abes
agence bibliographique
de l'enseignement supérieur

IA

Indexation RAMEAU assistée par IA

Expérimentation réalisée avec 12 établissements testeurs

Rapport final – Février 2025

abes.fr

Table des matières

Résumé

Ce rapport présente les résultats de l'expérimentation d'un service d'indexation sujet automatique RAMEAU assisté par IA, testé par 81 bibliothécaires dans 12 établissements d'octobre 2024 à janvier 2025. L'objectif était d'évaluer la pertinence des suggestions d'indexation et leur intégration dans WinIBW, l'outil central de signalement bibliographique de l'ESR.

Les retours des testeurs montrent que ce service améliore la qualité de l'indexation du sujet des documents signalés et peut faire gagner du temps aux catalogueurs, notamment grâce aux ajustements apportés à mi-parcours sur le service, par exemple sur la présentation des suggestions dans WinIBW.

Les performances varient selon les disciplines documentaires, et sont souvent moins bonnes pour les contenus à forte spécialisation. Au cours de la période de tests, la suppression des suggestions peu pertinentes et l'amélioration de l'ergonomie du service ont été bien accueillies. Les utilisateurs demandent l'ajout de concepts géographiques et chronologiques au service d'indexation automatique, ainsi que l'extension du corpus d'entraînement aux thèses afin d'améliorer la qualité des suggestions sur des documents spécialisés.

Une large majorité des testeurs souhaite non seulement le maintien du service, mais surtout qu'il continue d'évoluer. L'expérimentation a mis en lumière l'importance d'un accompagnement pour optimiser l'usage de l'IA dans le catalogage au sein de chaque établissement.

Des recommandations sont formulées pour le déploiement et l'amélioration future du service.

Remerciements

L'équipe du Labo de l'Abes adresse ses remerciements pour leur participation fructueuse à cette expérimentation aux 81 collègues des établissements suivants : SCD de l'Université d'Aix-Marseille, SCD de l'Université de Rennes, SCD de l'Université de Rennes 2, BAPSO de l'Université de Grenoble-Alpes, SCDI de Montpellier, SCD de l'Université de Montpellier, SCD de l'Université de Reims Champagne-Ardenne, SCD de l'Université de Tours, BIU Cujas, Bibliothèque de Sciences Po Paris, Bibliothèque du MNHN, Bibliothèque de l'Ecole Française de Rome.

Contexte et objectifs de l'expérimentation

D'octobre à début janvier 2025, 81 bibliothécaires de 12 établissements ont testé un service d'aide à l'indexation RAMEAU intégré dans WinIBW, l'outil de catalogage du réseau Sudoc. Ce rapport présente le bilan de cette expérimentation, élaboré à partir des retours directs des référents de chaque établissement et à partir des données d'usage recueillies automatiquement.

Contexte

En 2023, le Labo de l'Abes a mené un projet dont l'objectif était de démontrer la faisabilité d'une indexation RAMEAU de qualité satisfaisante au moyen d'une intelligence artificielle (IA), à partir du titre et du résumé d'une monographie en français[1].

La qualité d'une indexation RAMEAU par IA doit être mesurée et appréciée en fonction du cas d'usage envisagé. Selon qu'il s'agisse de faciliter l'indexation par un humain (aide à la décision) ou de générer une indexation entièrement automatique, les critères d'évaluation ne seront pas tout à fait les mêmes. Notre objectif prioritaire était et demeure l'aide à l'indexation pour les catalogueurs de monographies du Sudoc.

D'un point de vue informatique, l'indexation automatique s'appuyant sur un vocabulaire tel que RAMEAU est une tâche complexe et ambitieuse. Techniquement, on parlera d'une **classification multilabel extrême**. En effet, il s'agit bien de classer, c'est-à-dire de ranger les documents dans des cases prédéfinies, et non pas seulement de regrouper les documents semblables (ce qu'on appelle "*clustering*"). En second lieu, cette classification est dite "multilabel" car un même document peut appartenir à plusieurs classes, c'est-à-dire être indexé par plusieurs sujets RAMEAU à la fois. Enfin, cette classification multilabel est dite "extrême" car le nombre de classes est très important (de l'ordre de 100 000 concepts), ce qui complique considérablement la tâche. Les caractéristiques de cette tâche en font un réel défi pour une machine, mais aussi, et c'est essentiel, pour les humains.

L'expérience des catalogueurs, diverses études et nos évaluations démontrent en effet que, dans l'immense majorité des cas, deux indexeurs ne choisissent pas la même indexation pour le même document (en moyenne, nous avons observé une intersection de 50% d'une indexation

[1] Certaines restrictions au périmètre de l'étude ont été définies dès le départ :
- Suggestion des seuls noms communs/concepts RAMEAU (Td8 dans le Sudoc)
- Exclusions des autorités pré-construites, en raison de la réforme en cours de RAMEAU (susceptibles d'être scindées)
- Suggestion de concepts uniques, les chaînes d'indexation n'étant pas la cible prioritaire.

humaine à une autre). Il serait donc inexact de prendre comme point de référence absolu quelque indexation humaine que ce soit, *a fortiori* celle du Sudoc (l'œuvre de milliers de personnes différentes sur plus de vingt ans). Paradoxalement, il est donc insuffisant de prendre le corpus d'entraînement (l'indexation Sudoc seule) comme référence absolue, comme « LA vérité ». Pourtant, il s'agit là du postulat de base de la méthodologie classique d'évaluation des prédictions en *machine learning*. Cela signifie qu'il est nécessaire de concevoir et d'employer d'autres modes d'évaluation pour cette tâche d'indexation automatique.

Les particularités de cette tâche nous ont conduits à adopter plusieurs stratégies d'évaluation complémentaires :

1. **Évaluation des indexations machines avec les métriques classiques adaptées à la classification multilabe**l (= **Sudoc comme LA vérité**). Parmi toutes ces métriques, nous avons privilégié celle qui se rapproche le plus d'une évaluation humaine spontanée (l'idée étant d'évaluer la qualité d'une indexation document par document, et non concept par concept à l'échelle de tout le corpus).
2. **Évaluation des indexations machines en les comparant à plusieurs indexations humaines**, et pas seulement à l'indexation humaine du Sudoc (**= pluralité des vérités**). Pour ce faire, nous avons demandé à 6 collègues de l'Abes (que nous nommerons "réindexeurs") d'indexer une centaine de documents déjà indexés dans le Sudoc, sélectionnés de manière aléatoire.
3. **Évaluation qualitative de toutes les indexations, humaines et machines, au moyen d'une grille de notation** (= **notation comme la vérité**). Selon cette grille, noter une indexation, c'est, d'une part, noter chaque sujet retenu pour une notice donnée (on note l'**exactitude** et la **spécificité** de chaque sujet) et, d'autre part, noter le bloc des sujets retenus pour une notice donnée (on note la **complétude** et la **redondance** de chaque bloc).

Nous avons exploré les pistes suivantes :

- Interrogation d'un grand modèle linguistique pré-entraîné (LLM) comme GPT 3.5.
- Interrogation d'un LLM *fine-tuné* sur un corpus Sudoc (*fine-tuner*, c'est « ajuster finement » un modèle pré-entraîné de *machine learning* sur une tâche spécifique)
- *Deep learning*
- Recherche sémantique dans une base vectorielle générée par la vectorisation d'un corpus Sudoc par un modèle de *sentence embedding*.[2]
- Algorithmes intégrés dans le logiciel ANNIF

Nos tests ont conclu que les deux dernières pistes sont les meilleures. Plus précisément, nous avons retenu le modèle de sentence *embedding all-MiniLM-L6-v2* et l'algorithme intégré à ANNIF OMIKUJI, que nous avons appliqués sur un corpus de 130 000 notices de monographies en français.

[2] En traitement du langage naturel, un *sentence embedding* est la représentation d'une phrase à travers un vecteur de nombres qui parvient à « refléter » sa signification. Cette vectorisation d'une phrase permet, par exemple, de mesurer la similarité d'un texte avec un autre texte qui peut être très différent du point du vocabulaire. Dans notre cas, il s'agit de comparer le texte d'une notice sans indexation avec l'ensemble des vecteurs représentant chacun un concept RAMEAU. Le vecteur de chaque concept RAMEAU est construit à partir du contenu textuel de l'ensemble des notices Sudoc liées à ce concept.

Les expérimentations et les évaluations menées permettaient de conclure que l'indexation RAMEAU par une IA est aujourd'hui réalisable en garantissant un niveau de qualité suffisant, et un temps de traitement satisfaisant. Cependant, du fait de la grande difficulté à évaluer la qualité d'une indexation et de la nouveauté que présente l'assistance d'une IA, ce rapport recommandait le lancement d'une expérimentation *in vivo*, en situation réelle, c'est-à-dire dans l'environnement de travail du catalogueur Sudoc.

Objectifs

Dans le cadre de cette expérimentation, il s'agit d'évaluer à la fois la **qualité ressentie du service de suggestion d'indexations** et la **qualité ressentie de l'intégration** de ce service dans l'outil de travail quotidien. L'évaluation de ces 2 dimensions permettra de mesurer la **satisfaction globale** de l'utilisateur et d'envisager les améliorations pertinentes.

Nous avons demandé aux testeurs d'évaluer ce service d'aide à l'indexation selon 4 critères :

Critère 1 : configuration du service
Exemples de questions à se poser :
- le script est-il facile à installer ?
- ai-je rencontré des difficultés dans la configuration ou l'utilisation du script ? si oui, lesquelles ?

Critère 2 : ergonomie de la présentation des suggestions
Exemples de questions à se poser :
- le nombre de suggestions est-il suffisant ?
- les intitulés des blocs de suggestions doivent-ils être reformulés ?
- la présentation des suggestions doit-elle être améliorée ? Si oui, comment ?

Critère 3 : qualité des suggestions d'indexation
Exemples de question à se poser :
- les suggestions sont-elles pertinentes ?
- les suggestions sont-elles adaptées au type de ressource que je catalogue ?

Parmi les cas les plus flagrants de suggestions inutiles, voire erronées, les testeurs peuvent en noter quelques-uns, à titre d'exemple. Mais il n'est pas demandé de qualifier de façon exhaustive toutes les suggestions : il y aura forcément des suggestions inappropriées.

Critère 4 : satisfaction globale
Exemples de question à se poser :
- le service fait-il gagner du temps dans l'activité de catalogage ?
- le service s'insère-t-il facilement dans ma pratique habituelle d'indexation ?
- à quel point facilite-t-il mon travail ?

Organisation et déroulé

Principes

Se rapprocher autant que possible des conditions de travail habituelles.

Nous avons cherché à intégrer le service d'indexation de la manière la plus fluide possible dans WinIBW, l'outil de catalogage utilisé chaque jour par des centaines de professionnels du réseau Sudoc (potentiellement 3 000). Les suggestions se présentent sous la forme de zones UNIMARC 606 contenant :

- Les indicateurs ##
- Le terme RAMEAU, représenté à la fois par l'élément d'entrée du point d'accès autorisé (derrière un $a) et par son identifiant (derrière un $3)
- Le code du système d'indexation $2rameau

  `606 ##$3027233464$Coutumes alimentaires$2rameau`

L'action du testeur consiste alors à évaluer les différentes suggestions, puis à conserver celles qui lui semblent pertinentes. Il ajoutera probablement des indexations complémentaires. Il peut également ne retenir aucune des suggestions. Il peut enfin retenir un concept suggéré et y ajouter une subdivision de son cru (ou l'inverse). Ses actions sont entièrement libres, comme pour toute autre zone UNIMARC dans WinIBW (modulo les règles de validation à l'enregistrement).

Avant d'enregistrer la notice ainsi complétée, le testeur doit supprimer les informations contenues dans les suggestions incompatibles avec le format de catalogage (par exemple les éléments d'entrée des points d'accès autorisés) pour ne conserver qu'une zone structurée `606 ##$3<identifiant>$2rameau`, ce qui a le second mérite d'empêcher tout enregistrement accidentel.

Par défaut, seul le titre de la ressource (exprimée dans les zones 200 ae de la notice) est pris en compte pour proposer une indexation, mais le catalogueur peut sélectionner d'autres données (résumé, note de contenu) afin qu'elles servent également à la génération des suggestions.

Dans le contexte de l'expérimentation, le catalogueur ne peut pas paramétrer la liste des modèles appelés, ni le nombre de résultats par modèle.

Disposer des retours d'une réelle diversité d'établissements et d'agents

Nous avons sélectionné les 12 établissements testeurs dans l'optique de constituer un groupe le plus hétérogène possible en matière de taille d'établissement, de collections et de niveaux d'expertise au sein des équipes. Tous les établissements sollicités ont répondu positivement.
En outre, chaque établissement était invité à impliquer des agents aux profils variés dans l'expérimentation, sans fixer le nombre de participants par établissement.

Dialoguer avec les établissements via un référent

Nous souhaitions animer l'expérimentation en privilégiant un dialogue direct avec les établissements, notamment pour être capables de répondre très rapidement à leurs questions ou difficultés et de proposer des ajustements sans attendre la fin de l'expérimentation. Ce

dialogue s'est établi à travers la désignation par chaque établissement d'un seul interlocuteur. C'est ce référent qui avait pour mission de coordonner l'expérimentation en interne, de synthétiser les remontées de leurs collègues et de participer aux réunions de *debriefing* (réunissant tous les référents et l'équipe Abes).

Modalités

Pour les testeurs, l'expérimentation a commencé le 17 octobre 2024 et s'est terminée début janvier 2025, ce qui revient à une période de deux mois, si on exclut les vacances d'automne et de Noël. Une réunion de lancement avec les référents puis une autre réunion avec tous les testeurs ont été organisées en amont de cette phase.

Au bout d'un mois, les référents ont été invités à nous transmettre leurs retours, sous une forme assez libre, afin d'alimenter une réunion de *debriefing* le 21 novembre 2024, ayant pour objectif d'établir un bilan intermédiaire et de décider d'éventuels ajustements avant de poursuivre l'expérimentation.

Début janvier, chaque référent a transmis à l'équipe Abes de nouveaux et ultimes retours, en amont d'une réunion finale le 16 janvier 2025.

En parallèle de ces échanges, l'équipe Abes a mis en place une solution technique lui permettant d'observer aussi finement que possible les activités de chaque établissement (pas au niveau individuel). Ce dispositif a permis d'identifier quelles suggestions étaient conservées à l'enregistrement, après l'appel du service par le catalogueur. Ces données d'usage nous ont permis de calculer un **indicateur**, à savoir le ratio suivant : **nombre de suggestions conservées à l'enregistrement / nombre de concepts enregistrés**. Par exemple, si, pour une notice bibliographique donnée, le catalogueur ajoute quatre concepts RAMEAU à une notice, dont deux étaient présents dans les suggestions, le ratio est de 50%. Il est ensuite possible d'agréger ce ratio à différentes échelles : pour l'ensemble des établissements, par établissement, par modèle, par agrégation de modèles.

Configuration initiale du service

Configuration de l'IA

Concrètement, WiniBW appelle un **web service** qui prend en paramètres le titre (et sous-titre éventuel) et, en option, le contenu de n'importe quelle autre zone (le résumé, habituellement). Ce web service prévoit d'autres paramètres, fixés en dur dans le cadre de cette expérimentation :

- La liste des modèles
- La liste des « agrégations » (différentes manières de combiner les résultats des différents modèles, comme l'intersection de deux ou de tous les modèles)
- Le nombre de résultats par modèle.

Dans l'absolu, ces paramètres pourraient être modifiés par tout utilisateur du web service, depuis un programme ou une interface comme WinIBW.

Ce web service renvoie une réponse en JSON qui liste les suggestions correspondant à chaque modèle et chaque agrégation. Pour chaque suggestion, sont mentionnés un identifiant (IdRef) et un libellé.

Les modèles sélectionnés dans le cadre de l'expérimentation (*sentence embeddings* et OMIKUJI), sont les meilleurs modèles évalués en 2023 (des *sentence embeddings* et OMIKUJI). Cependant, un autre modèle de *sentence embedding* s'est ajouté depuis, qui fournit d'excellentes suggestions (précisons que ce modèle n'a pas été évalué comme les modèles de 2023,).

Ce nouveau modèle a été appliqué à notre corpus francophone mais également à un corpus anglophone. Il se trouve qu'il renvoie souvent de bonnes suggestions pour d'autres langues également, comme l'allemand ou le chinois. Nous avons donc invité les testeurs à essayer ce service sur des documents en langues variées, tout en rappelant que ses performances devaient être évaluées en priorité sur des documents en français.

Les agrégations sélectionnées dans le cadre de l'expérimentation sont à l'intersection entre deux modèles et le filtrage par *Large Language Model* (LLM). Cette dernière technique consiste à demander à un LLM de repérer des anomalies parmi les suggestions renvoyées par les différents modèles. Cette "agrégation" permet d'améliorer la qualité moyenne des suggestions et de réduire la liste des suggestions à consulter par le testeur. Cette méthode avait été évaluée en 2023.

A noter que le nouveau modèle d'*embedding* et l'agrégation par LLM ne peuvent tourner que sur un serveur GPU.

Configuration de la présentation des suggestions

Plusieurs questions se posaient : Comment présenter les résultats du service à l'utilisateur ? Faut-il les présenter de la manière la plus brute possible, permettant ainsi la « traçabilité » des suggestions, modèle par modèle, mais au risque d'une surcharge cognitive ? Ou bien faut-il « *designer* » l'affichage des résultats ? Dans ce cas, comment ? Aussi prosaïque qu'elle puisse paraître, la disposition de zones UNIMARC reste une question de design. A ce titre, elle induit fatalement un certain degré de confort d'utilisation et des incitations silencieuses (*nudges*).

Nous avons choisi de présenter les suggestions des modèles et agrégations en trois blocs :

1. Le bloc 1 des suggestions filtrées par LLM, mais sans les présenter ainsi. Nous avons préféré appeler ce bloc « Suggestions à analyser en premier », pour souligner qu'il était plus susceptible de proposer des suggestions pertinentes, sans toutefois le garantir.
2. Le bloc 2 des suggestions communes à au moins deux modèles (« Suggestions communes à plusieurs modèles »)
3. Le bloc 3… du reste des suggestions (« Suggestions complémentaires proposées »).

À noter que les deux premiers blocs peuvent se recouper partiellement.

Bilan d'étape et ajustements

Un mois après le lancement, nous avons procédé à un bilan d'étape en exploitant deux sources d'information : d'une part, des indicateurs d'usage et d'autre part, des retours d'expérience de la part des testeurs.

Ce bilan d'étape avait trois objectifs :

- Identifier et lever tout point de blocage qui aurait entravé la participation d'un établissement, qu'il s'agisse d'incompréhensions sur les objectifs ou les modalités, ou bien d'obstacles techniques rédhibitoires
- Recueillir les premières impressions sur la qualité du service
- Envisager et concevoir ensemble d'éventuelles améliorations à mettre en œuvre dans le cadre de l'expérimentation elle-même

Indicateurs d'usage

Nous avons mis en place des traces d'activité, ce qui a permis de savoir *quel établissement* (RCR) avait appelé le service, pour *quelle notice*, à *quelle date et quel horaire précis, à partir de quel titre*. Par ailleurs, nous avons conservé la réponse du web service associée à cet appel. Enfin, nous avons pu consulter l'historique des modifications des notices Sudoc pour identifier les zones d'indexation ajoutées après cet appel.

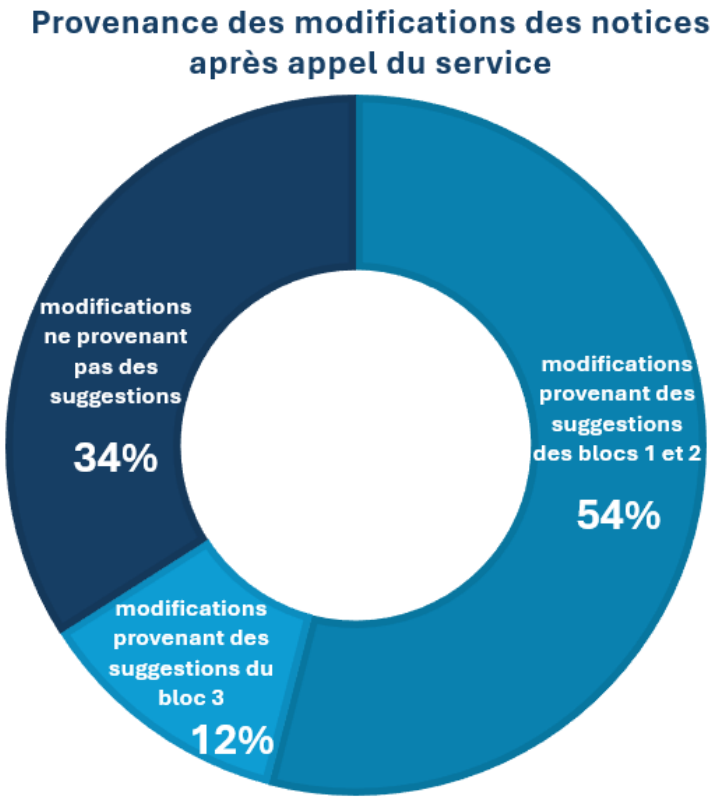
En croisant ces trois sources de données, nous avons pu identifier avec précision les concepts RAMEAU proposés par le service et les concepts enregistrés par l'utilisateur. Cela nous a permis de calculer **l'indicateur** suivant : **parmi les concepts enregistrés par l'utilisateur après l'appel du service, combien étaient présents dans la liste des concepts suggérés ?** Appelons cet indicateur « Ratio suggérés/enregistrés ».

Du 17 octobre au 17 novembre, 671 notices ont été modifiées (ajout de zones 606) à la suite d'un appel du service.

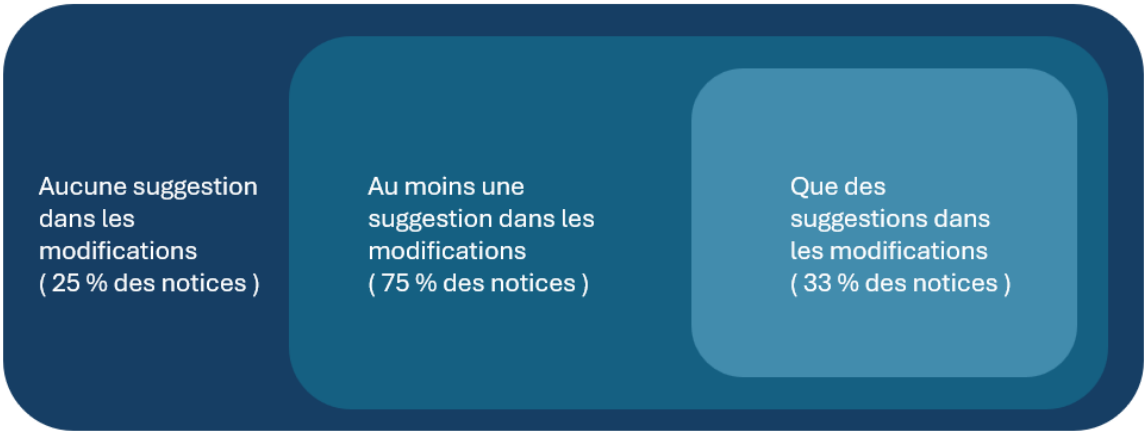
- 54 % des concepts RAMEAU ajoutés étaient présents dans les suggestions du bloc 1 ou 2 (dont 47 % dans le seul bloc 1).
- 12 % étaient présents dans le bloc 3 (qui de fait jouait le rôle de « poubelle », accueillant les concepts non éligibles aux blocs 1 ou 2)

Cela signifie que 34% des modifications ne provenaient pas des suggestions. C'est un pourcentage faible, qui doit sans doute être relativisé : dans le cadre d'une expérimentation dédiée aux suggestions par IA, il est naturel que ces dernières bénéficient d'une attention particulière de la part de l'utilisateur.

Même en prenant en compte la précaution interprétative précédente, on peut retenir de ces chiffres que **plus de la moitié des modifications effectuées par les utilisateurs concernaient des concepts présents dans les suggestions de l'IA**.



Par ailleurs, pour 207 notices modifiées (1/3 des notices), 100 % des concepts RAMEAU ajoutés étaient présents dans les suggestions. Pour 147 notices modifiées (25 % des notices), aucun concept RAMEAU ajouté n'était présent dans les suggestions. Et donc pour toutes les autres notices (75 % des notices), au moins un concept ajouté venait des suggestions.



Notices modifiées après appel du service

Enfin, en ventilant les données par bibliothèque (RCR), on constate que le nombre de notices modifiées après appel du service affiche des variations très importantes (entre 0 et 81) et que le ratio « suggérés/enregistrés » affiche des variations non négligeables mais moins fortes (la moyenne et la médiane sont à 63% pour les 23 RCR les plus actifs et à 49% pour les 10 RCR les plus actifs).

Agent	count_biblio	mean_suggestedAdded
RCR_751052119	81	50.3
RCR_352382102	73	42.2
RCR_751072303	66	35.5
RCR_751052304	65	56.7
RCR_991272301	53	41.7
RCR_352382101	45	47.1
RCR_385162101	31	40.8
RCR_514542102	30	62.2
RCR_130012101	26	60.0
RCR_372610011	23	75.3
RCR_341722103	17	60.8
RCR_130012305	14	62.2
RCR_511082201	13	85.1
RCR_384212101	13	68.5
RCR_384212103	13	63.4
RCR_372615209	12	70.0
RCR_352382106	11	78.7
RCR_384219901	11	69.1
RCR_352382210	9	85.1
RCR_372612103	7	76.1
RCR_130552106	7	90.0
RCR_341722106	5	64.0
RCR_301892102	5	78.2

[3]

Retours des testeurs

L'Abes a souhaité susciter des retours très libres, ainsi guidés :

Critère 1 : configuration du service
Exemples de questions à se poser :
- le script est-il facile à installer ?
- ai-je rencontré des difficultés dans la configuration ou l'utilisation du script ? si oui, lesquelles ?

Dans un certain nombre d'établissements, l'installation et surtout le bon fonctionnement du script dans WinIBW ont posé des difficultés importantes voire bloquantes. Le dysfonctionnement était dû à la dépendance du script VBScript à Internet Explorer, obstacle connu qui empêche déjà plusieurs établissements d'utiliser le script IdRef. Une collaboration étroite et persévérante entre les établissements impactés et l'Abes a permis de réduire petit à petit le nombre de situations bloquées, non sans avoir généré une perte de temps aux uns et aux autres, et différé d'autant l'expérimentation proprement dite. Début décembre 2024, une solution complètement indépendante d'Internet Explorer a été diffusée, ce qui a permis d'outrepasser les quelques blocages qui demeuraient.

[3] Ce tableau ne liste pas la totalité des bibliothèques (RCR)

Critère 2 : ergonomie de la présentation des suggestions

Exemples de questions à se poser :
- le nombre de suggestions est-il suffisant ?
- les intitulés des blocs de suggestions doivent-ils être reformulés ?
- la présentation des suggestions doit-elle être améliorée ? Si oui, comment ?

Pour la plupart des testeurs, le service apporte une réelle plus-value mais l'ergonomie choisie initialement n'est pas satisfaisante :

- Trop de suggestions, parfois redondantes d'un bloc à l'autre (blocs 1 et 2)
- Bloc 3 peu ou pas pertinent

Ce choix d'organisation des suggestions entraîne une perte de temps pour l'utilisateur.

Critère 3 : qualité des suggestions d'indexation

Exemples de question à se poser :
- les suggestions sont-elles pertinentes ?
- les suggestions sont-elles adaptées au type de ressource que je catalogue ?
Parmi les cas les plus flagrants de suggestions inutiles, voire erronées, les testeurs peuvent en noter quelques-uns, à titre d'exemple. Mais il n'est pas demandé de qualifier de façon exhaustive toutes les suggestions : il y aura forcément des suggestions inappropriées.

Ici, les retours sont assez contrastés, ce qui reflète la diversité des profils de testeurs et d'établissements :

- Les testeurs issus d'établissements dont les collections relèvent d'un domaine très spécialisé tirent moins ou peu bénéfice du service (la BIU Cujas, la bibliothèque du Muséum National d'Histoire Naturelle, la bibliothèque de Sciences Po Paris).
- Les déceptions tiennent parfois à des attentes élevées, voire extérieures au cadre annoncé (par exemple, sur les langues supportées, les suggestions de subdivisions, la nécessité de vérifier attentivement...)

À des degrés variés, le service s'avère peu pertinent pour :

- Les documents en langue gaélique, en arabe, en turc, russe et grec
- Les documents aux sujets très spécifiques, comme la classification des espèces en biologie
- Les documents précisément datés et localisés, comme en archéologie
- Les documents cartographiques
- Les œuvres de fiction
- Les CD audio
- Les documents en santé
- L'analyse des mots clés libres (présents en zone UNIMARC 610)

À des degrés variés, le service s'avère pertinent pour :

- Les documents en français, en allemand, en italien, en anglais, en espagnol, et même en chinois
- Les thèses d'exercice et les mémoires
- Les documents avec présence de résumé
- Les documents des disciplines de sciences « dures »

L’utilisation du résumé, recommandée mais pas toujours possible, améliore nettement la qualité des suggestions. Au contraire, la table des matières y nuit.

Souhaits pour compléter la qualité : autorités préconstruites, subdivisions géographiques et chronologiques, genre/forme, indexation LCSH (à convertir en RAMEAU), prise en compte par défaut du résumé lorsqu’il est présent.

Critère 4 : satisfaction globale
Exemples de question à se poser :
- le service fait-il gagner du temps dans l’activité de catalogage ?
- le service s’insère-t-il facilement dans ma pratique habituelle d’indexation ?
- à quel point facilite-t-il mon travail ?

Les testeurs attendent davantage de précision et d’ergonomie, pour gagner du temps.
Un des principaux enseignements est que la plus-value du service porte moins sur le gain de temps que sur la qualité de l'indexation finale : le service apporte des suggestions auxquelles les testeurs n'auraient pas forcément pensé, ou bien nourrit leur réflexion pour orienter leur recherche dans la base RAMEAU. Néanmoins, dans certaines configurations, notamment le catalogage rétrospectif, où l’indexation ne peut être très approfondie, il peut aussi faire gagner du temps.

Public préconisé : catalogueurs avec une bonne connaissance de RAMEAU

Ajustements

L'équipe de l’Abes avait pu prendre connaissance de ces retours avant la réunion de bilan d'étape du 21 octobre 2024, et proposer des améliorations de court terme. Début novembre, les testeurs ont pu disposer d'une nouvelle version du service qui, autant que possible, prenait en compte leur retour. Les utilisateurs ont apprécié que l’Abes ait fait son possible pour prendre en compte les premiers retours et proposer des évolutions, implémentées dans le temps de l’expérimentation.

Réforme de l’organisation de l’affichage des suggestions :

1. Suppression du bloc 3 et des suggestions qu’il contenait. Pour rappel, ce bloc contenait les suggestions qui n'étaient ni dans le bloc des intersections ni dans le bloc des suggestions retenues par le LLM.
2. Fusion et dédoublonnage des deux premiers blocs
3. Présentation différenciée des concepts RAMEAU qui ne peuvent être utilisés qu’en subdivision, pour éviter des erreurs et faciliter le travail.
4. Présentation hiérarchique des concepts RAMEAU quand la liste des suggestions contient des concepts qui entretiennent entre eux des relations hiérarchiques dans le vocabulaire RAMEAU (terme générique / spécifique)

Cette série de changements promettait d’offrir une présentation plus simple, plus courte et néanmoins plus riche (relations hiérarchiques), et visait à gagner en qualité d’indexation et en temps de travail.

Avant

```
... Suggestions à analyser en premier :
606##$3027805115$aVision par ordinateur$2rameau
606##$3033936242$aFisher, Équation de$2rameau
606##$3027940373$aApprentissage automatique$2rameau
606##$3030971098$aRéseaux neuronaux (informatique)$2rameau
606##$3223540633$aApprentissage profond$2rameau
606##$3027242307$aReconnaissance des formes (informatique)$2rameau
606##$3027234657$aTraitement d'images$2rameau
606 ##$3027234541$aIntelligence artificielle

... Suggestions communes à plusieurs modèles :
606##$3027805115$aVision par ordinateur$2rameau
606##$3223540633$aApprentissage profond$2rameau
606##$3027940373$aApprentissage automatique$2rameau
606##$3030971098$aRéseaux neuronaux (informatique)$2rameau
... Suggestions complémentaires proposées :

606##$3027569217$aCytopathologie$2rameau
606##$3027864383$aMesures optiques$2rameau
606##$3203035763$aReconnaissance multivues$2rameau
606##$3087833069$aPlaques d'immatriculation$2rameau
606##$3028185285$aFusion binoculaire$2rameau
```

Après

```
606 ##$3027234657$aTraitement d'images
     606 ##$3027805115$aVision par ordinateur
606 ##$3033936242$aFisher, Équation de
606 ##$3027234541$aIntelligence artificielle
     606 ##$3027242307$aReconnaissance des formes (informatique)
     606 ##$3027805115$aVision par ordinateur
     606 ##$3027940373$aApprentissage automatique
          606 ##$3223540633$aApprentissage profond
     606 ##$3030971098$aRéseaux neuronaux (informatique)
```

En outre, le modèle (OMIKUJI) qui proposait le plus de mauvaises suggestions (celles-ci se retrouvant dans le bloc 3), a été retiré du service, alors même que l'évaluation 2023 en faisait l'un des modèles les plus performants – ce qui témoigne des progrès réalisés par les modèles d'*embeddings* récents.

La prise en compte d'autres retours ne pouvait se traduire par des améliorations du service à court terme.

Des modèles spécialisés. Pour proposer des suggestions plus adaptées à des contextes particuliers, il semble logique de générer des modèles générés à partir de corpus spécialisés : corpus de domaine comme le droit ou la médecine, corpus de langue comme l'allemand ou l'italien, corpus de type de document comme les thèses de doctorat, etc. On pourrait même imaginer un corpus de collection, pour des bibliothèques au profil singulier, comme la bibliothèque du MNHN.

Plus de subdivisions. Certains de nos modèles ont été générés en conservant les subdivisions : au lieu de faire comme si telle notice avait une liste de concepts comme indexation (A et B), la structure de l'indexation Sudoc (A -- B) a été conservée. Cette méthode ne répond pas entièrement au souhait exprimé d'obtenir des suggestions avec subdivision(s) car elle ne permet pas d'obtenir des constructions inédites, associant deux concepts dont l'association est absente du Sudoc. La résolution de ce problème demanderait probablement une méthode nouvelle, différente de l'approche suivie par le service actuel.

Au-delà des concepts : les entités géographiques, les entités chronologiques et les genres-formes. Pour proposer une indexation géographique ou chronologique, on pourrait utiliser l'approche du service actuel mais il semble plus adéquat d'adopter une autre approche, qui s'appuierait sur les termes géographiques et chronologiques extraits du texte de la notice. Il s'agirait de procéder à une extraction d'entités puis à l'alignement sur des concepts RAMEAU correspondant.

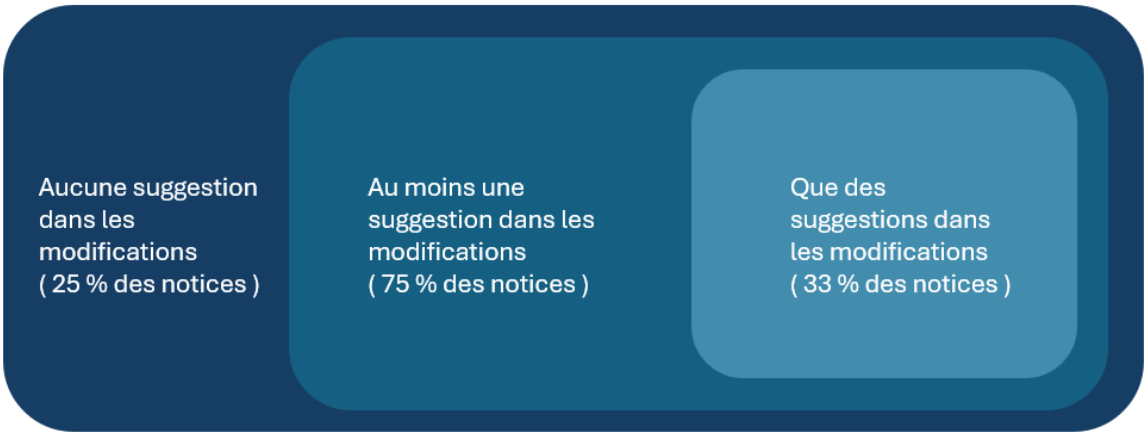
Bilan final

Synthèse de l'activité observée

Pendant la durée de l'expérimentation,

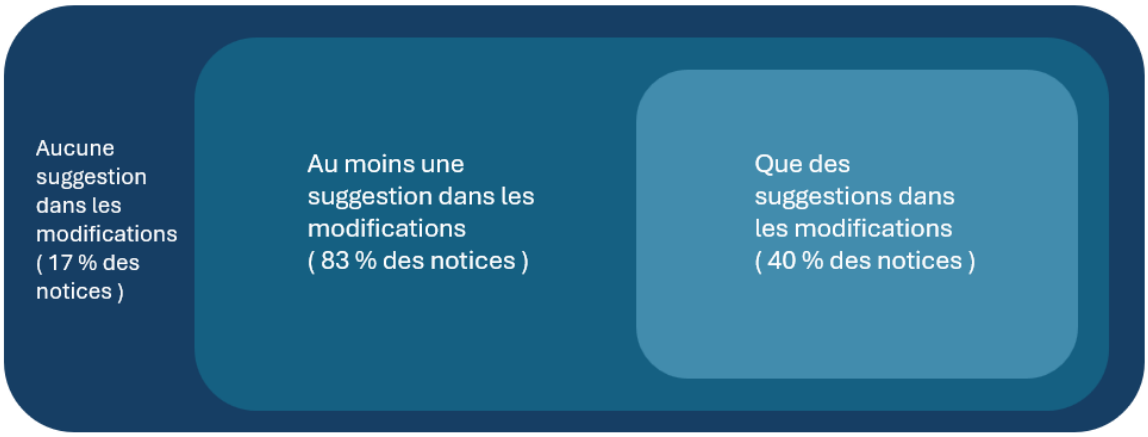
- ✓ 1 377 notices ont été modifiées (ajout de zones 606) à la suite d'un appel du service.
- ✓ 62 % des concepts RAMEAU enregistrés dans la base étaient présents dans les suggestions (moyenne pour tous les RCR). 58% pour les 15 RCR ayant modifié le plus de notices.
- ✓ Pour 513 notices modifiées (40% des notices), 100 % des concepts RAMEAU enregistrés étaient présents dans les suggestions.
- ✓ Pour 241 notices modifiées (17 % des notices), aucun concept RAMEAU enregistré n'était présent dans les suggestions. Ceci implique que, pour toutes les autres notices (83 % des notices), au moins un concept enregistré venait des suggestions.

Avant les ajustements



Notices modifiées après appel du service

Après les ajustements



Notices modifiées après appel du service

Cette amélioration du pourcentage des modifications de notice ayant retenu une ou plusieurs suggestions ne s'explique pas par une meilleure qualité des suggestions car les ajustements n'ont pas porté sur cet aspect du service : le retrait du modèle le moins performant n'a pas apporté de meilleures suggestions, mais seulement supprimé de moins bonnes ou mauvaises.

On peut donc faire l'hypothèse raisonnable **que c'est la meilleure lisibilité de l'affichage des suggestions qui a conduit les utilisateurs à retenir davantage de suggestions du service** (moins de concepts, mieux présentés).

Retours finaux

Pour chaque question du sondage final, nous produisons d'abord une synthèse des réponses sous la forme d'un tableau, puis notre interprétation des réponses et enfin quelques citations.

La présentation des suggestions (1 seul bloc de réponses hiérarchisées) répond-elle à vos attentes ?	
Oui	56
Non	1
Autre (réponses hors sujet)	3

La nouvelle présentation des suggestions est apparue comme un progrès majeur : elle est plus simple, plus claire, permet de gagner en efficacité. Il ne faut pas sous-estimer l'importance de l'ergonomie, de la satisfaction de l'utilisateur dans son expérience d'interaction avec l'IA (UX), y compris dans un environnement aussi contraint que WinIBW.

Non : « *L'emplacement du bloc n'est pas bon > il devrait s'intercaler dans la notice bibliographique, au niveau des 6XX, pour une meilleure ergonomie* »

Que pensez-vous finalement de la qualité des suggestions ?	
Satisfait (de plutôt satisfait à très satisfait)	37
Moyennement satisfait	17
Insatisfait	4
Autre	2

Le verbatim qui accompagne les réponses fait comprendre qu'il existe différents critères de satisfaction selon les utilisateurs :

- Pour certains, s'il y a de bonnes suggestions, c'est satisfaisant, malgré la présence des autres.
- Pour d'autres, la présence de suggestions non pertinentes suffit à discréditer le service dans sa globalité.

On peut voir là, plus généralement, deux attitudes psychologiques différentes face aux réponses d'une intelligence artificielle. Chacune a sa légitimité.

En outre, il existe différentes manières pour des suggestions d'être non pertinentes :

- Elles peuvent être manifestement hors sujet voire fantaisistes. Il s'agit là d'erreurs propre aux machines : un professionnel ne produirait jamais ce genre d'indexation. Là encore, soit la présence de ces suggestions déplacées est rapidement filtrée et négligée par l'utilisateur, soit elle provoque une sorte de blocage, tout le contraire d'une aide.
- Les suggestions sont souvent jugées trop imprécises, surtout pour certains établissements ou pour certains types de documents. Nous y reviendrons plus loin.
- Enfin, elles peuvent être incomplètes : la liste des suggestions ne représente pas tout le contenu du document.

Satisfait : « J'ai trouvé les suggestions dans l'ensemble pertinentes. Dans les résultats contenant le plus de bruits, j'ai tout de même trouvé des pistes me permettant de dégager les concepts appropriés. Je suis très satisfaite. »

Satisfait : « 8 sur 10 en moyenne ne sont pas adaptées, mais, paradoxalement, très satisfait. L'IA comme souvent donne ici des pistes auxquelles on n'aurait peut-être pas pensé. Rien que pour cela on aurait tort de s'en passer. »
Satisfait : « La qualité des suggestions est variable en fonction des éléments exploitables dans la notice, mais elles permettent au moins de conforter les choix ou parfois de donner des pistes auxquelles on n'avait pas pensé. »
Moyennement satisfait : « Généralement pertinentes, mais pas assez précises. »
Insatisfait : « Manque de pertinence, souvent pas adéquates, parfois fantaisistes »

Quel est votre niveau de satisfaction, à la fin de la période d'expérimentation ?	
Satisfait (de plutôt satisfait à très satisfait)	37
Moyennement satisfait	17
Insatisfait	3
Autre	1

Les jugements négatifs ou mitigés se concentrent autour de trois établissements les plus spécialisés (qui attendent plus de précision et de richesse) mais également deux établissements généralistes. Il est intéressant de constater que les avis sont assez convergents au sein d'un même établissement. On peut imaginer que la manière d'appréhender un service d'aide à la décision par IA est influencée par la culture professionnelle de l'établissement ou sa dynamique de travail collective.

Satisfait : « Il est bon. Les suggestions permettent de gagner du temps en donnant les PPN des vedettes matières. Plus besoin d'aller les chercher. La présentation des suggestions hiérarchisées en un bloc a aussi beaucoup aidé. »
Satisfait : « Je suis plutôt très satisfaite. Comme je travaille dans un service de périodiques, je ne suis pas spécialisée dans un domaine précis, je touche à tous les domaines de ma bibliothèque. Les suggestions permettent de m'aiguiller rapidement et sont un gain de temps important. »
Satisfait : « Je suis assez satisfaite de l'outil. Il ne fait pas gagner de temps dans la rédaction de notices bibliographiques, mais il est gage de qualité dans la description du document. »
Moyennement satisfait : « Correct, mais l'outil reste à développer. Il a beaucoup évolué pendant la période d'expérimentation, nos remarques et suggestions ont été prises en compte. C'est un bon début. »
Insatisfait : « Ne constitue pas une aide à la décision, ne constitue pas un gain de temps dans la pratique du catalogage. Peut-être cela est-il plus performant selon les ressources à traiter ? »

La citation suivante exprime bien l'opinion de la majorité des référents :

> « La limitation la plus gênante est due au service lui-même : seuls les concepts RAMEAU sont traités ; les disciplines, pour lesquelles l'indexation nécessite des constructions avec subdivisions ou d'autres accès 60x, sont négligées.
> De l'avis unanime de nos testeurs, le service est utile et était très attendu ; il doit être conservé, mais il est jugé encore incomplet : pour être pleinement fonctionnel, il doit pouvoir traiter la majorité des types de documents, et ne pas non plus se limiter aux concepts RAMEAU : personnes, titres, noms géographiques, subdivisions chronologiques, genre/forme doivent être pris en compte. »

Souhaiteriez-vous que le service reste disponible au-delà de l'expérimentation ?	
Oui	57
Non	2
Autre	1

La quasi-unanimité des sondés demande le maintien du service.

Si le service restait disponible, dans quel contexte l'utiliseriez-vous ?	
Catalogage courant ?	49
Rétrospectif	37
Chantiers spécifiques	24

Il était possible de sélectionner plusieurs de ces réponses, ce qui explique pourquoi le total des avis exprimés dépasse 60.

Si le service restait disponible, pour quelles langues de publication des documents souhaiteriez-vous l'utiliser ?	
Pour le français uniquement	24
Pour le français et d'autres langues	14
Autre réponse	20

Langues fréquemment mentionnées, dans l'ordre : anglais, allemand, espagnol, italien (+ mention du russe et du chinois)

Si le service restait disponible, à quelle catégorie de catalogueurs le destineriez-vous ? Catalogueurs débutants ou experts ? Les deux indifféremment ?	
Débutants seulement	4
Experts seulement	21
Les deux	34

Certains craignent légitimement qu'une personne peu expérimentée néglige la nécessité d'une analyse critique des suggestions et la nécessité d'effectuer des recherches complémentaires. Mais d'autres jugent que l'assistance de l'IA peut aider un débutant à découvrir le vocabulaire RAMEAU et l'indexation, à condition qu'il soit accompagné au sein d'une équipe.

Si le service restait disponible, pour quels domaines disciplinaires ou quels types de documents le préconiseriez-vous ?
. Surtout des monographies, mais également les périodiques . Ni la littérature, ni les cartes, ni les partitions . Plutôt en SHS, mais pas seulement, et cela dépendrait de l'évolution des modèles vers plus de précision (avec l'intégration thèses dans les corpus d'entraînement). . Pour un certain nombre d'établissements, pertinent également pour le niveau recherche des ouvrages, y compris les thèses de doctorat ou d'exercice, mais ces réponses sont difficilement conciliables avec le constat récurrent d'un défaut de précision. . Plutôt SHS

Les réponses à cette question sont très diverses, voire dissonantes, ce qui reflète encore une fois la diversité des contextes, des collections, des modes d'organisation, etc.

Bilan global

L'observation des pratiques, l'analyse du sondage et du contenu des échanges directs permettent de dessiner un bilan global de l'expérimentation du point de vue des utilisateurs.

Le service est considéré comme opérationnel et utile pour la grande majorité des membres d'établissements généralistes, qui constituent l'essentiel du réseau Sudoc : soit les suggestions sont pertinentes en l'état, soit elles constituent un bon point de départ pour l'exploration du vocabulaire RAMEAU, soit elles servent à confirmer ou consolider *a posteriori* les choix de l'utilisateur. De là découle le constat selon lequel **cette assistance conduit à une indexation de meilleure qualité, plus précise ou plus complète** – ce qui est un enseignement inattendu. Le gain de temps est également mentionné, mais surtout pour des activités spécifiques comme le catalogage rétrospectif. Ces gains de qualité et de temps sont augmentés grâce à la nouvelle version de l'affichage des suggestions. En résumé, on pourrait dire que **le service apporte bien une aide à la décision, mais également une aide à la délibération**, à l'exploration des solutions.

La qualité des suggestions demeure inégale, surtout **pour des documents pointus ou des établissements spécialisés,** pour lesquels **la précision est insuffisante**. Cela est inévitable car le corpus de travail ne contenant pas les thèses et leur indexation souvent très précise, de nombreux concepts RAMEAU n'était pas présents dans le modèle : par définition, ils ne pouvaient pas être proposés. Ces manques seraient en grande partie comblés par la prise en compte des thèses dans les corpus.

Ce n'est pas seulement par leur **silence**, mais également par leur **bruit** que les suggestions peuvent frustrer (surtout quand le titre est seul utilisé, est-il précisé). Malgré le filtrage *a posteriori* par LLM, qui élimine l'essentiel des suggestions fantaisistes, il est inévitable que figurent presque toujours des concepts inadéquats. Certains utilisateurs déclarent ne pas être gênés par ce bruit, car leur approche se concentre sur l'identification des « bons » concepts et l'élimination des autres. D'autres utilisateurs considèrent que ce bruit leur fait perdre du temps et peut desservir la lecture des suggestions dans leur ensemble. On observe ici une **distinction entre deux attitudes psychologiques, une approche « sélective » et une approche « globale »**, qu'on retrouve probablement dans d'autres interactions entre un humain et une intelligence artificielle. Est-ce à la machine de s'adapter aux différents profils d'utilisateurs ? ou bien à l'humain de s'adapter ? Il faut plutôt anticiper toutes sortes de modalités, de degrés et de rythmes de coévolution humain/machine, que seule la pratique quotidienne peut faire émerger, à l'échelle individuelle et collective.

Enfin, une majorité de testeurs regrettent **l'absence des concepts géographiques, des concepts chronologiques, des concepts de type genre ou forme et enfin de suggestions avec construction** (concept principal + subdivisions), tous ces types de suggestions ayant été d'emblée exclus du service. En creux, ces demandes témoignent des fortes attentes envers un service considéré comme capable d'aller plus loin, de proposer une indexation plus riche.

Conclusions et recommandations

L'expérimentation dans son ensemble et la réunion de bilan permettent de conclure qu'une grande majorité des testeurs souhaite la mise à disposition du service dans son état actuel. D'après leurs référents, quelques établissements, aux profils et besoins spécifiques, n'adopteraient le service qu'après l'enrichissement des capacités du service, avec pour priorités les suggestions de type géographique et chronologique et une montée en précision. Tous les autres établissements sont également et fortement demandeurs de ces évolutions.

Pour répondre à ces demandes, l'Abes pourrait :

1. Mettre en production le service actuel et l'ouvrir à l'ensemble du réseau, moyennant des adaptations techniques pour faciliter l'administration du service et son maintien en condition opérationnelle, ce qui implique d'automatiser la mise à jour régulière des modèles et le recours à un serveur GPU (en local ou pas). Ce passage en production aurait pour conséquence d'initier une montée en compétences sur les outils d'IA au sein de la DSI de l'Abes.

2. Mettre à disposition des ressources de formation sur les apports et limites de l'utilisation de l'IA dans un contexte d'indexation documentaire et sur la méthodologie de travail assisté par l'IA, afin de permettre une utilisation avertie du service par les catalogueurs du réseau Sudoc.

3. Publier le webservice de type REST pour permettre son utilisation ou son intégration dans des outils externes au SI de l'Abes. Le service actuel comprend à la fois le web service de type REST et le script WinIBW. En lui-même, le web service est totalement indépendant de WinIBW et peut être intégré dans n'importe quel autre outil, voire appelé par programme (à condition d'une infrastructure adaptée pour supporter la charge éventuelle). Cette publication nécessiterait un accompagnement spécifique auprès des développeurs et intégrateurs qui souhaiteraient utiliser l'API dans un système externe, comme un SGB par exemple.

4. Réentraîner les modèles afin de mieux prendre en compte les thèses ainsi que d'autres documents à contenu spécialisé.

5. Etudier les solutions opérationnelles pour mettre en place une indexation géographique et chronologique RAMEAU.

6. Poursuivre le travail de R&D sur l'indexation genre/forme RAMEAU engagé depuis octobre 2024. Cette indexation est très spécifique.

7. Etudier les solutions pour générer des suggestions construites. Cette indexation est très spécifique, et complexe. Néanmoins, les testeurs reconnaissent que cette évolution est moins prioritaire, et on peut supposer que la suggestion d'une indexation géographique et chronologique RAMEAU faciliterait grandement la construction par l'utilisateur.

8. Générer un modèle d'indexation en FMeSH, selon la même procédure que celle utilisée pour RAMEAU. Cette demande, revenue à plusieurs reprises, est d'autant plus pertinente que les catalogueurs sont moins familiers avec ce vocabulaire contrôlé, qui porte sur un domaine spécialisé (le médical).

Le déploiement de ce service nécessiterait un accompagnement spécifique, à la fois par l'Abes et au sein de chaque établissement. Côté Abes, outre les dispositifs d'accompagnement classiques à l'occasion d'un nouveau service (communication, documentation, webinaires de formation), il faudrait prévoir un investissement initial spécifique concernant le script WinIBW (s'il était proposé, en plus du web service). Mais dans le cadre de cette expérimentation, l'équipe Labo a conçu et implémenté une solution technique qui dispense de l'installation d'Internet Explorer sur les postes utilisateurs, source principale de blocage.

Côté établissement, l'accompagnement serait plus profond et progressif, et prendrait des formes différentes selon les établissements. Ces derniers considèrent par ailleurs que cet investissement constituerait une première étape de l'acculturation des pratiques professionnelles des catalogueurs à la collaboration avec une IA en général, jugée inévitable et, dans l'ensemble, potentiellement bénéfique.

Au cours du 1er semestre 2025, l'Abes analysera les ressources nécessaires à la mise en place de nouveaux services sur la base des recommandations du présent rapport.

abes

Le rapport a été réalisé dans le cadre du Labo de l'Abes

Agence bibliographique de l'enseignement supérieur
227 avenue du Professeur Jean-Louis Viala - CS 84308 - 34193 Montpellier Cedex 5 - Tél. 04 67 54 84 10

MINISTÈRE CHARGÉ DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité