs archives paraît prometteur. Des projets sont en cours et une certaine d'émulation et mutualisation des connaissances se dégagent.

À la Chambre des Députés du Grand-Duché, deux projets pilotes ont été poussés. Le premier est un projet de *speech to text* [25]. Il est en cours, en collaboration avec l'Université du Luxembourg. Le second, *InventAIre*, est à l'origine de ce mémoire. Ces deux projets ont parmi leurs objectifs de prouver l'efficacité de l'IA générative sur l'automatisation des tâches liées à l'information et au langage. Un autre projet a été lancé afin de recenser les besoins métier spécifiques qui pourraient être automatisés via l'IA. Ces projets permettent également de poser les bases pour des applications futures. Actuelle-

22. Les données du projet LlaMandement postées sur Gitlab sont accessibles via cet url : `https://gitlab.adullact.net/dgfip/projets-ia/llamandement`

23. Luís Kimaid, *Artificial Intelligence-Driven Archibot : Transforming Access to European Union Parliament Archives*, en, URL : `https://library.bussola-tech.co/p/artificial-intelligence-archibot-eu-parliament` (visité le 05/08/2024).

24. *Ibid.*

25. Technologie qui convertit la parole en texte écrit en temps réel à l'aide de systèmes de reconnaissance vocale.

ment, l'accent est mis sur le traitement de l'information, domaine où l'IA générative est particulièrement efficace. Les mises en production demeurent encore limitées et souvent davantage orientées vers la médiation avec le public qu'au service d'applications métier spécifiques. Cette phase d'expérimentation est l'occasion d'étudier les capacités de l'IA et d'identifier les prérequis nécessaires à une intégration plus large dans les administrations parlementaires.

Pour conclure ce chapitre, le contexte actuel est favorable au développement de projets d'intelligence artificielle dans le secteur public. Les ambitions des institutions, parfois élevées, se confrontent à la réalité de l'implémentation, nécessitant des cadres régulatoires clairs. Les initiatives récentes, comme l'*AI Act* de l'Union européenne et la rédaction de chartes internes, reflètent une volonté croissante de structurer l'usage de l'IA. Cette dynamique d'innovation et de régulation s'accompagne d'une exploration des applications spécifiques de l'IA, notamment dans la gestion des archives. Le Luxembourg, fait face à des défis archivistiques qui encouragent cette exploration de solutions d'automatisation afin de moderniser et optimiser leur gestion et conservation.

Chapitre 2

Les archives au Luxembourg : législation récente et traitements urgents qui poussent vers l'exploration de moyens d'automatisation

1. Un cadre légal récent

La réglementation sur la conservation des archives est récente au Luxembourg. Le sujet a longtemps été ignoré. Ce manque d'attention est lié à l'histoire du pays. Jusqu'à l'abolition du secret bancaire par une loi votée le 5 novembre 2014, le secret y prévalait sur la transparence. Cette culture persiste. Il faut encore aujourd'hui davantage justifier la conservation des documents que leur destruction. Cette situation a influencé la gestion des archives et nous l'avons perçu pendant notre stage. Par exemple, les archivistes doivent demander un accès aux données sensibles stockées sur les serveurs métier et la demande n'est pas toujours acceptée. Archiver des documents contenant des données sensibles est un défi malgré les exceptions prévues par l'article 89 du RGPD, qui octroie des dérogations pour le traitement « à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques [1] ». Pour illustrer la difficulté de l'archivage à la Chambre, nous pouvons mentionner le cas des archives de Fernand Etgen, président de la Chambre de 2018 à 2023, qui n'ont pas été récupérées.

Le consensus tient par ailleurs une place importante dans le processus politique

1. Parlement européen, *Règlement (UE) 2016/679 du Parlement européen*, et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données), Article 89, 4 mai 2016, URL : `https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A02016R0679-20160504` (visité le 13/07/2024).

luxembourgeois. Les lois doivent être discutées et approuvées collectivement, et non imposées. Le processus législatif se déroule comme suit : les textes de loi sont étudiés par une ou plusieurs Commissions parlementaires, qui peuvent les amender, et sont ensuite généralement transmis pour avis à des Chambres professionnelles et au Conseil d'État. Le texte est plus tard débattu en séance publique. Il peut être amendé à ce moment-là aussi si nécessaire, puis il est voté en séance plénière [2]. Ce besoin de consensus a pu contribuer à ralentir le processus de législation sur les archives.

En ce qui concerne l'histoire des archives au Luxembourg, la constitution d'un véritable fonds d'archives publiques remonte à la loi du 5 brumaire de l'an V (26 octobre 1796), quand la région était sous l'administration française. Ce n'est qu'avec la loi du 5 décembre 1958 que les « Archives de l'État » obtiennent une forme de base légale [3]. En 1988, rebaptisées « Archives nationales », elles reçoivent le statut d'« institut culturel » [4]. Une loi adoptée le 25 juin 2004 sur la réorganisation de ces instituts culturels détaille les missions des Archives nationales : elles ont non seulement un rôle de collecte et de conservation, mais aussi de sensibilisation, de conseil et d'encadrement des détenteurs d'archives, publiques ou privées. Elles ont également un rôle scientifique : elles doivent organiser des expositions ou des colloques. Elles doivent accepter des archives publiques ou « privées d'intérêt historique, scientifique, économique, sociétal ou culturel [5] ». Elles doivent enfin « contribuer au développement de l'archivistique au niveau national et au niveau international [6] ».

La plus grande avancée législative arrive avec la loi du 17 août 2018, qui établit pour la première fois un réel cadre légal pour les archives publiques au Luxembourg. Cette loi fixe des règles concernant « la gestion, la conservation, la communication, le versement et la destruction des archives publiques [7] ». Les archives publiques doivent être gérées de manière à garantir leur pérennité, accessibilité et lisibilité tout au long de leur cycle de vie. La loi attribue aux Archives nationales du Luxembourg (ANLux) une mission d'encadrement des producteurs d'archives publiques. L'article 4 distingue deux types de régimes pour les producteurs : le régime général et le régime dérogatoire. Les établisse-

2. *La Chambre des députés du Luxembourg : promouvoir la gouvernance démocratique*, fr, juin 2024, URL : `http://luxembourg.public.lu/fr/societe-et-culture/systeme-politique/chambre-deputes.html` (visité le 18/07/2024).

3. Chambre des Députés du Grand-Duché de Luxembourg, *Loi du 5 décembre 1958 ayant pour objet l'organisation de la Bibliothèque Nationale et des Archives de l'Etat*, 5 déc. 1958, URL : `https://www.stradalex.lu/fr/slu_src_publ_leg_mema/document/mema_1958A15511?access_token=a91bb2cd70dac12a291757375cb9de2f2520b197` (visité le 14/07/2024).

4. Id., *Loi du 28 décembre 1988 portant réorganisation des instituts culturels de l'Etat*, 28 déc. 1988, URL : `https://legilux.public.lu/eli/etat/leg/loi/1988/12/28/n1/jo` (visité le 14/07/2024).

5. Id., *Loi du 25 juin 2004 portant réorganisation des instituts culturels de l'Etat*, 25 juin 2004, URL : `https://legilux.public.lu/eli/etat/leg/loi/2004/06/25/n7/jo` (visité le 14/07/2024).

6. *Ibid.*

7. Id., *Loi du 17 août 2018 relative à l'archivage*, 17 août 2018, URL : `https://legilux.public.lu/eli/etat/leg/loi/2018/08/17/a706/jo` (visité le 14/07/2024).

ments soumis au régime dérogatoire gèrent et conservent eux-même leurs archives, alors que les établissements soumis au régime général doivent proposer le versement aux ANLux. Ces régimes ont été créés en vertu de la séparation des pouvoirs. Le régime général concerne les administrations et services de l'État. Les administrations qui représentent les autres pouvoirs sont soumises au régime dérogatoire. C'est donc le cas de la Chambre des Députés, représentante du pouvoir législatif. Cette législation est révélatrice d'une prise de conscience de l'importance des archives, mais surtout du pouvoir de l'information, qui fait aussi écho au passé de paradis fiscal du pays. Cette conscience devrait favoriser un meilleur traitement des archives, toutefois, une certaine peur de la diffusion d'informations sensibles subsiste. La transparence est promue pour lutter contre cette image du secret. Elle est mentionnée dans l'article premier de la loi.

L'État a à cœur de développer un esprit de transparence, qui passerait par une bonne conservation et communication des archives. Des efforts restent malgré tout à fournir pour atteindre cet objectif. En ce qui concerne la communication, la loi de 2018 fixe des délais de communicabilité pour les archives définitives. Les archives sont consultables par les citoyens passés ces délais. Des dérogations peuvent également être demandées pour avoir un accès aux documents avant qu'ils soient échus. Les délais de communicabilité sont exposés dans l'article 16. Ils sont les suivants :

Type de donnée	**Délai de communicabilité**
Données à caractère personnel	25 ans après le décès de la personne concernée ou 75 ans à compter de la date du document le plus récent inclus dans le dossier
Actes d'état civil	100 ans à partir de la date de l'acte
Actes notariés	75 ans à partir de la date de l'acte
Atteinte aux relations extérieures, à la sécurité du Grand-Duché ou à l'ordre public	50 ans à compter de la date du document le plus récent inclus dans le dossier
Affaires portées devant les instances juridictionnelles, extrajudiciaires ou disciplinaires	50 ans à compter de la date du document le plus récent inclus dans le dossier
Prévention, recherche de faits punissables	50 ans à compter de la date du document le plus récent inclus dans le dossier
Données commerciales et industrielles	50 ans à compter de la date du document le plus récent inclus dans le dossier
Secret fiscal	100 ans à compter de la date du document le plus récent inclus dans le dossier

L'inventaire des Archives nationales, dont notre stage avait pour but d'automatiser le remplissage, est la conséquence de l'établissement de ces délais. Chaque colonne correspond à un type de document listé dans la loi, qui ne donne que peu de précisions sur ces différentes typologies. Cela a parfois complexifié notre travail. Nous avons dû réaliser un effort plus approfondi de définition des typologies mais leur subjectivité est demeurée problématique. Elle est abordée plus en détail dans la partie 3.1.2. de la note méthodologique en annexe. Un travail plus approfondi de définition est à réaliser et a été commencé par les Archives nationales. Une nouvelle loi sur l'archivage est également en préparation, elle pourrait être l'occasion d'apporter davantage de précisions. Une consultation publique a été organisée en avril 2024. Le manque de précision des attentes réglementaires en matière d'archivistique peut être un obstacle en cas d'automatisation. Les résultats des tentatives d'automatisation seront réellement précis lorsque les définitions seront précises. Toutefois, la loi de 2018 fixe un cadre qu'il était nécessaire de mettre en place afin d'harmoniser les pratiques archivistiques publiques, mais surtout d'assurer la conservation et la communicabilité des archives, composante essentielle dans une démocratie qui met en avant la transparence.

Beaucoup de traitements sont à réaliser pour atteindre ces objectifs. Nous avons déjà abordé le remplissage d'inventaires, permettant d'obtenir une forme de description archivistique et de gérer les délais de communicabilité. L'article 6 de la loi de 2018 impose l'établissement de tableaux de tri, dont la rédaction est à la charge des établissements lorsqu'ils sont soumis au régime dérogatoire. D'autres traitements sont également prioritaires dans les administrations. Il faut classer les archives. A la Chambre des Députés, les projets prioritaires de la Cellule archives sont actuellement le tableau de tri et le plan de classement. Une première version du tableau sera bientôt publiée. Une ébauche de plan de classement a été réalisée pour un service, le Service des relations européennes, internationales et du protocole (SREIP). La Cellule archives de la Chambre a encore du travail à faire pour assurer une bonne conservation et communication de ses archives. C'est pourquoi les projets d'automatisation sont bienvenus. C'est également le cas pour d'autres services d'archives publiques. Au moment de la rédaction de ce mémoire, le seul régime dérogatoire dont le tableau de tri est disponible sur le site internet des Archives nationales est celui de la Commission consultative des Droits de l'Homme. [8]. Les services d'archives doivent également se rendre visibles et faciliter la recherche dans leurs fonds. Dans cette optique, des portails numériques d'archives émergent. Les archives communales de Differdange ont par exemple lancé le leur en 2022, permettant d'accèder à des registres numérisés et à des galeries thématiques d'images [9].

8. ANLux, « Tableaux de tri », URL : `https://anlux.public.lu/fr/gerer-ses-archives/tableaux-de-tri.html`, Consulté le 07/08/2024.

9. VLA, « Nouveau portail des archives communales de Differdange ! », URL : `https://www.archives.lu/media/Accueil/Nouveau%20site%20internet_Archives%20communales%20de%20Differdange_com.pdf`, Consulté en 24/08/2024.

La loi du 17 août 2018 mentionne par ailleurs la sous-traitance privée. Les producteurs ou détenteurs d'archives publiques peuvent en effet externaliser leur conservation à un sous-traitant privé. Ils doivent informer les Archives nationales de l'identité et de la durée du contrat du sous-traitant. Les régimes dérogatoires doivent conserver eux-mêmes leurs archives destinées à être conservées définitivement [10]. Le traitement des archives par des entreprises privées spécialisées est récurrent au Luxembourg. Ce recours au privé permet de combler le manque de personnel et de formation de ce dernier.

À la Chambre des Députés, la première personne chargée de la gestion des archives avec une formation archivistique est arrivée en 2008. Avant cela, une personne sans formation avait travaillé sur les dossiers parlementaires, le reste des documents n'avait pas été traité. Une deuxième archiviste a été recrutée en 2023. La Cellule archives est vouée à s'agrandir prochainement.

Il n'y a pas de formation archivistique au Luxembourg. La plupart des archivistes viennent de France, de Belgique ou d'Allemagne. Un concours existe pour les conservateurs des archives. Il est la plupart du temps destiné et obtenu par des historiens. Beaucoup d'archivistes luxembourgeois sont formés en France à l'INP (Institut national du patrimoine) via le Stage technique international d'archives, dont le rythme est de deux modules de deux heures par semaine pendant un mois à distance et d'environ quinze jours de cours [11]. L'archivistique luxembourgeoise n'est pas encore dans une période de maturité. Elle se nourrit de différentes pratiques. Le dialogue entre les pratiques des différents pays et le fait qu'il existe un recul sur ces dernières est un facteur intéressant. Le Grand-Duché a une opportunité de capitaliser sur l'expérience de différentes régions pour construire ses propres pratiques.

Les défis de gestion des archives au Luxembourg sont donc nombreux pour un domaine longtemps négligé. Ils mettent en lumière l'importance de l'exploration de moyens d'automatisation qui permettraient d'y répondre. La récurrence de la sous-traitance rend ces perspectives encore plus attrayantes pour les producteurs d'archives publiques, leur permettant de ne pas dépendre du privé, d'être ainsi davantage maîtres de leur données et de réaliser des économies.

2. Les enjeux des archives numériques : un territoire peu exploré et de nouvelles données à appréhender

Les archives sont définies dans la loi du 17 août 2018 comme « l'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur

10. *Ibid.*

11. INP, « Stage technique international d'archives (STIA) »,URL : `https://www.inp.fr/stage-technique-international-darchives-stia`, Consulté le 07/08/2024

forme matérielle et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité [12] ». Les documents et données numériques sont donc inclus dans la notion d'archive. Une loi a été consacrée aux documents électroniques. Il s'agit de la loi du 25 juillet 2015 relative à l'archivage électronique. Cette loi définit les règles pour garantir l'intégrité, la confidentialité et la valeur probante des copies numériques, équivalentes à celles des originaux [13]. Elle encadre l'activité des Prestataires de Services de Dématérialisation et de Conservation (PSDC), imposant leur certification et leur inscription auprès de l'ILNAS (Institut Luxembourgeois de la Normalisation, de l'Accréditation, de la Sécurité et qualité des produits et services) [14]. Cette loi est donc centrée sur la question de la valeur juridique des documents électroniques et non sur leur traitement et elle n'est pas orientée vers le secteur public, seulement vers le privé. L'idée du pouvoir de l'information prévaut de nouveau, les valeurs scientifiques ou patrimoniales des archives sont ici ignorées. Elles sont en effet d'autant plus complexes à faire reconnaître lorsque l'on traite de documents numériques. Les archives électroniques ne sont ici réglementées que lorsque et parce qu'elles ont un équivalent papier. Les archives nativement numériques restent ainsi à appréhender chez les producteurs d'archives publiques luxembourgeois. Les premiers systèmes d'archivage électronique (SAE) émergent, dont un SAE mis à disposition par le CTIE (Centre des Technologies de l'Information de l'Etat) pour les administrations du secteur public. L'inventaire qui a fait l'objet de notre stage est quant à lui lié au système d'information archivistique (SIA) des ANLux. Il permet d'automatiser la communicabilité des archives qui y sont rentrées. À la Chambre des Députés, il n'y a pas encore de système d'archivage électronique mais il est en projet. Comme évoqué précédemment, les archivistes travaillent aussi sur un plan de classement des documents numériques.

Le chantier du numérique est mené en parallèle d'autres chantiers dans les services d'archives publiques. Il faut d'abord sécuriser la collecte et la conservation des documents papier. Les documents anciens, et par conséquent souvent fragiles, ont été priorisés. Il est plus aisé de lancer des projets de traitement de ce type de documents parce que les personnes qui ne viennent pas du milieu archivistique y voient davantage d'intérêt historique. Les documents numériques ont toutefois autant de valeur historique. On pourrait de plus argumenter que le numérique aussi revêt une certaine fragilité. Malgré les possibilités de traçage des opérations, il est facile de supprimer, de reclasser et de modifier des documents stockés sur des serveurs sans que qui que ce soit ne s'en rende compte. Le passage d'un document d'un système informatique à un autre peut aussi entraîner des

12. *Ibid.*

13. Id., *Loi du 25 juillet 2015 portant création du système de contrôle et de sanction automatisés et modification de la loi modifiée du 14 février 1955 concernant la réglementation de la circulation sur toutes les voies publiques*, 25 juill. 2015, URL : `https://legilux.public.lu/eli/etat/leg/loi/2015/07/25/n2/jo` (visité le 14/07/2024).

14. *Ibid.*

pertes de données. Il existe des moyens de vérifier l'intégrité des documents, par exemple en calculant un *hash*, code en principe unique généré à partir du contenu du fichier via un algorithme, qui permettra de détecter toute modification par comparaison. Certains formats de fichiers peuvent également être difficiles à ouvrir parce que ce sont des formats obsolètes ou propriétaires. Les formats sont divers et les manières de les traiter peuvent varier : une boîte mail ne sera pas traitée de la même manière qu'une arborescence de fichiers bureautiques ou qu'une base de données. Certains supports sont par ailleurs fragiles. C'est par exemple le cas du CD-ROM qui se dégrade parfois en une dizaine d'années. Par ailleurs, un document numérique est soit lisible soit illisible [15]. Il n'y a pas d'entre deux, pas de dégradation lente qui pourrait sonner un signal d'alarme, contrairement au papier. Les liens entre les documents sont eux aussi fragiles. Un exemple concret est celui du déplacement d'un tableur d'un dossier numérique à un autre : si ce tableur contient des formules faisant référence à un autre document du même dossier, le lien peut être rompu. Les formules et macros posent aussi des soucis. Il faut donc prendre un certain nombre de précautions face aux documents électroniques. De plus, le numérique évolue vite. Une veille constante est à réaliser sur les technologies utilisées dans les administrations et sur les technologies de traitement des archives. De nouvelles données sont constamment ajoutées sur les serveurs et dans le *cloud* des administrations. Une autre question se pose face à cette mutabilité : quand archiver définitivement ? Cette question concerne par dessus tout les archives du web, dont les sites sont constamment actualisés [16]. La solution trouvée est la collecte régulière, ou bien le traçage des modifications. La Chambre doit encore décider si elle archive son propre site web malgré les opérations de capture effectuées par la Bibliothèque nationale du Luxembourg (BNL). Ce sera une possibilité à éventuellement explorer. Les documents numériques revêtent donc une certaine fragilité et leur traitement doit donner lieu à des réflexions spécifiques.

Les traitements ne sont pas les mêmes que pour des supports matériels. De nouvelles méthodologies sont à mettre en place et de nouvelles compétences sont à développer au sein des équipes [17]. De nouvelles normes seront à adopter d'après certains archivistes [18]. Les Archives nationales du Luxembourg tentent de guider les producteurs d'archives publiques sur l'archivage numérique, notamment sur les étapes de traitement de vracs. Les difficultés de traitement des vracs numériques ont été résumés dans un guide publié par l'AAF

15. « Scarcity or Abundance ? Preserving the Past in a Digital Era », *The American Historical Review* (, juin 2003), DOI : 10.1086/ahr/108.3.735.

16. Maranke Wieringa, « The Fragility of Digital Media Content : On Preservation and Loss : Sketching the Pilgrimage of Future Scholars to Recover Our Digital Vellum », *Junctions : Graduate Journal of the Humanities*, 2–2 (sept. 2017), p. 27, DOI : 10.33391/jgjh.33.

17. Françoise Banat-Berger, « La prise en charge des archives electroniques en France dans le secteur public », *Archives*, 40–1 (2008), p. 27-69, URL : https://www.archivistes.qc.ca/revuearchives/vol40_1/40_1_banat-berger.htm (visité le 30/08/2024).

18. David Rajotte, « La réflexion archivistique à l'ère du document numérique : un bilan historique », *Archives*, 42–2 (2010), p. 69-105.

(Association des archivistes français) [19]. En voici un résumé non exhaustif :

— Copies et versions multiples des documents

— Structuration anarchique et gros volumes à traiter

— Variété des formats de fichiers

— Existence de fichiers corrompus

— Valeurs dynamiques et accès restreints aux documents

— Présence de données sensibles

— Absence de traçabilité des opérations sur les données

Nous avons eu l'occasion de discuter avec plusieurs membres du service « Collecte, conseil et encadrement » des ANLux des traitements à fournir face à ces problématiques. Nous les avons regroupés en une illustration ci-dessous. Les ANLux travaillent sur des moyens d'automatisation de ces différentes étapes. Plus d'une trentaine de scripts *shell* ont été développés et sont diffusés auprès des producteurs d'archives publiques. Ils ont en partie été réalisés à l'aide de l'intelligence artificielle [20]. Les Grands modèles de langage sont en effet performants pour la génération de programmes informatiques. Les scripts interviennent pour classer, trier et vérifier l'intégrité des archives [21]. La figure 2 ci-dessous, extraite de la note méthodologique, liste divers moyens d'automatisation de traitements de vracs numériques, incluant des scripts créés par les ANLux.

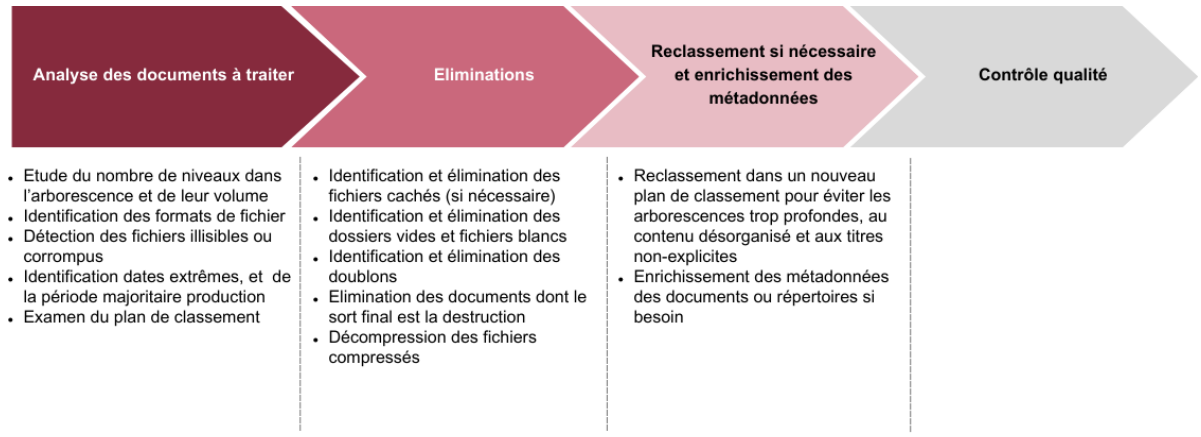

FIGURE 1 – Étapes de traitement des vracs numériques identifiées pendant le stage

La réalisation de l'inventaire vient après ces traitements. Une fois les archives prêtes à être conservées, elles doivent être décrites et les données sensibles doivent être repérées. L'ensemble de ce travail nécessite du temps et des moyens, mais ils sont de moins en moins considérables grâce aux outils d'automatisation. L'article présentant le travail des ANLux

19. Lorène Béchard, Lourdes Fuentes Hashimoto et Édouard Vasseur, *Les archives électroniques*, 2e éd., enrichie et mise à jour, Paris, 2020 (Les petits guides des archives).

20. Michel Cottin, Camille Forget et Richard Gaudier, « Traitement des vracs bureautiques et IA : un premier pas dans la porte », dir. Association des archivistes français, *Archivistes !*–147 (2024).

21. *Ibid.*

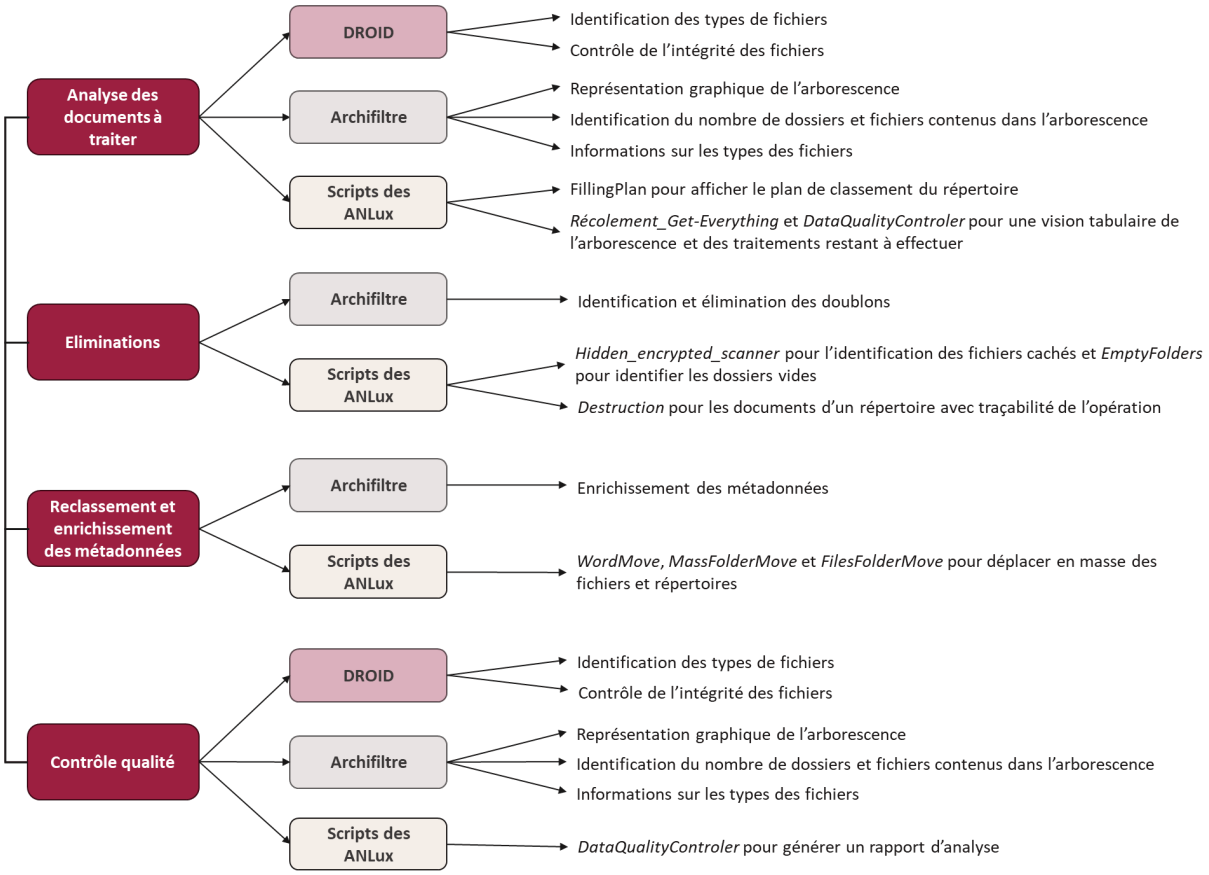

FIGURE 2 – Moyens d'automatisation par étape de traitement de vracs numériques [22]

en termes d'automatisation évoque qu' « en 2022, un vrac de 350 Go a nécessité près de 60 jours alors qu'en 2023, un vrac de 1,2 To n'en prenait que 20. [23] » En ce qui concerne l'inventaire, les archivistes de la Chambre avaient calculé qu'il faudrait sept ans à temps plein pour rédiger les inventaires de l'ensemble des fonds papier et numériques sans moyen d'automatisation. Un autre enjeu du traitement des archives numériques, au delà de sa complexité, est en effet leur masse. Le numérique fait partie intégrante du fonctionnement des administrations. Des quantités importantes de données sont dès lors produites. Le personnel de l'administration n'a pas de vision claire de l'ampleur des données générées et il est aisé de créer un nouveau document dans un système ou bien de générer de nouvelles données. Le fonds du Service des relations européennes et internationales et du protocole (SREIP), qui a été choisi comme base de données pour la recherche et le développement dans le cadre du projet InventAIre, contient par exemple 140 000 documents pour 80 000 répertoires. Il s'agit de l'arborescence complète des fichiers du service contenus sur les serveurs. Un bénéfice de la réalisation de l'inventaire des archives est de mieux s'y retrouver dans ces documents en fournissant des titres et des descriptions des différents

23. *Ibid.*

23. N.B. D'après l'expertise des ANLux, les chiffres sur les nombres de fichiers sont à prendre avec vigilance, Archifiltre, Droid et Windows n'obtenant parfois pas les mêmes chiffres pour une même arborescence.

répertoires.

Idéalement, une fois les vracs traités, la communication des documents doit être assurée. C'est à ce moment que se pose la question de l'identification données sensibles dans les documents. Il s'agit d'un processus long qui demanderait à un ou une archiviste d'examiner chacun d'entre eux. C'est un obstacle à une communication optimale des documents. Ces archives dont l'accessibilité est empêchée par les défis liés aux données sensibles ont été théorisées sous le nom de « *dark archives* ». [24]. D'après Jason Baron et Nathaniel Payne, chercheurs américains et canadiens, la communication des archives numériques nécessite un long travail d'identification des données sensibles. Les administrations publiques sont en quête de transparence mais l'accès à leurs archives serait menacé. Pour automatiser le repérage des données sensibles, ils proposent plusieurs solutions. L'utilisation d'expressions régulières (REGEX) pour repérer des informations confidentielles, telles que des numéros de sécurité sociale, est une première solution, même si son efficacité est limitée. Le recours au *machine learning* est présenté comme une perspective fructueuse. La classification automatique automatique et le deep learning sont envisagés [25]. L'article date de 2017, l'usage du *machine learning* dans les archives en était à ses débuts. L'architecture *Transformer*, nouvelle manière de concevoir des modèles d'IA ayant permis de rendre les modèles de langage plus rapides et plus puissants et donc le développement des Grands modèles de langage, a été proposée en 2017 [26]. Les auteurs avaient donc déjà saisi les bénéfices des futurs grands modèles de *deep learning* pour le traitement des archives. Le *cloud computing* est également évoqué comme une méthode efficace pour faciliter le traitement automatique des données sensibles. Il permet d'accéder à une puissance de calcul et de stockage massive via des serveurs distants, par rapport à un hébergement sur des ordinateurs en local. L'analyse des documents peut alors être réalisée beaucoup plus rapidement et avec des outils plus volumineux, comme des grands modèles d'intelligence artificielle. L'usage du *machine learning* pour automatiser la rédaction de l'inventaire des ANLux, contenant des colonnes sur les données sensibles, n'est donc pas sans fondement, ce type d'application a déjà été pensé. Toutefois, pour que l'automatisation fasse réellement gagner du temps aux équipes, elle doit être précise.

Ainsi, les défis auxquels sont confrontés les producteurs d'archives publiques luxembourgeois sont nombreux, qu'ils soient liés à la gestion des documents papier ou à la gestion des archives numériques. Ces dernières posent des problèmes spécifiques liés à la quantité massive de données à traiter et aux complexités techniques associées à ce trai-

24. Jason R. Baron et Nathaniel Payne, « Dark Archives and Edemocracy : Strategies for Overcoming Access Barriers to the Public Record Archives of the Future », dans *2017 Conference for E-Democracy and Open Government (CeDEM)*, Krems, Austria, 2017, p. 3-11, DOI : 10.1109/CeDEM.2017.27.

25. *Ibid.*

26. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser et Illia Polosukhin, *Attention Is All You Need*, arXiv :1706.03762 [cs], août 2023, URL : http://arxiv.org/abs/1706.03762 (visité le 16/08/2024).

tement. Pour relever ces défis, il convient d'explorer des solutions innovantes, de tenter d'automatiser des processus, et de tirer parti des pratiques développées ailleurs. Ces défis peuvent également être perçus comme des opportunités. Le fait de partir sur des bases récentes permet d'expérimenter de nouvelles approches. Cette situation rend les producteurs d'archives plus ouverts à tester de nouveaux outils et à initier des projets innovants, qui pourraient transformer la manière dont les archives sont gérées à l'avenir.

Chapitre 3

Prérequis et points d'attention pour le pilotage de projets d'automatisation via l'IA dans les archives

1. Des besoins métier multiples mais des cas d'usage à préciser

Les paragraphes précédents illustrent à quel point les besoins et possibilités en termes d'automatisation sont nombreux chez les producteurs d'archives publiques. Les services ont l'opportunité d'expérimenter avec l'IA. Il est plus aisé d'obtenir des financements publics pour des projets qui utilisent des technologies dans l'ère du temps. Les projets IA obtiennent d'autant plus de financements dans le cadre de leur promotion par les états mentionnée en première partie. Ce type de projets nécessite néanmoins une réflexion approfondie en amont pour assurer un pilotage efficace et garantir des résultats concrets.

Tout d'abord, pour que ces projets impliquant du *machine learning* réussissent, ils doivent non seulement répondre à des besoins métier, mais également avoir des cas d'utilisation bien précis. Le cas d'utilisation ou cas d'usage est défini par Alistair Cockburn, expert en méthodes agiles, comme « une description des séquences possibles d'interactions entre un système en question et ses acteurs externes, liées à un objectif particulier [1] » [Traduction libre]. L'informaticien suédois Ivar Jacobson aurait été le premier à introduire des cas d'usage à la fin des années 1960 [2]. C'est à partir des années 1980-1990 qu'ils ont été davantage formalisés. D'après Alistair Cockburn, pour rédiger un cas d'utilisation efficace, il faut commencer par définir clairement le périmètre du système et identifier tous les acteurs

1. Alistair Cockburn, *Writing effective use cases*, Boston, 2001 (The Crystal series for software development).

2. *Ibid.*

et leurs objectifs. Il faut ensuite rédiger le scénario de succès principal en décrivant chaque étape comme un objectif atteint, puis ajouter les alternatives et les échecs possibles [3]. Il s'agit donc d'une modélisation de processus qui place les acteurs et leurs interactions avec le système au centre. Les cas d'usage sont d'une réelle importance pour les projets IA. Ils permettent d'ancrer ces derniers dans des réalités concrètes et de s'assurer que les solutions développées répondent réellement aux besoins des utilisateurs finaux. Sans cas d'usage bien définis, les initiatives IA risquent de s'éparpiller et de ne pas apporter la valeur ajoutée escomptée pour le métier. Les problèmes liés aux cas d'utilisation sont parmi les cinq catégories de facteurs d'échec des projets IA, avec les attentes irréalistes, les contraintes organisationnelles, le manque de ressources et les problèmes techniques selon une étude réalisée par des chercheurs de l'université de Reutlingen et un consultant de la société EXXETA en 2021 [4]. Une bonne spécification de cas d'usage serait également un des éléments clés de la réussite économique des projets IA dans le secteur privé d'après une autre étude récente menée en Allemagne [5]. Les entreprises peineraient à faire évoluer les projets IA pilotes vers des environnements de production. Les cas d'usage permettent de se concentrer sur les points les plus complexes à traiter et d'exploiter les opportunités de valeur ajoutée, au lieu de se limiter à des projets réalisés sur des données parce qu'elles sont facilement accessibles mais sans but précis. La validation du bon fonctionnement d'un cas d'usage est une condition pour le passage d'un projet pilote à un déploiement à plus grande échelle. Les objectifs et les interactions avec les utilisateurs finaux doivent être clarifiés [6]. Cette approche assure que les projets répondent aux véritables besoins du métier. Il est en effet important de noter que l'IA n'est pas la solution la plus efficace dans tous les cas.

Une définition précise des cas d'usage permet en outre une implication des équipes dès la genèse du projet, un élément clé dans la conduite du changement. En suivant une approche similaire au *design thinking* [7], où l'utilisateur final est au centre du processus de conception, les cas d'usage permettent de visualiser comment l'IA peut s'intégrer dans les processus actuels. Cela favorise son acceptation en tant qu'outil collaboratif : la machine doit compléter et accélérer le travail humain au lieu de le remplacer. Cette idée est évoquée

3. *Ibid.*

4. Jens Westenberger, Kajetan Schuler et Dennis Schlegel, « Failure of AI projects : understanding the critical factors », *Procedia Computer Science*, International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021 196 (janv. 2022), p. 69-76, DOI : 10.1016/j.procs.2021.11.074.

5. Michael Grebe, Marc Roman Franke et Armin Heinzl, « Artificial intelligence : how leading companies define use cases, scale-up utilization, and realize value », *Informatik Spektrum*, 46–4 (août 2023), p. 197-209, DOI : 10.1007/s00287-023-01548-6.

6. *Ibid.*

7. Méthode de conception centrée sur l'utilisateur, qui le met parfois à contribution dans le processus de développement.

dans un article précédemment cité intitulé « Implementing AI in the public sector » : l'IA doit permettre au personnel de se concentrer sur des tâches décisionnelles complexes et la libérer des tâches répétitives. Au lieu de se concentrer sur l'idée de remplacement des humains par l'IA, il est suggéré aux acteurs publics de réfléchir à la manière dont l'IA augmentera les capacités humaines et dont humains et machines peuvent collaborer [8]. Cette approche collaborative constitue effectivement une stratégie pour atténuer les craintes des personnes qui redoutent d'être remplacées par l'IA, notamment dans un secteur où les recrutements risquent de diminuer et les externalisations de se multiplier. Bien que la sécurité de l'emploi soit généralement plus élevée dans le secteur public, l'accent mis sur l'intégration harmonieuse de l'IA permet de rassurer le personnel en mettant en avant la complémentarité entre les capacités de l'humain et de la machine. Les systèmes intégrant du *machine learning* sont presque constamment humanisés dans la littérature. Le terme « intelligence artificielle » est déjà un résultat de cet anthropomorphisme. Les prototypes d'automatisation sur les tâches complexes n'atteignent pas les 100 % de précision. Il ne s'agit donc pas réellement d'une automatisation complète de processus, mais l'IA serait une sorte d'agent qui réaliserait une partie du travail au sein du processus, il s'agirait davantage d'une augmentation que d'une automatisation. En mettant des mots sur l'organisation de cette collaboration homme-machine, l'efficacité des outils IA pourrait donc se voir maximisée, et les inquiétudes des équipes minimisées.

Dans le cas du projet InventAIre, l'usage a été défini par l'équipe : l'outil développé a pour but de produire automatiquement des inventaires d'archives au format *Excel* d'après le modèle fourni par les ANLux. Nous avons réalisé des diagrammes détaillant le processus technique mis en œuvre par l'outil produit. Il faudrait pousser ce travail plus loin en impliquant davantage les utilisateurs, précisant leurs interactions avec le système. Pour cela, la réalisation de diagrammes suivant des langages normés, tels que l'UML [9], est une possibilité à envisager. Le scénario précis d'utilisation de l'outil gagnerait à être davantage défini. Il semble nécessaire de clarifier comment la collaboration homme-machine se déroulera concrètement : l'inventaire produit sera-t-il utilisé tel quel, avec un processus de vérification en place, ou l'outil servira-t-il à accélérer le travail de l'archiviste en fournissant un document pré-rempli que celui ou celle-ci pourra ensuite compléter et ajuster ? Cette approche doit également tenir compte des risques associés à la précision des résultats. Par exemple, les colonnes de l'inventaire concernant les descriptions et titres présentent moins de risque en cas de manque que celles contenant les informations sur les données sensibles,

8. Ines Mergel, Helen Dickinson, Jari Stenvall et Mila Gasco, « Implementing AI in the public sector », *Public Management Review* (, juill. 2023), p. 1-14, DOI : 10.1080/14719037.2023.2231950.

9. *Unified Modeling Language*, un langage de modélisation standardisé utilisé en ingénierie logicielle pour visualiser, spécifier, concevoir, et documenter les éléments d'un système logiciel à travers différents types de diagrammes. Il aide à représenter les structures, les comportements, et les interactions d'un système de manière claire et compréhensible.

qui nécessitent une attention particulière. Un parallèle peut être tracé avec l'exemple des voitures autonomes exposé dans l'article sur les causes des échecs des projets IA : il est dit que « dans des cas d'utilisation spécifiques, comme la conduite autonome, une faible tolérance aux erreurs peut entraîner l'échec du projet. Ces cas d'utilisation dépendent de prévisions et de résultats précis et corrects, car une erreur peut avoir des conséquences fatales[10]. » [Traduction libre]. Dans le cas du repérage des données sensibles, si l'inventaire est réutilisé tel quel, sans vérification, la tolérance aux erreurs sera faible et si le modèle de *machine learning* donne des taux d'erreur qui ne sont pas assez proches de zéro, le projet sera un échec. Concernant la tolérance aux erreurs, d'après Lise Jaillant et Arran Rees , « le risque de divulguer des données potentiellement sensibles doit être comparé au risque de garder les archives confidentielles et inaccessibles[11] » [Traduction libre]. Une précision inférieure à 100 % peut être un risque choisi et mesuré par le service. Des réflexions supplémentaires sur ces questions seront par conséquent à mener en cas de poursuite du projet InventAIre. Les choix devront être guidés par une analyse de risque approfondie, la définition d'un niveau de précision acceptable et un processus rigoureux d'évaluation des résultats produits par les modèles de *machine learning* pour vérifier qu'ils sont conformes à ce niveau de précision. Si cette dernière n'est pas proche des 100%, nous préconisons une vérification des inventaires par les archivistes au moins sur les colonnes de l'inventaire traitant les données sensibles avant leur mise à disposition pour le public.

La définition précise des cas d'usage revêt donc une grande importance pour assurer la réussite des projets d'IA dans les archives, car elle permet de clarifier les objectifs et d'aligner les attentes tout en intégrant les besoins réels des utilisateurs. Cette étape est interdépendante avec les étapes de définition du périmètre et d'analyse des risques. L'approche à adopter présente des différences par rapport à la gestion des projets archivistiques traditionnels. Le notion de cas d'usages vient du domaine de l'informatique. La gestion des projets doit ainsi être réfléchie et adaptée pour une application de l'IA dans les archives.

2. Optimiser la gestion de projets IA dans les archives : réflexions et défis

Avant de mettre en place des projets IA, les services d'archives publiques doivent considérer les particularités de la gestion de projets impliquant des outils basés sur le *machine learning*. La gestion de projet peut se définir comme l'organisation, la planification et la coordination des ressources dans le but atteindre des objectifs spécifiques dans un délai donné. La gestion de projet comme outil managérial se serait rationalisée dans les

10. J. Westenberger, K. Schuler et D. Schlegel, « Failure of AI projects… ».
11. L. Jaillant et A. Rees, « Applying AI to digital archives… ».

années 1930 dans le secteur public et aurait été théorisée en tant que modèle à la fin des années 1950 [12]. Une bonne gestion de projet permettrait de mieux visualiser et de maximiser les résultats. Différents outils et méthodes de gestion de projet ont émergé. La plus connue est la méthode agile, popularisée suite à la publication en ligne du *Manifeste pour le développement agile de logiciels* en 2001 [13]. Les méthodes agiles sont des méthodes de gestion de projet qui privilégient l'adaptabilité, la collaboration et l'itération dans le développement de produits. Elles se basent sur des cycles de travail courts, nommés sprints, à l'issue desquels les équipes évaluent et ajustent leurs priorités en fonction du retour des acteurs du projet, en particulier des commanditaires ou utilisateurs. Les avantages incluent une meilleure réactivité aux changements, une bonne communication entre les acteurs et un produit final davantage aligné sur les besoins réels des utilisateurs. Ces méthodes ne sont pas forcément éligibles à tout type de projet mais sont particulièrement adaptées aux projets informatiques car elles permettent de livrer rapidement des versions fonctionnelles d'un produit et favorisent alors une amélioration continue en fonction des retours de ses utilisateurs.

La sociologue Camille Girard-Chanudet décrit la logique projet employée dans la pseudonymisation automatique sur des documents par l'IA à la Cour de Cassation comme s'inscrivant dans « le cadre de la "transformation de l'action publique" [14] ». Cette logique projet s'est manifestée à la Cour de Cassation par l'intégration d'une approche structurée autour d'objectifs clairs, d'un produit minimum viable (MVP) et d'un calendrier précis [15]. Cette approche favorise non seulement l'efficacité, mais s'inscrit également dans une volonté plus large de modernisation et de digitalisation des services publics, conformément aux objectifs de la « transformation de l'action publique ». Cette transition numérique a favorisé l'adoption de méthodes de gestion de projet, notamment les méthodes agiles, au sein des administrations publiques. Elles se sont souvent inspirées du secteur privé. Ces pratiques sont en effet adoptées par les entreprises de services numériques travaillant pour le secteur public. En France, un accent a été mis sur l'innovation et le dynamisme avec, au cours des dix dernières années, l'introduction de concepts comme les « entrepreneurs d'intérêt général » (EIG), recrutés de manière régulière depuis 2017, et le développement de « start-ups d'État ». Dans le domaine archivistique, *Archifiltre* est un exemple de start-up d'État. L'objectif est de moderniser le service public, d'améliorer la productivité et de favoriser l'innovation. Au Luxembourg, la création du *GovTech Lab* vise à soutenir

12. Gilles Garel, « Pour une histoire de la gestion de projet », *Annales des mines*, Gérer et comprendre–74 (déc. 2003).

13. K. Beck, M. Beedle, V. A. Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, *et al.*, *Manifesto for Agile Software Development*, 2001, URL : `https://agilemanifesto.org/` (visité le 28/07/2024).

14. Camille Girard-Chanudet, « Le travail de l'Intelligence Artificielle : concevoir et entraîner un outil de pseudonymisation automatique à la Cour de Cassation », *RESET. Recherches en sciences sociales sur Internet*–12 (mars 2023), DOI : `10.4000/reset.4731`.

15. *Ibid.*

l'innovation au sein de l'État. Ainsi, les états encouragent l'innovation en adoptant des méthodes courantes et performantes dans le secteur privé, ce qui explique la popularité croissante de ces pratiques dans le secteur public. À la Chambre des Députés, le service Technologies de l'information (TI) a recruté quatre chefs de projet depuis 2022, s'inscrivant dans cette même dynamique de modernisation. Les chefs de projet se multiplient aussi dans les autres administrations publiques.

En ce qui concerne les projets archivistiques, en 2018, Cyndi Shein, Hannah Robinson et Hana Gutierrez ont proposé d'introduire les principes agiles dans leur gestion, soulignant que les archivistes gèrent des projets mais n'accordent pas suffisamment d'attention à la théorie de la gestion de projet [16]. À la Chambre des Députés, les projets archivistiques sont formalisés de la même manière que ceux des autres domaines, avec des noms et des identifiants, c'est le cas du projet InventAIre qui porte l'identifiant P1134. Un chef de projet de l'équipe du service TI a été désigné pour le suivre, et les outils des méthodes agiles issus de la gestion informatique ont été utilisés. Un diagramme de Gantt [17] a été réalisé. Des comités de projet et de pilotage ont été organisés pour suivre l'avancement et valider différentes décisions. D'autres outils de gestion de projet qui sont quant à eux spécifiques à l'informatique, comme la méthode MoSCoW [18], ont été employés pour hiérarchiser les fonctionnalités de l'outil à produire. Ces outils ont largement contribué aux succès.

Cette hiérarchisation a permis la réalisation du minimum requis. Lorsque des difficultés sont apparues, les fonctionnalités non prioritaires ont été mises de côté. Le projet InventAIre a ainsi démontré l'importance de l'utilisation de méthodes de gestion de projets informatiques et d'une bonne définition du périmètre. En effet, l'ampleur des ressources et du personnel nécessaires à la réussite des projets IA est souvent sous-estimée. Camille Girard-Chanudet évoque en parlant du système basé sur le *machine learning* développé à la Cour de Cassation que « loin de l'image d'autonomie généralement associée à ce type de dispositifs techniques, la charge de travail humain dans le fonctionnement d'un tel outil est conséquente : celui-ci mobilise au quotidien plus de 20 personnes à la Cour [19] ». Le travail d'annotation des données à pseudonymiser demande en effet beaucoup de ressources humaines. L'annotation en *machine learning* consiste à étiqueter ou décrire des données (comme des images, du texte ou des sons) pour que les algorithmes puissent apprendre à

16. Cyndi Shein, Hannah E. Robinson et Hana Gutierrez, « Agility in the Archives : Translating Agile Methods to Archival Project Management », *RBM : A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 19–2 (nov. 2018), p. 94, DOI : 10.5860/rbm.19.2.94.

17. Planning des tâches à accomplir montrant leur durée et leur chevauchement sur une échelle de temps.

18. Hiérarchisation des objectifs de développement d'un outil en quatre catégories : *Must have* (minimum à développer), *Should have* (fonctionnalités que l'on devrait développer mais non prioritaires), *Could have* (pourrait avoir en mettant beaucoup d'efforts), *Won't have* (hors périmètre).

19. C. Girard-Chanudet, « Le travail de l'Intelligence Artificielle… ».

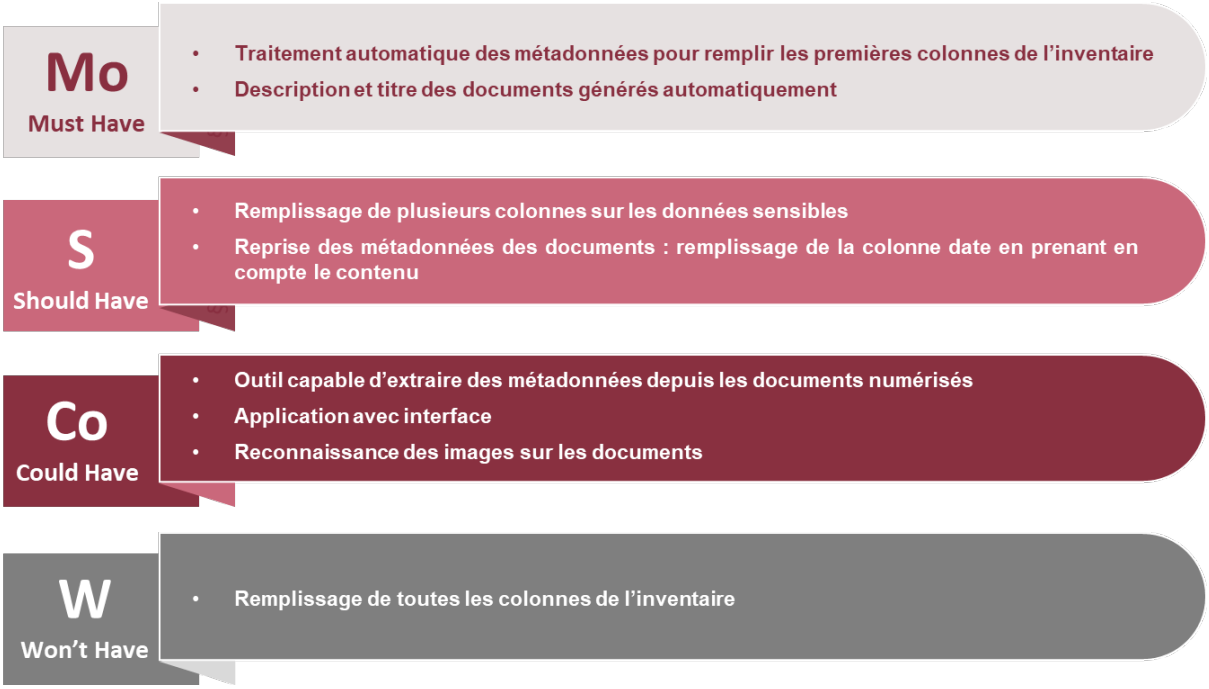

FIGURE 3 – Hiérarchisation des fonctionnalités à produire (MoSCoW)

reconnaître et à traiter ces informations de manière autonome. Cette étape est obligatoire pour le développement de modèles de *machine learning* maison. En plus des personnes pour réaliser les annotations, il faut également mobiliser du personnel pour la gestion de projet et pour le développement de l'outil. Dans le cas du projet InventAIre, nous n'avions pas le temps d'annoter des données suffisantes pour développer un modèle d'IA en quatre mois de stage. Il était donc impossible d'entraîner un modèle d'apprentissage supervisé. Nous avions ainsi comme option l'apprentissage non supervisé ou le choix d'un modèle pré-entraîné. Nous avons pu adapter le périmètre et l'approche. Néanmoins, même dans le cas de l'usage de modèles pré-entraînés, les investissements en temps et en ressources humaines restent non-négligeables. Il faut par exemple prévoir une longue période de tests afin de choisir le bon modèle pré-entraîné et mobiliser du personnel pour évaluer les résultats produits par l'IA. Pour un projet de *chatbot* à la BNL, d'après Yves Maurer, en charge du projet, le « plus chronophage a été de tester plusieurs alternatives pour chaque brique du projet : des modèles de langage ouverts, un autre créé par un groupe de recherche, celui de Meta et de Google… [20] ». Dans le cadre de notre projet la période de tests a pris environ deux mois avec une seule personne mobilisée.

D'autres différences notables existent entre les projets archivistiques et ceux impliquant l'IA. L'approche cyclique est encore plus importante pour l'IA : la précision des outils produits doit être évaluée à chaque étape. Les cycles sont de tailles différentes. Les étapes de test et de réflexion sont assez lentes. Le code de l'outil en lui-même a été as-

20. Clémence Jost, *Comment la BNL a développé son chatbot basé sur ChatGPT*, fr, 16 févr. 2024, URL : `https://www.archimag.com/bibliotheque-edition/2024/02/16/comment-bnl-developpe-son-chatbot-base-sur-chatgpt` (visité le 10/07/2024).

sez rapide pour le projet InventAIre. Les périodes d'évaluation et de reprise de l'outil en fonction des résultats de ces dernières sont quant à elles longues et on ne peut pas prévoir leur nombre d'itérations.

Estimation du temps sur les différentes « tâches IA » pendant le projet

Tâche	**Estimation de temps**
Définition du périmètre et de l'approche	15%
Tests pour le choix d'un modèle	40%
Code sur l'intégration du modèle	10%
Prompt-engineering	15%
Évaluation de l'outil*	15%
Autre	5%

*Ce temps a été réduit parce que le stage se terminait, l'évaluation doit idéalement être plus longue pour être davantage pertinente et mener à des reprises de l'outil

Un inconvénient de ces méthodes agiles est le temps que la gestion de projet exige. Pendant notre stage, les tâches de gestion de projet ont représenté environ 15 % de notre temps [21], incluant la définition du périmètre et du calendrier, la préparation et la tenue des réunions ainsi que la rédaction des comptes rendus. Pour des projets de plus grande envergure, il est essentiel de séparer les fonctions de recherche et développement de celles de gestion de projet entre plusieurs personnes. La constitution d'une équipe projet favorisant la collaboration entre archivistes et spécialistes des technologies de l'information est nécessaire.

Enfin, ces projets nécessitant d'importants investissements humains et financiers et l'intelligence artificielle étant un nouveau territoire à explorer pour les institutions publiques, il est nécessaire de commencer par des projets pilotes, des preuves de concept (POC) ou des études de faisabilité avant de se lancer dans de grands projets. La plupart des institutions suivent ces recommandations. C'est notamment le cas de la BNL (Bibliothèque nationale du Luxembourg) qui a lancé plusieurs projets pilotes ces dernières années, dont un projet d'amélioration de la transcription par *OCR (optical character recognition)* et un projet de *chatbot* permettant la recherche dans les fonds de presse numérisés [22]. À la Chambre des Députés, le projet InventAIre constitue la première étape d'un processus plus large. La phase suivante consiste à réaliser un POC, à partir duquel un outil serait développé et mis en production.

21. Un diagramme de répartition du temps est consultable dans la note méthodologique en annexe.
22. *Ibid.*

L'intégration des méthodes de gestion de projet, en particulier les approches agiles, est par conséquent une composante de la réussite des projets IA dans le secteur des archives. Le projet InventAIre illustre comment une gestion efficace, une collaboration interdisciplinaire et une approche itérative permettent d'être en mesure de naviguer avec succès entre les défis spécifiques à l'IA.

3. Gestion des risques et défis éthiques des systèmes basés sur le *machine learning*

Les risques liés à l'intelligence artificielle, en particulier dans le domaine des archives, sont nombreux, allant bien au-delà des simples considérations légales facilement identifiables. Outre les enjeux juridiques, il existe des risques sociaux et éthiques, qui seront détaillés dans le début du chapitre 8. Face à ces défis, une analyse de risques est à prévoir. A la Chambre des Députés, une mitigation des risques a été réalisée en amont en collaboration avec le *DPO (data protection officer)* et le Responsable sécurité des systèmes informatiques. C'est ainsi qu'il a été décidé qu'il n'était pas question de transmettre des données à un tiers, nous avons ainsi dû travailler avec des modèles pré-entraînés hébergés localement sur nos ordinateurs, et non dans le *cloud*. Cela n'est pas sans conséquence. Ceux que nous pouvions faire tourner étaient en effet moins précis que les grands modèles hébergés dans le *cloud* et très volumineux sur nos machines, donc relativement lents. Les autres risques identifiés concernaient un mauvais remplissage de l'inventaire, contenant des biais ou hallucinations. Les titres et descriptions en texte libre générés par des grands modèles de langage peuvent facilement contenir des erreurs, des informations non pertinentes ou des hallucinations. Au contraire l'IA pourrait aussi invisibiliser certaines informations jugées non pertinentes, entraînant une perte de données essentielles. Enfin, la gestion des données sensibles nécessite une attention particulière : une erreur dans ces données pourrait avoir des conséquences graves. Face à ces risques, il a été décidé qu'un contrôle qualité rigoureux serait effectué. Un tableau contenant les risques et contre-mesures identifiés au début du projet se trouve dans la partie 1.1.2. de la note méthodologique en annexe.

Le chercheur James Lappin propose dans sa thèse intitulée « The science of recordkeeping systems – a realist perspective », une distinction entre les applications « low-stakes » (à faible enjeu) et « high-stakes » (à fort enjeu) de l'IA dans le domaine du *record management*. Les applications « low-stakes » incluent l'utilisation de l'IA pour classer les résultats de recherche, personnaliser les recommandations, visualiser les contenus ou extraire des entités, sans altérer les règles d'accès ou de conservation des documents. En revanche, les applications « high-stakes » impliquent des décisions aux conséquences irréversibles, qui modifient les règles de conservation ou d'accès aux documents, par exemple

via des éliminations ou l'octroi de certains accès [23]. Les autres projets d'intelligence artificielle mis en place dans le domaine des archives et des bibliothèques au Luxembourg sont majoritairement des projet « low-stakes ». C'est par exemple le cas des *chatbots* de la Bibliothèque Nationale du Luxembourg (BNL) et du Parlement européen mentionnés précédemment, qui permettent une recherche dans des documents publics. Quant aux scripts générés par IA des ANLux, ils ont vocation à être partagés, c'est aussi une utilisation avec moins de risques, car sans données confidentielles.

Le projet InventAIre est précurseur en termes d'application « high-stake » de traitement automatique d'archives par IA. Il a ainsi constitué une occasion de tirer plusieurs enseignements précieux concernant la gestion des risques et leur atténuation. D'abord, il est apparu particulièrement intéressant d'impliquer divers acteurs dans le processus, tels que le responsable de la sécurité des systèmes informatiques (RSSI) et le *data protection officer* (DPO). Ces experts dans leur domaine apportent des compétences spécifiques complémentaires à celles des archivistes et des informaticiens. De plus, nous avons cherché à impliquer le plus possible l'humain. C'est le concept de l'« Human in the loop » : l'humain a été intégré dans le processus de développement et d'évaluation de l'IA. Cela est particulièrement pertinent et à développer dans le cadre des sprints agiles mentionnés précédemment, où chaque étape du développement était validée par des retours humains. Cette méthode contribue non seulement à une meilleure qualité des résultats, mais facilite également l'instauration d'une meilleure confiance des différents acteurs impliqués envers l'outil développé. Cette dernière facilite la conduite du changement. L'analyse des risques a joué un rôle dans la construction de cette confiance en garantissant que chaque décision prise soit justifiée. Cela a permis de créer un cadre sécurisé et le plus transparent possible, favorisant ainsi l'adoption et l'acceptation de l'IA dans un domaine aussi sensible que celui des archives.

Les projets d'intelligence artificielle dans le domaine des archives publiques offrent des perspectives prometteuses, mais ils posent également des défis importants. Avant d'initier de tels projets, il convient d'assurer une analyse des risques de manière proactive. Les exigences éthiques et techniques sont nombreuses. Les besoins et usages doivent être clairement identifiés et la gestion de projet réfléchie.

En conclusion de ce chapitre, nous pouvons dire que les producteurs d'archives publiques luxembourgeois font face à de nombreux défis qui, paradoxalement, offrent des opportunités d'innovation et d'expérimentation. Un contexte favorable au Luxembourg et en Europe encourage l'implémentation de projets IA dans le secteur public avec des ambitions élevées. Cependant, avant de lancer de tels projets, des réflexions sur les prérequis en termes de pilotage sont à mener. Une fois ces réflexions approfondies, les projets IA

23. James Lappin, *The science of recordkeeping systems - a realist perspective*, en, thèse, Loughborough University, 2024.

pourront aboutir à des résultats et avoir des apports concrets pour les services d'archives publics.

Deuxième partie

Les apports des projets IA dans les archives : perspectives, état des lieux et synergies

Chapitre 4

Des solutions légères de *machine learning* pour les archives

1. TAL (Traitement Automatique du Langage) pour la classification : *clustering* et *topic modelling*

Le premier travail effectué dans le cadre du stage a été de tester plusieurs moyens de remplir les différentes colonnes de l'inventaire. Nous avons commencé par tester des algorithmes légers d'apprentissage non supervisé. L'apprentissage non supervisé est une technique d'apprentissage qui permet de découvrir des structures ou des modèles dans des données non étiquetées, sans intervention humaine pour guider la machine. L'objectif est de révéler des relations ou des regroupements intrinsèques entre les données. Avant d'utiliser des grands modèles pré-entraînés, il est important d'étudier ce qu'il est possible de faire avec des moyens techniques plus légers. Le TAL (Traitement Automatique du Langage) est une discipline ancienne. Les premières recherches auraient été menées aux débuts de l'informatique, dès les années 1940 [1]. L'article le plus ancien détaillant les potentiels usages du TAL, ou *NLP (Natural langage processing)* dans les archives que nous ayons trouvé date de 1998 et a été publié dans la revue *The american archivist* [2]. Un rapport plus ancien de l'UNESCO intitulé « Regional Training Centre for Archivists, Accra : Africa - (mission). Project findings and recommendations » et publié en 1981, évoque l'idée qu'« en tant que banques de données de documents originaux, les archives ont beaucoup en commun avec les bibliothèques et les centres de documentation, et doivent de plus en plus utiliser des techniques automatisées de traitement des données, de recherche et d'exploitation de l'information, de résumé, d'indexation et de diffusion [3] » [Traduction libre].

1. Thierry Poibeau, « Le traitement automatique des langues pour les sciences sociales », *Réseaux*, 188–6 (2014), p. 25-51, DOI : 10.3917/res.188.0025.

2. Daniel Pitti, « Encoded Archival Description : The Development of an Encoding Standard for Archival Finding Aids », dir. Jackie Dooley, *The American Archivist*, 60–3 (juill. 1997), p. 268-283, DOI : 10.17723/aarc.60.3.f5102tt644q123lx.

3. UNESCO, *Regional Training Centre for Archivists, Accra : Africa - (mission). Project findings and recommendations*, en, 1981, URL : https://unesdoc.unesco.org/ark:/48223/pf0000044236

Le rapport n'évoque pas directement le TAL (Traitement Automatique du Langage) mais les « techniques automatisés » évoquées en découlent. La théorisation de l'usage du TAL dans les archives date donc au moins des années 1980. L'idée a eu le temps de mûrir en plus de quarante ans, et les technologies de s'améliorer.

Dans le cadre du projet InventAIre, nous avons commencé par expérimenter à l'aide d'algorithmes de classification automatique. L'usage de petits algorithmes de TAL par rapport à des grands modèles pré-entraînés a effectivement pour avantage de réduire les besoins en ressources informatiques, de diminuer le temps de calcul et d'éviter qu'un inventaire prenne trop de temps à se générer. La classification automatique est une technique d'apprentissage automatique qui consiste à générer automatiquement des regroupements d'objets en fonction de leurs caractéristiques. Pour que l'algorithme de classification puisse regrouper automatiquement les documents, une représentation mathématique de ces documents doit au préalable avoir été générée, c'est l'étape de vectorisation. Nous avons ainsi vectorisé les textes de chaque document. Nous avons utilisé la méthode TF/IDF (*Term Frequency-Inverse Document Frequency*).

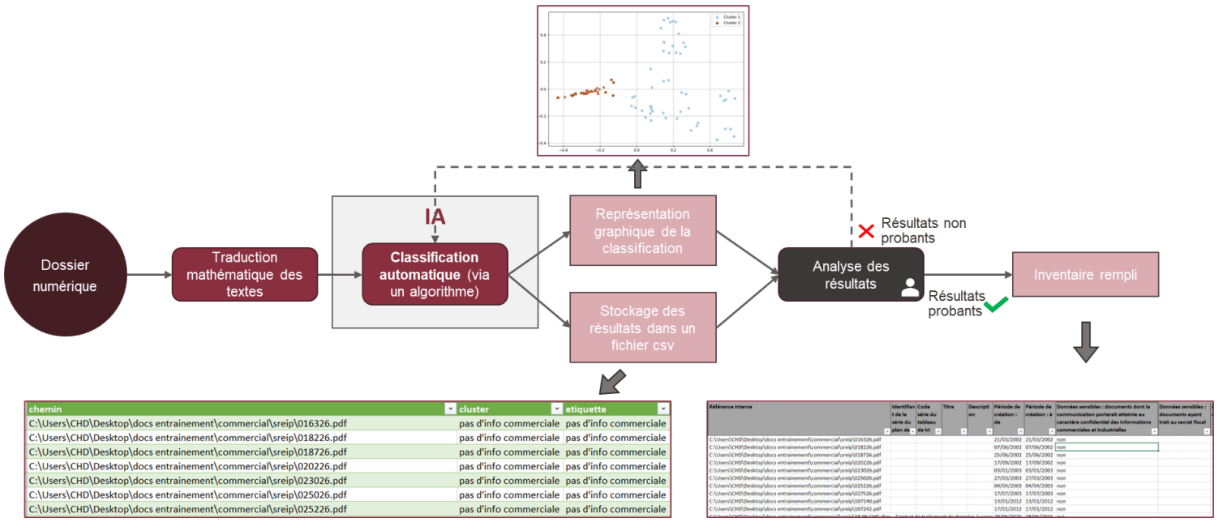

FIGURE 4 – Exemple de processus de classification automatique : Clustering sur des documents contenant des informations à caractère commercial

Il y a plusieurs méthodes de vectorisation. TF/IDF consiste à représenter un texte par un vecteur dont les composantes sont les fréquences des mots dans le document, pondérées par leur importance dans l'ensemble des documents. Cela met en évidence les mots qui ont le plus de poids dans chaque document. Nous avons également défini des *stop words*, qui sont des mots très courants dans une langue ou dans le corpus de document, tels que « le », « la », « les », etc. qui n'apportent pas d'information significative pour la classification. En les éliminant, nous pouvons améliorer la précision de cette dernière. Notre liste de *stop words* comportait les mots les plus fréquents de la langue française et les mots du champs lexical de la législation, qui se retrouvent dans la plupart des documents.

(visité le 06/08/2024).

Nous avons également ignoré les chiffres qui se trouvaient dans les textes. Les vecteurs obtenus sont des représentations numériques des documents, sous la forme de tableaux dont chaque valeur correspond à l'importance d'un mot dans le document.

Une fois ces vecteurs réalisés, il est possible de calculer le texte le plus proche d'un autre, d'étudier les mots qui ont le plus de poids pour les ajouter dans les *stop words* s'ils ne sont pas pertinents, ou d'afficher des groupes de documents les plus proches afin d'identifier d'éventuels sujets. Après cette étape de vectorisation, nous avons travaillé sur la classification des documents en appliquant un algorithme de classification automatique nommé *k-means*. Nous sommes parvenue, après plusieurs essais, à créer des groupes de documents à propos d'affaires juridiques et d'ordre commercial (factures, contrats). Cette méthode était dans une certaine mesure pertinente pour l'inventaire. Nous étions en mesure de détecter des documents qui remplissent la colonne « affaires portées devant des instances juridictionnelles, extra-judiciaires ou disciplinaires » et la colonne « informations commerciales ou industrielles ». Toutefois, les textes dans lesquels la mention de ces informations était plus subtile se voyaient ignorés. Par exemple, un rapport mentionnant de manière implicite une affaire juridique, sans utiliser de termes spécifiques, passait inaperçu. Inversement, une loi ou un document issu d'une question parlementaire sur la justice, donc des documents publics, étaient détectés comme faisant partie des documents sur des affaires juridictionnelles. Le multilinguisme pose également problème. Même si la langue de l'administration parlementaire est le français, le Luxembourg a trois langues officielles : le luxembourgeois, le français et l'allemand. De plus, nous avons travaillé sur un fonds relatif aux relations internationales, ce qui impliquait un certain nombre de documents en anglais. Il aurait fallu répéter l'ensemble du processus dans ces quatre langues. Cette méthode n'a donc pas été choisie pour remplir l'inventaire. Néanmoins, nous en avons retenu les avantages analytiques des différents tests. Nous avons en effet expérimenté la méthode avec des nombres différents de *clusters* à produire par l'algorithme, c'est à dire de groupements de documents générés automatiquement. Un ensemble de paramètres est modifiable dans l'algorithme de classification. L'algorithme *k-means* génère un nombre k prédéfini de *clusters*, en assignant chaque point au cluster dont il est le plus proche du centre (centroïde), lui-même déterminé à partir de la moyenne des coordonnées des points du *cluster*. À chaque itération, nous tentions de comprendre ce qui liait les documents regroupés. Nous avons pu en dégager des thèmes, des types de documents et identifier les différentes langues dans le fonds étudié, puisque les documents se regroupaient aussi par langue. Ces expérimentations de classification ont donc constitué une phase productive d'analyse du contenu du fonds. On pourrait imaginer cette expérimentation comme partie intégrante de la première étape des pré-traitements de l'archivage numérique, qui consiste à étudier comment est constitué le fonds, notamment pour visualiser ce qui est à éliminer. Les techniques de TAL (Traitement Automatique du Langage) sont expérimentées dans d'autres projets dans cette même optique analytique.

Par exemple, l'application Pêle-mél fournit un outil d'analyse des messageries grâce à un système de classification automatique [4].

D'autres méthodes de classification automatique existent. Plusieurs ont été testées par des chercheurs indiens sur des jeux de données d'articles de presse et de critiques de films. Leur objectif était d'analyser les résultats produits par les différentes méthodes sur des longs documents [5]. Dans cette étude, dix méthodes différentes ont été essayées pour classer automatiquement des textes longs. Certaines utilisent des modèles de langage déjà entraînés sur de grandes quantités de texte, qu'on adapte ensuite à la tâche spécifique de tri des documents, comme les méthodes *ULMFiT* et *USE (Universal Sentence Encoder)*. D'autres s'appuient sur des réseaux neuronaux, un type de modèle inspiré du cerveau humain, qui analysent les mots dans l'ordre où ils apparaissent et utilisent des mécanismes d'attention pour se concentrer sur les parties les plus importantes du texte. Il y a aussi des techniques plus classiques, qui comptent la fréquence des mots et les pondèrent en fonction de leur importance pour le document, comme *TF/IDF*, que nous avons utilisé, adopté dans l'étude avec l'algorithme de classification *Naive Bayes*. Une méthode hiérarchique a également été testée, analysant d'abord les mots, puis les phrases, pour comprendre le texte dans son ensemble [6]. Les méthodes de classification automatique sont donc très diverses. Les résultats de l'étude montrent que les gros modèles pré-entraînés tendent à offrir de meilleures performances. Cependant, pour la plupart des jeux de données, des modèles plus simples [7] peuvent fonctionner avec une perte minimale en précision.

Les modèles pré-entraînés testés dans le cadre de l'étude sont *BERT* et *DistilBERT*. Il existe en effet depuis quelques années des moyens plus précis de vectoriser des documents que les calculs mathématiques de méthodes comme TF/IDF. Il est possible de réaliser une vectorisation qui aurait une valeur sémantique grâce aux *embeddings* des grands modèles pré-entrainés tels que les *LLM (Large language models)*). Contrairement à des méthodes comme TF/IDF, ces *embeddings* capturent la signification contextuelle des mots et rapprochent les termes ayant des significations similaires dans un même espace vectoriel. Ces représentations sont générées par des grands modèles de langage entraînés sur de vastes corpus de textes, ce qui leur permet d'appréhender des relations complexes entre les mots et d'offrir une compréhension plus fine et plus nuancée du contenu sémantique des documents. Ils gèrent également le multilinguisme lorsqu'ils sont entraînés sur des corpus en plusieurs langues. Des exemples de modèles d'*embeddings* sont ceux de *BERT*, développé

4. Bénédicte Grailles, *Pêle-mél. Plate-forme d'exploration, de livraison et d'évaluation des méls. Rapport d'évaluation des usages*, 31 mars 2023.

5. Vedangi Wagh, Snehal Khandve, Isha Joshi, Apurva Wani, Geetanjali Kale et Raviraj Joshi, « Comparative Study of Long Document Classification », dans *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, arXiv :2111.00702 [cs], 2021, p. 732-737, DOI : 10.1109/TENCON54134.2021.9707465.

6. *Ibid.*

7. Ex : les embeddings de GloVe ou les réseaux LSTM peu profonds

par Google, ainsi que ceux des *LLM (Large language models)* tels que *GPT-3* et *GPT-4* par *OpenAI*, ou *LlaMa 2* par *Meta*. Une autre étude a été réalisée par des chercheurs et des chercheuses norvégiens pour tester l'efficacité de différentes méthodes de vectorisation avec de la classification pour regrouper des articles de presse à propos de mêmes événements [8]. Les meilleures performances étaient obtenues avec des *embeddings* de *LLM*, puis ceux de *BERT*, suivis par les algorithmes plus légers et classiques comme *TF/IDF* et *GloVE*[9] *(Global Vectors for Word Representation)* [10]. L'inconvénient principal de ces gros modèles est leur exigence en puissance de calcul, nécessitant souvent des équipements spécialisés comme des cartes graphiques (*GPU*). Pour répondre à ce défi, des versions optimisées des grands modèles de langage ont été développées. Par exemple, *DistilBERT* est une version allégée de *BERT* qui est 40 % plus petite, fonctionnerait 60 % plus rapidement mais qui conserverait environ 95 % des performances du modèle complet [11]. Ce type de modèles n'a pas été testé dans le cadre du projet *InventAIre*. Nous avons principalement utilisé les grands modèles de langage pour la génération de texte, sans tester leurs *embeddings* pour des tâches de classification. Il pourrait être intéressant de comparer leurs performances avec celles de la méthode *TF/IDF* sur des documents d'archives pour voir à quel point ils sont davantage précis. La fourniture de données annotées peut également améliorer significativement les performances des algorithmes, il s'agit dans ce cas d'un apprentissage supervisé. Un projet mené par des chercheurs brésiliens et américains sur la classification automatique de documents contenant des secrets d'État a par exemple montré des résultats prometteurs, avec une précision de 90 % et environ 11 % de faux positifs [12]. Avec de grandes quantités de données, leur méthode pourrait être applicable au remplissage automatique de l'inventaire des ANLux.

Une autre perspective réside dans l'extraction de groupes de mots significatifs à partir de groupements de documents. C'est le concept du *topic modelling*. Cette méthode permet d'identifier des thèmes ou sujets sous-jacents dans un ensemble de textes en regroupant des mots qui apparaissent fréquemment ensemble. La technique la plus couramment utilisée pour le *topic modelling* est *LDA* (*Latent Dirichlet Allocation*) [13]. *LDA* fonctionne en attribuant chaque document à une combinaison de sujets et en représentant

8. Adane Nega Tarekegn, *Large Language Model Enhanced Clustering for News Event Detection*, arXiv :2406.10552 [cs], juill. 2024, DOI : 10.48550/arXiv.2406.10552.

9. Technique d'apprentissage automatique pour créer des représentations vectorielles d'entités textuelles en capturant des relations sémantiques entre mots à partir de leurs co-occurrences dans un corpus de texte

10. *Ibid.*

11. Huggingface, *DistilBERT*, URL : https://huggingface.co/docs/transformers/model_doc/distilbert (visité le 12/08/2024).

12. Renato Rocha Souza, Flavio Codeco Coelho, Rohan Shah et Matthew Connelly, *Using Artificial Intelligence to Identify State Secrets*, 1er nov. 2016, DOI : 10.48550/arXiv.1611.00356, arXiv : 1611.00356[cs].

13. Adrian Velonis, « Topic Modeling, Named-Entity Recognition, and Network Analysis of Literary Corpora » (, 2022), URL : http://hdl.handle.net/10066/24507 (visité le 15/07/2024).

chaque sujet comme une distribution de mots. De la même manière qu'avec l'algorithme *k-means*, l'utilisateur doit définir au préalable le nombre de groupes, donc le nombre de sujets souhaités. *LDA* distribue alors les documents et les mots entre ces sujets de manière à maximiser la cohérence sémantique des groupes formés. Cela permet de dégager les thèmes principaux abordés dans un corpus de manière automatisée, et sert ainsi à l'analyse de grands corpus de texte.

Ces méthodes de TAL peuvent fournir des résultats mais nécessitent une série de tests et d'évaluations pour sélectionner la méthode la plus appropriée, ce qui peut rapidement devenir déroutant. Pour automatiser une tâche complexe, comme la rédaction d'un inventaire, avec des ressources limitées, il est souvent nécessaire d'utiliser une combinaison de plusieurs méthodes ou modèles. Pour obtenir une bonne précision, il faudra affiner les modèles sur des sous-tâches. Dans le cas du projet InventAIre, une colonne aurait donc correspondu à un algorithme ou modèle avec des paramètres qui lui étaient propres.

2. La reconnaissance d'entités nommées

Nous avons par ailleurs durant le stage eu l'occasion de tester la reconnaissance d'entités nommées pour mesurer son efficacité dans la détection des données à caractère personnel. La reconnaissance d'entités nommées ou *NER (Named Entity Recognition)* est une méthode qui consiste à identifier automatiquement les entités dans un texte. Elles peuvent être, entre autres, des noms de personnes, de lieux, d'organisations ou encore des dates. Des bibliothèques existent en langage Python pour implémenter la reconnaissance d'entités nommées, telles que *SpaCy* ou *NLTK*, très utilisées dans le domaine du traitement automatique du langage naturel. Ces bibliothèques sont basées sur des modèles pré-entraînés capables d'identifier et de classifier automatiquement les entités dans un texte en analysant le contexte des mots dans la phrase pour déterminer leur rôle. Elles ont l'avantage d'être gratuites, ouvertes et faciles d'utilisation. Il est possible d'obtenir rapidement des résultats avec relativement peu de code. De plus, pour des besoins spécifiques, il est envisageable de ré-entraîner les modèles de *NER* sur des corpus personnalisés. Cela améliore la précision des résultats dans des domaines particuliers ou sur des types d'entités spécifiques. Nos expérimentations de *NER* ont suggéré qu'elle était efficace sur la reconnaissance des noms mais malgré ses atouts, elle s'est révélée insuffisante pour remplir la colonne intitulée « données à caractère personnel » dans l'inventaire. En effet, la loi luxembourgeoise définit une donnée à caractère personnel comme :

> Toute information de quelque nature qu'elle soit et indépendamment de son support, y compris le son et l'image, concernant une personne identifiée ou identifiable (« personne concernée ») ; une personne physique ou morale est réputée identifiable si elle peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs

> éléments spécifiques, propres à son identité physique, physiologique, génétique, psychique, culturelle, sociale ou économique [14].

La définition est donc très vaste et s'étend bien au-delà des noms, des adresses e-mail ou encore des numéros d'identification personnelle.

Nous avons également testé la *NER* dans une perspective de pseudonymisation de documents. L'objectif était de pseudonymiser des noms dans les titres de répertoires issus des dossiers RH afin d'être en mesure de sélectionner des documents que nous pourrions utiliser pour la réalisation de nos tests d'apprentissage machine ou *machine learning*. Nous avons développé un script qui remplaçait automatiquement les noms de personnes, les adresses, les adresses mail, les noms d'organisations, mais il n'a pas été utilisé. La définition d'une donnée à caractère personnel illustre bien le fait que pour pseudonymiser efficacement, il faut détecter et modifier des informations plus subtiles que des noms, adresses mail, etc. De nombreux outils basés sur des modèles plus complexes ont malgré tout été lancés pour anonymiser les décisions de justice en Europe. C'est le cas par exemple à la Cour de Cassation en France dans le cadre d'un projet mentionné précédemment [15]. Ce type de modèle a nécessité une grande quantité de données d'entraînement annotées pour être performant.

L'usage principal de la *NER (Named Entity Recognition)* dans les archives ne concerne pas la recherche de données personnelles ni la pseudonymisation, mais l'indexation. C'est un objectif du projet *NER4Archives* porté par les Archives nationales de France et l'Inria (Institut national de recherche en sciences et technologies du numérique) débuté en 2020. Ses objectifs sont la « conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD [16] ». Le projet part du constat que l'indexation des inventaires EAD français est souvent insuffisante. Or, une bonne indexation facilite la recherche dans les archives et la mise en relation des instruments de recherche. L'architecture de *NER* choisie est un affinage de celle offerte par la librairie *Spacy*. Cette dernière utilise *CamemBERT* comme base, un modèle de langage pré-entraîné basé sur l'architecture *BERT* mentionnée précédemment, optimisé pour la compréhension et le traitement du texte en français. Le modèle a par la suite été affiné dans le cadre du projet sur des instruments de recherche *EAD* annotés. Pour obtenir des résultats précis en *NER*, il est effectivement souvent né-

14. Chambre des Députés du Grand-Duché de Luxembourg, *Loi du 2 août 2002 relative à la protection des personnes à l'égard du traitement des données à caractère personnel.* 2 août 2004, URL : `https://legilux.public.lu/eli/etat/leg/loi/2002/08/02/n2/jo` (visité le 17/07/2024).

15. C. Girard-Chanudet, « Le travail de l'Intelligence Artificielle… ».

16. Florence Clavaud, Laurent Romary, Pauline Charbonnier, Lucas Terriel, Gaetano Piraino et Vincent Verdese, « NER4Archives (named entity recognition for archives) : Conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD. » Dans *Atelier Culture-INRIA*, Pierrefitte sur Seine, France, 2022, URL : `https://hal.science/hal-03625734` (visité le 01/07/2024).

cessaire de travailler sur des données annotées. Dans le cadre du projet *NER4Archives*, ce processus d'annotation, qui a duré quatre mois, a mobilisé quatre annotateurs. Il a permis d'obtenir un F1 score moyen de 0,91 sur l'architecture la plus performante, en l'occurrence celle de SpaCy [17]. Il s'agit d'une bonne performance. Cette précision accrue grâce aux données annotées est également mise en avant dans le projet *NewsEye*, dont le but était de produire un je de données multilingue annoté pour la reconnaissance automatique d'entités nommées [18].

Dans le processus de traitement de *NER4Archives*, le processus de *NER* est suivi d'un processus de *NEL (Named Entity Linking)* pour désambiguïser les entités identifiées et ainsi éviter les doublons. La *NEL* peut effectivement permettre d'associer chaque entité à une entrée unique dans un thesaurus ou une base de données. Au delà des tâches d'indexation, les traitements sur les entités nommées peuvent offrir la possibilité de nettoyer les thesaurus. C'est un besoin qui se manifeste à la Chambre des Députés et dans beaucoup d'administrations. À la Chambre, par exemple, le thesaurus aurait besoin d'être actualisé. Le personnel chargé de l'indexation des documents législatifs ajoute régulièrement de nouveaux termes en fonction des documents traités, mais cela peut entraîner la création de doublons. Par exemple, le terme « Covid-19 » a dû être ajouté parce qu'il était absent du thesaurus. Une autre personne pourrait quant à elle ajouter « Covid » ou « Coronavirus », ce qui crée des entrées équivalentes dans le thésaurus et complique la recherche et l'organisation des documents.

La reconnaissance d'entités nommées semble être un outil puissant, en particulier en ce qui concerne l'indexation. En complément de la classification automatique et du *topic modelling*, la *NER (Named Entity Recognition)* peut contribuer à fournir une meilleure description et de meilleurs possibilités de recherche par entité dans les archives. Ces outils techniques ont également du potentiel en termes d'évaluation de gros corpus pour détecter des documents à éliminer. La *NER* est par exemple intégrée au logiciel de traitement des mails *ePADD* en partie dans cette optique de tri [19]. Malgré leur complexité, ces méthodes offrent une approche analytique approfondie des fonds.

3. Le traitement automatique sur les images

Le dernier type de traitement automatique via des algorithmes plus légers que les grands modèles de langage est le traitement des images. La pratique de la reconnaissance

17. *Ibid.*

18. Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno et Antoine Doucet, « A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers », dans *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2021 (SIGIR '21), p. 2328-2334, DOI : 10.1145/3404835.3463255.

19. Christopher A. Lee, « Computer-Assisted Appraisal and Selection of Archival Materials », dans *2018 IEEE International Conference on Big Data (Big Data)*, 2018, p. 2721-2724, DOI : 10.1109/BigData.2018.8622267.

automatique de caractères est l'usage du *machine learning* le plus répandu dans les services d'archives, parce qu'elle nécessite peu de connaissances techniques, étant intégrée à de nombreux logiciels. Les Archives nationales du Luxembourg et la BNL travaillent par exemple sur des projets d'*HTR (Handwritten Text Recognition)* à l'aide du logiciel *Transkribus*. Nous avons intégré une étape d'*OCR (optical character recognition)* à notre traitement de données pour générer l'inventaire. A chaque nouveau document traité, si le texte n'est pas disponible, l'outil développé tente de l'extraire. Pour cela, nous transformons le document en image dans notre environnement Python, avant d'utiliser la librairie *Tesseract* pour en extraire le texte. Ce traitement excluait les documents contenant des écritures manuscrites, qui sont loin d'être majoritaires dans les vracs bureautiques, mais on pourrait imaginer la reconnaissance automatique sur ces fichiers dans le futur si le projet continue.

Nous n'avons pas réalisé d'autres traitements sur les images dans le cadre du projet InventAIre, mais nous y avons réfléchi. Nous avons par exemple émis l'hypothèse que les actes notariés pourraient être reconnus automatiquement grâce à la présence d'un tampon de notaire, de même pour les actes d'état civil. Un modèle de détection de tampons a été mis en place par un groupe de chercheurs. Il fonctionne avec l'algorithme de détection d'objets *YOLO (You Only Look Once)* comme base. Ce modèle est décrit comme « efficace, contenant un petit nombre de paramètres et [pouvant] être exécuté rapidement sur des appareils mobiles[20] »[Traduction libre]. Les modèles basés sur l'analyse d'image sont en effet moins volumineux que les modèles de traitement automatique du langage car ils opèrent sur des pixels, donc des données en deux dimensions minimum (pixel noir ou blanc). Des filtres peuvent permettre de réduire le nombre de dimensions en cas de traitement d'images en couleur. Les modèles de TAL doivent quant à eux modéliser des relations complexes dans des séquences de texte, et possèdent donc bien plus de paramètres. Nous aurions pu imaginer le développement d'un modèle détectant les tampons liés à l'État civil ou de notaires, qui, pour être plus précis, aurait été associé à du TAL. Ces modèles qui mélangent les deux types de traitement sont des modèles multimodaux. Des chercheurs ont travaillé sur ce type d'outils pour la classification de documents d'archives turques, développant un algorithme de classification qui classifie à la fois les images et le contenu textuel après océrisation[21]. La précision de l'algorithme serait supérieure à 96 %[22], ce qui est prometteur.

20. A. Gayer, D. Ershova et V. Arlazarov, « Fast and Accurate Deep Learning Model for Stamps Detection for Embedded Devices », *Pattern Recognition and Image Analysis*, 32–4 (déc. 2022), p. 772-779, DOI : 10.1134/S1054661822040046.

21. Gürcan Durukan, Meryem Tuğba Nar, Abdullah Özcan, Lütfü Çakıl et Hüseyin Kara, « Multimodal Classification Algorithm for Turkish Document Archiving : Improving Digital Document Storage by Unifying Image and Text-Based Classifiers », dans *Innovative Methods in Computer Science and Computational Applications in the Era of Industry 5.0*, dir. D. Jude Hemanth, Utku Kose, Bogdan Patrut et Mevlut Ersoy, Cham, 2024, p. 1-12, DOI : 10.1007/978-3-031-56322-5_1.

22. *Ibid.*

Le traitement des images et la détection d'objets sont par ailleurs utiles à l'indexation des documents graphiques. Un projet a été lancé sur l'indexation des photos du gouvernement par le Service information et presse du gouvernement du Luxembourg dans le cadre de l'initiative *AI4Gov* mais nous avons trouvé peu d'informations sur ce dernier [23]. Ce type de projets se multiplie dans le monde des bibliothèques et des archives. En France, le prototype de *GallicaPix* a été lancé en 2021. Basé sur la détection d'objets, il sert à effectuer des recherches par contenu iconographique dans les contenus de la BnF. Ce genre de projets pourrait être aussi mené dans les archives pour indexer des images par contenu par exemple. L'intelligence artificielle pour le traitement des images a été exploitée dans le domaine de l'*OCR (optical character recognition)* et de l'*HTR (Handwritten Text Recognition)*, mais reste encore à explorer pleinement, notamment pour l'indexation automatique de contenu audiovisuel.

4. Des usages sur les tâches aux impacts moins élevés : recherche, indexation et découvrabilité

Comme nous l'avons vu, de nombreuses méthodes de traitement automatique peuvent être appliquées sur des fonds numérisés ou nativement numériques. Elles ont chacune leurs avantages et leurs inconvénients. La classification automatique et le *topic modelling* sont par exemple complexes à appréhender techniquement. Ils reposent sur des logiques mathématiques dont il faut être en mesure de comprendre les bases. Des précautions sont par exemple à prendre au moment d'appréhender les graphiques représentant des résultats de classification automatique. Les vecteurs peuvent être des vecteurs de plusieurs milliers de dimensions, or, ils sont réduits à deux dimensions pour être représentés sur un graphique. Des métriques sont par conséquent à calculer pour vérifier la qualité de la réduction de dimension, pour s'assurer que les groupes formés sont réellement aussi groupés qu'ils ne l'apparaissent sur le graphique. Un autre inconvénient de la classification est le fait que les petits groupes sont plus difficiles à identifier par la machine. Un certain bagage technique ou des recherches sont nécessaires avant de prendre en main ce genre d'algorithmes et des précautions sont à prendre avant de diffuser des graphiques de classification. Les limites sont également d'ordre matériel : les besoins en termes de puissance de calcul augmentent plus les modèles sont gros et plus les corpus à traiter sont grands. On peut s'interroger sur la nécessité d'utiliser les *embeddings* de gros modèles pré-entraînés pour la classification si l'objectif est analytique. En effet, des groupes se dégagent avec des méthodes classiques. Dans le cadre du projet *InventAIre*, la méthode *TF/IDF* a permis d'identifier des groupes pertinents et de mieux appréhender le contenu du fonds à traiter. Sur le plan analytique, une meilleure précision dans les groupes n'aurait peut-être pas eu un grand impact. Il

23. Ministère de la Digitalisation, *L'initiative AI4Gov*, fr, text, mars 2021, URL : `http://mindigital.gouvernement.lu/fr/dossiers/2021/AI4Gov.html` (visité le 24/07/2024).

s'agit potentiellement d'une piste à explorer.

La classification automatique présente par ailleurs l'avantage d'être une méthode rapide à coder. La partie code de la classification nous a demandé deux heures tout au plus. Le plus long est de tester avec différents paramètres, d'examiner les différents groupes à chaque itération et de créer des sauvegardes quand les résultats sont pertinents. Il faut tâtonner pour obtenir des résultats. Cet aspect a été souligné par Seth van Hooland et Mathias Coeckelbergs dans un article explorant les possibilités et limites du *topic modelling* et du *word embedding*. Ils évoquent que les paramètres et les termes inclus en tant que *stop words* ont un impact non négligeable sur les résultats [24]. Ils parlent du caractère « boîte noire » de ces méthodes [25] et concluent en disant que l'application des techniques d'apprentissage automatique présente une « nature semi-automatisée » et qu'« à des étapes cruciales du processus, les experts en archivistique doivent encore prendre des décisions stratégiques et intervenir manuellement » [26][Traduction libre]. Il en va de même pour le *topic modeling* : les « résultats [sont] satisfaisants mais obscurs [27] »[Traduction libre]. Les limites et les avantages sont similaires à ceux de la classification. Les deux servent au *disant reading* [28], c'est à dire à découvrir des motifs ou des tendances dans des grands corpus sans avoir à les explorer dans leur intégralité, mais la production de résultats satisfaisants demande des expérimentations et un grand travail d'analyse.

Les perspectives archivistiques des modèles de *NER (Named Entity Recognition)*, de *NEL (Named Entity Linking)* et de la classification automatique sont diverses. Néanmoins, comme évoqué en première partie, il est important de définir des cas d'usage pertinents pour ces technologies, de s'interroger sur la nature des usages qui offrent le meilleur équilibre entre facilité de mise en production et valeur ajoutée. Comme nous avons pu le voir, la classification automatique et le *topic modelling* ne sont pas une solution miracle pour automatiser un traitement aussi complexe que le classement d'unités de description dans l'inventaire des ANLux. Ils sont utiles néanmoins pour visualiser les documents similaires dans les archives, identifier les mots significatifs et ainsi différents sujets dans le corpus, et constituent des outils d'analyse de ce dernier dans sa globalité. Ils peuvent être également intéressants dans une optique de découvrabilité, par exemple par la proposition du document le plus proche d'un autre dans l'espace vectoriel à un lecteur. Cet usage a un impact moins important en cas d'erreur que des données sensibles non repérées.

24. Seth Van Hooland et Mathias Coeckelbergs, « Unsupervised machine learning for archival collections : Possibilities and limits of topic modeling and word embedding », *Revista catalana d'arxivística*, 41 (2018), p. 73, URL : `https://arxivers.com/wp-content/uploads/2018/10/1.4_-Dossier_SVHooland_MCoeckelbergs.pdf` (visité le 06/07/2024).

25. *Ibid.*

26. *Ibid.*

27. Elijah Meeks et Scott B. Weingart, « The Digital Humanities Contribution to Topic Modeling », *Journal of Digital Humanities*–1 (2012), URL : `https://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/` (visité le 17/07/2024).

28. *Ibid.*

Ces techniques peuvent également servir à la recherche, en produisant des *dashboards* de visualisation de documents ou de groupements thématiques. Il en va de même pour le traitement automatique des images, qui permet une meilleure indexation et description de leur contenu, mais qui nécessite plus de recherches et de développement avant d'être efficace sur l'automatisation de tâches archivistiques complexes. La question de la découvrabilité et de l'efficacité des outils de recherche n'est pas prioritaire au Luxembourg mais demeure un défi archivistique. De même, en France, la question se pose dans les bibliothèques, mais concernant les Archives nationales, Bruno Ricard a rappelé lors de la Journée des archivistes luxembourgeois le 7 juin 2024, que la priorité était de bâtir un système d'information archivistique solide, avant de se pencher plus précisément sur ce sujet et celui de l'intelligence artificielle. Il est cependant possible de s'interroger sur le fait qu'une excellente découvrabilité et des moteurs de recherche efficaces puissent compenser une description archivistique insuffisante.

D'un point de vue technique, les limites matérielles et intellectuelles se manifestent rapidement en ce qui concerne le TAL (Traitement Automatique du Langage) et la classification d'images. Comme mentionné précédemment, il est souvent nécessaire de diviser les grandes tâches en sous-tâches plus spécifiques, chacune étant traitée par un modèle distinct. Bien que l'utilisation d'algorithmes légers permette d'automatiser, ou au moins de gagner du temps sur certaines tâches, elle complexifie la chaîne de traitement des archives. Il serait donc pertinent de mesurer l'efficacité de cette approche par rapport à l'utilisation de modèles plus généraux, comme les *LLM (Large language models)*, plus volumineux mais potentiellement plus efficaces et générant moins de complexités dans la chaîne de traitement.

Chapitre 5

Les grands modèles de langage : un moyen efficace d'automatisation de tâches archivistiques ?

1. Les promesses des *LLM (Large language models)*

Les promesses des *LLM (Large language models)* sont nombreuses. Ces derniers sont entraînés sur des corpus considérables de texte dans le but de pouvoir prédire des combinaisons de mots en réponse à une question ou une affirmation. Pour donner un ordre d'idée de ces quantités, Llama 3 serait entraîné sur un corpus de plus de quinze trillions de *token*[1]. Un token est une unité de texte traitée comme une séquence distincte par le modèle, il peut correspondre à un mot, une partie de mot ou un symbole. Il est généralement admis qu'un *token* est en moyenne équivalent aux trois quarts d'un mot. Les données d'entraînement de la base du modèle Llama 3 contiendraient donc environ onze trillions de mots, soit environ 21 000 milliards de fois le contenu des Misérables de Victor Hugo, ou bien 1,5 milliard de fois le wikipédia français[2]. Ces grandes quantités de données, provenant majoritairement du web, permettent aux modèles d'avoir des connaissances générales dans des domaines très divers. Ils sont davantage capables de générer des réponses en prenant en compte des contextes, qui sont alors plus pertinentes et cohérentes que des modèles spécialisés sur une tâche. Ils analysent la relation entre les mots et leur signification dans une phrase ou un paragraphe et le contexte de cette dernière. Par exemple, si une question est posée dans un contexte juridique, le modèle pourra adapter sa réponse en fonction de ce domaine spécifique. Par ailleurs, certains grands modèles de langage sont simples à utiliser grâce à des interfaces de discussion, on parle alors de *chatbot*s ou d'IA générative conversationnelle. C'est le cas de *Chagpt* d'*OpenAI.* En plus de

1. Meta, *Introducing Meta Llama 3 : The most capable openly available LLM to date*, 18 avr. 2024, URL : `https://ai.meta.com/blog/meta-llama-3/` (visité le 02/08/2024).

2. Au 21 août 2024, à 16h30, le wikipédia français est composé de 2 630 307 articles d'une moyenne de 2 758 mots d'après cette source : `http://fr.wikicount.net/`

cette interface, il possède une API (*Application Programming Interface*), permettant aux développeurs d'intégrer ses capacités dans leurs applications et services, pour réaliser des automatisations de plus grande envergure. Il est également possible d'installer localement des modèles *open source*, pour éviter de passer par des API et ainsi de transmettre des données. Les avantages de l'IA générative conversationnelle sont nombreux. Le code est beaucoup plus léger que pour l'entraînement de modèles maison. Il nécessite moins de compétences techniques et il est plus aisé de réaliser des expérimentations. Les utilisateurs peuvent tester ou affiner leurs prompts, c'est à dire les questions posées au *LLM*, directement en ligne pour certains modèles et souvent sans frais. Les prix pour l'inférence via l'API des modèles payants ne sont pas forcément aussi élevés qu'on pourrait le croire. Au moment de la rédaction de ce mémoire, le prix d'un million de *token*s (environ 750 000 mots) est de 30 $ en entrée (c'est à dire dans le prompt) et 60 $ pour un million en sortie pour gpt-4 [3].

Des réponses précises peuvent être générées par ces modèles en un ou plusieurs prompts et il est possible de les guider pour les améliorer. C'est le concept du *few-shots learning* : l'utilisateur fournit des exemples de la tâche à remplir par le modèle dans son prompt pour guider ce dernier. Ces possibilités évitent d'avoir à développer des modèles maison spécialisés. Si les réponses ne sont pas assez précises, on peut aussi *fine-tuner*, c'est à dire affiner les modèles sur des tâches spécifiques en les entraînant sur un corpus de prompts et de réponses à ces derniers mais cela demande beaucoup de temps et de capacités matérielles [4]. De nouvelles recherches sont constamment menées sur les manières d'améliorer les réponses générées. Une discipline, le *prompt engineering*, a émergé : il s'agit de l'optimisation des prompts afin d'obtenir les meilleures réponses possibles, via l'ajout d'exemples dans le *prompt*, la contextualisation et beaucoup d'autres techniques. Les modèles évoluent aussi rapidement que la recherche avance donc il faut constamment être à jour sur les dernières publications et recherches menées, d'où la présence de nombreuses références à des *preprints* dans ce mémoire. Les plus grands modèles ont par ailleurs l'avantage de gérer le multilinguisme et peuvent réaliser des tâches de traduction. 5 % des données d'entraînement de Llama 3 sont « des données de haute qualité non anglophones couvrant plus de trente langues [5] » [traduction libre]. Des modèles multimodaux capables de traiter du texte et de l'image, voire du son et de la vidéo font en outre leur apparition.

Les possibilités d'usage de ces modèles semblent infinies. En ce qui concerne le traitement des archives, un article de chercheurs de l'université de Wuhan et de Loughborough

3. Les tarifs d'OpenAI sont indiqués ici : `https://openai.com/api/pricing/`

4. Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai et Quoc V. Le, *Finetuned Language Models Are Zero-Shot Learners*, arXiv :2109.01652 [cs], févr. 2022, DOI : `10.48550/arXiv.2109.01652`.

5. Meta, *Introducing Meta Llama 3 : The most capable openly available LLM to date…*

liste des perspectives d'usage [6], que nous résumerons en plusieurs catégories :

— Contenu des documents : océrisation, correction d'océrisation

— Description : repérage et génération de métadonnées (titres, dates), indexation, résumé

— Communication : repérage de données sensibles, détermination de durées de rétention, gestion des accès utilisateurs

— Médiation : réponses aux questions des chercheurs sur les archives et le domaine de l'archivistique, aide à la recherche, recommandations

— Reporting : statistiques sur les documents conservés, sur les interactions avec les usagers, prédictions sur les documents qui seront consultés

Les *LLM (Large language models)* peuvent donc être utilisés à différentes étapes du traitement des archives. Les perspectives sont séduisantes, mais qu'en est-il de la pratique ? Les *LLM* sont des bons outils pour les tâches simples liées au langage mais ne le sont pas pour les tâches complexes de raisonnement [7]. Les *LLM* fonctionnent effectivement en se basant sur des raisonnements mathématiques pour prédire les combinaisons de mots les plus probables en réponse à une entrée donnée. Leur « intelligence » apparente résulte d'une analyse mathématique de vastes ensembles de données textuelles, sans processus de compréhension ou réflexion. Cela limite leur efficacité sur des tâches nécessitant un raisonnement complexe. Les *LLM* sont donc avant tout des outils puissants pour générer du texte, mais ils restent au fond des systèmes de prédiction. Face à des modèles généraux qui seraient incapables de véritablement raisonner, et donc dépourvus de l'intelligence artificielle qui leur est attribuée, quelles contributions concrètes les *LLM* peuvent-ils réellement offrir dans le domaine des archives ? Quelles sont les perspectives concrètes d'automatisation ?

2. La réalité archivistique : les possibilités d'automatisation dans les processus métier

Il convient ici d'étudier les réelles possibilités d'automatisation de processus métier grâce aux *LLM*, en commençant par présenter leurs apports dans le cadre du projet InventAIre. Pour ce projet, après des tests sur plusieurs modèles, présentés dans la note méthodologique en annexe, notre choix s'est porté sur l'usage du *LLM Llama 3 7B* de *meta*, quantisé [8], c'est-à-dire avec une précision réduite. Nous ne l'avons pas *fine-tuné*.

6. Shitou Zhang, Siyuan Peng, Ping Wang et Jingrui Hou, « Archives Meet GPT : A Pilot Study on Enhancing Archival Workflows with Large Language Models », *iConference 2024 Proceedings* (, mars 2024), URL : `https://hdl.handle.net/2142/122806` (visité le 30/08/2024).

7. Spencer Torene, *Do LLMs Reason ?*, en, 9 oct. 2023, URL : `https://medium.com/@spencertorene/do-llms-reason-d33fa885872f` (visité le 10/08/2024).

8. cf. quantisation dans le glossaire.

Nous posions un *prompt* par colonne à remplir dans l'inventaire en fournissant du contexte et le texte d'un document ou d'une partie de document en entrée. Un exemple de prompt se trouve dans le chapitre 10.

Les statistiques de précision réalisées sur l'inventaire d'une arborescence de 420 répertoires et documents sont les suivantes :

Qualité de la réponse	**Ensemble des colonnes**	**Titre et description**	**Données sensibles**
Correcte	77%	71%	79%
Incomplète/sujette à débat	10%	22%	6%
Incorrecte	13%	6%	15%

Beaucoup de titres et descriptions sont incomplets ou sujets à débat (22%), mais il s'agit des colonnes comportant le moins de réponses totalement incorrectes. L'usage de ce modèle s'est donc révélé particulièrement pertinent sur la génération de titres et de descriptions. Il s'agit d'une tâche de traitement de langage qui ne demande pas de capacités de raisonnement et sur laquelle les performances des *LLM* sont bonnes. Ces titres et descriptions peuvent permettre de faciliter la recherche dans les archives et d'interpréter des documents ou unités de description dont le nommage serait obscur. Il s'agit là de l'automatisation d'une base du travail de l'archiviste. Cela permet de gagner beaucoup de temps. La rédaction des descriptions peut en effet être longue. Pour que cette description soit optimale, un *fine-tuning* est envisageable. Nous avons précédemment abordé le projet LlaMandement de la Direction générale des Finances publiques. Le *fine-tuning* de Llama 2 a permis de produire des résumés davantage neutres. Les performances sont décrites comme satisfaisantes sur le point éthique[9]. Le système développé réduit significativement la charge de travail des agents, qui n'ont que des tâches de vérification à faire.

Dans les *prompts* de génération des titres et descriptions, nous demandions au modèle d'intégrer les mots importants pour l'indexation des documents ou dossiers. L'objectif était d'obtenir de meilleurs résultats en cas de recherche plein texte sur une personne, une organisation, un mot matière ou un événement. Cette approche a produit de meilleures descriptions avec une valeur ajoutée pour la recherche. Les *LLM* sont performants sur l'indexation automatique grâce à leur capacité à traiter le langage. Un projet récemment mené en collaboration entre l'université de Stanford et l'université de North Texas avait pour but d'étudier le potentiel de *ChatGPT* sur la génération de métadonnées[10]. L'étude,

9. J. Gesnouin, Y. Tannier, C. G. Da Silva, *et al.*, *LLaMandement…*

10. Peter Chan, Mark E. Phillips, Jessica Cebra et James Jacobs, *Leveraging ChatGPT for Efficient Metadata Creation of Government Reports*, en, [Note : Cet article n'a pas encore été publié au moment de la rédaction de ce mémoire], 2024.

basée sur un échantillon de cent rapports gouvernementaux, a révélé que *ChatGPT* est capable de produire des métadonnées avec une précision variable, affichant des résultats prometteurs pour des attributs tels que les titres et les descriptions, mais rencontrant des difficultés avec des normes spécifiques tels que les *Library of Congress Subject Headings (LCSH)* et les *Library of Congress Name Authority Files (LCNAF). ChatGPT* n'est pas encore pleinement performant et ne remplace pas les catalogueurs et catalogueuses. Le grand modèle de langage reste cependant une aide qui permet d'accélérer le travail de ces derniers.

Un autre projet de catalogage par LLM a été mené dans le but d'automatiser la génération de *LCSH*, notices normées d'indexation matière, sur des thèses et des articles [11]. Les conclusions sont les mêmes : *ChatGPT* fait gagner du temps mais ne remplace pas les catalogueurs qui doivent vérifier ce qui a été généré et apporter des améliorations. L'article présentant les résultats du projet mentionne la possibilité que des données telles que des notices MARC et des *LCSH* soient présentes dans les données d'entraînement de *ChatGPT*. Envisager l'automatisation de la description des fonds en XML EAD pourrait ainsi être une voie prometteuse. En posant des questions à Llama 3 et ChatGPT sur le format nous avons constaté qu'ils avaient une certaine familiarité avec ce dernier, ce qui suggère la présence d'instruments de recherche EAD dans leurs données d'entraînement. En fournissant aux modèles de langage de la documentation bien structurée sur la norme à suivre, par exemple le dictionnaire des balises de l'XML EAD et/ou des exemples d'instruments de recherche, avec le texte d'un document ou groupe de documents, une potentielle automatisation de la description archivistique semble envisageable. Les archivistes pourraient alors se concentrer sur la correction et l'amélioration des résultats, ce qui permettrait de gagner du temps. Cependant, il est important de noter que les capacités de traitement des *LLM* sont limitées par leur fenêtre de contexte, c'est à dire la quantité maximale de texte qu'ils peuvent analyser en une fois.

En ce qui concerne le repérage des données sensibles, qui représentent neuf colonnes dans notre inventaire et qui est primordial afin d'assurer la communicabilité des archives, nos tentatives avec Llama 3 révèlent un certain manque de précision, avec 15 % d'erreurs dans notre évaluation. Des colonnes sont plus aisées à remplir que d'autres : la colonne pour laquelle la précision est la meilleure est celle sur la « Prévention, recherche ou poursuite de faits punissables » avec 93% de bonnes réponses, 5% de réponses sujettes à débat et 2% de réponses incorrectes. Les données personnelles et les informations sur des « affaires portées devant les instances juridictionnelles, extrajudiciaires, disciplinaires » étaient quant à elles plus difficiles à repérer. Cela est lié à la subjectivité de la définition,

11. Eric H. C. Chow, T. J. Kao et Xiaoli Li, *An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations*, arXiv :2403.16424 [cs], juill. 2024, DOI : 10.48550/arXiv.2403.16424.

au fait que nous ayons utilisé une version réduite du modèle et au fait que le repérage de ces données demande une certaine capacité d'analyse.

Nous n'avons pas trouvé mention d'autres projets basés sur des *LLM* ayant abouti à la détection de données sensibles. Nous avons seulement trouvé mention de projets d'anonymisation. Nous citerons ici deux articles. Le premier, « Large Language Models are Advanced Anonymizers » évoque le fait que les *LLM* ont pour avantage de générer une anonymisation qui laisse le texte compréhensible [12]. L'anonymisation est en effet une tâche complexe. Cette complexité est illustrée dans le second article, « Robust Utility-Preserving Text Anonymization Based on Large Language Models » par l'exemple d'un « joueur de tennis », « tennis athlete », à anonymiser en « athlete », que l'on remplacerait sûrement par « sportif » en français [13]. D'après les auteurs, le modèle développé dans le cadre de leur étude surpasserait les modèles existants en matière d'anonymisation [14]. Le modèle développé est un affinage d'un *LLM* grâce à la méthode de la *DPO (Direct Preference Optimization)* [15]. Si le modèle est performant sur la tâche d'anonymisation, il n'est pas impensable qu'il le soit sur la détection de données personnelles, puisque ces dernières sont modifiées par le processus d'anonymisation.

Les *LLM* performent donc bien sur des tâches de description et d'indexation, surpassant les approches traditionnelles de TAL (Traitement Automatique du Langage) lorsqu'elles ne sont pas approfondies et affinées sur des corpus spécifiques. Cependant, la question de l'ampleur exacte de ces performances est difficilement analysable. Elle nécessiterait des recherches plus poussées afin d'en évaluer pleinement la portée. L'intégration des différentes tâches archivistiques, telles que la description, l'indexation, et potentiellement d'autres fonctions, dans un outil basé sur un seul *LLM* représente une perspective prometteuse. Un outil tout-en-un permettrait de réaliser des économies en ressources et en processus. C'est ce qu'ont tenté de réaliser des chercheurs de l'université de Wuhan et de Loughborough avec *ArcGPT*, un modèle de sept milliards de paramètres spécifiquement développé pour réaliser des tâches archivistiques [16]. Le développement ou le *fine-tuning* d'un *LLM* sur des tâches archivistiques précises pourrait ainsi permettre d'alléger les modèles tout en les rendant plus précis et utiles pour les archivistes.

Par ailleurs, une dernière perspective permettant des économies de moyens est l'annotation par les *LLM* pour générer les données d'entraînement de modèles plus légers, basés par exemple sur les méthodes traditionnelles du TAL (Traitement Automatique du

12. Robin Staab, Mark Vero, Mislav Balunović et Martin Vechev, *Large Language Models are Advanced Anonymizers*, en, févr. 2024, URL : `https://arxiv.org/abs/2402.13846` (visité le 31/08/2024).

13. Tianyu Yang, Xiaodan Zhu et Iryna Gurevych, *Robust Utility-Preserving Text Anonymization Based on Large Language Models*, arXiv :2407.11770 [cs], juill. 2024, DOI : `10.48550/arXiv.2407.11770`.

14. *Ibid.*

15. Pour plus d'informations sur la DPO : `https://arxiv.org/pdf/2305.18290`

16. S. Zhang, J. Hou, S. Peng, Zuchao Li, Qibiao Hu et P. Wang, *ArcGPT : A Large Language Model Tailored for Real-world Archival Applications*, arXiv :2307.14852 [cs], juill. 2023, DOI : `10.48550/arXiv.2307.14852`.

Langage). Un article intitulé « Large Language Models as Annotators : Enhancing Generalization of NLP Models at Minimal Cost » illustre cette approche. Il aborde les difficultés de généralisation des modèles de TAL (Traitement Automatique du Langage) et le besoin constant de fournir de nouvelles données d'entraînement [17]. Il propose l'annotation automatisée par des *LLM* pour améliorer la généralisation, notamment dans les domaines où les données sont rares. Le principal défi réside dans le choix de données appropriées pour renforcer le modèle sans nuire à sa précision. Avec une automatisation des annotations, des économies peuvent être réalisées sur les coûts en personnel et sur les moyens matériels par le développement des modèles légers, plus aisés à déployer sans intermédiaire.

Pour conclure cette section, les possibilités d'automatisation des processus métier dans le domaine des archives grâce aux *LLM* sont vastes. Néanmoins, les applications concrètes restent limitées, notamment en raison des difficultés liées à leur intégration dans les processus existants et à des incertitudes quant à la précision des résultats produits. L'IA ne peut pas remplacer l'archiviste mais elle pourrait accélérer son travail. Le secteur archivistique luxembourgeois n'est pas encore mature en matière de déploiement de ces technologies et un travail de recherche et de développement important reste à accomplir. Les *LLM* étant des technologies récentes, des recherches et des tests devraient continuer afin d'explorer leurs apports, mais le recul sur ces dernières est pour l'instant limité.

3. Au delà des usages métier : RAG, recherche et médiation

De la même manière que pour le traitement automatique du langage et des images détaillés dans le chapitre précédent, les usages les plus simples à mettre en production sont ceux qui ont les impacts les moins importants en cas d'erreur en termes de sécurité et de transparence. Plusieurs services d'archives et bibliothèques mettent à profit les *LLM* pour des tâches de médiation avec leurs usagers. Ces projets ont l'avantage de demander moins de temps et de budget que ceux à grand impact.

Une des contraintes des *LLM* réside dans la limitation de leurs données d'entraînement. Elles sont certes volumineuses, mais ne permettent pas de répondre à toutes les questions. Ce sont des modèles faits pour générer du langage et non des réponses factuelles. Une solution a été développée face à ce problème. Il s'agit du RAG (Retrieval Augmented Generation), technique permettant de récupérer des informations pertinentes au sein d'une base de connaissance pour répondre au *prompt* d'un utilisateur. Cette récupération se fait par mesure de similarité entre un document et le prompt. Les documents et le prompt peuvent être vectorisés avec des algorithmes tels que TF/IDF pour les versions légères, mais pour avoir un résultat le plus précis possible, ce sont souvent les *embeddings*

17. Parikshit Bansal et Amit Sharma, *Large Language Models as Annotators : Enhancing Generalization of NLP Models at Minimal Cost*, juin 2023, URL : `https://arxiv.org/abs/2306.15766` (visité le 10/08/2024).

de grands modèles de langage [18] qui sont utilisés. Le ou les passages pertinents pour répondre au *prompt*, donc les plus proches de ce dernier dans l'espace vectoriel, sont ajoutés à la fenêtre de lecture du *LLM*, qui peut se baser dessus afin de générer sa réponse et peut être en mesure de citer ses sources [19].

L'application principale du RAG (Retrieval Augmented Generation) dans le domaine des archives est la recherche d'informations ou de documents en langage naturel. En intégrant une base de connaissance contenant des fonds numérisés ou nativement numériques, l'usager peut poser des questions en langage naturel sur les fonds, ou éventuellement sur le fonctionnement du service. On peut s'interroger sur l'efficacité de cette méthode pour le chercheur par rapport à un moteur de recherche classique. Un moteur de recherche classique offre davantage de réponses en moins de temps, plus rapidement analysables par le chercheur, qui peut aussi jouer avec des filtres. Il doit néanmoins aller lire chaque document, ce qui peut lui faire perdre du temps. Le *RAG* peut donc paraître plus efficace qu'un moteur de recherche pour la recherche d'une seule information précise, mais ne l'est pas forcément pour la recherche de documents.

Au Luxembourg, quelques outils basés sur le *RAG* ont récemment été mis à disposition du grand public. A la BNL, une version beta d'un *chatbot* a été mise en ligne fin 2023. Ce *chatbot* offre la possibilité de poser des questions sur l'histoire luxembourgeoise en se basant sur des articles de presse anciens [20]. L'interface est similaire à celle de ChatGPT, avec une zone pour écrire un message, un historique conversationnel et des réponses en langage naturel. L'architecture technique est aussi basée sur le modèle GPT. La seule différence est la section « Ma réponse se base sur les articles suivants », où sont citées les sources. Le *chatbot* est décrit comme un « service complémentaire et expérimental [21] » à la barre de recherche. Il sert à obtenir des réponses rapides et sourcées sur des faits historiques au Luxembourg et identifier des documents qui en parlent. C'est assez efficace pour une recherche rapide, mais un chercheur devra sans doute aller plus loin, ce qui est toutefois une composante naturelle de son travail. Il s'agit d'un bon outil de médiation pour le grand public, qui impressionne par ailleurs. Dans ce domaine, le Parlement européen a lancé *Archibot 3.0*, agent conversationnel auquel il est possible de poser des questions sur l'histoire législative européenne et dont la base de connaissance est composée d'archives numérisées de l'institution.

18. Ces méthodes ont été détaillées dans le chapitre 4.

19. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, *et al.*, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, en, mai 2020, URL : `https://arxiv.org/abs/2005.11401v4` (visité le 30/08/2024).

20. BNL, *Eluxemburgensia.lu s'est doté d'un nouveau chatbot*, fr, oct. 2023, URL : `http://bnl.public.lu/fr/a-la-une/actualites/communiques/2023/chatbot-eluxemburgensia.html` (visité le 10/07/2024).

21. *eluxemburgensia chat*, URL : `https://chat.eluxemburgensia.lu/info` (visité le 10/07/2024).

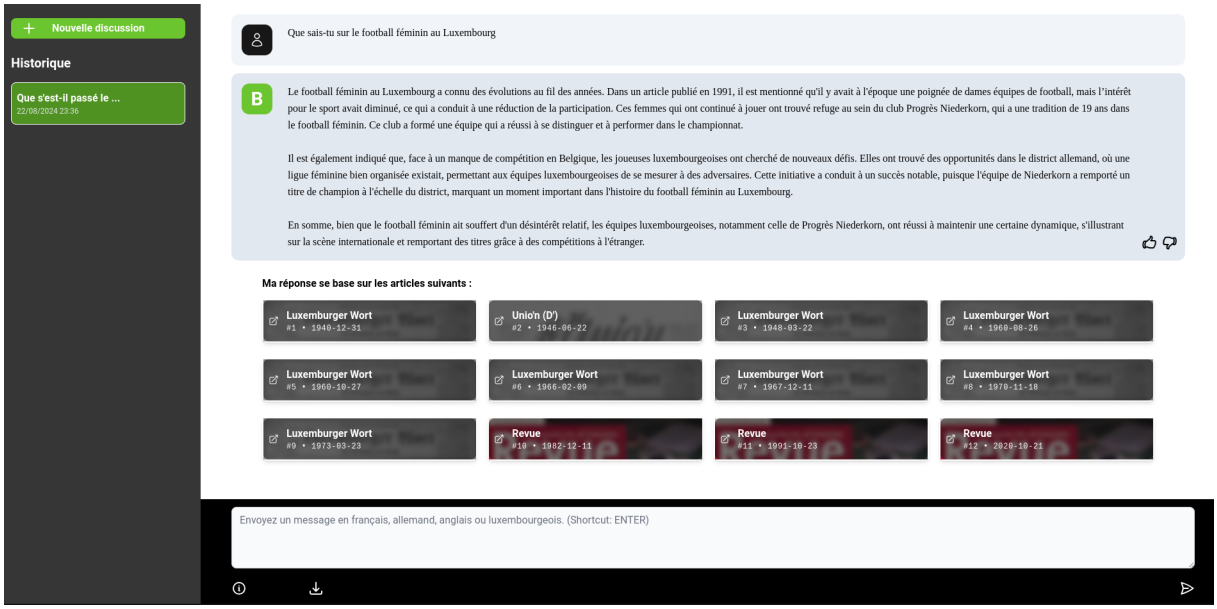

FIGURE 5 – Capture d'écran de l'interface du chatbot de la BNL *eLuxemburgensia*

Le *RAG* est une perspective intéressante en ce qui concerne la médiation dans le domaine des archives. Ce type d'agents joue un rôle dans le développement d'une forme de transparence des administrations avec le public mais ne remplace pour l'instant pas la description et l'indexation dans un moteur de recherche. Néanmoins, ces initiatives contribuent à procurer une visibilité plus accrue aux services d'archives, les éloignant de l'image stéréotypée de lieux poussiéreux, sombres et peu accessibles. Elles les présentent comme des lieux modernes et innovants. Elles montrent que les archives sont ouvertes à tous et à toutes, y compris au grand public, et pas seulement aux chercheurs ou aux passionnés de généalogie. Cette amélioration de l'image des services et cette démarche de médiation peuvent servir à mieux faire saisir l'importance des archives, tant au public qu'aux administrations dans leur ensemble et à encourager les investissements.

Les usages des grands modèles de langage et en particulier de l'intelligence générative conversationnelle sont donc divers dans les archives. Il est plus aisé de mettre en place ces technologies lorsqu'elles ont un faible impact en cas d'erreur. Il y a pour l'instant peu de recul sur ces dernières. Les *LLM* ont l'inconvénient d'être très généraux, nécessitant d'ajuster les prompt avec des informations précises, des sources ou de les affiner sur des données spécifiques au domaine archivistique. Leur implémentation demande des connaissances techniques et des investissements matériels. Les recherches à venir les rendront peut-être plus performants sur des tâches archivistiques complexes ou plus légers. L'IA a un grand potentiel dans les archives mais nous en sommes encore au début de ses applications dans le domaine. Même si les projets ne conduisent pas encore à l'automatisation de processus complexes, ils peuvent avoir des apports connexes divers.

Chapitre 6

Au-delà de l'automatisation, des apports connexes : exploration des données, collaboration et visibilité

1. La quête des données d'entraînement : exploration en profondeur des données et découverte de silos à décloisonner

Dans la première partie du projet InventAIre, avant de choisir une approche et de réaliser les premier tests de *machine learning*, nous avons dû regrouper des données d'entraînement et de tests. Nous avons déjà abordé le fonds bureautique du Service des relations européennes et internationales et du protocole (SREIP) qui a été mis à disposition. Nous avons tenté d'obtenir d'autres documents auprès du service Technologies de l'information, qui nous a alors présenté les bases de données de plusieurs applications. Nous avons obtenu les fichiers pdf des documents publiés sur le site web de la Chambre [1]. Il s'agit de documents parlementaires publics tels que des projets de lois, des questions parlementaires [2], les réponses à ces questions, des procès-verbaux publics de réunions, etc. Une extraction de la base de données du Courrier électronique de la Chambre, dont les documents contenant des données confidentielles ont été exclus, nous a aussi été fournie au format csv. Le Courrier électronique est un système d'expéditions numériques interne pour le personnel et les députés. Les expéditions contiennent des pièces-jointes stockées dans un système de fichiers et leurs métadonnées sont stockées dans une base de données. Les pièces-jointes peuvent être des convocations, des invitations ou encore des projets de loi. Ce sont ces documents qui ont été mis à disposition avec certaines de leurs métadonnées. Nous avons donc eu un aperçu d'une partie des données stockées sur les serveurs de la Chambre et de leur organisation. Cet aperçu nous a été révélateur de l'existence d'un

1. URL : `https://www.chd.lu/fr/searchactivities`
2. Questions posées par les députés aux membres du gouvernement sur l'actualité politique ou bien d'intérêt général.

grand nombre d'applications métier différentes et de silos de données.

Les silos de données sont des systèmes de gestion de l'information isolés, où les données sont stockées de manière indépendante sans communication fluide avec d'autres systèmes ou applications au sein d'une même organisation. Les inconvénients de ces silos auraient été mis en avant à partir des années 1980 quand a commencé à se poser la question de la collaboration entre les services [3]. Ils peuvent en effet compliquer les projets impliquant différentes équipes qui ne travaillent pas avec les mêmes applications et n'ont pas accès aux mêmes données. Pour les mêmes raisons, ils peuvent néanmoins parfois constituer une forme de sécurité. Les données sensibles de chaque équipe restent dans les mains de ces dernières ou dans celles de l'équipe qui gère les systèmes d'informations. La question de rassembler ces données est source d'inquiétude dans un contexte national où, comme évoqué précédemment (chapitre 2), les préoccupations sur la divulgation de données sensibles sont importantes. À la Chambre des Députés, l'affaire des « *Chamber-Leaks* » a contribué à accroître ces dernières. Il s'agit d'une faille de sécurité qui avait permis la consultation de documents internes via de simples URL en 2018. Toutefois, les archivistes devraient idéalement pouvoir archiver les différentes données produites par les services lorsqu'elles ont une valeur archivistique. Par exemple, à la Chambre des Députés, le contenu du Courrier électronique peut avoir un intérêt patrimonial, voire juridique. Des documents officiels transitent entre le personnel de l'administration parlementaire et les députés par cette voie. Or, son contenu est stocké dans une base de données complexe à appréhender et les archivistes n'y ont pas accès. Les archivistes de la Chambre n'ont pas de vue d'ensemble des différents silos et c'est également rarement le cas dans les autres administrations. Iels ont accès aux applications et aux données produites par leur service. En général, ce ne sont que quelques personnes qui travaillent côté systèmes d'information qui y ont accès et en ont une vision globale. Par ailleurs, ces silos peuvent causer une réplication des données, ce qui augmente par la suite le risque de désynchronisation des informations entre les différents systèmes. Cette réplication engendre une augmentation des besoins de stockage. La diversité des technologies et des bases de données utilisées complique leur maintenance, rendant les processus de gestion plus complexes et coûteux. Enfin, avant la création d'une nouvelle application, il est souvent nécessaire de recommencer l'étape de modélisation des données, ce qui ralentit les projets et génère des redondances inutiles. Le décloisonnement des silos peut donc permettre des gains financiers. D'après Gauthier Poupeau à propos des silos qui existaient dans les données de l'INA, « maintenir un système et maîtriser les données est plus difficile et plus coûteux lorsque les silos se multiplient les uns à côté des autres, car une telle architecture dilue les moyens [4] ».

3. Bendik Bygstad, Ole Hanseth et Dan Truong Le, « From IT Silos to Integrated Solutions. A Study in E-Health Complexity », *ECIS 2015 Completed Research Papers* (, mai 2015), DOI : 10.18151/7217283.

4. Gautier Poupeau, « Le « lac de données », une infrastructure technique pour déployer la gouvernance des données à l'Ina », dans *Les nouveaux paradigmes de l'archive*, dir. Claire Scopsi, Clothilde

A la Chambre des Députés du Grand-Duché, il existerait entre dix et quinze silos [5]. Ils sont à la fois applicatifs et organisationnels. En effet, chaque service a ses applications. Il en existe par exemple une pour gérer les procès-verbaux de commission, une pour les travaux en commission, une pour les motions, etc. Ils sont organisationnels car chaque application répond à un besoin métier d'un service et applicatifs parce qu'ils sont liées à une seule application. Chaque application est basée sur des technologies propres et a sa base de données. Il y a donc plus d'une dizaine de bases de données différentes. Il existe toutefois une base de données métier sous la forme d'un grand référentiel qui est connectée aux différentes applications. Elle contient par exemple un référentiel personne, des moyens de contact, l'organigramme de la Chambre, etc. Cette base de données connectée évite ainsi certaines réplications de données, mais constitue aussi une application en elle-même. Face à ces silos, plusieurs solutions existent. À la Chambre, l'utilisation de référentiels communs à différentes applications est une première étape afin de créer des connexions entre silos au sein du système d'information. Un projet de *data warehouse* est également en cours. Un *data warehouse* est un regroupement de données structurées provenant de différentes sources dans le but de faciliter leur analyse et la prise de décision au sein d'une administration [6], dans un souci d'amélioration de la *business intelligence* [7]. Le projet de la Chambre des Députés vise à répliquer les données de chaque application en mutualisant les différentes bases de données. Ce projet a trois objectifs principaux. La mutualisation des bases devrait faciliter le travail de *reporting*, c'est à dire de production de rapports afin de monitorer la qualité des données. En effet, le *data warehouse* devrait permettre de réaliser différentes datavisualisations via le logiciel d'analyses et visualisation de données *Power BI* avec des modèles de données simplifiés. Le deuxième objectif est d'octroyer aux différents services un moyen de visualisation de leurs données, toujours via *Power BI*. Enfin, ces données réorganisées pourront être connectées sur la plateforme d'*open data* de l'État luxembourgeois [8] de manière automatique afin d'être publiées et téléchargeables par l'ensemble des citoyens. Les *data warehouses* et les projets de décloisonnement des silos en général offrent ainsi des avantages considérables pour l'open data et l'intelligence artificielle. En centralisant et en structurant les données, ils permettent une meilleure accessibilité et transparence des informations pour le public. Ils permettent également une meilleure disponibilité de jeux de données de qualité pour l'entraînement et les tests

Roullier, Martine Sin Blima-Barru et Édouard Vasseur, Pierrefitte-sur-Seine, 2024 (Actes), URL : `https://books.openedition.org/pan/7253` (visité le 22/07/2024).

5. Le contenu de ce paragraphe est basé sur un entretien avec un membre du service Technologies de l'information en charge du projet de refonte du système d'information de la Chambre des Députés.

6. *Qu'est-ce qu'un Data Warehouse*, fr, URL : `https://www.oracle.com/fr/database/data-warehouse-definition/` (visité le 20/08/2024).

7. Ensemble des technologies, outils et pratiques permettant de collecter, analyser et transformer des données brutes en informations pertinentes pour prendre des décisions éclairées en entreprise.

8. Plateforme `data.public.lu`. URL des jeux de données publiés par la Chambre des Députés : `https://data.public.lu/fr/organizations/chambre-des-deputes-du-grand-duche-de-luxembourg/`

de modèles d'IA. Ils apportent des gains de temps sur cette première phase de collecte et préparation de données. L'analyse facilitée par le *data warehouse* garantit aussi une meilleure qualité des données pour entraîner des modèles.

Un second projet est en marche à la Chambre de Députés. Il s'agit d'un projet de refonte des systèmes d'information. Cette refonte a pour but de regrouper les différentes applications métier en une seule. L'inconvénient d'une refonte des systèmes face au problème des silos est son coût et le temps qu'elle demande. Dans d'autres administrations publiques, la refonte peut être simplifiée par un nombre plus réduit d'applications et de données à restructurer. Cette refonte est par ailleurs une occasion de mieux organiser le *records management* au sein de l'administration. À la Chambre des Députés, le projet a donné lieu à l'identification de quatre types principaux de données stockées dans les systèmes. Il s'agit des données sur les acteurs, des dossiers traités, des documents de ces dossiers, et des regroupements de personnes (réunions, évènements). Des dossiers sont effectivement traités par des ensembles d'acteurs, qui produisent différents documents pendant ce processus de traitement. Les documents, et idéalement l'ensemble des données ayant une valeur archivistique devraient être conservées. Le projet de refonte donne un regard sur ces données à archiver et fait émerger des réflexions quand à leur archivage. Aux archives municipales de la ville d'Amsterdam, une refonte du système informatique a récemment été l'occasion de réfléchir à une meilleure gestion et conservation des archives numériques. L'équipe a développé une intégration des fonctions d'archivistiques dès la création du système d'information [9]. Il s'agit de la méthode nommée *archiving-by-design*. Cette refonte a également permis de revoir les formats de métadonnées afin de faciliter l'adoption du modèle de description archivistique RiC-O [10]. La reprise des systèmes d'information face au problème des silos est donc une opportunité d'optimiser les pratiques d'archivage et de *records management*.

Il existe d'autres manières de décloisonner les silos. Un deuxième exemple est celui de l'INA (Institut national de l'audiovisuel) en France. Un « lac de données » a été développé dans un souci de « recentralisation des systèmes d'information constitués en silos [11] ». Il s'agit d'un système de stockage de données brutes et hétérogènes provenant de diverses sources au sein du système d'information. Plusieurs projets IA ont été menés à l'INA. Le décloisonnement des silos grâce au lac de données y est également un moyen d'avoir accès à des données propres et centralisées et à des systèmes performants pour faciliter les automatisations via l'IA : « disposer d'une infrastructure centralisée pour toutes les données

9. Elisenda Cristià, *Information management function at the Amsterdam City Archives*, en, janv. 2020, URL : `http://arxivers.com/ladada/information-management-function-at-the-amsterdam-city-archives/` (visité le 14/08/2024).

10. *Records in Contexts (11) : versie 1.0 gelanceerd !*, nl, 19 mars 2024, URL : `https://www.amsterdam.nl/stadsarchief/organisatie/blog-bronnen-bytes/records-contexts-(11)-versie-1-0/` (visité le 14/08/2024).

11. G. Poupeau, « Le « lac de données », une infrastructure technique pour déployer la gouvernance des données à l'Ina »…

de l'INA [lui] ouvre également une nouvelle perspective : celle de pouvoir automatiser, en s'appuyant sur des technologies d'intelligence artificielle, la production d'un certain nombre de données [12] ». La question des silos s'est aussi beaucoup posée en bibliothèque. Nous avons trouvé moins de littérature l'abordant du point de vue des archives.

Pour conclure sur ces problématiques de silos et de décloisonnement, la phase de collecte et de préparation des données des projets IA donne lieu à une exploration des données présentes dans les systèmes de l'administration qui les initie. Cette exploration peut mettre en évidence les failles de leur organisation. Elle a ainsi des apports pour les personnes gérant les systèmes d'information. La collaboration avec les services gérant ces derniers dans les administrations paraît également inévitable pour l'archivage d'une grande partie des données qu'ils produisent. La recherche de données pour notre projet IA a révélé une partie de la face cachée de cet iceberg des données cloisonnées des systèmes d'information. Au delà des apports en termes d'automatisations, les projets IA ont aussi des apports d'ordre stratégique. Le décloisonnement des silos identifiés offre quant à lui une opportunité de produire des données de meilleure qualité et mieux extractibles pour entraîner ou tester des modèles de *machine learning*. Si les archivistes sont inclus dans le processus, cela peut avoir un impact positif sur la gestion des archives produites par les systèmes. La découverte des silos fait aussi ressortir le silo dans lequel se trouvent souvent les archivistes. Cantonnés à des rôles trop précis, ils n'ont parfois pas une vision globale sur les systèmes d'information. En effet, les archivistes ont souvent été présentés comme des « généralistes de l'information [13] ». D'après Françoise Banat-Berger dans un article publié en 2012 dans la Gazette des archives intitulé « "Un métier à part entière, l'archiviste un généraliste de l'information" : qu'en est-il en 2012 dans le nouvel environnement numérique ? » ,

> L'archiviste est un professionnel bien singulier aux confins de la science de l'information, de l'archivistique, du juridique, de la qualité, des sciences administratives et historiques [14].
>
> La phrase de Gérard Naud : "un métier à part entière, l'archiviste un généraliste de l'information" ou, comme l'écrit aujourd'hui Jean-Michel Salaün un "architecte de l'information", reste par conséquent tout à fait opérante avec une nécessité de travailler avec un périmètre de plus en plus large de partenaires [15].

Les projets IA appliqués aux archives numériques soulignent l'importance de ce rôle

12. *Ibid.*
13. F. Banat-Berger, « « Un métier à part entière, l'archiviste un généraliste de l'information » : qu'en est-il en 2012 dans le nouvel environnement numérique ? », *La Gazette des archives*, 226–2 (2012), p. 117-126, DOI : 10.3406/gazar.2012.4901.
14. *Ibid.*
15. *Ibid.*

généraliste des archivistes, mettant en évidence la nécessité de collaborer étroitement avec divers acteurs, dont les informaticiens et responsables de projets numériques, pour optimiser la gestion des documents et l'automatisation des processus archivistiques.

2. Des apports moins techniques : des projets qui renforcent les liens entre les services et apportent une visibilité sur le monde des archives

La visibilité du monde des archives et sa collaboration avec d'autres domaines est importante pour plusieurs raisons. La question de la collaboration est abordée dans le code de déontologie de l'*ICA (International Council on Archives)*. Le dixième article stipule que :

> Les archivistes travaillent en collaboration avec leurs collègues et les membres des professions voisines afin d'assurer universellement la conservation et l'exploitation du patrimoine documentaire. Les archivistes cherchent à stimuler la collaboration et à éviter les conflits avec leurs collègues, en résolvant les difficultés par l'encouragement à respecter les normes archivistiques et l'éthique professionnelle. Les archivistes coopèrent avec les représentants des professions parallèles dans un esprit de respect et de compréhension mutuelle [16].

La collaboration est donc une valeur fondamentale dans le domaine archivistique. D'après Sylvie Forastier dans un article publié dans la *Gazette des archives* sur son expérience de *records manager* pour un cabinet d'avocat international au Luxembourg, « il serait réducteur d'envisager un service d'archives comme étant uniquement à vocation administrative interne, un service d'archives n'est pas totalement déconnecté du monde externe et des activités générales de l'entreprise [17] » et il « ne fonctionne pas en autarcie mais participe à la dynamique de l'entreprise [18] ». Se faire connaître auprès des autres services est effectivement essentiel pour que les données et documents à archiver de chacun d'entre eux soient mis à disposition. Une bonne compréhension du fonctionnement des ces services est obtenue après des discussions avec leur personnel et permet de mieux connaître les documents produits et leur contexte. L'archiviste est alors au fait des différents processus de travail et peut adapter la stratégie d'archivage. Avec l'émergence des documents numériques, l'archivistique n'est aujourd'hui plus limitée aux trois âges théorisés d'Yves Pérotin [19], mais il s'agit d'une « archivistique des flux » [20]. Les archivistes

16. ICA, *Code de déontologie de l'ICA*, fr, URL : `https://www.ica.org/fr/resource/code-de-deontologie-de-lica/` (visité le 12/08/2024).

17. Sylvie Forastier, « Archiviste : un métier protéiforme ? », *La Gazette des archives*, 240–4 (2015), p. 305-311, DOI : `10.3406/gazar.2015.5310`.

18. *Ibid.*

19. Yves Pérotin, « L'administration et les "trois âges" des archives », *Seine et Paris*, 20, (octobre 1961), p. 31-33

20. Céline Guyon, « Une archivistique sous influence », *Revue d'histoire culturelle. XVIIIe-XXIe*

doivent idéalement être impliqués dès la création des documents numériques, « dès l'âge courant, c'est-à-dire dès la conception des systèmes d'information [21] ». Comme évoqué précédemment, les archivistes sont désormais des « généralistes de l'information [22] » et ils traitent tout type de données. Or, d'après Jenny Bunn, *Head of Archives Research* aux Archives nationales britanniques, la tendance serait à une spécialisation du personnel, qui peut générer des silos [23]. Les administrations et les archives pourraient donc tirer parti de collaborations interdisciplinaires plus récurrentes [24].

Par ailleurs, pour mener à bien des projets d'intelligence artificielle, une collaboration étroite avec les acteurs informatiques et les responsables de la sécurité des systèmes d'information (RSSI) est indispensable. Les projets peuvent exploiter l'expertise des chefs de projets informatiques, de développeurs et autres professionnels de l'informatique, qui sont par la même occasion sensibilisés aux enjeux archivistiques, en général assez flous pour eux. Ils peuvent alors prendre plus naturellement en compte les préoccupations archivistiques au moment du développement informatique. La collaboration entre archivistes et informaticiens, bien que stimulante, est complexe, car elle nécessite une compréhension mutuelle entre deux mondes parfois perçus comme très différents, les archives étant parfois considérées comme une préoccupation supplémentaire non primordiale pour les personnes issues du monde de l'informatique. La communication peut également être délicate à mettre en place. Cette idée est évoquée dans un article datant de 2021 intitulé « How AI Developers Overcome Communication Challenges in a Multidisciplinary Team : A Case Study [25] ». Les équipes n'ont pas les mêmes connaissances et les personnes qui ne maîtrisent pas l'informatique peuvent avoir de trop hautes attentes. Il est donc important de travailler avec des personnes capables de faire le lien entre les acteurs, qui maîtrisent à la fois le langage du monde de l'informatique et des sciences de l'information. Des collaborations récurrentes pourront mener à une meilleure littératie numérique chez les archivistes et une compréhension plus approfondie des besoins de chaque partie prenante.

Une collaboration avec d'autres services est également envisageable pour obtenir les données d'entraînement nécessaires. Cette collaboration renforce alors les liens en impliquant les services qui fournissent eux-mêmes leurs données, devenant ainsi proactifs dans le travail archivistique. Dans le cadre du projet InventAIre, nous avons eu des liens avec les services RH et des relations européennes, internationales et du protocole. Ils ont dès

siècles–5 (oct. 2022), DOI : 10.4000/rhc.3466.

21. *Ibid.*

22. F. Banat-Berger, « « Un métier à part entière, l'archiviste un généraliste de l'information »… ».

23. L. Jaillant et Katherine Aske, « Are Users of Digital Archives Ready for the AI Era ? Obstacles to the Application of Computational Research Methods and New Opportunities », *J. Comput. Cult. Herit.* 16–4 (janv. 2024), 87 :1-87 :16, DOI : 10.1145/3631125.

24. *Ibid.*

25. David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller et Felix Portnoy, « How AI Developers Overcome Communication Challenges in a Multidisciplinary Team : A Case Study », *Proc. ACM Hum.-Comput. Interact.* 5–CSCW1 (avr. 2021), 131 :1-131 :25, DOI : 10.1145/3449205.

lors été mis au courant de l'existence de notre projet. Les projets d'IA apportent en effet également une certaine publicité aux archives, même s'il convient de rester vigilant face aux risques de mauvaise publicité en cas d'échec. Ils démontrent que le domaine peut aussi pousser l'innovation et est utile. Nous avons par ailleurs tenté une collaboration avec les Archives nationales du Luxembourg pour obtenir des données de test. La collaboration peut s'étendre au-delà d'une administration et favoriser les synergies entre administrations.

Les contributions des projets d'intelligence artificielle dans le domaine des archives vont bien au-delà de simples automatisations. En effet, l'interaction avec le personnel d'autres services et l'exploration de leurs données peuvent être profitables pour toutes les parties impliquées. Ces projets IA offrent par conséquent des bénéfices variés dans des secteurs tout aussi variés.

Chapitre 7

Ce que les archives peuvent apporter à l'IA

1. Maîtrise des normes, données structurées et classées

L'IA a beaucoup à apporter aux archives et aux autres domaines dans les administrations, mais le monde des archives a également beaucoup à apporter à l'IA. Comme évoqué précédemment, des données bien structurées dans un système d'information efficace sont un atout pour le développement de modèles de *machine learning*. De par les grandes quantités de documents conservés par les services d'archives et leurs habitudes de structuration de ces données, les relations peuvent être bilatérales entre IA et archives. Non seulement l'intelligence artificielle offre la possibilité d'automatiser des traitements archivistiques, mais le monde des archives peut aussi être un fournisseur privilégié de données d'entraînement structurées. Tout d'abord, les données sont disponibles en grandes quantités. Le site internet des Archives nationales du Luxembourg mentionne quarante-cinq kilomètres linéaires de documents d'archives papier et vingt-cinq mille microfilms répartis entre onze groupes de fonds décrits [1]. Il n'y a pas d'information sur les archives nativement numériques. En France, aux Archives nationales, « 253 applications informatiques, 94 114 910 fichiers de bureautique générés par des traitements de texte, 94 382 photos numériques ou enregistrements sonores et audio numériques » auraient été collectés entre 1983 et 2015 [2]. Le site internet des Archives nationales américaines mentionne quant à lui plus de trente-trois milliards de documents d'archives électroniques, soit huit cent trente-sept tera [3].

1. ANLux, « Fonds et collections », URL : `https://anlux.public.lu/fr/rechercher/fonds-collections.html`

2. FranceArchives, « Panorama des archives électroniques conservées aux Archives nationales », URL : `https://francearchives.gouv.fr/findingaid/b4bac9c0ed8ffa3f2451fbe6f6cb4c65f5a4e753/`

3. National Archives, « National Archives by the numbers », URL : `https://www.archives.gov/about/info/national-archives-by-the-numbers`

Les archives conservent non seulement de grandes quantités de documents classés, mais le domaine a également une longue tradition de description, d'abord sous la forme d'inventaires. Des normes de description archivistique ont ensuite émergé à partir des années 1980, notamment dans le monde anglo-saxon, sous l'influence des bibliothèques. En 1989, le Conseil international des archives (*ICA*) a convoqué des experts pour établir un plan d'action international sur la normalisation des pratiques archivistiques. Au cours des années 1990, l'intérêt pour ces normes a progressivement augmenté, donnant lieu à l'introduction d'une première norme internationale, l'ISAD(G), en 1993 [4]. Elle avait pour but d'améliorer l'efficacité des pratiques archivistiques, de professionnaliser le métier d'archiviste, en permettant notamment d'uniformiser les méthodes d'enseignement [5]. L'introduction de normes visait également à offrir une présentation homogène des archives pour le public, à faciliter la coopération entre institutions et à rationaliser le travail des archivistes [6]. Cette approche de rationalisation est comparée au taylorisme et au fordisme par les archivistes Bénédicte Grailles et Laurent Ducol dans un article publié dans la *Gazette des archives* en 2012 [7]. Par ailleurs, d'après Céline Guyon, maîtresse de conférence associée à l'ENSSIB, « la pratique archivistique s'est nourrie des innovations technologiques qui, en retour, ont contribué à sa normalisation [8] ». Les outils informatiques de traitement auraient entraîné cette normalisation et une homogénéisation des pratiques [9]. Nous avons pu voir que la normalisation répond aussi à des problématiques d'efficacité, néanmoins, les formats numériques normés sont surtout nés par besoin de communication entre les systèmes. Ils ont permis d'accroître l'interopérabilité entre services. L'*EAD* (*Encoded Archival Description*) a été développé dès 1993 en s'inspirant des bibliothèques et de leur format MARC et a permis de répondre à des besoins de diffusion d'instruments de recherche sur internet [10]. Depuis, d'autres normes sont apparues, telles que *RiC (Records in Context)*, dont la version 1.0 a été publiée en mai 2024, ou le SEDA (Standard d'Échange de Données pour l'Archivage), norme française d'interopérabilité utilisé pour le traitement d'archives nativement numériques. Ces normes sont basées sur des langages de structuration de données spécifiques. Les plus récurrents sont l'XML et le JSON. L'EAD est par exemple un format basé sur l'XML.

L'interopérabilité est une préoccupation dans le domaine archivistique. En France, le format EAD est poussé par la plateforme *FranceArchives* pour garantir cette dernière

4. Christine Nougaret, « Vers une normalisation internationale de la description des archives. La norme ISAD(G) du Conseil international des archives », *La Gazette des archives*, 169–1 (1995), p. 274-292, DOI : 10.3406/gazar.1995.3353.

5. Alice Motte, « La normalisation de la description archivistique : enjeux et actualités », *La Gazette des archives*, 238–2 (2015), p. 121-128, DOI : 10.3406/gazar.2015.5262.

6. *Ibid.*

7. B. Grailles et Laurent Ducol, « Les enjeux de la normalisation dans les services d'archives », *La Gazette des archives*, 228–4 (2012), p. 9-22, DOI : 10.3406/gazar.2012.4980.

8. C. Guyon, « Une archivistique sous influence »…

9. *Ibid.*

10. D. Pitti, « Encoded Archival Description… ».

et ainsi produire un portail unique contenant des instruments de recherche de plus d'une centaine partenaires. Les Archives nationales du Luxembourg produisent également des instruments de recherche en EAD pour certains de leurs fonds, notamment numériques. Cependant, les normes n'existent pas partout. À la Chambre des Députés, l'adoption de normes est encore en réflexion. L'inventaire des Archives nationales peut constituer une forme de normalisation, même si un fichier Excel n'est pas très interopérable.

La production de données structurées dans les archives facilite l'entraînement et les tests de modèles d'IA. La question de l'entraînement des grands modèles de langage sur les données des bibliothèques en ligne a déjà été débattue, notamment autour de la problématique du droit d'auteur. Elle ne s'est pas encore posée de manière significative pour les archives. Néanmoins, avec les vastes quantités de texte disponibles dans les fonds, qu'ils soient numérisés ou nativement numériques, et la normalisation de la description archivistique, cette question pourrait se poser davantage à l'avenir. Il faudra déterminer dans quelle mesure il est souhaitable que les archives contribuent à l'entraînement des modèles d'IA, notamment ceux développés par des grandes entreprises privées hors UE. Une des premières questions à se poser concerne l'accès à ces données. Faut-il les intégrer dans des initiatives d'*open data*, permettant ainsi un accès libre et gratuit pour le développement de modèles d'IA, ou au contraire, envisager une monétisation de cet accès, en négociant avec les entreprises souhaitant utiliser ces données pour entraîner leurs modèles ? Cette décision aura un impact direct sur la visibilité et l'utilisation des données nationales dans des outils largement utilisés par le grand public. Si ces données ne sont pas intégrées dans les corpus d'entraînement des modèles les plus répandus, il y a un risque que le patrimoine documentaire du pays soit sous-représenté, voire invisible. Une ouverture totale des données pourrait offrir plus de visibilité sur les archives. Cependant, cela nécessite de faire la balance entre les enjeux économiques, culturels et éthiques. La protection des droits d'auteur, la confidentialité des données sensibles et l'équilibre entre partage et exploitation sont autant de facteurs à prendre en compte. Le monde des archives a ainsi du potentiel pour enrichir celui de l'intelligence artificielle grâce à ses données structurées et normalisées, dans un monde où l'information prend de plus en plus de valeur.

2. Une certaine maîtrise de la description : l'opportunité de renforcer la transparence des données produites par les algorithmes

D'après Herbjørn Andresen, professeur d'archivistique à l'université d'Oslo, dans un article intitulé « A discussion frame for explaining records that are based on algorithmic output » publié en 2019, le contenu généré par des algorithmes est difficilement explicable. Il est complexe à analyser dans le cadre du *records management*, pourtant, son

explicabilité est en partie une obligation légale [11]. En effet, l'article 13 du RGPD oblige la transparence « des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée » en cas de prise de décision automatisée dans le traitement des données personnelles [12]. Herbjørn Andresen aborde également dans son article l'émergence de la question de l'éthique algorithmique. Les algorithmes nécessitent des explications contextuelles. L'auteur propose la réalisation de recherches sur les concepts de description provenant du *records management* qui resteraient applicables et ceux qui seront à développer [13]. Des réflexions sont donc à mener sur la description archivistique des documents et données produites par des modèles d'intelligence artificielle.

Une documentation technique est en général établie lors du développement de systèmes informatiques. Elle sert à la correction de bugs par les équipes de développement, à réinstaller les applications sur d'autres machines ou à adapter les systèmes à de nouveaux environnements. C'est ce que nous avons cherché à accomplir avec la documentation de l'application codée dans le cadre du projet *InventAIre* en produisant une documentation technique aussi précise que possible. Cependant, avec le recul, il apparaît que cette documentation aurait pu être encore plus détaillée, notamment pour améliorer la transparence de l'outil. Il est également important d'informer le lecteur ou la lectrice que les inventaires ont été produits par IA. Cela pourrait inclure des métadonnées précisant qu'ils ont été générés par une IA, des filigranes ou des informations intégrées directement dans les fichiers JSON et Excel produits. Nous aurions pu mettre à profit nos connaissances archivistiques et celles de notre équipe. La description archivistique peut en effet être complémentaire à la documentation technique pour ce genre de données. C'est également l'avis de Jenny Bunn, qui prône une collaboration entre *records managers* et informaticiens afin de réfléchir à une « intelligence artificielle explicable » [14]. D'après la chercheuse, des concepts se recoupent entre intelligence artificielle explicable et *records management*, même s'ils ne sont pas définis exactement de la même façon. C'est le cas par exemple des concepts de la transparence, de la confiance et de la responsabilité [15]. Un des objectifs des services à travers la communication des archives est effectivement de garantir la transparence de l'administration. Cette sensibilité s'étend naturellement aux données ouvertes. Par exemple, les Archives nationales du Luxembourg publient leurs tableaux de tri sur la plateforme

11. Herbjørn Andresen, « A discussion frame for explaining records that are based on algorithmic output », *Records Management Journal*, 30–2 (nov. 2019), p. 129-141, DOI : 10.1108/RMJ-04-2019-0019.

12. *Le règlement général sur la protection des données - RGPD*, fr, URL : https://www.cnil.fr/fr/reglement-europeen-protection-donnees (visité le 13/08/2024).

13. Id., « A discussion frame for explaining records that are based on algorithmic output »…

14. Jenny Bunn, « Working in contexts for which transparency is important : A recordkeeping view of explainable artificial intelligence (XAI) », *Records Management Journal*, 30–2 (1er janv. 2020), p. 143-153, DOI : 10.1108/RMJ-08-2019-0038.

15. *Ibid.*

nationale *open data*, renforçant ainsi l'accès et la transparence des informations. La mise à disposition des corpus d'entraînement des modèles développés ou affinés sur ce genre de plateformes peut être un pas vers plus de transparence en permettant une meilleure compréhension de ces derniers et l'identification de leurs éventuelles failles. Certaines peuvent effectivement être dues à des biais dans les données. Cette mise à disposition nécessite néanmoins que les modèles soient entraînés sur des données non-confidentielles.

L'article de Jenny Bunn mentionne par ailleurs que les *records managers* ont l'habitude de décrire précisément les producteurs et les contextes de production [16]. La théorisation du principe de producteur émerge en relation avec la notion de fonds au XIX^{e} siècle avec Natalis de Wailly, historien et archiviste [17]. Il théorise en effet le concept de fonds et définit le respect des fonds dans une circulaire du 24 avril 1841 : il s'agit de « rassembler les documents par fonds, c'est-à-dire réunir tous les titres qui proviennent d'un corps, d'un établissement, d'une famille ou d'un individu [18] ». Les archives doivent donc rester rassemblées par producteur. Or, dans le cas des archives numériques, cette notion de producteur est vouée à évoluer. Les principes fondamentaux de l'archivistique française tels que le respect des fonds et la théorie des trois âges sont à questionner [19]. Les pratiques luxembourgeoises ont l'opportunité de se nourrir de ces habitudes de description au sein des modèles français, mais également de les questionner et de se nourrir d'autres pratiques. L'archivistique canadienne, à travers les réflexions de Terry Cook, prône par exemple l'idée que le fonds est un concept bien lié à la provenance des documents, mais que la provenance n'est pas seulement liée au producteur, d'autant plus en ce qui concerne les données numériques, parfois manipulées par plusieurs producteurs : la provenance est liée à un processus métier [20]. La description du processus métier est donc une perspective intéressante en ce qui concerne la description des archives numériques, des archives produites par IA et des systèmes IA eux-mêmes.

De nombreuses questions se posent et les théories archivistiques françaises sont vouées à connaître des évolutions. En ce qui concerne les pratiques luxembourgeoises, il y a là une opportunité de se nourrir de ces pratiques françaises, mais surtout de les questionner, et de tirer parti d'autres approches, notamment des approches anglo-saxonnes. Le monde des archives et de l'informatique peuvent s'apporter mutuellement. Les professionnels des archives peuvent apporter leur sensibilité et expertise sur les questions

16. *Ibid.*

17. C. Guyon, « Théorie, technique et pratique archivistique en environnement numérique », dans *Colloque International sur le Document Electronique : Document et archivage : pratiques formelles et informelles dans les organisations*, Grenoble, France, 2023, URL : `https://hal.science/hal-04371177` (visité le 04/08/2024).

18. Michel Duchein, « Le respect des fonds en archivistique : principes théoriques et problèmes pratiques », *La Gazette des archives*, 97, (1977), p. 71-96.

19. *Ibid.*

20. Terry Cook, « Mind over Matter : Towards a New Theory of Archival Appraisal », in Barbara L. Craig (dir), *The Archival Imagination*, 1992, p. 38-70.

d'explicabilité et de transparence au monde de l'informatique. Il est cependant important de noter que les pratiques de description et de classement des archives numériques ne sont pas encore fixes et peu théorisées à ce jour. Les discussions seront davantage prolifiques quand elles l'auront été. Inversement, la meilleure compréhension des outils développés et les pratiques de documentation technique issues du monde de l'informatique peuvent enrichir les approches archivistiques. La combinaison entre procédure de documentation technique et méthodologie archivistique, avec par exemple une description du contexte, du producteur et du processus métier, est un pas vers plus de transparence des systèmes IA. Mettre à disposition les données utilisées pour l'entraînement et l'affinage des modèles, selon les principes de l'*open data*, est également une voie à explorer, bien que cela pose des défis en ce qui concerne les modèles pré-entraînés et les questions de confidentialité des données. L'explicabilité de l'IA ne sera jamais parfaite en raison de la complexité inhérente à ces technologies mais l'intégration de principes archivistiques dans le développement et la gestion des systèmes IA pourrait contribuer à améliorer la transparence et la confiance.

Pour conclure cette deuxième partie, l'intelligence artificielle a beaucoup de potentiel dans le domaine des archives, tant pour l'automatisation de processus que pour la recherche et la médiation avec le public. Les usages sont très divers et il est facile de s'y perdre. Archives et IA peuvent s'apporter mutuellement : la technologie peut répondre à des besoins métier et les archives sont des sources de données structurées non négligeables, avec des intérêts similaires pour la transparence. En revanche, la complexité de mise en place d'outils pérennes et utiles pour les archivistes basés sur de l'IA est parfois sous-estimée car le potentiel séduisant de ces technologies a tendance à occulter les défis qu'elles impliquent.

Troisième partie

Les défis éthiques et techniques du déploiement de systèmes basés sur l'IA

Chapitre 8

Problématiques éthiques et environnementales

1. Des risques sociaux et éthiques

Les risques sont loin d'être nuls en ce qui concerne les systèmes basés sur le *machine learning*. L'identification des risques est d'autant plus complexe que la réglementation a souvent un temps de retard par rapport aux avancées technologiques. À ce jour, il n'existe pas de cadre spécifique pour l'IA dans les archives. Se pose alors la question de la responsabilité en cas de problème : doit-elle incomber à l'utilisateur, au régulateur, au développeur, ou à la machine elle-même (si l'on considère la possibilité d'une personnalité juridique pour les systèmes d'IA) ? Les IA génératives, en particulier, soulèvent plusieurs préoccupations. Nous avons déjà abordé la nature souvent opaque des systèmes d'IA, qualifiés de « boîtes noires », qui pose des défis en termes de transparence.

Les phénomènes d'hallucinations, c'est à dire la production d'informations incorrectes ou fictives par le modèle, sont l'un des principaux dangers. Rappelons ici une nouvelle fois que les grands modèles de langage sont des systèmes de génération de texte, et non de transmission d'informations factuelles. Les résultats ne sont donc pas toujours bons, les modèles peuvent « halluciner ». Certains *prompts* sont plus susceptibles de générer des hallucinations, tels que des *prompts* dans une langue que le modèle maîtrise moins ou avec des fautes d'orthographe ou de syntaxe. Bien qu'un *prompt-engineering* efficace puisse réduire ce risque, il ne l'élimine pas complètement. En outre, des informations « périmées » peuvent être données par les modèles. Afin d'éviter ce risque, ces derniers doivent être évalués et réactualisés régulièrement sur des données plus récentes.

Loin de l'idée de neutralité parfois prônée en ce qui concerne les outils numériques [1], les biais, avec les risques de discrimination ou de partialité constituent un danger important. Les systèmes de *machine learning* qui ne sont pas des IA génératives partagent ces risques, causés en général par des biais dans les données d'entraînement, qui peuvent

1. C. Girard-Chanudet, « Le travail de l'Intelligence Artificielle… ».

mener à des décisions injustes ou incorrectes. Une attention particulière doit ainsi être accordée à ces données en cas de développement d'un modèle maison. Un exemple de décision opaque et potentiellement basée sur un biais est le calcul d'un crédit vingt fois supérieur pour une femme par rapport à son mari par Apple en 2019 [2]. Dans le cas d'un usage en contexte archivistique, ces biais peuvent entraîner l'invisibilisation de certaines communautés, dont les archives pourraient être jugées de moindre valeur car sous-représentées dans les données d'entraînement, ou bien, à défaut d'être éliminées, l'indexation de ces archives peut s'avérer insuffisante. Cette invisibilisation de communautés minoritaires constitue un enjeu archivistique plus large. Les archives de ces dernières ont parfois été négligées ou détruites, contribuant à un effacement de leurs mémoires, ou leur description n'est pas assez précise, le vocabulaire qui leur est relatif n'étant pas inclus dans les thésaurus. L'un des cas les plus documentés en France est celui des archives des luttes contre le SIDA et des luttes LGBTQIA+. Les Archives nationales ont collecté les archives de l'association Act Up Paris mais de nombreuses archives d'autres associations ont été détruites [3]. Certains modèles d'IA semblent moins sujets aux biais que d'autres. Par exemple, *Claude*, développé par la société Anthropic, a été conçu avec davantage de considérations éthiques. Ce modèle est utilisé par le Parlement européen pour son *Archibot*. Claude est un agent multimodal capable de traiter une grande quantité de texte. La version 2.1 lancée en novembre 2023 accepte 200 000 *tokens* en entrée, soit environ 150 000 mots. Anthropic a développé le concept d'« IA constitutionnelle ». Cette dernière repose sur des principes éthiques clairs, une forme de constitution inspirée par des textes tels que la *Déclaration universelle des droits de l'Homme*. Pendant deux phases d'apprentissage, le modèle évalue ses propres réponses en fonction de cette constitution et essaie de réduire le contenu potentiellement nocif de ses réponses [4]. L'avantage de ces deux phases d'entraînement spécialisées est une perte de qualité moindre par rapport à un entraînement général pour générer du contenu inoffensif, puisque les réponses sont évaluées et modifiées une fois générées. Bien qu'il soit présenté comme plus éthique et performant, Claude n'est pas complètement exempt de biais ni d'hallucinations [5]. Développer des modèles véritablement impartiaux est un grand enjeu et revêt de grandes difficultés : compte tenu de l'ampleur des données nécessaires à l'entraînement de ces modèles, elles ont de grandes probabilités de présenter des biais. Les biais linguistiques constituent une autre source de préoccupation. Les corpus d'entraînement des gros modèles d'IA génératives sont souvent

2. Neil Vigdor, « Apple Card Investigated After Gender Discrimination Complaints », *The New York Times* (, nov. 2019), URL : `https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html` (visité le 13/08/2024).

3. Patrick Comoy, « Archives LGBTQI+ en France : de la « déplacardisation » à l'autonomie », *La Gazette des archives*, 255–3 (2019), p. 141-152, DOI : `10.3406/gazar.2019.5836`.

4. Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, *et al.*, *Constitutional AI : Harmlessness from AI Feedback*, arXiv :2212.08073 [cs], déc. 2022, DOI : `10.48550/arXiv.2212.08073`.

5. Aman Priyanshu, Yash Maurya et Zuofei Hong, *AI Governance and Accountability : An Analysis of Anthropic's Claude*, arXiv :2407.01557 [cs], mai 2024, DOI : `10.48550/arXiv.2407.01557`.

dominés par des contenus en anglais, rendant ainsi d'autres langues moins visibles. Le luxembourgeois et les données sur le Luxembourg sont en général extrêmement minoritaires. À titre d'exemple, le modèle Llama 3 de meta est entraîné sur 95 % de données en anglais [6]. Cette domination linguistique peut rendre certains pays ou communautés, et ainsi cultures, invisibles dans les modèles généraux. Dès lors, les états peuvent être tentés de rendre leurs données facilement accessibles sur le web pour être visibles et prises en compte par ces modèles. Toutefois, il s'agit d'un choix stratégique et politique complexe : ils peuvent ne pas souhaiter contribuer à l'entraînement de gros modèles développés par de grandes entreprises étrangères.

En outre, la génération automatique de contenu par les IA soulève des questions juridiques concernant le droit d'auteur et le plagiat. Les grand modèles d'IA générative sont entraînés sur du contenu disponible sur le web, qui peut par exemple contenir des données personnelles, ou bien des œuvres originales soumises au droit d'auteur. En effet, toutes les données qui peuvent être collectées le sont, dans l'optique d'améliorer les performances des modèles [7]. Ce contenu sensible pourrait ressortir au moment de l'inférence.

Pour atténuer ces risques, plusieurs solutions ont été proposées au fil de ce mémoire, telles que la documentation rigoureuse des projets, l'analyse préalable des risques, et des évaluations régulières des modèles pour détecter et corriger les informations périmées.

Un dernier risque concerne les modèles hébergés dans le *cloud*. Leur usage s'accompage d'une transmission de données à un tiers, la société qui héberge le modèle. Il est donc important de veiller à éviter la transmission de données confidentielles sans vérifications préalables. Si un projet IA traite des données sensibles, le modèles devront être installés localement dans l'administration ou chez un tiers de confiance. Cette installation locale demande des moyens matériels importants et a par conséquent un impact écologique non-négligeable.

2. Besoins matériels et impact écologique

Pendant notre stage, des limitations matérielles se sont rapidement faites ressentir, notamment dès les premiers tests de classification automatique lorsque nous travaillions avec un trop grand nombre de documents à la fois. La Chambre des Députés a dû investir dans un ordinateur et une carte graphique (GPU) pour que nous puissions travailler avec des modèles de *machine learning* volumineux. Les spécifications de la machine achetée sont les suivantes :

OS : Windows
RAM : 64 GB

6. Meta, *Introducing Meta Llama 3 : The most capable openly available LLM to date…*

7. Kate Crawford, *Contre-atlas de l'intelligence artificielle : les coûts politiques, sociaux et environnementaux de l'IA*, trad. par Laurent Bury, 18 cm. Bibliogr. et webliogr. p. 311-362. Index., Paris Veules-les-roses, 2023 (Z a).

GPU : Nvidia RTX A4000

Avec cet ordinateur et cette carte graphique, nous étions en mesure de réaliser des inférences avec des moyens et grands modèles de langage. Pour éviter que l'inférence ne soit trop lente, nous avons appliqué une quantisation au modèle LlaMa 3. C'est une technique de compression qui réduit la précision des calculs internes pour accélérer les inférences tout en maintenant des performances acceptables. L'outil développé pendant notre stage intègre une quantisation 4 bits pour les titres et les descriptions et 8 bits pour le repérage des données sensibles. La quantization 4 bits réduit la représentation numérique des poids des modèles, c'est-à-dire les paramètres internes du réseau de neurones qui sont ajustés pendant l'entraînement, à 16 niveaux distincts (2^4), tandis que la quantization 8 bits utilise 256 niveaux (2^8). Nous avons observé que le temps d'inférence pour chaque prompt était d'environ deux secondes avec une quantisation à 4 bits et d'une dizaine de secondes avec une quantisation à 8 bits. Ces chiffres sont cependant relativement arbitraires puisqu'ils dépendent de facteurs très divers, notamment la longueur du prompt et de la réponse à produire. Un calcul plus précis du temps de traitement par *token* aurait permis d'obtenir des mesures moins approximatives. Pour les tâches de génération de titres et descriptions, nous avons privilégié la quantisation à 4 bits, car la perte de précision par rapport à la quantisation à 8 bits était négligeable. En revanche, pour les prompts de repérage des données sensibles, nous avons utilisé la quantisation à 8 bits. Pour faire tourner des très grands modèles de langage à pleine puissance, un seul exemplaire de notre carte graphique n'aurait pas suffi, de même que pour le *fine-tuning* de *LLM (Large language models)*. Le développement de modèles maison peut être moins demandant en ressources mais l'investissement dans un GPU reste indispensable.

Cette nécessité de recourir à des machines puissantes rend l'impact environnemental de l'IA non négligeable, tant en termes de consommation d'énergie pour l'inférence qu'en raison de la nécessité d'investir dans des ordinateurs puissants. Le *Contre-atlas de l'Intelligence artificielle* écrit par la chercheuse australienne Kate Crawford fournit un détail des enjeux écologiques de l'IA [8]. Les ressources nécessaires à la fabrication de matériel informatique sont tout d'abord très importantes. Il faut du lithium pour construire les batteries des ordinateurs et les batteries de secours pour l'alimentation des centres de données. Ce métal provient de mines situées à des endroits géographiques divers, comme le Congo, la Bolivie ou la Mongolie. D'autres minéraux entrent dans la composition de matériel en plus du lithium. Dix-sept composants fossiles rares utilisés dans la production de matériel informatique sont en effet identifiés par Kate Crawford [9]. L'extraction de ces minéraux fossiles pose de nombreux problèmes éthiques. Au Congo par exemple, l'extraction du lithium est source de conflit. Des milices armées se battent pour le contrôle des mines. Les personnes chargées de l'extraction sont par ailleurs souvent exploitées et effectuent un

8. *Ibid.*
9. *Ibid.*

travail dangereux. Le matériel informatique utilisé pour produire des modèles d'IA ou les faire tourner est fabriqué à partir de matériaux nombreux quoi doivent être acheminés. Leur extraction et le processus d'assemblage sont également sources de pollution. A titre d'exemple, la chaîne d'approvisionnement de l'entreprise Intel comporterait plus de onze mille fournisseurs dans plus de quatre-vingt-dix pays [10].

Le développement de modèles maison plus légers a un impact écologique moins important que le développement de grands modèles de langage mais cet impact demeure. Les processus d'étiquetage des données, de développement et d'inférence nécessitent des ressources matérielles. L'entraînement des modèles, en particulier, requiert des machines qui fonctionnent pendant des heures, voire des jours. Dans le cas des *LLM (Large language models)*, il se fait dans de grands *datacenters* qui consomment beaucoup d'électricité et d'eau, principalement pour assurer le refroidissement des ordinateurs. Les entreprises qui produisent des gros modèles ont à cœur de « maximiser les cycles computationnels pour améliorer la performance » [11], or, ces cycles consomment énormément d'énergie. Dans des pays comme la Chine, les *datacenters* sont approvisionnés par de l'électricité majoritairement produite à partir du charbon [12], ce qui génère beaucoup de pollution.

Une des solutions envisagées pour limiter cet impact est l'utilisation de modèles avec le moins de paramètres possibles. De nombreuses entreprises travaillent actuellement à la réduction de la taille des grands modèles de langage, cherchant à développer des modèles plus compacts mais tout aussi performants. Cette approche pourrait réduire les coûts énergétiques et matériels, mais elle comporte aussi un risque d'effet rebond [13] ou de paradoxe de Jevons : la facilité d'utilisation et la légèreté de ces modèles pourraient inciter davantage d'acteurs à les déployer sur leurs systèmes, entraînant une augmentation globale de la consommation de ressources. Bien que certains tentent de mettre en avant des impacts écologiques positifs de l'IA, notamment pour la surveillance du changement climatique ou l'amélioration de l'efficacité des processus, ces avantages sont souvent contrebalancés par leur consommation en ressources ou des potentiels effets rebond. Lorsque des tâches sont optimisées, de nouvelles tâches émergent par ailleurs pour remplacer les précédentes. Les avantages potentiels de l'IA ne compensent pour l'instant pas son empreinte carbone, en particulier celle de l'IA générative.

Face à ces enjeux éthiques et écologiques, Yannick Meneceur propose un ralentissement : « les impacts sociétaux et environnementaux majeurs de la transformation numérique de notre société nous imposeraient donc, de manière raisonnable, de commencer

10. *Étude de cas d'Intel*, en, mai 2022, URL : `https : / / europeanpartnership - responsibleminerals . eu/blog/view/d3e23bef - 4b79 - 4f54 - 8338 - 8bf17cfc4b35/etude - de - cas - dintel` (visité le 30/08/2024).

11. Id., *Contre-atlas de l'intelligence artificielle…*

12. *Ibid.*

13. Thomas Guillory, Cyprien Tilmant, Alexis Trécourt et Lucie Gaillot-Durand, « Impacts environnementaux du numérique et de l'intelligence artificielle, à l'heure de la pathologie digitale », *Annales de Pathologie* (, juin 2024), DOI : `10.1016/j.annpat.2024.05.006`.

à ralentir au lieu de chercher à tout prix à accélérer[14] », selon lui, il faudrait « réserver le recours aux algorithmes à des besoins sectoriels très déterminés, avec une forte valeur ajoutée sociétale[15] ». On peut argumenter que les archives ont une grande valeur ajoutée sociétale, surtout dans un monde où la gouvernance de l'information prend de plus en plus de place mais au Luxembourg et dans bien d'autres pays, cette valeur n'est pas forcément reconnue. Dans le cadre d'une course à l'IA et d'un grand engouement, les projets se multiplient et leur objectif est de produire des résultats, peu importe leur forme et les cas d'usage. Il serait bénéfique pour les administrations de réfléchir à des moyens de mener ces derniers de manière responsable. Il est important de sensibiliser les acteurs aux impacts écologiques de l'IA et du numérique en général. L'organisation de *cleaning days* se fait de plus en plus dans les archives. Une journée mondiale « du nettoyage numérique », le *Digital Cleanup Day*, a été créée en 2019. Lorsqu'ils sont organisés par les archivistes, les *cleaning days* permettent une double sensibilisation du personnel des administrations, à la fois à l'archivage et à l'empreinte environnementale du numérique. La sensibilisation individuelle est-elle toutefois suffisante ? Est-ce qu'il faut attendre des innovations technologiques plus vertes et éthiques ou une régulation par les pouvoirs publics pour diminuer les impacts causés par l'IA et le numérique ? Les réponses à ces questions demanderaient bien plus de recherches que celles nous avons eu le temps de mener.

14. Yannick Meneceur, « Les trois grands défis posés par la gouvernance de l'intelligence artificielle et de la transformation numérique », *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, 23 (déc. 2021), DOI : 10.4000/ethiquepublique.6323.

15. *Ibid.*

Chapitre 9

Un processus d'évaluation long et nécessitant des connaissances techniques

1. Un processus itératif complexe à mettre en place

L'évaluation des modèles est un passage obligé dans la phase de recherche et développement de systèmes IA. Elle se fait à plusieurs étapes du projet. D'abord, il est nécessaire de sélectionner le modèle ou l'algorithme le plus adapté à la tâche d'automatisation à réaliser. Ce choix doit être basé sur une série de tests comparatifs. Cette phase de test est un processus long et exigeant en ressources puisqu'il convient d'évaluer différentes solutions afin de voir laquelle est la plus performante. C'est ce que souligne Yves Maurer, responsable de la division informatique et de l'innovation numérique à la BNL, à propos du *chatbot* développé par l'institution dans un entretien pour le magazine *Archimag* : « Le plus chronophage a été de tester plusieurs alternatives pour chaque brique du projet : des modèles de langage ouverts, un autre créé par un groupe de recherche, celui de Meta et de Google… [1] ».

Pour effectuer ces tests, il faut avoir à disposition un corpus de test étiqueté ou bien évaluer manuellement chaque modèle. Les problématiques d'évaluation seront détaillées dans la sous-partie suivante. Avant le choix, des recherches sur les modèles ou algorithmes les plus adaptés est à réaliser. Il existe en effet un grand nombre de modèles sur le marché, et ce nombre devrait s'accroître encore davantage dans le futur. Il est aisé de se perdre face au grand nombre de possibilités existantes, d'autant plus qu'il est difficile de prévoir la solution ou le modèle le plus performant sur la tâche à réaliser. Cela dépend des données disponibles, de leur volume, ainsi que de la nature de la tâche à accomplir. Le choix optimal peut donc sembler légèrement aléatoire, nécessitant une évaluation précise pour trouver la solution la mieux adaptée aux besoins spécifiques du projet. Une deuxième phase

1. C. Jost, *Comment la BNL a développé son chatbot basé sur ChatGPT…*

de tests et d'évaluation se fait une fois la méthode choisie, centrée sur l'optimisation des paramètres du modèle. Un grand nombre de paramètres est en effet modifiable. Nous avons déjà évoqué les paramètres mathématiques et les « *stop words* » dans le cas des algorithmes de classification automatique dans le chapitre 4 du présent mémoire. Dans le cas des *LLM (Large language models)*, de nombreux paramètres sont également modifiables et influent sur les performances. En voici une liste non exhaustive :

- **Température :** Ce paramètre contrôle la créativité des réponses générées. Une faible température favorise les choix de mots les plus probables, tandis qu'une température élevée encourage la génération de texte plus varié et créatif, mais potentiellement moins cohérent.
- ***Top-k* :** Limite la sélection des mots suivants à un sous-ensemble des k mots les plus probables, réduisant ainsi la complexité et augmentant potentiellement la cohérence des réponses.
- ***Top-p* ou *Nucleus Sampling* :** Ce paramètre permet aussi de sélectionner les mots suivants en fonction de leur probabilité, cette fois-ci non selon leur rang mais selon un niveau de probabilité cumulée minimal défini.
- **Pénalité de fréquence *(Frequency Penalty)* :** Une pénalité proportionnelle au nombre de fois où il a déjà été utilisé dans la réponse est appliquée au prochain *token*. Elle réduit donc la probabilité de répétition des mêmes mots, augmentant ainsi la diversité des réponses.
- **Pénalité de présence *(Presence Penalty)* :** Une pénalité est appliquée au prochain *token* s'il a déjà été utilisé dans la réponse. Elle est la même pour tous les *tokens* réutilisés, peu importe le nombre de fois où ils l'ont été. Elle réduit la redondance dans les réponses tout en aidant le modèle à rester concentré sur le sujet principal, évitant ainsi les digressions.
- **Nombre maximal de *token*s :** Définit la longueur maximale de la réponse à générer. Dans le cadre du projet InventAIre, par exemple, où seules des réponses booléennes ou des pourcentages de probabilité étaient nécessaires pour remplir les colonnes sur les données sensibles, nous avons limité le nombre maximal de tokens à 3 afin de n'obtenir que ces réponses. Cela permet surtout de gagner du temps et de réduire l'utilisation des ressources informatiques.
- ***Stop Sequences* :** Permet de définir des séquences de mots qui, une fois rencontrées, arrêtent la génération de texte, assurant ainsi un contrôle plus précis sur la sortie du modèle. Nous aurions également pu utiliser cette méthode dans le code de notre outil pour le projet InventAIre.
- **quantisation :** Technique de compression qui réduit la précision des calculs internes pour accélérer les inférences tout en maintenant des performances acceptables. L'outil développé pendant notre stage intègre une quantisation 4 bits pour les titres et

les descriptions et 8 bits pour le repérage des données sensibles. La quantisation 4 bits réduit la représentation numérique des poids des modèles, c'est-à-dire les paramètres internes du réseau de neurones qui sont ajustés pendant l'entraînement, à 16 niveaux distincts (2^4), tandis que la quantisation 8 bits utilise 256 niveaux (2^8).

Après avoir ajusté ces paramètres, l'outil doit être soumis à une évaluation rigoureuse. Les résultats de cette évaluation détermineront les conditions d'utilisation du modèle : si la précision est proche de 100%, une vérification humaine minimale sera nécessaire, tandis qu'une précision moyenne exigera un contrôle plus strict par l'humain.

Par la suite, des évaluations régulières du modèle sont à réaliser afin de garantir sa pertinence au fil du temps. La réactualisation d'un modèle d'intelligence artificielle peut impliquer un ré-entraînement pour améliorer ses performances ou l'adapter à de nouvelles données. Si le modèle a été développé à partir de zéro, il peut nécessiter un ré-entraînement complet pour intégrer des améliorations ou des changements dans les données. Si le modèle a été affiné, il devra l'être de nouveau. Dans le cas des systèmes basés sur du RAG (Retrieval Augmented Generation), cette réactualisation inclut l'injection de nouvelles données dans la base de connaissance et leur vectorisation pour enrichir le modèle avec des informations plus récentes et/ou pertinentes. Les *LLM* conversationnels nécessitent également une actualisation régulière. Cela peut inclure la modification des prompts pour mieux refléter les nouvelles données ou les exigences spécifiques du projet. Il est aussi possible d'adopter une version plus récente du modèle choisi, qui peut ainsi amener à des améliorations en termes de performance et de capacité, ou bien de sélectionner un autre modèle plus récent et plus puissant, capable de mieux répondre aux besoins du moment et d'intégrer les avancées technologiques les plus récentes. Dès lors, les prompts devront être actualisés selon les exigences des modèles. Une longue phase de tests et éventuellement de ré-entraînement est donc aussi à prévoir pour en cas de changement de modèle. Les réflexions sur l'actualisation des outils mis en place doivent être accompagnées d'une veille sur l'actualité des méthodes et des modèles de *machine learning*. Une partie du processus est automatisable. Par exemple, si les résultats fournis par la machine sont constamment vérifiés par l'humain, des statistiques peuvent être aisément réalisées sur le nombre de données produites ayant dû être corrigées par le personnel au fil du temps. Si les réponses des modèles sont utilisés telles quelles, sans processus de vérification, il faudrait penser à mettre en place des pratiques de test par échantillonnage réguliers.

Ce processus itératif est à penser en amont. Le temps qu'il demande et son coût peuvent être sous-estimés, en particulier quand le choix technique se porte sur les *LLM* conversationnels qui semblent être des outils « clé en main » alors que leur usage nécessite des compétences techniques non négligeables, en particulier sur l'automatisation de tâches complexes. Cette nature itérative a toutefois l'avantage de bien s'inscrire dans des pratiques de gestion de projet agiles cycliques. Des statistiques d'évaluation et des visualisations de données les plus neutres possibles et faciles à comprendre sont utiles

pour valider les résultats à la fin des cycles et orienter les décisions tout au long du projet. Néanmoins, dépendant de métriques élaborées, ces statistiques peuvent parfois être difficiles à établir.

2. Quelles métriques d'évaluation pour des tâches liées au langage ?

Pour que le processus d'évaluation soit précis, des métriques doivent avoir été définies en amont. Les métriques de base en apprentissage machine ou *machine learning* sont la précision et le rappel. La précision est le ratio de prédictions correctes par rapport à l'ensemble des prédictions et le rappel correspond au ratio de prédiction correctes par rapport à l'ensemble des entités qui devraient être identifiées. La précision mesure donc la capacité du modèle à éviter les faux positifs, tandis que le rappel évalue sa capacité à identifier toutes les entités pertinentes. A partir de cette précision et ce rappel peut être établi le F1 score, qui calcule une moyenne entre les deux située entre 0 et 1.

Le calcul de cette précision et ce rappel sont relativement simples dans le cas d'une classification ou de réponses booléennes. Par exemple, concernant les colonnes sur les données sensibles de notre inventaire, la réponse est soit vraie, soit fausse. Toutefois, en ce qui concerne de la génération de langage, cette analyse est plus subtile. Elles est donc plus complexe à automatiser. En effet, il est plus aisé de faire tourner différents modèles ou bien le même modèle dont les paramètres ont été modifiés sur des données préalablement étiquetées avec des « vrai » ou « faux » et de regarder le taux d'erreur. Des métriques spécifiques d'évaluation du langage généré par comparaison avec un texte de référence existent. Ces dernières permettent d'automatiser l'évaluation et donc de lutter contre les inconvénients liés à la subjectivité humaine, même si cette dernière reste présente dans le texte de référence, rédigé par l'humain. En voici une liste non-exhaustive :

- ***WER (Word Error Rate)*** : métrique au départ utilisée pour évaluer la précision des systèmes de reconnaissance vocale. Elle mesure le pourcentage d'erreurs, telles que les substitutions, les insertions et les suppressions, dans le texte transcrit par rapport au texte de référence.
- **Approches basées sur des modèles pré-entraînés** : mesurent la similarité entre le texte généré et le texte de référence en utilisant les *embeddings* de modèles pré-entraînés. L'évaluation est donc basée sur une similarité sémantique et contextuelle. Un exemple de métrique basée sur un modèle pré-entraîné est le *BERTScore*, avec le modèle *BERT* de *Google*.
- **Métriques développées pour l'évaluation de traductions** mais restant applicables à l'analyse de résumés.
    - ***BLEU (BiLingual Evaluation Understudy)*** : compare les n-grammes (mots ou séquences de mots) du texte généré aux n-grammes du ou de textes de

référence, en utilisant une formule qui mesure une précision avec un ajustement selon la longueur des phrases.

  - ***METEOR (Metric for Evaluation of Translation with Explicit ORdering)*** : mesure la qualité des traductions/résumés en comparant les n-grammes, en incluant leurs synonymes et la correspondance de leur radical, entre le texte généré et les textes de référence. Il s'agit d'une moyenne entre précision et rappel.

- ***ROUGE (Recall-Oriented Understudy for Gisting Evaluation)*** : développée pour mesurer la qualité de la génération de résumés, cette métrique mesure la similarité entre un résumé généré automatiquement et un ou plusieurs résumés de référence. Pour cela, les scores de précision et de rappel sont calculés en comparant les n-grammes présents dans le résumé généré avec ceux des résumés de référence. Une plus grande attention est accordée au score de rappel, qui évalue la quantité d'informations importantes des résumés de référence capturées par le résumé généré. [2].

Les métriques sont donc très diverses et ne sont pas forcément aisées à saisir. Elles nécessitent un travail de réflexion en amont, et parfois des moyens techniques conséquents lorsqu'elles sont basées sur des algorithmes complexes ou du *machine learning*. Il est nécessaire de bien les maîtriser pour fournir une évaluation la plus précise possible.

Des *benchmark*s plus généraux existent par ailleurs pour évaluer les grands modèles de langage. De nombreux sont disponibles en ligne [3]. Ils permettent d'avoir une idée générale de leur qualité. Dans le cadre de notre stage, pour choisir le modèle, nous avons expérimenté plusieurs méthodes d'évaluation. La première consistait en un système de notation. Un système de notation a également été utilisé dans le cadre de la sélection d'une méthode pour la réalisation du projet LlaMandement mentionné précédemment [4]. En ce qui concerne le projet InventAIre, le système de notation s'est rapidement avéré subjectif mais a permis d'éliminer les modèles les moins précis. Face à cette question de la subjectivité, les personnes chargées de l'évaluation sur le projet LlaMandement avaient dans leur corpus des résumés rédigés par des humains. La note moyenne octroyée à ces derniers s'est élevée à 16,5, ce qui donne plus de crédibilité à l'évaluation et donne une cible à atteindre par l'IA [5]. Après avoir expérimenté avec des notes, nous avons mis au point une évaluation comparative. Nous comparions les titres et descriptions générés par plusieurs modèles sur une même unité de description afin d'établir un classement [6].

2. Blog ML Explained, *Large Language Model (LLM) Evaluation Metrics – BLEU and ROUGE*, en, juill. 2023, URL : `https://mlexplained.blog/2023/07/08/large-language-model-llm-evaluation-metrics-bleu-and-rouge/` (visité le 30/08/2024).

3. Voir par exemple ceux disponibles ici : `https://huggingface.co/collections/open-llm-leaderboard/the-big-benchmarks-collection-64faca6335a7fc7d4ffe974a`

4. J. Gesnouin, Y. Tannier, C. G. Da Silva, *et al.*, *LLaMandement…*

5. *Ibid.*

6. Plus d'informations sur l'évaluation dans la note méthodologique en annexe (partie 2.3.1. et

Ce système garde une forme de subjectivité mais il est moins subjectif qu'un système de notation. Nous avons effectué le travail à l'envers par rapport à ce qui se fait normalement en *machine learning*. Un corpus de documents préalablement étiqueté est normalement séparé en deux parties, une pour l'entraînement, une pour la validation. Adopter cette approche, et ainsi avoir préalablement étiqueté des données de test, aurait poussé l'équipe à réfléchir davantage en amont afin de se mettre d'accord sur le contenu des différentes colonnes. Une forme de consensus est effectivement nécessaire pour l'étiquetage de données ou l'évaluation de performances. Par exemple, une donnée peut être considérée comme sensible par un archiviste et non par un autre. Sans lignes directrices précises, il peut y avoir autant d'inventaires que d'archivistes. Face à ce souci de subjectivité, nous avons mis en place un système d'évaluation finale en trois possibilités : « correct », « incorrect » et « sujet à débat/insuffisant ». À la Cour de cassation française, pour mener à bien le projet de pseudonymisation, une « élaboration théorique des catégories et [une] définition des types de contentieux [7] » a été réalisée par des juristes, malgré tout, l'exécution des conseils donnés par les juristes est complexe dans la pratique. Certains termes peuvent appartenir à plusieurs catégories ou bien à aucune des catégories définies par les juristes [8]. Il y a alors un dialogue qui se crée et certaines catégories sont précisées. Le fait d'étiqueter en amont est donc bénéfique pour amener la discussion et préciser les définitions des différentes catégories ou données à repérer, ou bien pour préciser la forme et le contenu du texte à rédiger. Un étiquetage des résumés et descriptions en amont aurait rendu l'exécution du projet InventAIre davantage productive.

La précision n'est pas la seule métrique à laquelle nous nous sommes intéressée dans le cadre de notre stage. Nous avons également examiné le temps de génération, en calculant des moyennes par prompt. Il aurait été davantage pertinent de mesurer le temps par *token* dans le prompt, puisque ces derniers peuvent être de tailles différentes selon la taille du document qu'il inclut. Il y a une inférence par colonne de l'inventaire, et parfois plus d'une inférence par ligne dans le cas de documents longs qui, dépassant la capacité de la fenêtre du *LLM*, seraient divisés en plusieurs parties. Ce temps s'accumule vite et la génération d'un inventaire peut prendre des jours entiers si l'on n'y fait pas attention. Nous aurions aussi pu évaluer les ressources dépensées pour chaque inférence par la machine.

Nous nous sommes enfin intéressée à l'évaluation des biais. Cette dernière est complexe. Il existe des *benchmarks* publics pour estimer à quel point les modèles les plus utilisés sont sujets au biais, ainsi que des *datasets* publics pour tester les modèles maison ou affinés. C'est ce qui a été fait dans le cadre du projet LlaMandement. Le modèle fine-tuné a été testé sur le jeu de données spécialisé *BOLD (Bias in Open-ended Language Generation Dataset)*. La réalisation de ce test suppose que les biais en français sont les

annexe 4 et 5)

7. C. Girard-Chanudet, « Le travail de l'Intelligence Artificielle… ».
8. *Ibid.*

mêmes qu'en anglais [9], ce qui n'est pas forcément vrai. Des métriques de mesurant les biais ont par ailleurs été calculées. Il s'agit de *Regard* et *Honest*. *Regard* évalue la polarité du langage et les perceptions sociales de divers groupes démographiques (par exemple selon le genre, l'appartenance ethnique et l'orientation sexuelle), en identifiant les biais dans le ton ou avec une analyse des sentiments qui pourraient influencer la représentation de ces groupes [10]. *Honest* mesure la fréquence des complétions de phrases offensantes dans les modèles de langage, en utilisant un lexique multilingue pour fournir un score de toxicité et détecter les disparités potentielles entre différents groupes démographiques [11].

Les réflexions sur l'évaluation seront à pousser en cas de suite du projet InventAIre. Il s'agit d'une étape qui se répète tout au long du projet. Elle doit être ainsi optimisée, voire rationalisée, pour éviter de perdre trop de temps. Ces évaluations sont déterminantes dans le succès des projets et sont indispensables à la mise en production de systèmes basés sur de l'IA.

9. J. Gesnouin, Y. Tannier, C. G. Da Silva, *et al.*, *LLaMandement…*
10. *Ibid.*
11. *Ibid.*

Chapitre 10

Problématiques de mise en production

1. Une chaîne de traitement à mettre en place

Le choix d'un algorithme ou d'un modèle pour accomplir une tâche donnée n'est que la première étape vers le développement d'une solution d'automatisation basée sur du *machine learning*. Le modèle sélectionné traite des données en entrée et génère une sortie, qui doit ensuite être intégrée dans une chaîne de traitement plus large. Ces données de sortie ne sont en effet pas directement exploitables en l'état. Dans cette section, nous décrivons la chaîne de traitement mise en place pour le projet InventAIre, afin de prendre du recul, en analysant ses forces et ses faiblesses.

Le schéma ci-dessous résume ce processus général de traitement mis en place, qui sera expliqué plus en détail, étape par étape.

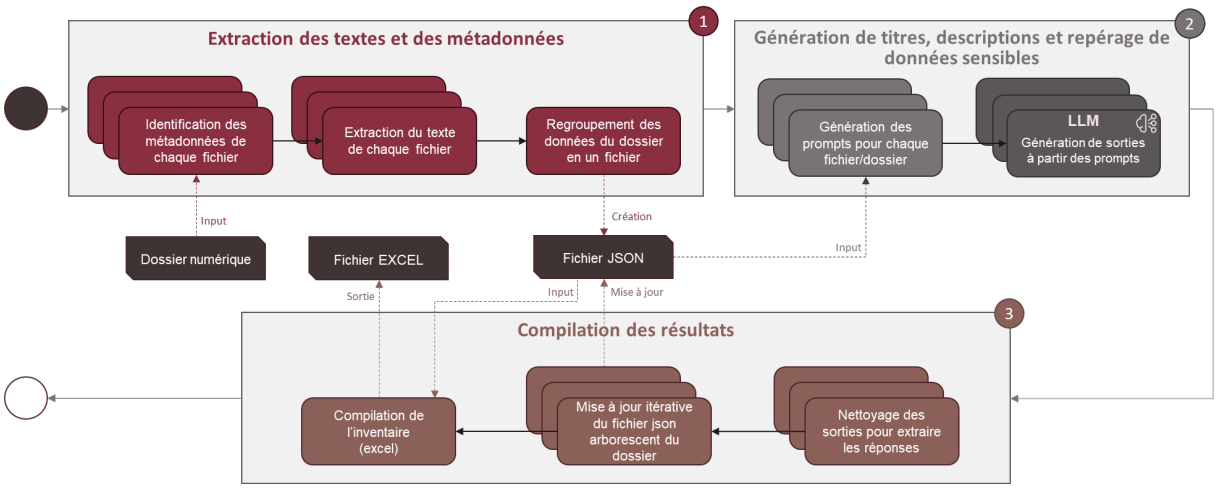

FIGURE 6 – Processus de traitement mis en place afin de générer l'inventaire

Étape 1 : Extraction des textes et des métadonnées des documents

L'application prend en entrée un chemin de répertoire numérique. Les métadonnées extractibles de l'intégralité du contenu de ce dernier sont compilées en un seul un fichier au format JSON, avec le texte de chaque document. Si le texte n'est pas disponible, une tentative d'océrisation est réalisée avec le logiciel Tesseract. Dans le cas où les textes des fichiers seraient trop grands et dépasseraient la fenêtre de Llama 3, ils sont divisés en différentes parties, que nous avons nommées « chunks », de 10 000 caractères. Ce découpage est relativement arbitraire, nous aurions pu couper au niveau de la fin d'une phrase pour que cela perturbe moins notre LLM. Nous aurions également pu tester avec des chevauchements entre les différents « chunks ». Ci-dessous se trouve un exemple de fichier JSON arborescent produit pour un dossier qui contiendrait un seul document.

Exemple de fichier JSON arborescent pour un dossier contenant un document


```
{
    "path": "C:\User\Chemin\du\dossier_exemple",
    "folderName": "dossier_exemple",
    "type": "directory",
    "content": [
    {
        "path": "C:\User\Chemin\du\dossier_exemple\document.txt",
        "type": "file",
        "fileName": "document.txt",
        "creationDate": "Wed May 20 10:26:26 2024",
        "modificationDate": "Wed May 29 10:35:06 2024",
        "mimeType": "text/plain; charset=UTF-16LE",
        "texte": "Ceci est le texte d'un document txt.",
        "generatedTitle": "to be defined"
    }
    ],
    "generatedTitle": "to be defined"
}
```


Étape 2 : Génération de titres, descriptions et repérage de données sensibles

A partir du document JSON, des prompts sont ensuite générés pour chaque niveau dans l'arborescence et chaque donnée sensible. Le LLM est interrogé avec ces différents prompts. L'encadré ci-dessous en montre un exemple. En rose se trouvent les éléments mutables selon le type d'entrée (document, dossier, chunk) et en vert se trouvent les

éléments mutables selon la colonne à remplir.

Exemple de prompt

Input :


```
{
    "path": "C:\User\Chemin\du\dossier_exemple\document.txt",
    "type": "file",
    "fileName": "document.txt",
    "creationDate": "Wed May 20 10:26:26 2024",
    "modificationDate": "Wed May 29 10:35:06 2024",
    "mimeType": "text/plain; charset=UTF-16LE",
    "texte": "Ceci est le texte d'un document txt.",
    "generatedTitle": "to be defined"
}
```


Context : You are an achivist at the Chambre des Députés of Luxembourg.
Question : You have been given a full document and some metadata formatted as json as input. You are working on an archival inventory project. Please generate one title (max 20 words) in french for this document for the inventory.
In your title, avoid the word 'fonds'. Avoid generic words such as 'divers' or 'variés'. Also avoid the word 'documents' without further precisions : add the types or names of the documents being exhaustive (all of the major types). It would be better if you could format it this way : 'subject, topic - action/type of document', only do it if there is a clear topic or subject and a clear action or only one clear types/names of documents. If there is a chronological indication in your title (not madatory, only if you are sure and if it is pertinent, must be with the topic if it is the date of the event that constitutes the topic) the format is day (number) month (letters, no abbreviation) year (number). If there are several parts in your title, separate them with commas. Be precise and exhaustive in your title (remember : topic/subject - action(s)/precise type(s) of documents if you can) but do not put information you are not sure about. If the document's text does not give you information, do not make it up. Make sure to mention the important people, organizations, locations, dates, ... if you identify some.
Examine well the whole content for this. Please only write the title in french, nothing else, so I can directly reuse it in my inventory. Do not explain your answer or tell me about alternatives in your answer. I only need the title. Do not invent information, only use the ones you have. It's ok if your answer is minimalist because you do not have enough info. Be careful, sometimes a file contains several types of document.
Make sure the syntax and spelling in french in your answer are correct.
title in french :

Quand le document est un dossier, c'est une arborescence plus grande qui est donnée en entrée, avec des titres et descriptions déjà générés, parce que l'ensemble des textes serait trop long et dépasserait la fenêtre du *LLM*. Il en va de même pour la génération d'un titre ou d'une description pour un document composé de « chunks » : un résumé de chaque partie du document est généré et c'est à partir de ces résumés que sont générés les descriptions et titres de ce dernier.

Étape 3 : Compilation des résultats obtenus par le *LLM*

Les réponses sont extraites des résultats fournis par le *LLM* : selon le prompt, on extrait le « oui »/« yes », le « non »/« no » ou un pourcentage de probabilité généré pour remplir les colonnes sur les données sensibles avec des « oui » ou des « non ». Le fichier JSON est par la suite mis à jour de manière itérative après chaque inférence du modèle, qui correspond à un prompt. La mise à jour du fichier JSON part du bas de l'arborescence pour remonter vers la racine. Cela permet de décrire chaque niveau en fonction de son contenu. S'il y a un grand nombre de niveaux, les niveaux supérieurs seront décrits en fonction du titre et de la description générés pour les dossiers et fichiers qu'ils contiennent. L'avantage est la description de chaque niveau. Les données sensibles repérées sont également compilées au fur et à mesure par niveau. Un inconvénient réside dans le fait que la description est de moins en moins précise en remontant dans l'arborescence. Les erreurs sur les données sensibles s'ajoutent également, le niveau supérieur d'une grosse arborescence contient donc en général beaucoup de données sensibles alors que ce n'est pas forcément le cas.

Une fois l'inférence réalisée sur l'ensemble du fichier JSON et pour toutes les colonnes de l'inventaire, ce dernier est compilé en un fichier Excel avec une ligne par niveau dans l'arborescence.

Path	Type	Titre	Description	Période de création : de	Période de création : à	Données à caractère personnel	Atteinte au relations extérieures, à la sécurité ou à l'ordre public	Affaires portées devant les instances juridictionnelles, extrajudiciaires, disciplinaires	Prévention, recherche, poursuite de faits punissables	Données commerciales et industrielles	Documents ayant trait au secret fiscal
2005 CANADA	directory	Christos Sirro	Ce dossier coi	08/06/2006	18/07/2024	non	non	non	non	non	non
2006 Québec Sirros Christons représentant	directory	"Christos Sirrc	Ce dossier coi	19/06/2006	18/07/2024	non	oui	oui	non	non	non
2007 Canada ambassadrice Glasgow	directory	"L'ambassadri	Ce dossier coi	25/01/2007	18/07/2024	non	oui	oui	non	non	non
2008 Visite ambassadeur de la pêche SEM SU	directory	"Visite d'ambi	Dossier conte	06/10/2008	18/07/2024	non	non	non	non	non	non
2008 Visite ambassadeur de la pêche SEM SU	directory	"Visite de l'am	Dossier conte	30/10/2013	18/07/2024	non	non	oui	non	non	non
2012 Visite parlementaire\CV	directory	"Visite parlem	Ce dossier coi	11/01/2012	18/07/2024	non	oui	non	non	non	non
2012 Visite parlementaire\Documentation	directory	"Politique can	Dossier conte	25/01/2012	18/07/2024	non	oui	non	non	oui	non
2012 Visite parlementaire\Notes MAE	directory	"Relations coi	Ce dossier coi	19/01/2012	18/07/2024	non	non	non	non	non	non
2012 Visite parlementaire\Programme	directory	"Visite parlem	Ce dossier coi	25/01/2012	18/07/2024	non	oui	oui	non	non	non
2012 Visite parlementaire	directory	"Visite parlem	Ce dossier coi	25/01/2012	18/07/2024	non	oui	oui	non	oui	oui
2015 1617 04 comité exécutif de l'Associatio	directory	"Biographies (	Dossier conte	08/04/2015	18/07/2024	non	non	non	non	non	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Relations Ca	Ce dossier coi	14/04/2015	18/07/2024	non	oui	non	non	oui	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Comité exéc	Ce dossier coi	05/05/2015	18/07/2024	non	non	non	non	non	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Procès-verba	Ce dossier coi	15/12/2016	18/07/2024	non	non	oui	non	non	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Reconnaissa	Ce dossier coi	16/06/2015	18/07/2024	non	non	non	non	non	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Relations Ca	Ce dossier coi	08/04/2015	18/07/2024	non	non	non	non	non	non
2015 1617 04 comité exécutif de l'Associatio	directory	"Comité exéc	Ce dossier coi	15/04/2015	18/07/2024	non	oui	oui	non	oui	oui
2016 AUDET Michel Visite de courtoisie	directory	"Visite de cou	Dossier conte	28/09/2016	18/07/2024	non	oui	non	non	oui	non
2016 Pierre Marc Johnson\Revue de presse	directory	"Étude de pre	Ce dossier coi	21/01/2016	18/07/2024	non	non	non	non	non	non
2016 Pierre Marc Johnson	directory	"Pierre Marc J	Ce dossier coi	13/01/2016	18/07/2024	non	oui	oui	oui	oui	non
2017 Visite Assemblée du Québec	directory	"Visite de la d	Ce dossier coi	24/05/2017	18/07/2024	non	oui	oui	non	oui	oui
2018 Audet entretien Angel 18 janvier 2018	directory	"Biographie e	Dossier conte	15/01/2018	18/07/2024	non	non	non	non	non	non
2019 Québec Pierre-Luc Desgagné	directory	"Pierre-Luc De	Ce dossier coi	26/09/2019	18/07/2024	non	oui	non	non	non	non
2022 Orateur Penny QUébec	directory	"Politique scie	Ce dossier coi	26/10/2022	18/07/2024	non	non	oui	non	non	non
Ambassade\ 2009 Canada ambassadeur De Li	directory	Accréditation	Dossier conte	15/06/2011	18/07/2024	non	oui	oui	non	non	oui

FIGURE 7 – Exemple de tableau Excel contenant les résultats d'une inférence

Avec un peu de recul, nous pouvons dire que cette chaîne de traitement et le code en général sont assez complexes. Elle n'est pas composé de beaucoup d'étapes très distinctes

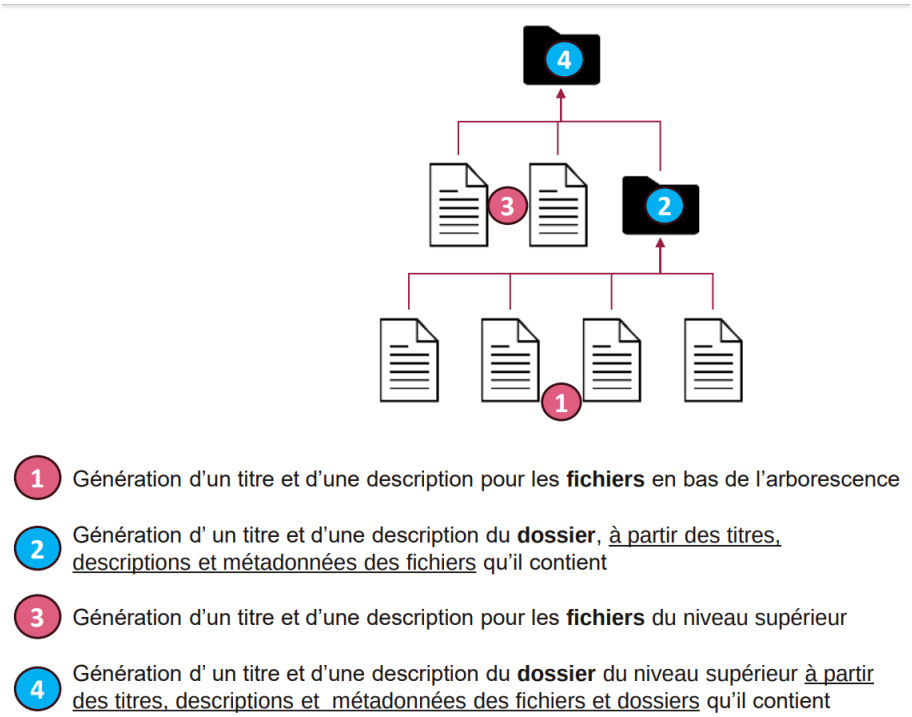

FIGURE 8 – Schéma du fonctionnement de l'inférence pour traiter chaque niveau d'arborescence dans le cas de la génération de titres et descriptions

les unes des autres, mais le traitement itératif pour chaque niveau de l'arborescence et chaque colonne à remplir est long, composé de beaucoup de variables et peut donc être difficile à saisir. La mise en place d'un outil de traitement qui contient une composante IA demande des capacités techniques et peut être un réel défi intellectuel, comme cela a été le cas pendant notre stage. Nous avons eu du mal à réaliser un outil capable de traiter tous les niveaux et la solution choisie n'est peut-être pas la meilleure. On pourrait la qualifier de « bricolage ». Il en va de même pour les prompts. Nous avons réalisé de nombreux tests pour voir lesquels fonctionnaient le mieux. Au final, cela a produit des prompts très longs et peut-être trop précis parfois.

Nous avons souhaité qu'il y ait une grande granularité dans la description. Elle comporte plusieurs niveaux et se répète parfois, ce qui est contraire à certaines normes de description archivistique, comme la norme ISAD(G). Ce n'est pas forcément utile, cela peut générer du bruit en cas de recherche ou bien générer des descriptions peu précises et ainsi inutiles. Il y a un juste milieu à trouver entre ce qu'il est utile de décrire et ce qu'il est possible de faire grâce à l'IA. En France, à l'INA, un travail a été réalisé pour trouver cet équilibre. Les pratiques de description ont même évolué grâce à l'IA. D'après Eléonore Alquier, directrice adjointe Data & Technologies, « l'enjeu du recours à l'IA n'est plus de poursuivre à l'identique la production documentaire "classique", mais de générer, de manière industrielle et sur la durée, des clés d'entrée nouvelles dans les collections,

sur un mode analytique et synthétique[1] ». Et si l'IA n'était pas seulement une solution d'automatisation de tâches métier mais le moyen d'ajouter autre chose, de faire évoluer les pratiques archivistiques ?

Lorsque les tâches à réaliser demandent beaucoup de réflexion ou d'étapes, les chaînes de traitement contenues dans les outils développés s'alourdissent et les automatisations peuvent perdre en précision. L'IA, bien que puissante, doit être intégrée avec précaution et réflexion dans les pratiques archivistiques existantes pour en maximiser l'efficacité tout en évitant les écueils liés à des tentatives d'automatisation excessives ou mal adaptées.

2. Interface et intégration à la chaîne archivistique

Ce long processus doit par la suite s'intégrer à la chaîne archivistique. La chaîne archivistique peut se définir comme « l'ensemble des activités de l'archiviste, depuis la collecte jusqu'à la communication éventuelle des documents, en passant par le traitement physique et intellectuel des documents[2] ». Ces activités sont des processus en eux-mêmes et peuvent être traités à l'aide de différents outils. Par exemple, il existe des logiciels de gestion et de description des archives.

Les systèmes IA développés doivent pouvoir être facilement utilisables par les équipes afin de s'intégrer efficacement à cette chaîne. Pour cela, la production d'un logiciel avec une interface est indispensable : tous les archivistes ne maîtrisent pas les outils qui fonctionnent en ligne de commande. Ce dernier ne doit pas non plus exiger une formation trop longue. Il doit être intuitif et engageant. Un exemple d'interface de système basé sur le *machine learning* qui pourrait être améliorée est celle de Pêle-mél. Bien que fournissant des visualisations utiles sur les contenus des boîtes mail, elle peut être jugée comme trop complexe, ce qui peut freiner l'adoption d'un outil malgré sa qualité. Toutefois, elle a l'avantage de contenir beaucoup de moyens de personnalisation. Le bilan du projet mentionne que « les participant · es ont compris et admis le côté expérimental du projet qui se traduit dans des interfaces austères, et aimeraient bien évidemment des interfaces plus conviviales et surtout un développement de l'interface de classification sous Windows[3] ». Une interface simple permet donc de faciliter l'adoption des outils et par la même occasion la conduite du changement pour les équipes.

L'outil développé pendant notre stage est une application web construite à l'aide du *framework* Flask en langage Python. Hébergée localement sur l'ordinateur acheté par la Chambre, l'application peut être lancée via une URL dans un navigateur web. Elle peut

1. Eleonore Alquier, « L'intelligence artificielle à l'INA : de l'expérimentation à l'industrialisation », dir. Association des archivistes français, *Archivistes !*–147 (2024).

2. Elisabeth Bellion, « Journée d'études : Les revues et leurs archives. Méthodologie d'archivage », carnet Hypothèses *ArchiSHS*, URL : `https://archishs.hypotheses.org/tag/chaine-archivistique`

3. B. Grailles, *Pêle-mél. Plate-forme d'exploration, de livraison et d'évaluation des méls. Rapport d'évaluation des usages…*

être également accessible à distance en se connectant à l'ordinateur comme serveur. Les applications web ont l'avantage d'être familières aux équipes, qui en utilisent couramment dans leur travail et leur vie quotidienne. Elles sont davantage ergonomiques et facilement personnalisables grâce aux langages HTML, CSS et JavaScript. Toutefois, cette personnalisation ajoute des couches de complexité, rendant le code plus lourd. Notre application comporte en effet plusieurs niveaux de complexité :

- Une couche en Python, composée elle-même de plusieurs couches rendues invisibles :
  - *Apache Tika* pour l'extraction des texte contenus dans les documents, codé en Java
  - *Tesseract* pour l'océrisation, codé en C++
  - Une couche de *machine learning* codée dans d'autre langages tels que le C, C++, CUDA

  Le langage Python présente l'avantage de permettre l'intégration de nombreux outils, même si cela peut complexifier l'architecture des applications.
- Une couche interface, dans les langages de développement web HTML/CSS/JavaScript

Bien que Flask ne soit peut-être pas l'outil idéal en raison de cette complexité, sa syntaxe est relativement simple et nous le maîtrisions assez bien. L'absence de base de données derrière l'application simplifie par ailleurs son développement, bien que, dans l'idéal, un historique des inventaires générés serait utile. Nous avons enfin utilisé Bootstrap, *framework* qui fournit des outils pour concevoir des sites web responsive et aux visuels modernes grâce à un ensemble de composants CSS et JavaScript préconçus. Tous ces outils nous ont permis de réaliser une interface simple et *responsive* rapidement (environ deux jours de travail pour l'intégration du processus de traitement dans une application web et le code du *front-end*) et en peu de lignes de code. L'application réalisée permet à l'utilisateur d'entrer un chemin de fichier, de cliquer sur un bouton pour lancer le processus de traitement. Une fois généré, l'inventaire est visualisable sous forme de tableau. Deux boutons permettent aussi de le télécharger au format Excel ou JSON. L'interface est simple, avec seulement deux pages. L'archiviste n'a qu'à sélectionner une arborescence, lancer le processus, et peut revenir lorsque l'inventaire est prêt afin de le télécharger. Une fonction a été intégrée pour qu'une inférence lancée se poursuive même en cas de mise en veille de l'ordinateur.

Le développement d'interfaces utilisateur attrayantes a des apports annexes. Il est bénéfique pour les démonstrations et la valorisation des projets. Les interfaces contenant des datavisualisations sont par exemple intéressantes afin de montrer des volumes de données traitées et les capacités du *machine learning* dans une optique de médiation.

Les interfaces de visualisation pour les processus d'analyse doivent avoir une fonctionnalité réelle et ne pas constituer une charge mentale supplémentaire pour l'archiviste.

La simplicité est bénéfique pour les outils d'automatisation des processus archivistiques. Moins il y a de fonctionnalités et d'options compliquées, aux valeurs ajoutées moindres, moins il y a de risque de confusion. Il est possible d'aller droit au but. Dans notre cas, il n'y a qu'une seule fonctionnalité. Il serait toutefois possible d'en ajouter si elles ont une réelle valeur ajoutée. Par exemple, il peut être envisageable d'intégrer les fonctionnalités des scripts shell développés par les ANLux d'automatisation des étapes de prétraitement des vracs numériques. Un exemple d'interface prenant bien en compte les enjeux évoqués est celle d'*Archifiltre*. L'interface est épurée et intuitive. Une sensation de sécurité informatique est fournie par le fait qu'il s'agisse d'un logiciel installé localement. Les manipulations sur les dossiers paraissent également plus transparentes parce que l'interface les rend visible. Cette perception de transparence est un atout pour assurer l'usage d'un logiciel, bien que le code sous-jacent soit complexe et invisible pour l'utilisateur. La transparence n'est donc qu'illusoire. Cela peut être un inconvénient. Les interfaces actuelles tendent à rendre les processus computationnels lourds invisibles. La recherche dans le domaine des interactions homme-machine a théorisé la réduction de la visibilité des processus informatiques en faveur d'une meilleure expérience utilisateur [4]. De nombreux niveaux de complexité sous-jacents sont masqués par des interfaces épurées et des temps de traitement qui sont poussés à l'optimisation. Pour l'archiviste, le fait que l'outil puisse s'intégrer efficacement dans chaîne archivistique est un avantage, mais peut donc également nuire à la transparence des systèmes basés sur du *machine learning*. Il n'a pas forcément de visibilité sur les technologies employées. Dans le cas du prototype d'application que nous avons développé, il faudrait que le fait qu'il y a un *LLM* en arrière-plan soit davantage explicite dans l'interface de l'application. Cela doit être écrit, et, les utilisateurs ne lisant pas forcément les textes, l'ajout d'icônes qui font écho à l'IA peut être complémentaire, pour les aider à en prendre conscience, tout en veillant à ce que le design soit esthétiquement plaisant.

Par ailleurs les interactions homme-machine sont souvent qualifiées avec un champs lexical humain : la machine est vue comme un partenaire [5]. Nous avons déjà évoqué cette théorisation des IA comme des assistants dans le troisième chapitre de ce mémoire. Des IA qui paraîtraient trop humaines peuvent être sources de danger. C'est aussi le cas du format conversationnel des *chatbots* tels que de *ChatGPT*, épuré et facile à utiliser, avec une seule fonctionnalité : la conversation. La simplicité de l'interface et son aspect anthropomorphe, via des conversations en langage naturel, rendent les réponses fournies davantage crédibles pour le grand public, laissant peu de place au doute sur les informations contenues. Sur le même principe, une interface peut influencer la perception de l'utilisateur en limitant les possibilités de personnalisation des paramètres et les possibilités de décision [6]. Cela

4. David Pucheu, « Effacer l'interface : Une trajectoire du design de l'interaction hommemachine », *Interfaces numériques*, 5–2 (mai 2018), p. 257-276, DOI : 10.25965/interfaces-numeriques.3044.

5. *Ibid.*

6. *Ibid.*

contribue à réduire le sentiment d'intentionnalité et de responsabilité, et peut constituer un risque en ce qui concerne les systèmes IA.

Ainsi, des interfaces bien conçues améliorent l'accessibilité des outils basés sur du *machine learning*, permettant une intégration plus fluide dans les processus de travail de l'archiviste, mais elles peuvent masquer de l'information. Elles ne garantissent également pas forcément une meilleure littératie numérique des archivistes. Néanmoins, un nombre très limité d'archivistes maîtrise la programmation informatique aujourd'hui au Luxembourg donc il semble préférable, dans un premier temps, de privilégier la simplicité. Dans ce cas, il paraît essentiel d'assurer une bonne formation des utilisateurs sur les outils et leurs enjeux en cas de mise en production.

3. De l'informatique recherche au déploiement

Nous avons abordé les chaînes de traitement potentielles intégrées dans les outils d'IA ainsi que la question de l'intégration dans les processus spécifiques au métier de l'archiviste. Il s'agit ici de réfléchir à l'incorporation dans l'architecture informatique plus globale de l'administration.

Nous avons pu examiner les différents niveaux de complexité de l'application développée durant notre stage dans la sous-partie précédente. Ce prototype, issu d'une réflexion informatique recherche, n'est pas optimisé. Sa complexité est élevée, avec de nombreuses boucles et conditions. Le traitement centré autour d'un grand fichier JSON hiérarchique n'est probablement pas le plus efficace. De plus, les dépendances sont nombreuses, avec plusieurs libraires Python, le logiciel Tesseract et un grand modèle de langage à installer. Les performances de l'ordinateur n'ont pas été optimisées non plus. Une amélioration serait nécessaire pour traiter efficacement les modèles de *machine learning* très exigeants. Il aurait été pertinent d'améliorer le code qui vient avant et après le traitement par le modèle, ainsi que d'explorer des méthodes de parallélisation. La parallélisation permet d'exécuter plusieurs tâches simultanément, ce qui peut grandement améliorer l'efficacité du traitement. Pour le *LLM*, cela pourrait inclure le traitement par lots (ou *batch processing*) pour l'inférence, leur permettant de traiter plusieurs entrées en une seule fois au lieu de les traiter individuellement.

Bien que le code soit documenté et commenté, ce qui facilite sa reprise, pour un déploiement potentiel, il faudra réfléchir à la manière dont l'application pourrait s'intégrer dans le système d'information de la Chambre. Il sera nécessaire de vérifier sa compatibilité avec les technologies existantes, de déterminer son emplacement sur les serveurs. L'intégration dans l'architecture globale de l'institution nécessitera une réflexion sur la gestion de son éventuelle base de données et la mise en place d'une équipe pour sa maintenance et son évaluation. Cela implique des ressources et des compétences spécifiques, qu'il faut parfois aller chercher dans le secteur privé.

Ces étapes sont encore loin d'être achevées, car notre démarche était axée sur la recherche : l'objectif était de réaliser des tests et de produire une application qui marche. Le code pour la recherche est conçu pour être fonctionnel, tandis que le code orienté vers la production doit être optimisé, lisible et compréhensible. Il a ses propres codes esthétiques [7]. Les bonnes pratiques en programmation logicielle ont été établies dans les années 1970 pour éviter les échecs dans les projets et pour éviter des logiciels difficiles à maintenir lors des changements d'équipe [8]. Des guides et des normes existent. Nous avons par exemple suivi la norme de Google [9] pour le commentaire de nos fonctions en Python et le guide de style du langage *PEP8* [10]. Ce dernier couvre des aspects tels que l'indentation, les espaces, le nommage des fonctions et variables, ainsi que les commentaires. Nous avons également veillé à typer nos fonctions, c'est-à-dire à spécifier les types de données attendus en entrée et en sortie pour chacune d'entre elle. Un exemple de fonction typée et commentée ci-dessous donne une idée de ce à quoi peut ressembler la documentation dans le code.

```
def format_dates(df: pd.DataFrame) -> None:
"""
Reformate les dates dans un DataFrame au format jj/mm/aaaa.

Args:
df (pd.DataFrame): Le DataFrame contenant les dates à reformater.
"""
date_columns = ["Période de création : de","Période de création : à"]
for col in date_columns:
    df[col] = pd.to_datetime(df[col],
       errors='coerce').dt.strftime('%d/%m/%Y')
```

Malgré ces efforts, notre code complexe et non optimisé. Le code des logiciels doit être efficace et durable. Il doit être lisible et bien commenté pour faciliter la maintenance, le débuggage et son éventuelle amélioration. Les lignes superflues doivent être supprimées ou réduites, un accent doit être mis sur la rapidité d'exécution et la réduction des ressources utilisées. Ces économies sont souvent motivées par des logiques commerciales. Cependant, il est important de noter que ce que l'on pourrait qualifier de « bidouillage » ou « bricolage » reste présent dans la programmation orientée production [11]. Des discussions avec un consultant de développement à la Chambre des Députés ont suggéré que la qualité du code, en termes d'optimisation et de documentation, peut varier en fonction

7. Pierre Depaz, *The role of aesthetics in understanding source code*, These de doctorat, Paris 3, 2023, URL : `https://theses.fr/2023PA030084` (visité le 12/08/2024).
8. *Ibid.*
9. Plus de détails ici : `https://github.com/google/styleguide/blob/gh-pages/pyguide.md#383-functions-and-methods`
10. Disponible ici : `https://peps.python.org/pep-0008/`
11. *Ibid.*

des équipes et des pratiques institutionnelles.

Le passage de l'informatique recherche, qui a davantage à cœur d'expérimenter et dont le but est de produire quelque chose de fonctionnel, à un outil maintenable et optimisé pour la production est un processus long. Un exemple de déploiement à grande échelle de l'IA dans le domaine des sciences de l'information est celui de l'INA en France. Eléonore Alquier note à propos de la segmentation automatique des journées de diffusion archivées qu'« elle a nécessité plusieurs années de tests et d'itérations, mais aussi d'approbation des enjeux de l'automatisation pour répondre aux attendus fonctionnels [12] ». Même si une application marche et répond à un besoin, un long processus de mise en production est à prévoir. L'informatique recherche est l'occasion d'expérimenter mais il est nécessaire de garder toutes ces questions en tête pour ne pas faire un travail qui ne sera pas maintenable ni réutilisable.

Malgré leurs apports potentiels, le déploiement de systèmes IA est complexe et doit s'inscrire dans une réflexion plus globale. Des enjeux éthiques et écologiques sont à prendre en compte. Des réflexions d'ordre technique sont également à développer. Ces systèmes posent des défis significatifs en matière d'explicabilité, d'évaluation et d'intégration dans les processus de travail.

12. E. Alquier, « L'intelligence artificielle à l'INA : de l'expérimentation à l'industrialisation »…

Conclusion

Le contexte archivistique et le contexte public luxembourgeois incitent à l'expérimentation de moyens d'automatisation. De vastes volumes d'archives papier ou numériques sont à traiter et décrire. Les personnes extérieures au domaine n'ont pas forcément une grande conscience de l'importance que revêtent les archives publiques, d'autant plus quand elles sont numériques. Les projets d'intelligence artificielle dans le secteur des archives nécessitent des moyens et des connaissances techniques spécifiques. Pourtant, l'expérimentation de ces technologies a plusieurs avantages. Les projets à faible impact aboutissent plus facilement à des résultats exploitables et peuvent apporter des bénéfices en termes de médiation et de visibilité des services d'archives. Les initiatives d'automatisation de tâches plus complexes permettent quant à elles d'avoir une vision plus précise de certaines données à archiver et d'amorcer ou renforcer la collaboration avec les producteurs de ces données. Le déploiement de ces technologies soulève également des questions sur l'évolution du rôle des archivistes. Ces professionnels sont pour l'instant loin d'être remplacés par des machines mais leur métier pourrait être amené à se transformer. Les tâches de description pourraient par exemple évoluer vers des tâches de validation d'instruments de recherche produits par IA. Il sera nécessaire de réfléchir à la valeur ajoutée des archivistes par rapport à celle des machines. Iels ont notamment beaucoup à apporter en matière de contextualisation et d'identification de métadonnées pertinentes [13] afin d'éviter l'effacement de certaines mémoires. Les systèmes IA, souvent décrits comme des « boîtes noires », soulèvent ainsi de nombreuses problématiques éthiques, notamment en termes de consommation énergétique, de pollution et à travers la question des potentiels biais dont ils peuvent être porteurs. L'archiviste a un rôle à jouer dans l'approche critique de ces technologies. Le domaine des archives dispose d'une certaine expertise face à ces enjeux, notamment en ce qui concerne la transparence et la mise à disposition de données structurées.

L'intelligence artificielle est un sujet de polarisation, oscillant entre peurs, réticences et idéalisme. Nous espérons avoir apporté dans ce mémoire une vision plus nuancée de son utilisation dans les archives. L'usage de ce type de technologies peut sembler pertinent face aux nombreux défis archivistiques au Luxembourg. Les grands modèles pré-entraînés, comme observé dans le projet InventAIre, ont un grand potentiel pour les tâches liées au

13. *Ibid.*

langage. Cependant, les tâches archivistiques nécessitant une réflexion approfondie restent difficiles à automatiser. Les projets à faible impact, tels que ceux visant à améliorer la médiation ou la découvrabilité des fonds, sont davantage aisés à mettre en œuvre, surtout en l'absence de systèmes d'information archivistiques performants et de normes bien définies. Notre expérience a démontré qu'avec des moyens limités, l'automatisation d'un inventaire intellectuellement complexe n'est pas encore réalisable, bien que les *LLM (Large language models)* montrent des capacités non négligeables en matière de résumé et d'indexation. Notre projet a également montré que ces grands modèles de langage ne sont pas des outils « clé en main ». Leur utilisation pour des tâches complexes nécessite une réflexion approfondie. Les bénéfices de l'IA pour répondre aux défis des producteurs d'archives publiques au Luxembourg sont peut-être à chercher davantage du côté de la publicité et la médiation avec le public qu'elle peut faciliter pour les services. Il s'agirait de capitaliser sur l'engouement actuel et l'image idéalisée de ces technologies. Des bénéfices se trouvent également dans l'expérimentation que ce type de projets motivent. L'arrivée des technologies IA est peut-être l'occasion de faire évoluer les pratiques archivistiques. Au Luxembourg, elles ne sont pas encore rigoureusement normées. Cette situation peut être vue comme une opportunité. Les projets IA sont aussi un moyen d'améliorer la littératie numérique des équipes et de les sensibiliser aux spécificités des données traitées ou produites par des systèmes basés sur de l'apprentissage machine ou *machine learning*. Les transformations majeures promises par ces technologies ne sont pas encore intégrées dans les processus métiers des archivistes via des automatisations, mais elles devraient commencer à se manifester dans les données à traiter par ces derniers.

En somme, l'intelligence artificielle n'est pas une solution miracle aux défis archivistiques luxembourgeois. Pour être efficace sur des tâches complexes, elle nécessiteraient un cadre archivistique solide, des données d'entraînement ou de test bien définies et une réflexion éthique rigoureuse. L'expérimentation dans ce domaine a toutefois des avantages certains.

Les investissements réalisés dans ces technologies peuvent avoir des apports intellectuels et les archives ont leur rôle à jouer dans les transformations à venir, en constituant un réservoir de connaissance et par leur capacité à contribuer à la lutte contre les fausses informations.

Glossaire

HTR (Handwritten Text Recognition) Technologie similaire à l'*OCR*, mais spécifiquement conçue pour reconnaître et convertir le texte manuscrit en format numérique. 10, 48

LLM (Large language models) Modèle d'intelligence artificielle entraîné sur de vastes quantités de texte pour comprendre, générer et manipuler du langage naturel. 42, 43, 51–53, 80, 81, 84, 104, 142

NEL (Named Entity Linking) Processus en TAL qui consiste à associer les entités nommées identifiées dans un texte à des entités spécifiques dans une base de données ou un référentiel de connaissances. 46

NER (Named Entity Recognition) Technique en traitement automatique du langage (TAL) qui identifie et classe automatiquement les entités nommées (comme les personnes, les organisations, les lieux) dans un texte. 44

NLP (Natural langage processing) Voir TAL (Traitement Automatique du Langage). 39

OCR (optical character recognition) Technologie qui lit et convertit des images de texte dactylographié ou imprimé en texte numérique. 10, 32, 48

chatbot Programme informatique conçu pour simuler une conversation avec des utilisateurs humains, généralement via une interface de messagerie ou vocale, en utilisant des règles préprogrammées ou des modèles d'intelligence artificielle. 10, 31, 51, 58, 83, 99

cloud computing Fourniture de services informatiques (serveurs, stockage, bases de données, etc.) à distance via une connexion réseau, permettant un accès flexible et évolutif aux ressources numériques. 22

embeddings Représentations vectorielles de mots ou de phrases dans le but de capter leurs significations et relations contextuelles dans un espace de grande dimension. 42, 58, 86

fine-tuning Processus d'ajustement d'un modèle d'intelligence artificielle pré-entraîné sur un ensemble de données spécifique afin d'améliorer ses performances pour une tâche particulière. 9, 54, 57, 80

speech to text Technologie qui convertit la parole en texte écrit en temps réel à l'aide de systèmes de reconnaissance vocale. 10

stop words Mots courants dans une langue (comme "le", "et", "de" en français) qui sont souvent filtrés ou ignorés lors du traitement de texte en *machine learning* car ils apportent peu de valeur informationnelle pour l'entraînement ou l'analyse. 40, 49, 84

token Une unité de texte traitée comme une séquence distincte par le modèle. Il peut correspondre à un mot, une partie de mot ou un symbole. 51, 52, 78, 80, 84, 88

topic modelling Technique de machine learning utilisée pour découvrir automatiquement les thèmes ("topics") latents présents dans un ensemble de documents. 39, 43, 142

apprentissage machine ou *machine learning* Branche de l'intelligence artificielle qui permet d'entraîner un système à partir de données, d'améliorer ses performances dans le but de lui faire faire des prédictions ou prendre des décisions sans être exactement programmé pour chaque tâche. xvi, 3, 4

apprentissage non supervisé Type d'apprentissage où un modèle est entraîné sur des données non étiquetées, c'est-à-dire sans indication préalable des réponses ou catégories. L'algorithme doit identifier des structures, des motifs ou des regroupements dans les données de manière autonome.. 31, 39

apprentissage supervisé Type d'intelligence artificielle dont les modèles ou algorithmes apprennent à partir de données étiquetées, où chaque entrée est associée à une sortie considérée correcte. L'algorithme utilise ces exemples pour identifier des motifs et générer des prédictions ou des décisions sur de nouvelles données similaires. 31, 43

benchmark Une référence ou un ensemble de critères utilisés pour évaluer la performance, la qualité ou l'efficacité d'un système, d'un modèle, d'un produit, en le comparant à un standard ou à d'autres systèmes similaires. 87, 88

chaîne de traitement Séquence d'étapes ou d'opérations par lesquelles les données ou les tâches passent afin d'être transformées, analysées ou traitées jusqu'à obtenir un résultat final. 50, 91, 143

classification automatique Processus par lequel un algorithme attribue automatiquement des catégories ou des étiquettes à des données en fonction de leurs caractéristiques. 79

conduite du changement Processus d'accompagnement des individus et des organisations dans la transition d'un état actuel vers un état futur désiré, en minimisant les résistances et en maximisant l'adhésion aux nouvelles méthodes, outils ou structures. 26, 97

deep learning Sous-domaine du machine learning utilisant des réseaux de neurones profonds pour modéliser et résoudre des problèmes complexes à partir de grandes quantités de données. xvi, 22

découvrabilité Facilité avec laquelle les utilisateurs peuvent trouver et accéder à des informations, des produits ou des services, souvent en lien avec la conception de l'interface utilisateur ou l'optimisation des moteurs de recherche. 48, 50, 104, 142

expressions régulières (REGEX) Ensemble de motifs utilisés pour identifier, rechercher, ou manipuler des chaînes de caractères selon des règles syntaxiques définies. xvi, 22

F1 score Le F1-score est une mesure de performance qui combine précision et rappel. Il s'agit d'une moyenne entre les deux se situant entre 0 et 1, 1 étant la meilleure performance possible. 46

fenêtre Quantité de texte (souvent mesurée en tokens ou caractères) qu'un modèle de langage peut traiter ou dont il peut se souvenir à un moment donné, influençant la cohérence et la pertinence de ses réponses. 58, 88, 92

hallucination En machine learning, une hallucination se produit lorsqu'un modèle, comme un LLM, génère des informations incorrectes, inventées ou non fondées, souvent présentées comme étant factuelles. 33, 77

IA générative Branche de l'intelligence artificielle dont les modèles créent de nouvelles données, telles que du texte, des images ou de la musique. 3, 6, 51, 77

inférence Processus par lequel le modèle génère des réponses ou des prédictions en s'appuyant sur son entraînement préalable pour traiter et interpréter des données qui lui ont été transmises. 52, 79, 84, 95

littératie numérique Capacité à utiliser de manière critique, sécurisée et efficace les technologies numériques pour accéder à l'information, communiquer, créer du contenu et résoudre des problèmes. 104

modèle pré-entraîné Modèle d'intelligence artificielle qui a été initialement formé sur un grand ensemble de données pour accomplir une tâche générale, avant d'être affiné ou adapté pour une tâche spécifique avec moins de données. Ce processus permet de bénéficier de l'apprentissage préalable, accélérant le développement et améliorant les performances du modèle sur des applications spécialisées. 9, 31

multimodal En intelligence artificielle, le terme désigne la capacité d'un modèle à intégrer et traiter plusieurs types de données ou de médias (texte, image, audio, etc.) pour effectuer des tâches complexes ou fournir des réponses plus précises. 78

méthodes agiles Ensemble de pratiques de gestion de projet et de développement logiciel centrées sur l'adaptabilité, la collaboration, l'itération et la réponse rapide aux changements de besoins. 25, 29

prompt Instruction ou texte d'entrée donné à un modèle d'intelligence artificielle génératif pour générer une réponse ou accomplir une tâche spécifique. 52, 54, 58, 77, 80, 88, 92

précision Proportion des prédictions positives correctes parmi toutes les prédictions positives faites par un modèle, mesurant ainsi sa capacité à éviter les faux positifs. 46, 86

quantisation La quantisation consiste à réduire le nombre de bits utilisés pour représenter les poids (para- mètres ajustables qui déterminent la force des connexions entre les neurones) et les activations (valeurs intermédiaires produites par ces neurones à chaque étape du traitement) d'un modèle de langage. Par exemple, au lieu d'utiliser des nombres à virgule de 32 bits (float32), on peut utiliser des entiers de 8 bits (int8). Cela permet de diminuer la mémoire nécessaire pour stocker le modèle et d'accélérer les calculs nécessaires pour faire des inférences. 80, 84

RAG (Retrieval Augmented Generation) Technique permettant de baser les réponses d'un LLM sur des informations contenues au sein d'une base de connaissance externe. Des informations pertinentes pour générer une réponse à un prompt sont récupérées dans la base et ajoutées à la fenêtre du LLM, qui s'appuie dessus pour générer sa réponse. Cette technique permet des réponses plus précies et sourcées. 10, 58, 85

rappel Proportion des prédictions positives correctes parmi tous les cas réellement positifs dans les données, indiquant ainsi la capacité du modèle à identifier les vrais positifs. 46

responsive Terme utilisé pour décrire des interfaces web ou des applications capables de s'adapter automatiquement à différentes tailles d'écran et dispositifs (ordinateurs, tablettes, smartphones) pour offrir une meilleure expérience utilisateur. 98

TAL (Traitement Automatique du Langage) Le traitement automatique du langage ou *natural language processing* en anglais, est un domaine de l'intelligence artificielle qui se concentre sur l'interaction entre les ordinateurs et le langage humain, en permettant l'analyse, la compréhension et la génération de texte ou de parole. 39–41, 50, 56, 57, 142

vectorisation Processus de transformation de données (textuelles ou non) en vecteurs numériques afin qu'elles puissent être traitées par des algorithmes de *machine learning*. 40, 42, 85

Table des matières