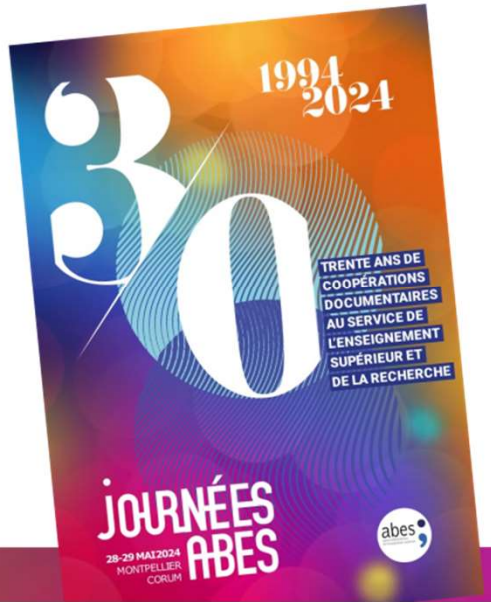
JOURNÉES ABES

L'IA à notre service

2023
2024
2025



Yann NICOLAS / Labo / Abes

abes
agence bibliographique de l'enseignement supérieur

La R&D et l'IA à l'Abes
Quoi de neuf depuis les Jabes 2023 ?

1. Un nouveau **projet d'établissement** (2024-2028)

2. Un service d'**indexation sujet** mûr pour être testé *in vivo*
   - Une **expérimentation** dans l'outil de catalogage actuel du Sudoc (2024 T4)

3. Quel regard rétrospectif sur ce premier service IA en test ? Quels enseignements et anticipations pour la suite ?

Projet d'établissement

1. Le projet définit des priorités générales pour l'agence pour 5 ans

   ➔Les priorités R&D doivent s'aligner sur ces priorités globales

2. Le projet définit une ambition nouvelle et des principes clairs pour la R&D à l'Abes

   ➔ Concentration sur l'IA

3. Le projet de définit pas de feuille de route

Objectifs prioritaires 2024-2025

« Faciliter le signalement »

Indexation Rameau en test dans l'outil de catalogage du Sudoc

Exploiter le texte intégral des thèses (PDF)

1. Extraire des métadonnées de la page de titre

2. Extraire la bibliographie

Signalement

Divers chantiers entamés, possibles, futurs…

Génération d'une indexation sujet	**. Aide à la sélection des sujets Rameau dans l'outil de catalogage (WinIbw aujourd'hui)** . Indexation automatique d'articles et de chapitres . Extraction et identification des entités géographiques . Extraction et identification des entités chronologiques
Structuration de zones	. Structuration de zones de note . Génération de points d'accès auteur à partir de la mention de responsabilité (zone de transcription, ie du texte)
Génération de liens	. Alignement vers des autorités Collectivité . Regroupement des manifestations d'un auteur dont les titres sont la traduction l'un de l'autre, pour lier à l'entité Œuvre
Conversion (imports, dérivation)	. Automatisation de la conversion du format X vers le format Y . Traduction automatique de zones textuelles (dérivation)

PDF des thèses

Divers chantiers entamés, possibles, futurs…

Extraction d'informations et de connaissances	. Segmentation de la thèse . Conversion en TEI **. Extraction des jurys des PDF** **. Extraction des références bibliographiques** . Extraction de la table des matières . Extraction et formalisation de connaissances (from LLM to knowledge graphs) . Extraction d'images, tableaux etc
Recherche sémantique	. Recherche "intelligente" qui capture le « sens », et pas seulement des mots. => pour retourner un document ou des extraits précis
Génération de texte	. Chat sur/avec une thèse . Chat sur/avec un ensemble de thèses . Résumé par section, pour lecture rapide dans theses.fr

Indexation Rameau

Tester un service d’indexation sujet par IA dans l’environnement du catalogueur

Objectifs

A partir d’un titre et d’un résumé, générer un ensemble de concepts Rameau suffisamment pertinent pour aider le catalogueur à indexer ce document

Résultats

Un service d’indexation Rameau

- Fonctionnel
- Qualifié
- Maîtrisé par l’Abes
- Paramétrable
- Intégrable dans les workflows et les outils

Déontologie de la recherche et intégrité scientifique par Olivier Leclerc

Pas dans le Sudoc

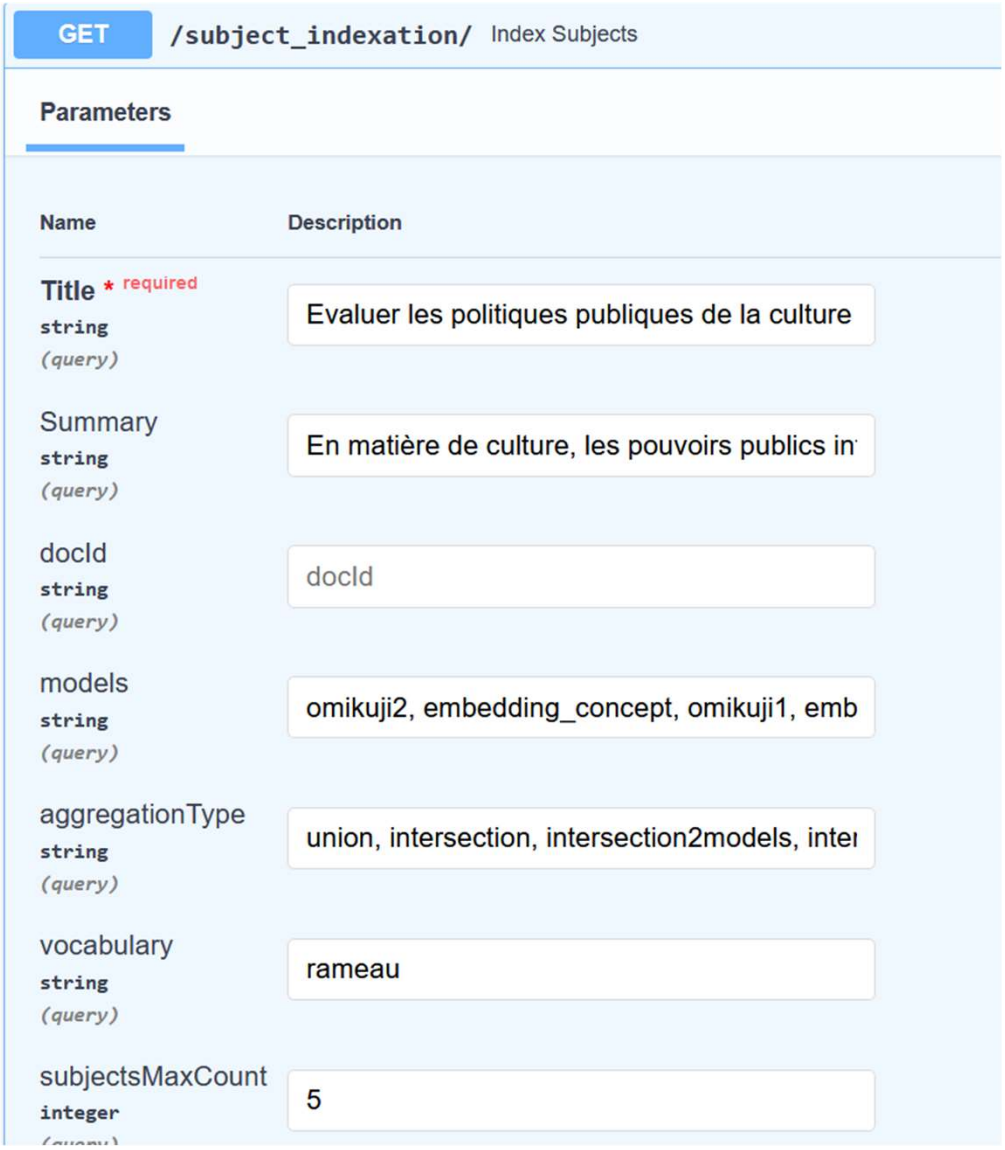
	Modèle 1	Modèle 2	Modèle 3
Concepts (6)	Information scientifique Rédaction scientifique et technique Données de la recherche Chercheurs Fraude scientifique Édition scientifique	Recherche Méthodologie Sciences Vulgarisation scientifique Information scientifique Chercheurs	Sciences et droit Fraude scientifique Liberté de la recherche Oeuvres scientifiques Politique scientifique Chercheurs

Web service flexible

A intégrer dans les workflows et les outils

- A l'Abes
- Chez vous

- Web service **paramétrable** :
  - Quel(s) modèle(s) ?
  - Combien de propositions Rameau ?
  - Comment agréger les propositions des modèles ?
    - Union
    - Intersections (différentes variantes)
    - IA qui propose, IA qui dispose
  - Tenter d'exploiter le titre seulement, en l'absence de résumé ?




```
Document ID:                    null
Prediction model by model:
  omikuji2:
    Result:
      0:
          label:                "Diffusion de la culture"
          id:                   "027224929"
          score:                0.10438302904367447
      1:
          label:                "Politique culturelle"
          id:                   "027416593"
          score:                0.03864680230617523
      2:
          label:                "Aspect politique"
          id:                   "027792102"
          score:                0.014079321175813675
      3:
          label:                "Effets des innovations technologiques"
          id:                   "028695151"
          score:                0.013222274370491505
      4:
          label:                "Médias"
          id:                   "027237524"
          score:                0.012716062366962433
      Response Time:            "2.13 secondes"
  embedding_concept:
    Result:
      0:
          label:                "Évaluation des politiques publiques"
          id:                   "029960835"
```

Déontologie de la recherche et intégrité scientifique par Olivier Leclerc

Pas dans le Sudoc

	Modèle 1	Modèle 2	Modèle 3
Concepts (6)	Information scientifique Rédaction scientifique et technique Données de la recherche Chercheurs Fraude scientifique Édition scientifique	Recherche Méthodologie Sciences Vulgarisation scientifique Information scientifique Chercheurs	Sciences et droit Fraude scientifique Liberté de la recherche Oeuvres scientifiques Politique scientifique Chercheurs
Intersection	**Chercheurs**		
Intersection2models	**Information scientifique, Fraude scientifique, Chercheurs**		
Intersection2models1best	**Information scientifique, Fraude scientifique, Chercheurs**		
Union			

Evaluation de la qualité des propositions Rameau

Qu'est-ce qu'une bonne indexation ?

Par un humain…

Par une IA

- Question conceptuellement difficile
- Gros travail d'évaluation :
  - Mesures classiques en machine learning insuffisantes
  - Produire de nouveaux lots de référence au-delà des notices Sudoc : **réindexer des notices**
  - Evaluation qualitative des indexations des humains et modèles IA : **noter les indexations (grille)**

Mesurer la qualité des propositions

En prenant l'intersection de deux IA :

- Entre 80 et 90% de propositions Rameau « vraies » (Hum. 95%)
- Entre 90 et 100% de documents avec au moins une proposition
- Autour de 4,5 propositions par document
- 63% du contenu du document est couvert par les propositions (Hum. 80%)

Minimiser le bruit
➔ compléter

- En prenant l'union de nos deux meilleurs IA :
  - 60% de propositions « vraies » (Hum. 95%)
  - 100% de documents avec au moins une proposition
  - Plus de 10 propositions par document (dont des redondances)
  - Plus de 90% du contenu du document est couvert par les propositions (Hum. 80%)

Minimiser le silence
➔ sélectionner

➔ Différentes préférences selon les utilisateurs : service paramétrable

Mesurer la qualité des propositions (suite)

En prenant l'intersection de deux ou plusieurs IA et en demandant à chatGPT de ne conserver que les bonnes propositions:

- Entre 95 et 99% de propositions Rameau « vraies » (Hum. 95%)
- Entre 61 et 69% de documents avec au moins une proposition
- Autour de 2 propositions par document
- ??% du contenu du document est couvert par les propositions

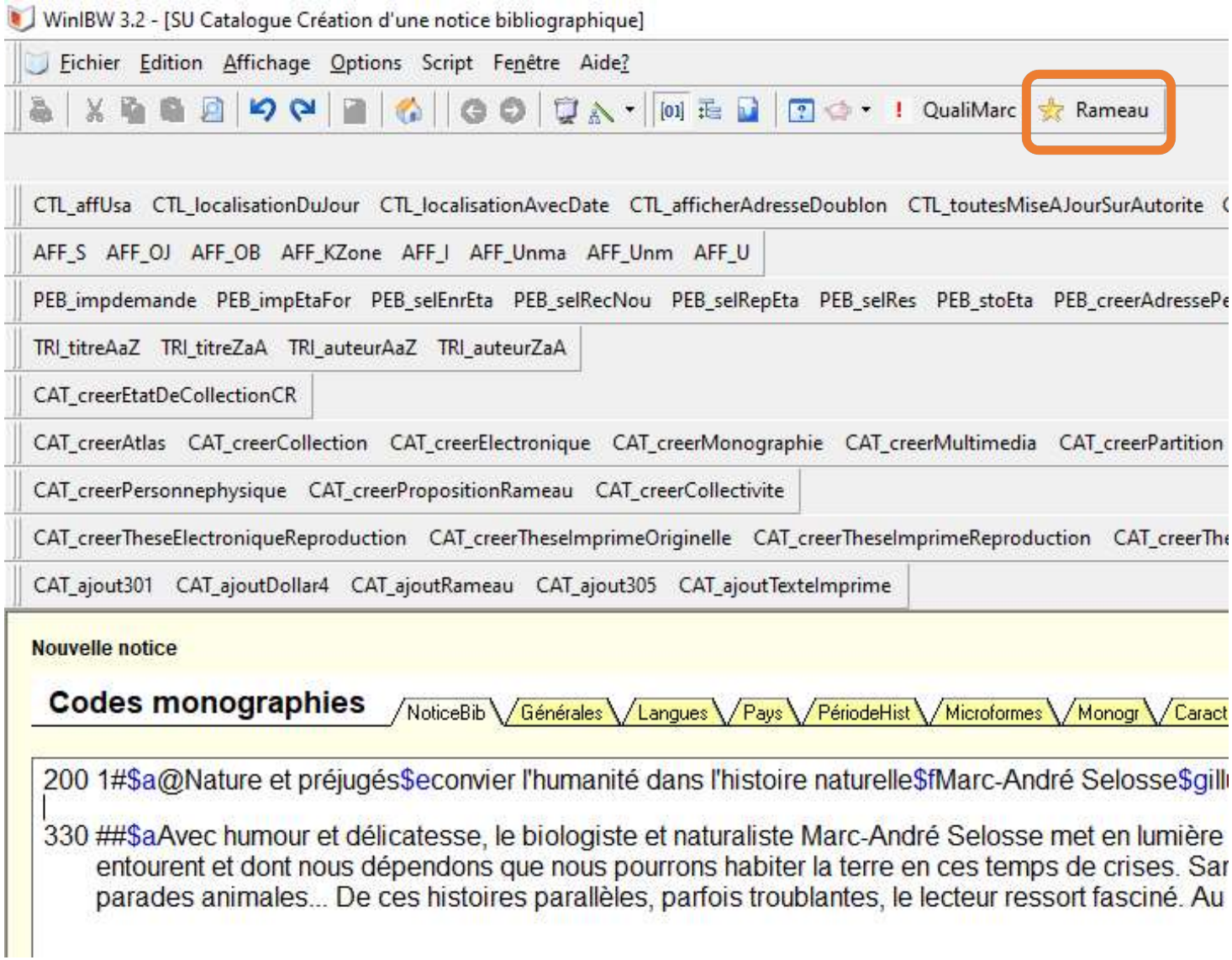
Maximiser la vérité

➔ Configuration acceptable pour une automatisation totale ?

Et maintenant ?
Test de Winnie+IA

- **Expérimentation dans l'outil de catalogage WinIBW (2024 T4)**

- Pourquoi expérimenter et non passer en production ?
  - Nos évaluations en chambre sont-elles confirmées en situation réelle ?
  - Quelle appropriation par les utilisateurs ?
  - Quel réglage du service (Paramétrage) ?
  - Quelle ergonomie de l'intégration ?



Nouveau Bouton

Nouvelle notice

Codes monographies NoticeBib | Générales | Langues | Pays | PériodeHist | Microformes | Monogr | CaractPhys | MatAnciens

```
200 1#$a@Nature et préjugés$econvier l'humanité dans l'histoire naturelle$fMarc-André Selosse$gillustrations d'Arnaud Rafaelian$g[Préface d' Erik Orsenna]
|
606##$3027311198$aÉcologie humaine$2rameau$9omikuji2
606##$3033366004$aÉvolution (biologie)$2rameau$9omikuji2
606##$3027271013$aHomme$2rameau$9omikuji2
606##$3027314197$aVirus$2rameau$9omikuji2
606##$3027568490$aChangements climatiques$2rameau$9omikuji2
606##$3027255646$aVie (biologie)$2rameau$9omikuji2
606##$324359934X$aCovid-19$2rameau$9omikuji2
606##$3027308464$aPhilosophie de la nature$2rameau$9omikuji2
606##$3145355039$aDéfenseurs de l'environnement$2rameau$9omikuji2
606##$3027225534$aÉvolution sociale$2rameau$9omikuji2

606##$3028718445$aRelations homme-animal$2rameau$9embedding_concept
606##$3027257002$aNature$2rameau$9embedding_concept
606##$302780397X$aHumanité$2rameau$9embedding_concept
606##$3027256421$aZoologie$2rameau$9embedding_concept
606##$3027235084$aInstinct$2rameau$9embedding_concept
606##$3027308464$aPhilosophie de la nature$2rameau$9embedding_concept
606##$3027243753$aComportement animal$2rameau$9embedding_concept
606##$3034791280$aRelations homme-plante$2rameau$9embedding_concept
606##$3027243745$aAnimaux$2rameau$9embedding_concept
606##$3027219380$aPhilosophie de l'homme$2rameau$9embedding_concept

606##$3027308464$aPhilosophie de la nature$2rameau$9Intersection2Models1Best
```

Quelles propositions Rameau injecter dans l'éditeur ?

Quelle expérience utilisateur pour le catalogueur ?

Expérimentation IA Rameau dans WinIBW

Dernier trimestre 2024

Quelles modalités d'expérimentation ?

- Ouvrir à tous ?
- Catalogage courant ? Chantiers ciblés ?
- Quel paramétrage initial du service ?
- Quels modes d'évaluation ?

Qu'en pensez-vous ?

Quel regard rétrospectif
sur ce premier service IA en test ?

Quels enseignements et anticipations
pour la suite ?

Aide à la décision ou Automatisation ?

Pour des opérations que les humains font déjà aujourd'hui, déployer une IA comme aide à la décision

IA assistant

Pour des opérations que les humains ne font pas aujourd'hui (soit que ces opérations soient déjà faites par des machines, soit qu'elles ne soient pas faites du tout), déployer une IA entièrement automatique

IA complément

Des services pour quels outils ?
Des outils avec quels services ?

- Approche traditionnelle à l'Abes :
  - Proposer des outils prêts à l'emploi, avec services intégrés
  - Proposer des web services tout crus (API)

- Quid de la réinformatisation de l'Abes ?
  - Quels services IA intégrés dans le logiciel acheté ?
  - Quelle liberté pour intégrer d'autres services IA ?
  - Quels droits sur les données générées ?
  - Etc.

Que changent les grands modèles de langue (LLM) comme GPT, Mistral, etc. ?

Modèles ad hoc

- Entraînés sur nos données
- Pour une tâche précise

Ex : indexation Rameau

LLM

- Entraînés sur … tout
- Pour diverses tâches
- Instructions en langage naturel

Structurer une zone de note avec un LLM, en le guidant

Structure ce texte : "Numérisation. Paris : Observatoire de Paris, 2019" en appliquant le modèle suivant : 325 ##$b{type de la reproduction}$c{Lieu de la reproduction}$d{Organisme responsable de la reproduction}$e{date de la reproduction}$u{URL de la reproduction}

Llama 3 70b

Voici la structure du texte :

325 ##$bNumérisation$cParis$dObservatoire de Paris$e2019$u[non spécifié]

Note: Le champ $u est souvent utilisé pour fournir l'URL de la reproduction, mais comme elle n'est pas spécifiée dans le texte original, je l'ai laissé vide.

Mixtral 8x22b

325 ##$bNumérisation$cParis$dObservatoire de Paris$e2019

Structurer la mention de responsabilité (zone 200 en UNIMARC)

Je voudrais que tu extraies les couples (nom, code de fonction bibliographique) à partir de citations bibliographiques.

Par exemple, dans "écrit par Jean Nicolas", extraie le couple (« Jean Nicolas, "070") et structure le ainsi en UNIMARC :

```
700 #1$aNicolas$bJean$4070
```

Par exemple, fais la même chose pour : "illustré par Christophe Lowe". Et aussi pour : "par John Bill".

Ne réponds qu'avec l'information en UNIMARC selon l'exemple ci-dessus..

Pour faire ça, appuie-toi sur cette liste de fonctions. Chaque fonction a un code et un libellé.

"003" = "Encadrant académique"

****

"005" = "Acteur"

La fonction "Acteur" concerne ce type de document : Enregistrements sonores

Images animées.

Table des codes de fonctions UNIMARC

Que changent les grands modèles de langue (LLM) comme GPT, Mistral, etc. ?

Programmation en langage naturel pour générer des métadonnées structurées

- En guidant le LLM (exemples)
- En lui fournissant des sources d'information prioritaires
- En le réentraînant à la marge (*finetuning*)

Quelle coopération avec ~~les réseaux~~ vous ?
Quelle puissance d'agir (*agency*) pour chacun ?

- IA subie ?
- IA immersive (si explicite, fluide et fiable) ?
- IA personnalisable ?
- IA coconçue ?
- IA coproduite ?
- IA produite par chacun ? (Prompts)

Microfonctionnalités ou *killer app* ?

Des questions ?
Des suggestions ?

Maintenant ou plus tard : slab@abes.fr