*Actes de CORIA-TALN 2026*
*Nantes, France, 29 juin au 3 juillet 2026*

# Un journalisme augmenté par une IA souveraine L'expérience Ouest-France

Victor Klötzer[1] Thomas Girault[1] Michel Le Nouy[1]
Julien Perron[1] Cédric Jézéquel[1] Laurent Amsaleg[2]
(1) Ouest-France, 10 rue du Breil, 35000 Rennes, France
(2) Univ. Rennes, CNRS, Inria, IRISA - UMR 6074, Rennes, France
prenom.nom@ouest-france.fr, laurent.amsaleg@irisa.fr

RÉSUMÉ

Cet article présente les travaux menés par le groupe de presse *SIPA Ouest-France* pour valoriser son patrimoine éditorial. L'objectif est d'explorer comment l'intelligence artificielle, le traitement automatique des langues, les modèles de langue et les techniques de génération augmentée par récupération peuvent répondre au besoin quotidien des journalistes et des lecteurs : celui de s'informer. Nous décrivons l'implémentation de ces technologies pour la mise en œuvre d'assistants conversationnels, confrontés à la réalité d'indexer un patrimoine historique composé de plus de 110 millions de contenus hétérogènes. Les premiers défis relèvent de contraintes industrielles, de souveraineté et de viabilité économique. Notre collaboration avec l'IRISA au sein du laboratoire commun Synapses fait apparaître que nombre des difficultés rencontrées sont encore de véritables verrous scientifiques, notamment ceux concernant l'entraînement, l'explicabilité et la robustesse des modèles utilisés.

ABSTRACT

**Augmented Journalism Powered by a Sovereign AI – The *Ouest-France* Experience.**

This article presents the work carried out by the *Ouest-France* press group to leverage its editorial heritage. The goal is to explore how AI, NLP, LLMs and RAG techniques can meet the daily needs of journalists and readers : staying informed. We describe the implementation of these technologies for building conversational assistants that must cope with the reality of indexing a historical archive comprising more than 110 million heterogeneous contents. While the initial challenges stem from industrial and sovereignty concerns, as well as economic viability, our collaboration with IRISA within the joint "Synapses" laboratory reveals that many of the difficulties we encounter remain genuine scientific bottlenecks, particularly those related to model training, explainability, and robustness.

MOTS-CLÉS : LLM, RAG, Base de connaissances, Classification, EN, Agent conversationnel.

KEYWORDS: LLM, RAG, Knowledge Base, Classification, NER, ChatBot.

## 1 Contexte et objectifs

*Ouest-France* est le premier quotidien francophone en diffusion papier (500 000 exemplaires par jour) et le leader numérique de l'actualité (230 millions de visites, 1er site d'actualité, classement ACPM, janvier 2026). Avec 700 journalistes et 2 400 correspondants, *Ouest-France* publie quotidiennement plus de 4 700 articles et 1 200 pages, enrichis de plusieurs milliers de photos, vidéos et podcasts, représentant près de 10 000 nouveaux contenus, couvrant l'actualité locale jusqu'à l'international.

Actes de la session industrielle de CORIA-TALN 2026, pages 20–29.

Le quotidien appartient au groupe *SIPA Ouest-France*, qui édite notamment *Actu.fr*, *Le Maine Libre*, *Le Courrier de l'Ouest*, *Presse Océan*, *Le Marin*, ainsi que des magazines *Voiles et Voiliers*, *Bretons*… Le groupe *SIPA Ouest-France* est lui-même rattaché à l'ASPDH – Association pour le Soutien des Principes de la Démocratie Humaniste.

Le patrimoine éditorial du groupe rassemblé au sein de la « banque de contenus » constitue une collection unique, de plus de 110 millions de documents, de 1899 à aujourd'hui (50 M articles, 40 M photos, 18 M pages, 135 K vidéos, …). Un projet de valorisation de cette banque de contenus a été lancé en 2014. Il s'articule autour de la préservation de la souveraineté du groupe sur son exploitation, c'est-à-dire conserver la propriété totale sur ce patrimoine éditorial et aussi l'entière maîtrise de sa valorisation. Pour cela, une équipe pluridisciplinaire (*data scientists*, ingénieurs, documentalistes, journalistes spécialisés) a été constituée pour se pencher sur les défis techniques que doivent prendre en compte les algorithmes d'analyse automatique face à la volumétrie de ce patrimoine, sa nature profondément hétérogène (multimodalité, mais aussi diversité informationnelle) ou encore la variété des connaissances portées par l'ensemble des documents archivés et produits quotidiennement.

Cette équipe a graduellement maîtrisé les technologies de traitement automatique des langues pour ensuite développer des outils permettant de fouiller la banque et proposer aux journalistes des réponses pertinentes aux recherches qu'ils et elles effectuent. Ces outils ont été bouleversés par l'avènement des grands modèles de langue à la suite du lancement de ChatGPT en 2022. Parallèlement, cela a accéléré les réflexions sur l'évolution des métiers et des services dans les entreprises, y compris chez *Ouest-France*. En effet, les métiers du journalisme et les médias de presse font face à de multiples mutations, notamment l'explosion du nombre des sources et des contenus informationnels, originaux ou générés, mais aussi une bascule des usages pour l'accès à l'information vers les agents conversationnels.

Pour *Ouest-France*, emprunter la voie de l'IA ne se limite pas à optimiser nos activités, c'est un levier d'innovation pour ouvrir des perspectives de services destinés aux publics auxquels nous nous adressons déjà et à ceux qui restent éloignés de nos publications. C'est aussi une opportunité pour diversifier les supports et les formats, de mieux diffuser nos contenus en passant de l'écrit à l'audio, de l'audio à la vidéo, de la vidéo à l'écrit et vice versa. C'est aussi et surtout un ensemble de nouveaux défis posés par les possibilités et les limites de ces nouvelles technologies, par les contraintes industrielles, et par les contenus et données à notre disposition.

Parmi les défis posés aux agents conversationnels : comment converser avec 110 millions de documents archivés et 10 000 contenus ajoutés quotidiennement, que l'on soit journaliste ou lecteur ?

# 2 Contraintes techniques et caractéristiques des données

La transition numérique des industries des médias impose une reconfiguration des méthodologies de gestion, d'archivage et de valorisation des contenus éditoriaux. Dans ce contexte, l'intégration des technologies d'IA générative au sein du groupe *SIPA Ouest-France* ne relève pas d'une simple optimisation opérationnelle, mais constitue un impératif stratégique pour la souveraineté des données et la pérennité des services d'information. Pour cette étude de cas, nous examinons les implications techniques et éthiques de l'exploitation de patrimoines éditoriaux massifs et hétérogènes *via* des modèles de TAL. Au regard de ces questions, nous développons ici l'expérience acquise par l'équipe pour faire face aux réalités des données accumulées, autant qu'à la limitation des ressources nécessaires pour y répondre, qu'elles soient de temps ou de calcul.

## 2.1 Contraintes industrielles

La problématique industrielle du groupe *SIPA Ouest-France* se déploie autour de trois contraintes visant à garantir l'autonomie stratégique et la fiabilité des systèmes déployés. En premier lieu, la souveraineté sur les données constitue un prérequis absolu, impliquant le refus de toute délégation de contenus à des acteurs tiers. Cette posture vise à prévenir l'entraînement non consenti de modèles externes et à protéger le patrimoine éditorial, une préoccupation croissante dans le paysage du TAL où les questions de propriété intellectuelle sont centrales.

Cette exigence de souveraineté induit, en second lieu, de constituer une équipe interne ayant les compétences et la maîtrise totale des concepts et des processus techniques. Elle se traduit aussi par le déploiement d'une infrastructure matérielle internalisée, reposant notamment sur des unités de calcul graphique (GPU) dédiées et l'exploitation de modèles ouverts *(open weights)* tels que Llama, Mistral ou Qwen en mode *on-premise*. Une telle architecture garantit une étanchéité totale des flux de données, éliminant les risques de fuite inhérents aux solutions *cloud* propriétaires.

Enfin, le processus intègre une dimension éthique et collaborative fondamentale : le principe de l'humain dans la boucle. Ce paradigme, concrétisé par la fréquente synergie entre développeurs, chercheurs et journalistes, permet une validation itérative et pragmatique des outils. Cette concertation assure l'alignement des technologies avec les besoins métiers et le respect de la charte éthique régissant l'usage de l'IA au sein des rédactions. Si la gouvernance définit le cadre éthique, l'efficacité des traitements reste tributaire de la qualité et de la structure des sources de données exploitées.

## 2.2 Caractérisation et hétérogénéité des archives de presse

La faisabilité technique des projets de TAL est conditionnée par la nature même des ressources documentaires disponibles. Les archives du groupe *SIPA Ouest-France* présentent une hétérogénéité tant qualitative que quantitative, représentant un défi majeur pour la standardisation des traitements.

Le fonds textuel numériquement natif, constitué de 50 millions d'articles, couvre principalement la période contemporaine (années 2000). Ces documents présentent une variabilité notable, s'étendant de brèves d'une dizaine de mots à des reportages de plusieurs milliers de mots. Les métadonnées attachées sont très diverses. Certaines sont automatiquement produites par le processus éditorial (articles publiés dans la page « Une » ou « Bretagne »). D'autres, peut-être plus sémantiques, sont soit explicitement fournies par les journalistes au cours de leur rédaction, soit implicitement ajoutées par de nombreux algorithmes d'analyse du langage. Les labels associés aux articles sont régis par une taxonomie évolutive. Celle-ci reflète les dynamiques de l'actualité, l'évolution de nos procédés de publication et se caractérise par une forte hétérogénéité sémantique, mêlant thématiques générales, sujets d'actualité, entités nommées et *tags* destinés au référencement naturel pour accroître la visibilité et l'audience de nos publications sur les moteurs de recherche tiers. Globalement, ces annotations sont souvent bruitées (trop générales, peu informatives, pauvres), redondantes et parfois absentes. La réalité de cette très grande hétérogénéité des annotations sur chaque article, en qualité et en nombre, complexifie leurs exploitations ultérieures.

Parallèlement, un fonds de 40 millions de photos, principalement constitué depuis les années 2000, illustre les défis du traitement multimodal. Seulement 30% de ces ressources disposent de métadonnées structurées (date, lieu, mots-clés, description, …), fournissant ainsi une information contextuelle. L'indexation effective de ces images nécessite par conséquent un alignement sémantique robuste avec le texte périphérique, englobant métadonnées ainsi que légendes et corps des articles.

Enfin, l'analyse diachronique des pages révèle une fracture technologique déterminante pour les stratégies de numérisation. Si la couverture chronologique s'étend de 1899 à nos jours, seules les dix millions de pages postérieures aux années 2000, issues des flux de production numérique natifs, offrent une qualité de structuration optimale. À l'inverse, les huit millions de pages antérieures, issues des campagnes de numérisation par reconnaissance optique de caractères (OCR) initiées en 1999, sont affectées par un bruit textuel. La qualité de ces données dégradées est corrélée à des facteurs variables tels que l'état de conservation des microfilms sources, l'évolution des maquettes de mise en page et la diversité des polices typographiques employées au fil des décennies. Extraire automatiquement et individuellement les articles contenus dans ces huit millions de pages reste aujourd'hui une difficulté. La qualité des résultats obtenus par nos modèles n'atteint pas un niveau acceptable par rapport aux coûts des calculs nécessaires.

## 2.3 Enjeux techniques d'indexation et de multimodalité

Avant d'envisager des mécanismes de recherche ou d'interaction avancés, la nature de nos données journalistiques soulève des problèmes complexes de représentation des contenus et des connaissances qui ont un impact fort sur les processus d'indexation à grande échelle. Les articles combinent en effet des formats d'objets hétérogènes, incluant notamment textes, images ou tableaux, qui participent conjointement à la construction du sens éditorial.

Dans ce contexte, les bases de connaissances constituent également un élément clé pour exploiter les contenus journalistiques dans des systèmes de recherche d'information, plusieurs centaines de milliers d'entités nommées faisant notre identité locale. Leur construction et leur maintenance restent toutefois complexes, en raison entre autres de problèmes de désambiguïsation d'entités ou de forte dynamique temporelle de l'actualité.

Plusieurs problématiques structurantes en découlent. Les spécificités de la presse régionale exigent d'aligner et d'affiner les modèles de vision (ou multimodaux) et les modèles textuels sur des catégories rares, peu dotées. Les meilleurs modèles sont en échec face aux allées couvertes, aux calvaires, aux lieux de vie et autres éléments très locaux, fondamentaux à notre identité. Constamment créer et faire évoluer des classes, forger des représentations précises et discriminantes est difficile et coûteux.

Le niveau de détail d'information que l'on souhaite pouvoir rechercher pose ensuite des questions de granularité de représentation des données. Comment représenter et capturer différents niveaux de détail dans du texte ou des images ? Comment avoir à la fois une représentation du sujet général de l'article ainsi que des différents éléments amenés dans le corps du texte ? Comment indexer à la fois l'évènement représenté par une photo tout comme les différents monuments et personnes présents sur l'image ? Outre la complexité de découpage et de structuration de l'information, ce sujet pose aussi des problématiques techniques de volumétrie et de capacité de recherche.

L'hétérogénéité sémantique des contenus (narratifs, descriptifs ou structurés) complexifie également la mise en correspondance entre une requête textuelle de recherche et les documents qui lui seraient associés. Elle nécessite des représentations capables de dépasser la similarité lexicale. Dans le contexte de presse locale de *Ouest-France*, ces difficultés sont accentuées par la forte ambiguïté de certaines entités nommées (les lieux et les personnes principalement), par exemple les communes de Plouhinec dans le Finistère et Plouhinec dans le Morbihan, ce qui nécessite des mécanismes de désambiguïsation robustes pour garantir la pertinence des résultats. De plus, la variété des titres publiés par le groupe *SIPA Ouest-France* – du langage général du quotidien régional aux jargons très spécialisés de titres comme *Le Marin* ou *Voiles et Voiliers* – engendre des domaines de langage distincts qui exigent des

modèles capables de gérer des registres lexicaux et terminologiques très différents, ce qui accroît encore la difficulté de l'indexation et de la recherche.

Par ailleurs, la présence d'une structuration dans les données textuelles ou non textuelles, notamment des tableaux, pose des difficultés liées à la préservation des relations internes à un document lors de leur transformation pour les indexer. Le passage à l'échelle impose aussi des exigences fortes en matière de performance et de pertinence, notamment dans le cadre de recherches vectorielles sur de larges collections dynamiques. Enfin, l'évolution inexorable des modèles, des technologies, des représentations, interroge l'apprentissage continu mais aussi la possible cohabitation de multiples représentations, vectoriser à nouveau toute la banque de contenus apparaissant hors de propos.

Ces différents enjeux structurent les choix méthodologiques et techniques des systèmes de recherche d'information journalistique. Afin d'en proposer une première mise en application, la section suivante introduit une approche de type *Retrieval-Augmented Generation* (RAG), expérimentée sur un corpus restreint d'articles.

# 3 Mise en œuvre d'une approche RAG sur un corpus restreint : le cas des 24 Heures du Mans

Un système de recherche d'information vise à restituer des documents pertinents en réponse à une requête généralement textuelle. Les approches classiques, lexicales ou vectorielles, permettent d'identifier des contenus proches, sans pour autant produire de réponses synthétiques. Les modèles de langues génératifs ont permis de s'intéresser à cette notion de synthèse, mais au prix de limitations en termes de fiabilité. Le RAG vise justement à articuler ces deux dimensions : (1) récupérer des documents pertinents dans un corpus indexé, puis (2) générer une réponse en s'appuyant explicitement sur ces sources. Concrètement, les documents sont segmentés en unités *(chunks)*, vectorisés, puis interrogés *via* une recherche par similarité. Les segments les plus pertinents sont ensuite injectés dans le contexte du modèle génératif. Ce cadre permet de contraindre la génération à des sources identifiées, améliorant la pertinence et la traçabilité des réponses tout en essayant de limiter les hallucinations.

## 3.1 Cas d'étude : le dispositif Topo pour les 24 Heures du Mans

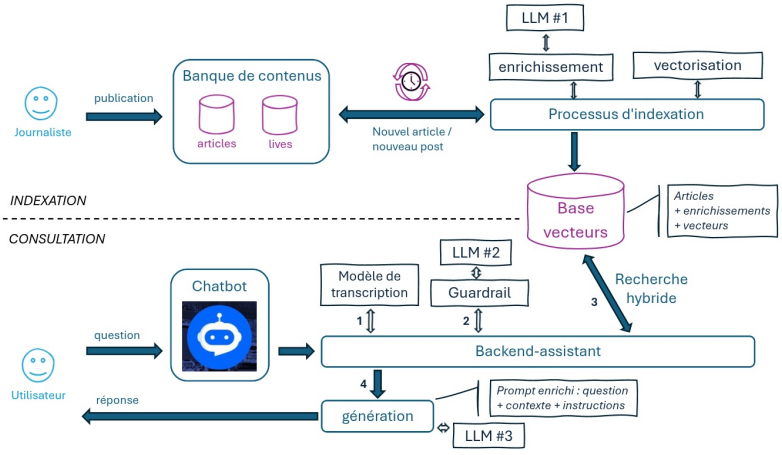



FIGURE 1 – Dispositif technique de l'agent « Topo »

Afin d'évaluer cette approche en conditions réelles, un prototype nommé « Topo » a été déployé à l'occasion des 24 Heures du Mans 2025 et ouvert au grand public sur le site `ouest-france.fr`. Ce système repose sur un RAG appliqué à un corpus restreint et spécialisé. Le corpus comprenait environ 700 articles récents, enrichis en continu pendant l'événement, ainsi que des archives historiques. Cette configuration a permis de tester le système dans un contexte dynamique, caractérisé par des mises à jour fréquentes au fil de l'évènement et une grande variété de requêtes utilisateurs.

Un principe clé a été introduit pour essayer de garantir la fiabilité : en l'absence de documents pertinents, le système adopte une stratégie explicite de non-réponse, évitant toute génération d'information. En pratique, ce mécanisme repose sur une étape finale de génération contrôlée par *prompt*, incitant explicitement le modèle à ne pas répondre en l'absence de contexte pertinent pour la question utilisateur. Bien qu'encore difficile à évaluer rigoureusement, cette approche apparaît empiriquement efficace. En complément, un mécanisme de *guardrail*, fondé sur un appel à un autre modèle de langue, est exécuté en parallèle du traitement principal afin de qualifier le type de question posée et de prévenir d'éventuels usages inappropriés. Cette exécution concurrente permet de renforcer le contrôle des réponses sans dégrader la fluidité ni les performances de l'agent conversationnel (voir figure 1). Ce mécanisme s'est montré particulièrement utile, filtrant de nombreuses requêtes malignes.

## 3.2 Enseignements

Plusieurs enseignements techniques et opérationnels émergent de cette expérience. Sur le plan de la performance, environ 70% des requêtes ont donné lieu à des réponses jugées satisfaisantes par les utilisateurs, avec un temps moyen de complétion de 5,86 secondes. L'outil a notamment suscité l'intérêt des 18-34 ans, 43% des utilisateurs, confirmant l'attrait grandissant pour ces interfaces conversationnelles. Ces résultats reposent sur une expérimentation de 7 jours, impliquant environ 2 750 utilisateurs et 430 retours d'usage collectés.

Les expérimentations soulignent également qu'un système de récupération d'information fondé uniquement sur la similarité vectorielle atteint des limites face à la complexité des requêtes réelles et à l'augmentation du volume documentaire. Certaines questions, notamment lorsqu'elles impliquent des contraintes temporelles, des entités nommées ou des données structurées, nécessitent de combiner plusieurs stratégies comme le filtrage, la recherche par entités ou le re-classement.

Enfin, une approche de validation itérative appuyée sur le retour des journalistes s'est révélée essentielle pour ajuster le système et enrichir progressivement le corpus en fonction des usages observés. En particulier, la réactivité de la rédaction sportive a alimenté la base des documents vectorisés, avec les éléments de réponse ayant fait défaut aux premières questions du public.

## 3.3 Leçons

Cette approche ouvre la voie à une évolution des moteurs de recherche éditoriaux vers des interfaces conversationnelles, permettant une interaction directe avec les contenus. Il s'agit ainsi de passer d'une logique de restitution documentaire à une interaction sémantique avec les articles et les archives journalistiques. Cela marque la mutation d'une recherche documentaire mécanique à une véritable interaction sémantique avec le patrimoine de notre banque de contenus.

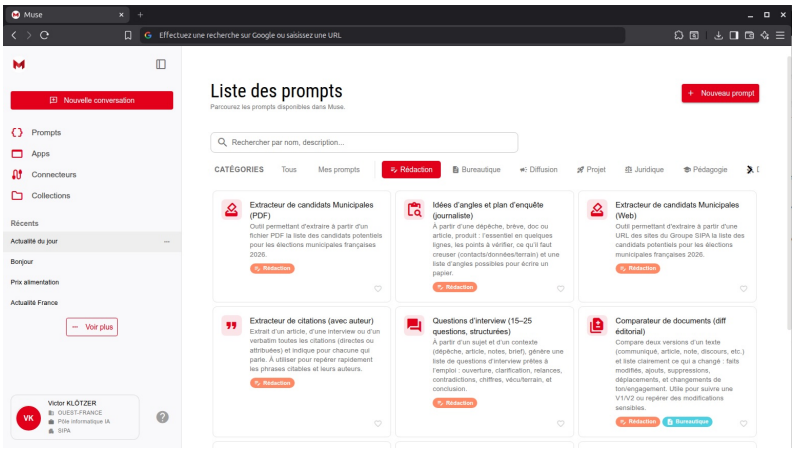
Si un système de RAG permet d'améliorer significativement la pertinence et la traçabilité des réponses, il ne constitue toutefois qu'un premier niveau d'interaction avec les contenus. Le passage à des systèmes conversationnels plus complets soulève de nouveaux défis, notamment en matière de

volumétrie des contenus, de gestion du contexte de la conversation, de pilotage des interactions et d'orchestration de multiples sources d'information. La section suivante s'intéresse à ces enjeux à travers le développement d'agents conversationnels appliqués aux contenus éditoriaux.

# 4 Défis de l'agent conversationnel sur les contenus éditoriaux

Dans la continuité des approches de type RAG, l'objectif n'est plus uniquement de restituer une réponse ponctuelle, mais de proposer une nouvelle forme d'interaction avec les contenus éditoriaux. Cette évolution vise à transformer le moteur de recherche traditionnel en un système conversationnel capable d'explorer, synthétiser et contextualiser l'information.

Dans ce cadre, une plateforme interne, baptisée « Muse », a été développée afin de fournir aux journalistes un accès unifié à des agents conversationnels spécialisés ainsi qu'à des outils et prompts spécialisés (voir figure 2). Ces agents permettent d'interroger le patrimoine éditorial *via* des requêtes en langage naturel, tout en s'appuyant sur une architecture RAG garantissant l'ancrage des réponses dans des sources vérifiées.




Muse est un agent conversationnel généraliste reposant sur l'orchestration d'un ensemble d'agents et de *prompts* spécialisés pour répondre aux requêtes utilisateurs. L'exemple présenté illustre un catalogue de *prompts* dédiés à des tâches de rédaction journalistique.

FIGURE 2 – Interface de la plateforme « Muse »

Un principe fondateur de cette approche réside dans l'intégration de l'humain dans la boucle, qui conserve le dernier mot sur les décisions du système dans un cadre d'usage assisté, contribue à l'enrichissement des ressources mobilisées entre autres par l'apport de contenus éditoriaux complémentaires, et participe à l'amélioration continue des modèles à travers les retours d'usage, notamment au travers d'un mécanisme d'acceptation ou de rejet des messages. La plate-forme orchestre un ensemble d'agents conversationnels souverains, conçus pour assurer la protection des données éditoriales. L'utilisateur conserve un rôle décisionnel central dans l'interaction et l'usage : il peut explorer, affiner et évaluer les réponses produites par le système.

L'interface assure une transparence explicite des sources mobilisées, associant chaque élément généré à ses documents d'origine, garantissant ainsi une traçabilité indispensable à la vérification. Cette exigence de transparence est critique dans le contexte journalistique, où la validation et la contextualisation demeurent sous la responsabilité de l'expert humain.

Au-delà des usages exploratoires, plusieurs outils spécialisés ont été développés, notamment pour la génération de résumés ou l'analyse de données sportives (classements, performances). Ces cas d'usages illustrent le potentiel d'un système hybride, à mi-chemin entre moteur de recherche avancé et agent conversationnel.

## 4.1 Typologie des requêtes et implications techniques

L'un des principaux défis réside dans la diversité des questions adressées au système. Contrairement à un moteur de recherche classique, un agent conversationnel doit être capable d'interpréter des intentions variées et d'adapter dynamiquement sa stratégie de traitement. Une taxonomie des requêtes a ainsi été établie pour orienter la conception du système :

- les questions à réponse absolue simples, qui requièrent l'identification d'un fait unique présent dans un document ;
- les questions à réponse absolue complexes, nécessitant l'analyse et la consolidation d'informations issues de plusieurs sources ;
- les questions d'analyse temporelle, impliquant la sélection et la synthèse de documents sur une période donnée ;
- les questions de résumé, visant à produire une synthèse d'un ensemble de contenus récents ou historiques ;
- les questions de méta-recherche, proches de requêtes structurées sur un moteur (filtrage, comptage, tri) ;
- les questions ouvertes ou créatives, pour lesquelles il n'existe pas de réponse unique mais qui sollicitent des capacités de suggestion ou de structuration.

Cette diversité implique que la chaîne de traitement ne peut être uniformisée. Elle impose au contraire une adaptation dynamique des stratégies de recherche, de sélection et de génération, conditionnée par une classification fine de l'intention utilisateur. Dans cette perspective, « Muse » vise une forme d'utopie d'abstraction des mécanismes internes, afin de rendre la complexité technologique transparente pour les rédacteurs et de se rapprocher au plus près des pratiques journalistiques. Cette démarche souligne ainsi l'indispensable rôle de l'humain dans la boucle, garant de l'adéquation entre les productions du système, les attentes éditoriales et les exigences de qualité informationnelle.

## 4.2 Chaîne de traitement et orchestration des composants

Pour répondre à ces exigences, l'architecture du système repose sur une chaîne de traitement modulaire et orchestrée. En amont, une phase d'ingestion assure la préparation et l'indexation des contenus, incluant un traitement différencié selon les modalités (texte, image), un choix de granularité de découpage *(chunking)*, ainsi qu'une vectorisation *via* des modèles de plongement adaptés.

Lors de l'exécution d'une requête, un module d'interprétation analyse la question pour en déterminer la typologie et extraire des contraintes (filtres temporels, entités nommées). Cette étape conditionne l'aiguillage vers un ou plusieurs modules de recherche spécialisés *(retrievers)*, pouvant combiner recherche vectorielle sémantique, recherche lexicale ou filtrage structuré. Les résultats obtenus font ensuite l'objet d'une phase de sélection et de reclassement *(reranking)*, visant à identifier les documents les plus pertinents. Les contenus sont alors segmentés et compressés afin de constituer un contexte exploitable par le modèle génératif qui formule la réponse du système.

La génération de la réponse s'accompagne d'un mécanisme de citation explicite des sources, ainsi que de la gestion des cas limites (absence d'information, ambiguïté). Une attention particulière est portée à la mise en forme, afin de distinguer clairement les éléments générés des sources référencés, et de faciliter leur consultation.

## 4.3 Gestion du contexte conversationnel

La gestion du dialogue constitue un enjeu technique distinct de la simple génération de réponses isolées. L'agent doit permettre des interactions itératives où l'utilisateur affine ou reformule ses demandes.

Ces dynamiques conversationnelles introduisent plusieurs scénarios. Certaines requêtes de relance nécessitent de réutiliser le contexte initial (par exemple : « Peux-tu développer le premier point ? »), tandis que d'autres relèvent d'une simple transformation du contenu généré (« Peux-tu reformuler ce passage ? »). À l'inverse, certaines formulations *a priori* dépendantes du contexte (« Et en 1984 ? ») impliquent en réalité une nouvelle phase complète de recherche.

Cette variété d'interactions impose un arbitrage entre différents modes de fonctionnement : RAG standard, génération pure ou approches hybrides. Cet arbitrage est déterminant pour assurer la cohérence sémantique et la robustesse des échanges.

Le passage d'un système RAG à un agent conversationnel complet complexifie significativement l'architecture logicielle. L'interprétation fine des requêtes, l'orchestration des composants de récupération et la gestion du contexte conversationnel représentent des verrous scientifiques et techniques majeurs. Ces défis sont centraux pour l'élaboration de systèmes robustes adaptés aux usages éditoriaux, et leur résolution fiable demeure, à l'heure actuelle, un chantier ouvert.

# 5 Discussion et interaction recherche-industrie

Les solutions présentées mettent en évidence l'écart persistant entre les performances observées en environnement de recherche contrôlé *(in-vitro)* et les contraintes inhérentes à un déploiement industriel à grande échelle *(in-vivo)*. Le projet Synapses [1], laboratoire commun financé par l'ANR associant le groupe *Ouest-France* et l'IRISA (notamment pour ses expertises en traitement automatique du langage, en vision par ordinateur et en visualisation de données), a pour vocation d'explorer cette transition méthodologique et technique.

Notre travail conjoint au sein de Synapses nous permet d'identifier des verrous sur lesquels se penchent actuellement les membres de diverses communautés scientifiques. Le premier verrou concerne la fiabilité des systèmes génératifs, notamment la maîtrise des hallucinations et la traçabilité des réponses, dans un contexte où l'exigence de vérification est centrale. Le passage à l'échelle, de corpus restreints à plusieurs dizaines de millions de documents, constitue ici un défi supplémentaire.

Le second verrou porte sur la multimodalité, avec pour enjeu central l'alignement sémantique entre contenus textuels et visuels. S'y ajoutent la dynamique du corpus, impliquant des mises à jour continues, ainsi que l'évolution des modèles d'analyse, nécessitant des mécanismes d'indexation incrémentale efficaces, voire des campagnes de re-vectorisation à grande échelle. Un groupe de presse comme *Ouest-France* produit un flux continu de données : 10 000 nouveaux contenus sont ajoutés quotidiennement à la banque de contenus.

Enfin, la variabilité structurelle des documents et les spécificités éditoriales du corpus soulignent les limites des approches purement vectorielles et la nécessité de stratégies de recherche hybrides.

1. https://synapses.cnrs.fr

# 6 Conclusion

Gérer un patrimoine éditorial couvrant plus de 125 ans d'histoire – de l'affaire Dreyfus aux crises climatiques actuelles – n'est pas qu'un défi technique, c'est pour *SIPA Ouest-France* une responsabilité démocratique. La manipulation de 50 millions d'articles et 40 millions de photos, ne se résume pas au traitement de « données », mais engage la conservation et l'activation d'une mémoire collective partagée avec les lecteurs.

Dans ce contexte, l'explicabilité des modèles d'IA que nous employons n'est pas qu'une série de *leaderboard* pour ingénieurs curieux : c'est la condition dictée par notre charte éditoriale. L'explicabilité est en effet notre priorité absolue car la confiance que nous construisons avec nos journalistes et nos lecteurs est primordiale. Si un agent conversationnel répond à un lecteur ou un journaliste en s'appuyant sur nos contenus, il doit être capable de dire pourquoi il affirme tel propos ou tel fait.

Il est important d'être transparent sur les références et sources : un LLM « brut » est une boîte noire qui hallucine parfois avec aplomb. Le RAG sert entre autres de garde-fou. L'explicabilité doit permettre de remonter le fil : « Voici la réponse, et voici les trois articles de 1924 dont elle est issue… ». Sans cela, nous ne relatons pas des faits, nous émettons une opinion probabiliste.

Nous devons aussi prendre en compte les biais historiques : nos archives de 1899 ne reflètent pas les valeurs de 2026. L'IA doit pouvoir expliciter si elle retranscrit un point de vue d'époque ou si elle analyse un fait. L'explicabilité doit mettre en évidence si le modèle favorise certains segments de notre banque de contenus au détriment d'autres.

Nous sommes juridiquement engagés car toutes nos publications mentionnent un Directeur de la publication pénalement responsable de tout ce qu'on y affirme. Si l'IA génère un contenu diffamatoire en mélangeant deux faits historiques, nous devons comprendre le cheminement logique du modèle pour corriger le tir et répondre de nos actes.

Si les LLM sont optimisés pour la persuasion plutôt que pour la vérité factuelle, le défi industriel réside dans la capacité à fournir une explication détaillée en temps réel sur une base de 110 millions d'entrées (articles, pages et photos), sans engendrer de coûts calculatoires prohibitifs ni de latence excessive.

Le partenariat Synapses illustre ainsi comment la recherche académique peut s'emparer de problématiques industrielles concrètes pour faire progresser l'état de l'art en TAL, tout en dotant un acteur de la presse d'outils économiquement viables pour préserver sa souveraineté numérique.

# Remerciements

Ce travail a été partiellement financé par le laboratoire commun CNRS/Ouest-France Synapses, soutenu par l'Agence Nationale pour la Recherche sous la référence ANR-23-LCV2-0007.