BULAC

[도서관] [शिक्षक] [ဘာသာစကား] [ሥልጣኔ]

Bibliothèque universitaire
des langues et civilisations

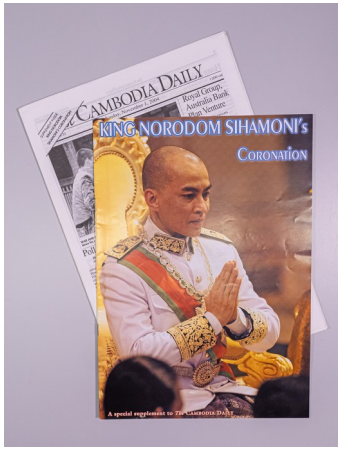
EFEO

École française d'Extrême-Orient

Projet de numérisation du fonds de périodiques khmers de l'EFEO déposé à la BULAC

Visite de S.M. le roi Norodom Sihamoni

05 décembre 2025



GIP BULAC
65 rue des Grands Moulins
F-75013 Paris

T + 33 1 81 69 18 00
contact@bulac.fr
www.bulac.fr

Sommaire

1. De la collecte à la conservation
2. Une collection patrimoniale
3. Le projet de numérisation des périodiques khmers
4. Un thésaurus exceptionnel pour répondre aux enjeux de l'Intelligence artificielle en langue khmère

BULAC
[도서관] [ቤተ-መጽሐፍት] [ဘာသာစကား] [ᓄᓇᕗᑦ]

1. De la collecte à la conservation

- Collecte systématique depuis 2005 par l'EFEO des périodiques publiés au Cambodge (quotidiens, hebdomadaires)
- 4 opérations de transport organisées : 2009, 2011, 2017 et 2024
- Rassemblement de collections privées couvrant les périodes antérieures





1. De la collecte à la conservation

- Partenariat entre l'EFEO et la BULAC pour l'acheminement et le traitement des fascicules (catalogage, conditionnement)
- Collections de l'EFEO en dépôt à la BULAC, où elles sont mises à disposition du public dans les salles de lecture





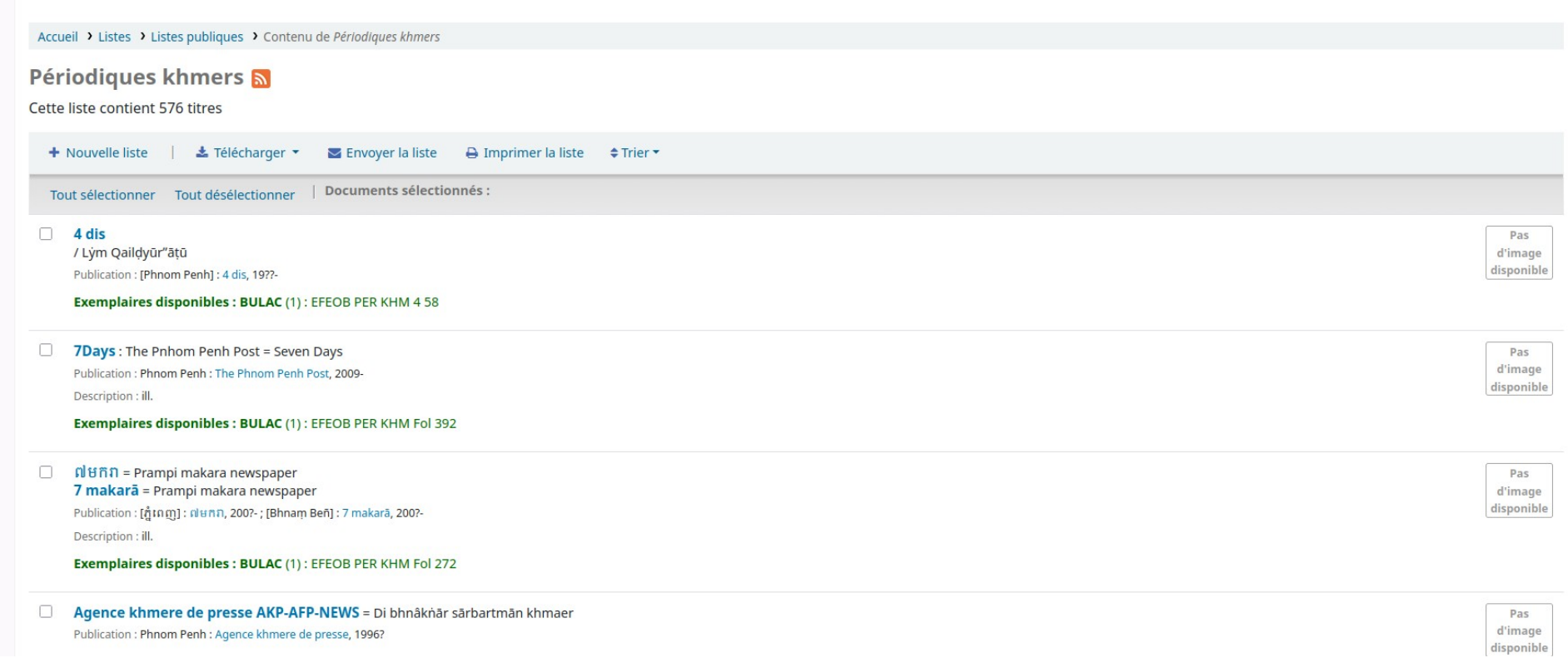
1. De la collecte à la conservation

- Une collection complète, riche de 576 titres, dont 10 toujours collectés
- Les titres sont visibles dans les catalogues nationaux (Koha, SUDOC) et internationaux (ISSN, WorldCat)
- Un catalogage multi-écritures



1. De la collecte à la conservation

- La collection est visible dans le catalogue de la BULAC, où elle est accessible à tous les publics.



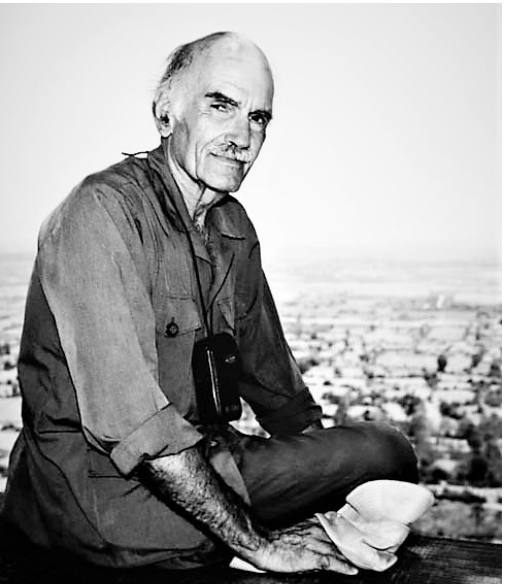
BULAC

Volumétrie de la collection

- 576 titres signalés au catalogue
- Dates extrêmes : de 1957 pour les plus anciens fascicules jusqu'à nos jours
- 499 mètres linéaires de collections conditionnées dans 2 145 boîtes de conservation
- Une volumétrie globale de la collection estimée à 2 850 000 pages

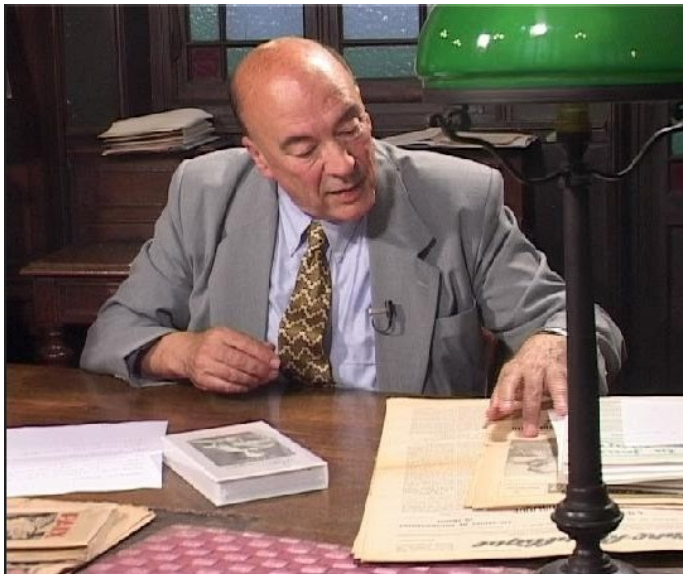


2. Une collection patrimoniale



Des segments rassemblés grâce à la collecte effectuée par Olivier de Bernon.

Cette collecte poursuit l'œuvre de deux chercheurs :



- Michael Vickery, qui fit don à l'EFEO de ses collections de périodiques cambodgiens de la période 1979-1989 et d'un important fonds de publications khmères des années 1960 conservé dans les locaux de l'EFEO à Phnom Penh.
- Gérard Brissé, qui légua à l'EFEO sa collection inestimable de tous les périodiques francophones parus au Cambodge entre 1963 et 1971.

2. Une collection patrimoniale

- Une collection d'intérêt pour l'histoire moderne du Cambodge
  - Le Sangkum (1965-1970)
  - Kambuja (1965-1969)
- Une collection d'intérêt également pour la conservation de la littérature khmère
  - អាថ៌កំបាំងបញ្ញាពិភពគុន | ʻʻarth kampāṃng pañjā bibhabgun
  - អ្នកនិពន្ធខ្មែរ | Qnak nibandh khmaer (1995-1997)
- Une collection pour enrichir les ressources en ligne disponibles en langue khmère



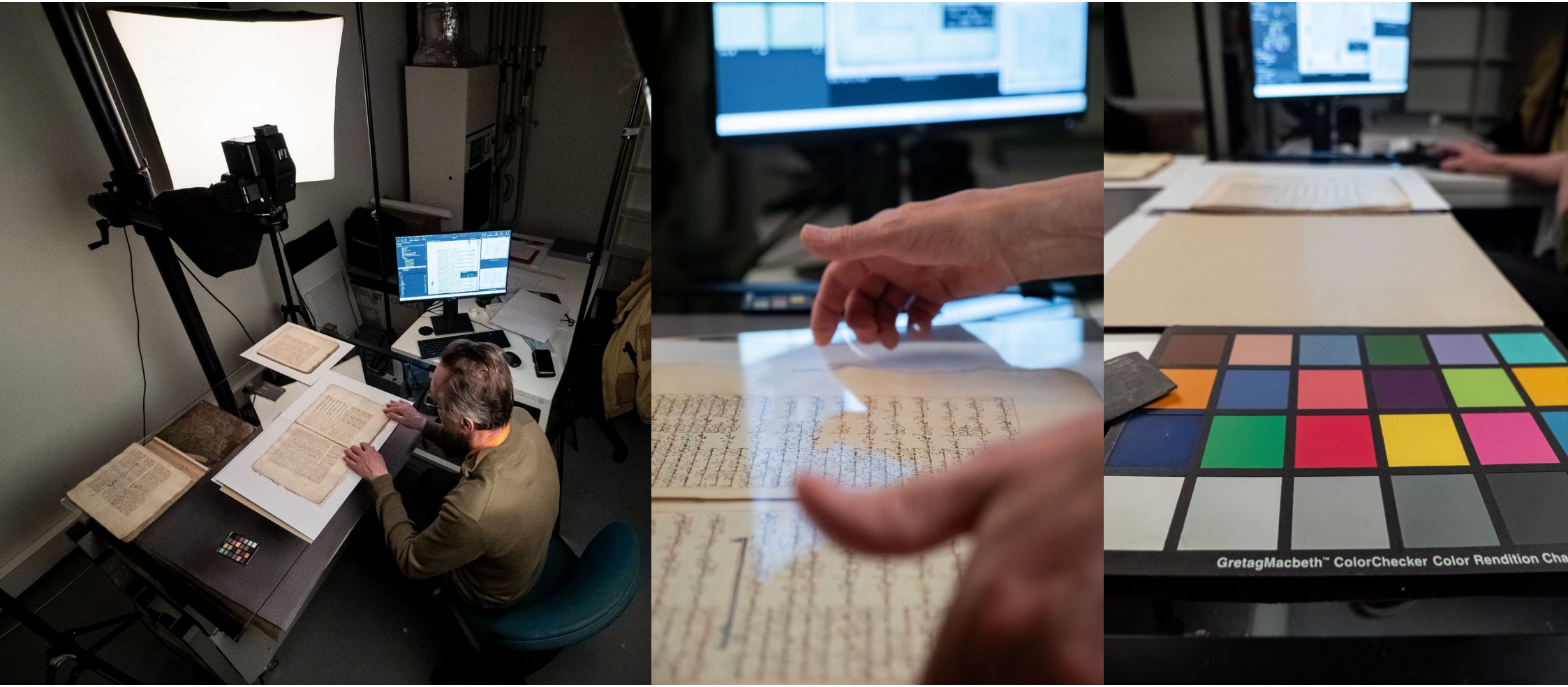


3. Le projet de numérisation des périodiques khmers

Domaine	Avancement
Compréhension du fond	Terminée
Vérification physique des périodiques en magasin	Environs 48%
Correction des titres / Translittération / Bi-écriture	Avancement important (environs 278 titres réalisés, soit 48% de la collection)
Enrichissement du tableur de description	en cours
Indexation RAMEAU	Reprise depuis le début
Etat de collection	Reprise depuis le début



3. Le projet de numérisation des périodiques khmers



Fonctionnement d'une station de numérisation *in situ* à la BULAC (Photos M. Ruscio/BULAC 2025)

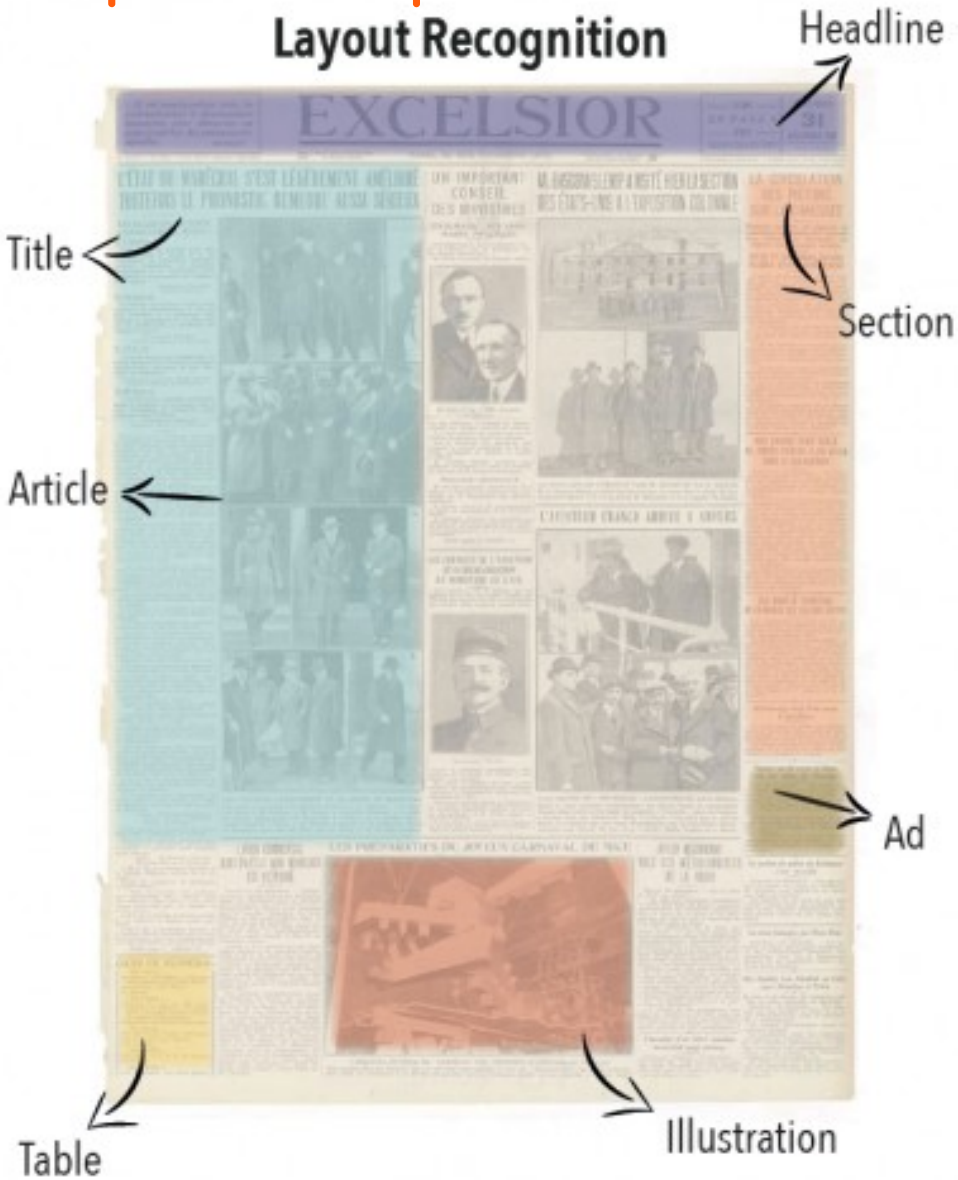
3. Le projet de numérisation des périodiques khmers

Les périodiques numérisés peuvent bénéficier de plusieurs traitements complémentaires :

- OCR : pour permettre la recherche dans le texte du document
- OLR : pour reconnaître la segmentation de la page de presse
- ENR : la reconnaissance des entités nommées

(noms de personnes, de lieux, d'organisations notamment) pour enrichir les recherches possibles

Ces traitements supposent l'utilisation d'outils dédiés et s'ajoutent au coût de la numérisation.



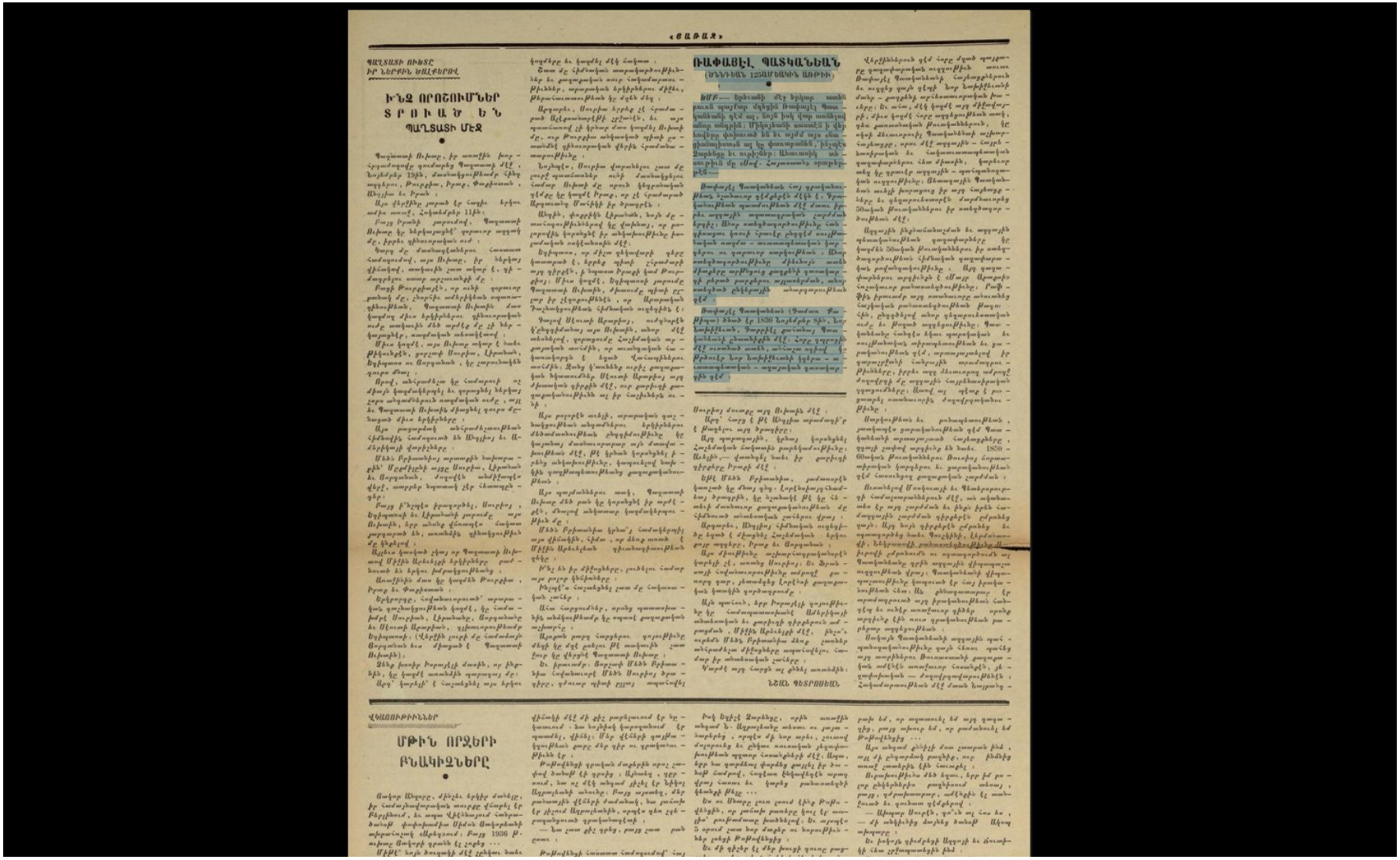
3. Le projet de numérisation des périodiques khmers

La presse numérisée est diffusée sur :

- la bibliothèque numérique de la BULAC (la BiNA)
- Internet Archive, plate-forme avec une visibilité à l'échelle internationale, avec plus de 100 000 visites par an et près de 320 000 pages vues par an en moyenne ces dernières années pour les documents que la BULAC diffuse par ce biais dans des formats d'export multiples (JPEG, PDF avec ou sans OCR, fichier texte)
- D'autres sites externes grâce au standard international IIIF de diffusion des images

Un exemple de valorisation avec le parcours thématique sur le quotidien arménien Haratch : accès au corpus accompagné de présentations et mise en contexte de l'histoire du journal

3. Le projet de numérisation des périodiques khmers



HARATCH (1925-2009) : UNE VOIX ARMÉNIENNE DANS LA DIASPORA PARISIENNE

Haratch

Collection

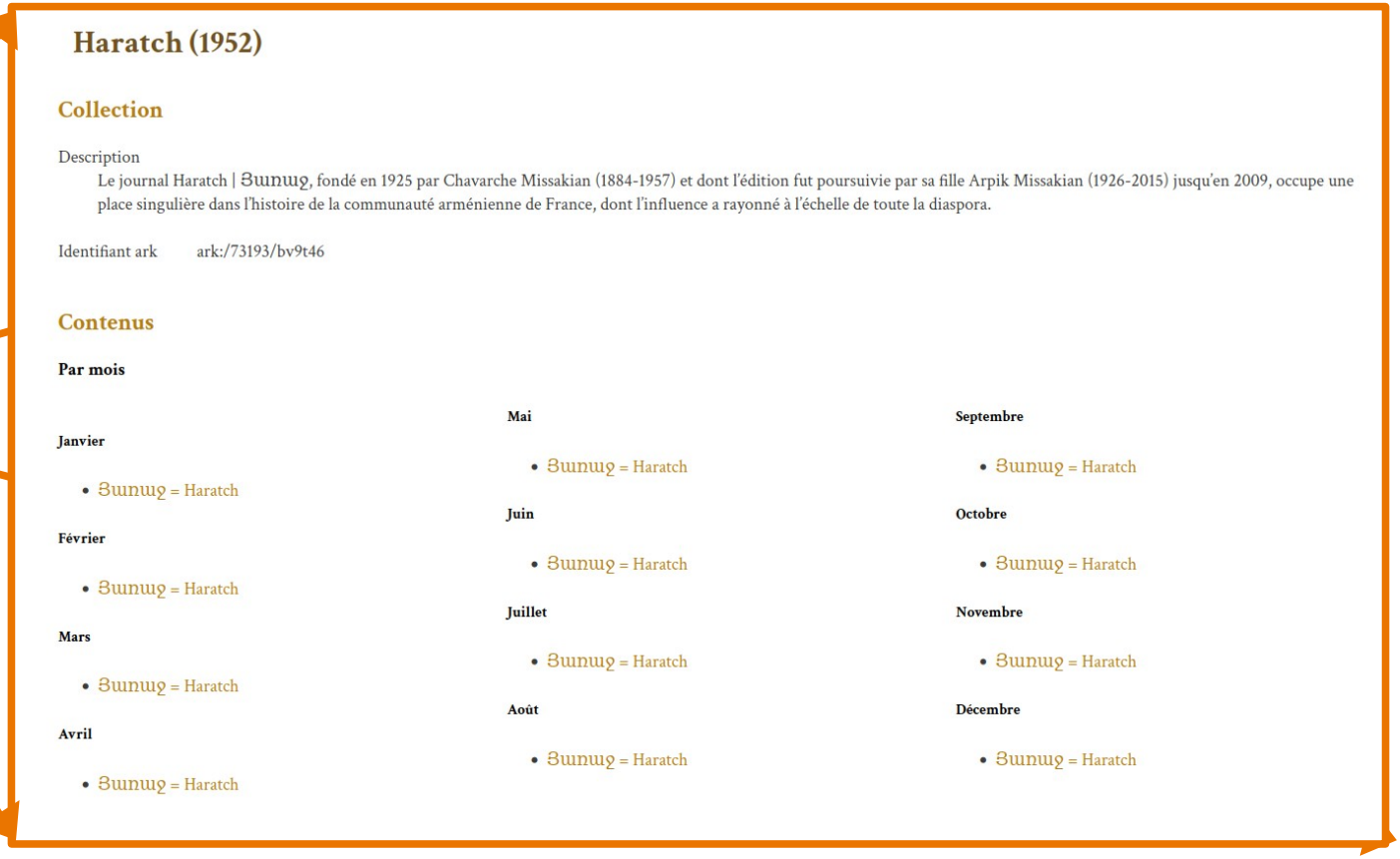
Description

Le journal Haratch | Յառաջ, fondé en 1925 par Chavarche Missakian (1884-1957) et dont l'édition fut poursuivie par sa fille Arpik Missakian (1926-2015) jusqu'en 2009, occupe une place singulière dans l'histoire de la communauté arménienne de France, dont l'influence a rayonné à l'échelle de toute la diaspora.

Identifiant ark ark:/73193/bv9t46

Contenus

- 1925
- 1926
- 1927
- 1928
- 1929
- 1930
- 1931
- 1932
- 1933
- 1934
- 1935
- 1936
- 1937
- 1938
- 1939
- 1940
- 1945
- 1946
- 1947
- 1948
- 1949
- 1950
- 1951
- 1952
- 1953
- 1954
- 1955
- 1956



Haratch (1952)

Collection

Description

Le journal Haratch | Յառաջ, fondé en 1925 par Chavarche Missakian (1884-1957) et dont l'édition fut poursuivie par sa fille Arpik Missakian (1926-2015) jusqu'en 2009, occupe une place singulière dans l'histoire de la communauté arménienne de France, dont l'influence a rayonné à l'échelle de toute la diaspora.

Identifiant ark ark:/73193/bv9t46

Contenus

Par mois

Janvier
- Յառաջ = Haratch

Février
- Յառաջ = Haratch

Mars
- Յառաջ = Haratch

Avril
- Յառաջ = Haratch

Mai
- Յառաջ = Haratch

Juin
- Յառաջ = Haratch

Juillet
- Յառաջ = Haratch

Août
- Յառաջ = Haratch

Septembre
- Յառաջ = Haratch

Octobre
- Յառաջ = Haratch

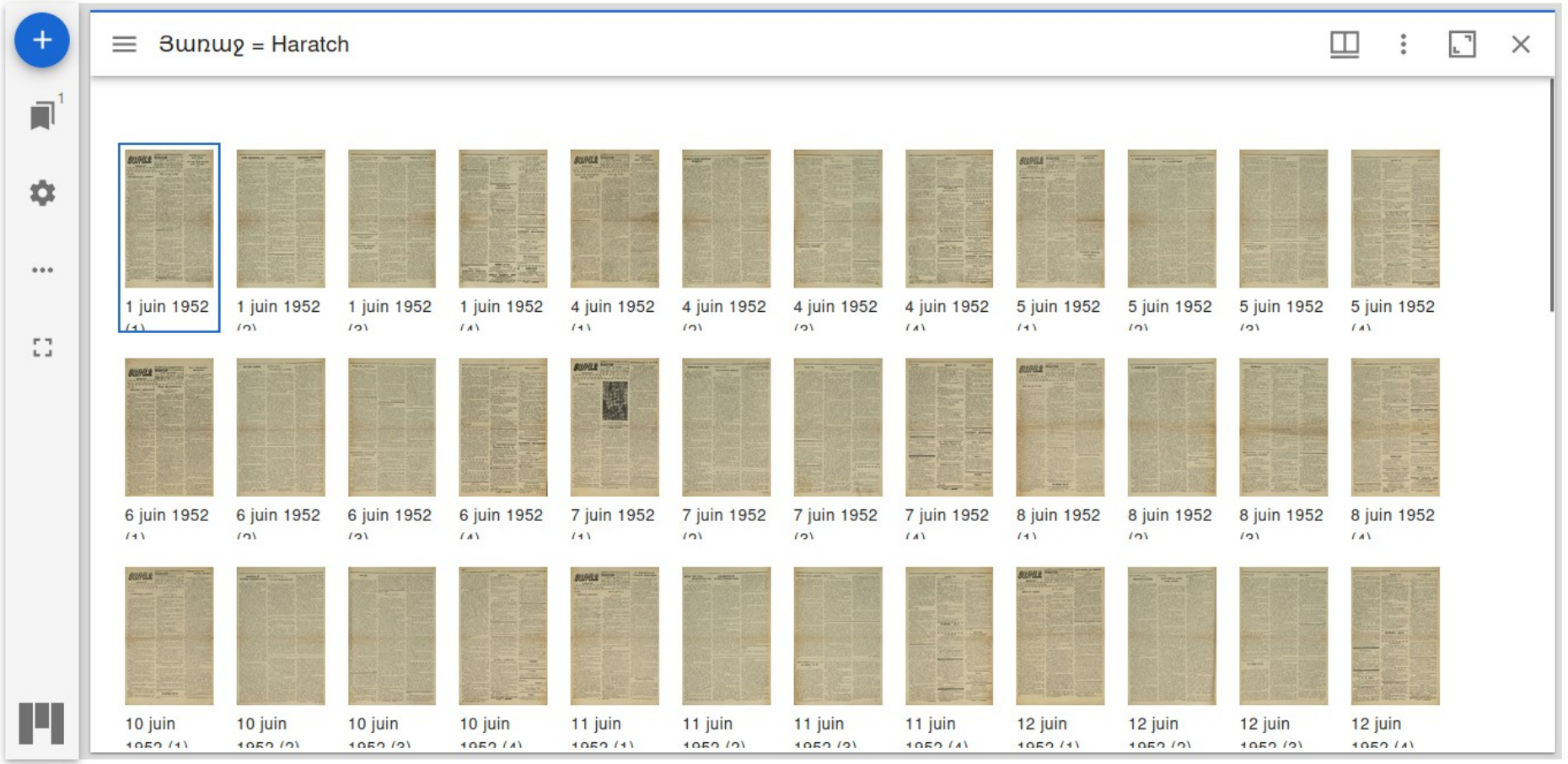
Novembre
- Յառաջ = Haratch

Décembre
- Յառաջ = Haratch

Un exemple de valorisation, le journal *Haratch*

Յառաջ = Haratch

Contenu



B U L A C
[도서관] [ቤተ-ፍቅር] [ספרייה] [ᓂᐲᑲᓇᐎᒃ]

Յառաջ = Haratch

Contenu



Haratch (1925-2009) : une voix arménienne dans la diaspora parisienne

« Avant tout, notre travail va consister à défendre toutes les choses sacrées qui donnent vie aux peuples et raniment leur foi pour de meilleurs jours. »

Chavarche Missakian, *Haratch*, numéro 1, 2 août 1925

Haratch (en arménien Յառաջ, littéralement « En Avant ») est un quotidien en langue arménienne, publié à Paris de 1925 à 2009. Il est reconnu comme le premier quotidien arménien en Europe.

Le journal fut fondé le 2 août 1925 par Chavarche Missakian (1884–1957) qui en resta le rédacteur en chef jusqu'à sa mort en janvier 1957. Sa fille, Arpik Missakian, lui succéda de 1957 jusqu'à la publication du dernier numéro, le 30 mai 2009. La parution fut suspendue le 9 juin 1940, en raison de l'occupation nazie de la France. Le 8 avril 1945, il devient le premier journal en langue étrangère à obtenir l'autorisation de paraître à la Libération.

Dans un premier temps trihebdomadaire, *Haratch* devient quotidien à partir du 1er janvier 1927. On dénombre 22 214 numéros du journal.

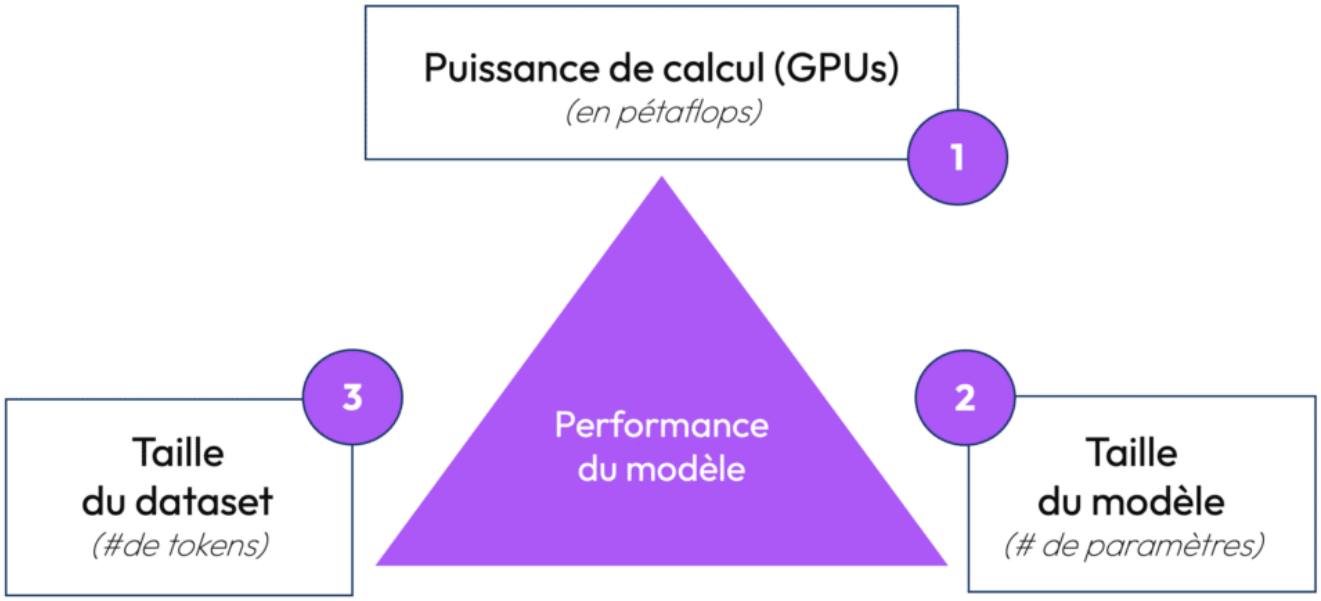
Haratch a été un lieu d'expression littéraire et journalistique de référence, attirant des plumes telles que Shahan Shahnour, Chavarche Nartouni, Zareh Vorpouni, Hrant-Samuel, Nichan Béchiktachlian, Jiraïr Missakian, Zoulal Kazandjian, Nigoghos Sarafian, Vazken Chouchanian, Khachig Tölölyan, Marc Nichanian, Haroutioun Kurkjian, Krikor Chahinian, Harout Gosdantian, Bedros Terzian ou encore Yves Ternon.



4. Un thésaurus exceptionnel pour répondre aux enjeux de l'Intelligence artificielle en langue khmère

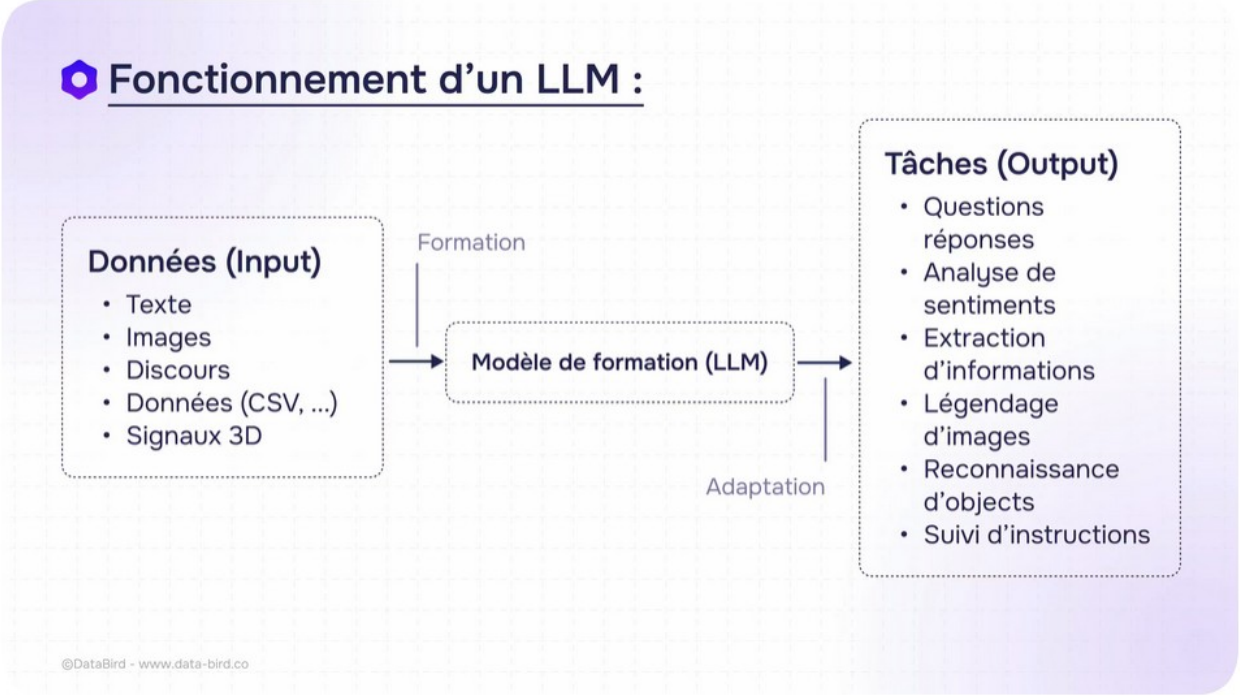
Les modèles de langue (LLM) ont besoin d'une grande masse de données de qualité pour être entraînés.

Le volume de données nécessaire pour entraîner un *Large Language Model* (LLM) de qualité dépend de plusieurs facteurs, notamment la taille du modèle, sa complexité, et les performances visées.



4. Un thésaurus exceptionnel pour répondre aux enjeux de l'Intelligence artificielle en langue khmère

Le corpus numérique public en khmer est estimé de l'ordre de quelques Giga octets à quelques dizaines de Giga octets de texte brut, contre des Téra octets pour d'autres langues plus dotées comme l'anglais, le français ou le chinois.



4. Un thésaurus exceptionnel pour répondre aux enjeux de l'Intelligence artificielle en langue khmère

Les articles de presse sont une source riche en mots différents d'une même langue pour constituer un corpus d'entraînement d'un tel outil d'IA.

La collection de l'EFEO déposée à la BULAC avec plus de 2 850 000 pages constitue un réservoir important de caractères et d'entités nommées dans la langue khmère et couvre une large période chronologique et une grande diversité de niveaux de langue.

Une fois numérisée, cela permet d'envisager sa réutilisation dans la perspective de créer un modèle de langue pour la langue khmère.

Merci à Votre Majesté et à vos Excellences pour leur attention

