ISIDORE 2030 : adapter les IA aux besoins de la recherche de documents et de données en SHS

Stéphane Pouyllau, ingénieur de recherche au CNRS
co-fondateur d'Huma-Num.
Responsable du HN Lab et en charge d'ISIDORE

Huma-Num IR* est une infrastructure de recherche dédiée aux pratiques numériques de la recherche en sciences humaines et sociales. Elle fédère des communautés scientifiques nationales et internationales et développe avec elles des services et outils numériques, ainsi que des méthodes pour les projets de recherche et leurs données.

Huma-Num IR* est une infrastructure de recherche « étoile », du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, mise en œuvre par le CNRS avec le Campus Condorcet et Aix-Marseille Université. Elle est, avec son entrepôt de données NAKALA l'un des Centres de référence de l'écosystème national Recherche Data Gouv pour la science ouverte. Engagée dans la construction de l'European Open Science Cloud, elle porte la participation de la France dans l'European Research Infrastructure Consortium DARIAH.

Poser une question sur l'IR* Huma-Num ?

Demander l'ouverture de services ?

Consulter la documentation et les ressources ?

Rendez-vous sur Huma-Num.fr

Huma-Num IR* – UAR 3598

Campus Condorcet
Bâtiment de recherche Nord
14, cours des humanités
93322 Aubervilliers cedex
Contact@huma-num.fr

Huma-Num IR*

Infrastructure de recherche et services numériques pour les communautés SHS

Cycle de vie des données

Huma-Num IR* met à disposition un ensemble de **services et outils numériques** pour les projets de recherche à toutes les étapes du cycle de vie des données : stockage sécurisé, hébergement, calcul, traitements, travail collaboratif, écriture scientifique… et accompagne ses utilisateurs en vue d'améliorer la **qualité des données.**

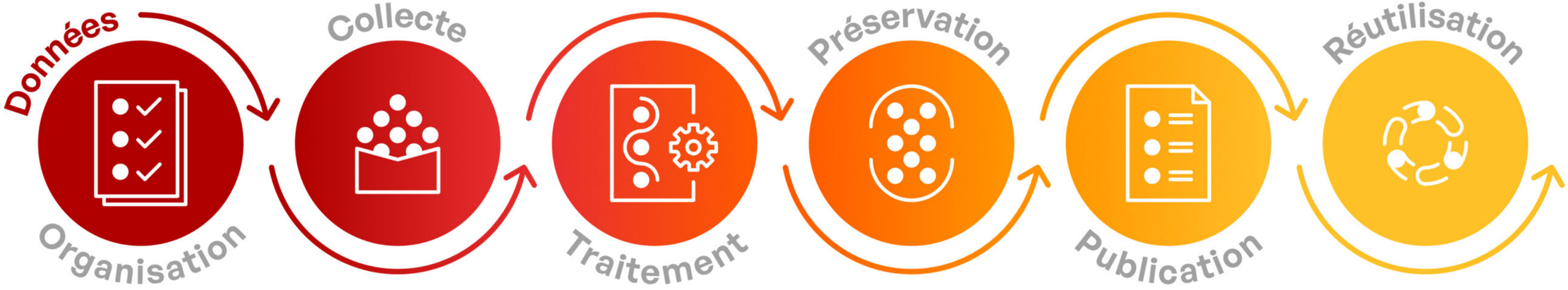
Ces services et outils sont construits sur un ensemble de **technologies d'infrastructure** et de systèmes informatiques sécurisés.

L'entrepôt de données **NAKALA** et l'assistant de recherche **ISIDORE** permettent d'assurer la pérennisation, la diffusion et le signalement des données et productions scientifiques selon les pratiques académiques des sciences humaines et sociales.

Ces services et outils s'articulent avec un dispositif de concertation collective animé par les **communautés de recherche, nationales et internationales, au sein des Consortiums-HN**, du **HN Lab** pôle de **recherche et d'innovation** et de différents partenaires, dont les Maisons des Sciences de l'Homme.

Écosystème IR* Huma-Num

Des services pour les données en sciences humaines et sociales



Des services pour organiser le travail collaboratif autour de vos données.

– ShareDocs
– GitLab
– Kanboard
– Mattermost

Des services de stockage sécurisé pour la collecte et la création de vos données.

– ShareDocs
– Huma-Num Box

Des services et outils spécifiques pour le traitement et l'analyse de vos données.

– Calcul statistique et environnements R
– Logiciels d'enquête et d'analyse de données
– Reconnaissance de caractères
– Puissance de calcul (+ CC-IN2P3)

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

– Nakala
– Huma-Num Box
– Préservation à long terme (+ CINES)

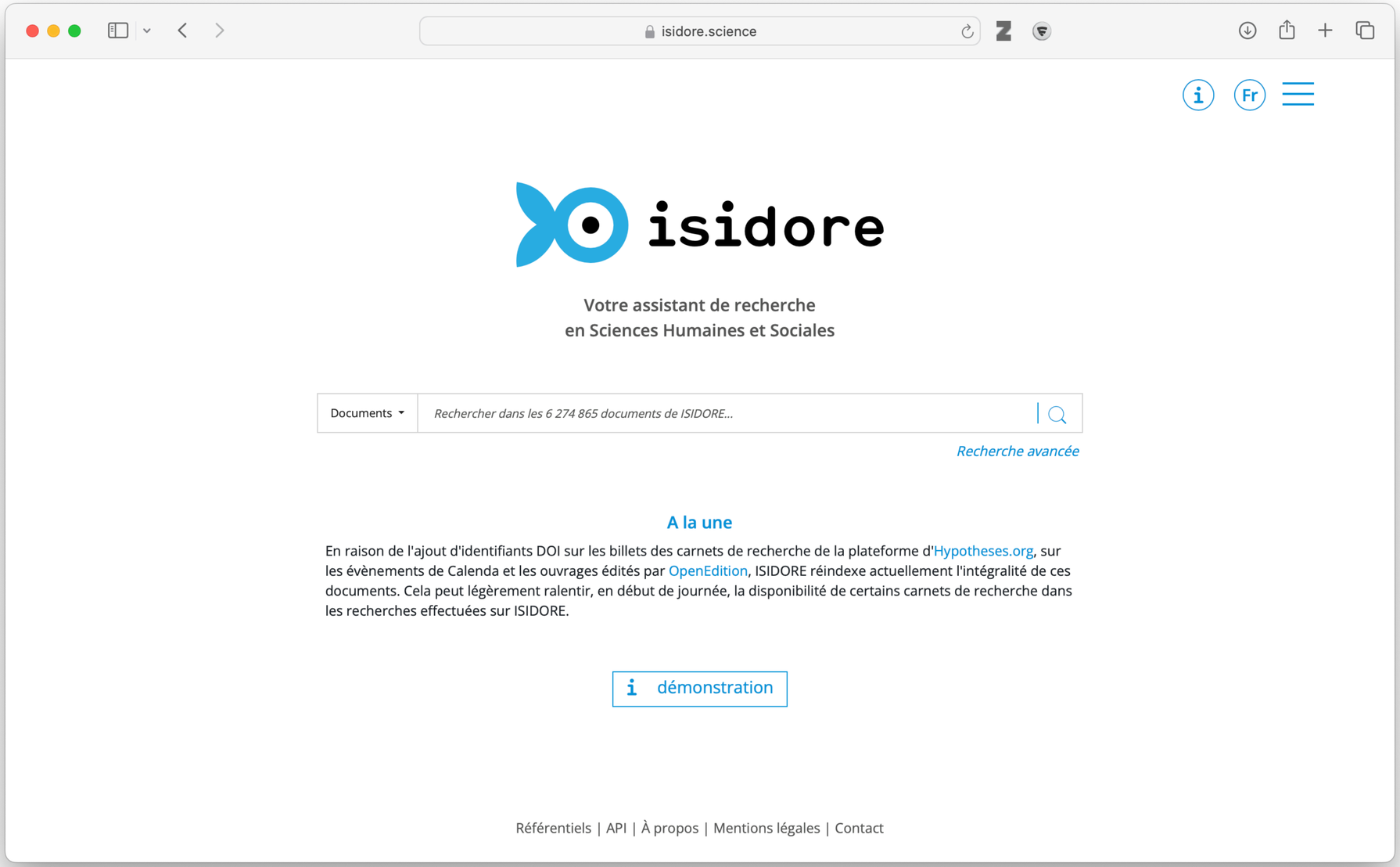
Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

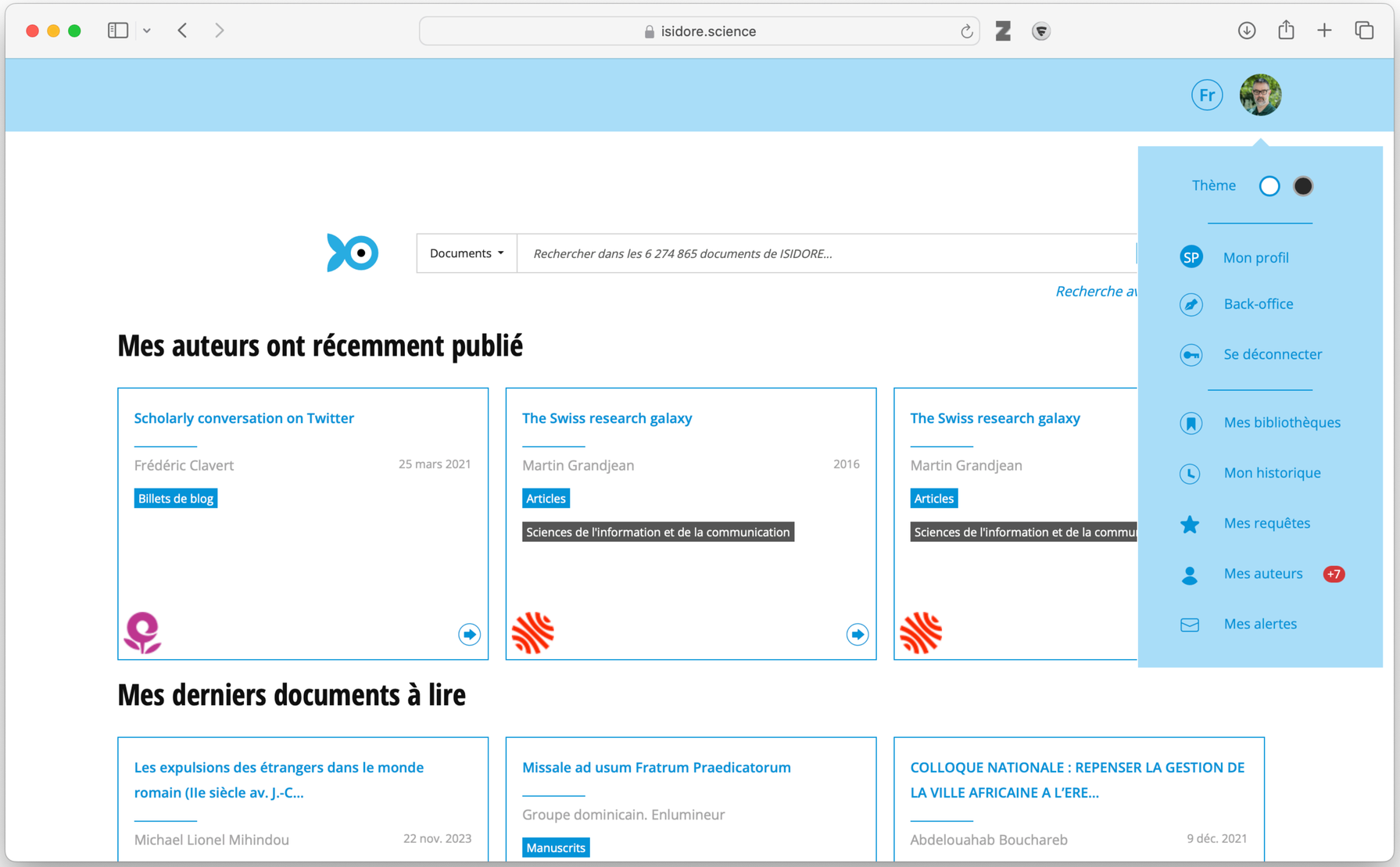
– Hébergement Web
– Machines Virtuelles
– Nakala
– Isidore

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

– Portail web
– API
– Triplestore
– OAI-PMH

Huma-Num IR*


isidore.science
Fr
isidore
Votre assistant de recherche
en Sciences Humaines et Sociales
Documents
Rechercher dans les 6 274 865 documents de ISIDORE...
Recherche avancée
A la une
En raison de l'ajout d'identifiants DOI sur les billets des carnets de recherche de la plateforme d'Hypotheses.org, sur les évènements de Calenda et les ouvrages édités par OpenEdition, ISIDORE réindexe actuellement l'intégralité de ces documents. Cela peut légèrement ralentir, en début de journée, la disponibilité de certains carnets de recherche dans les recherches effectuées sur ISIDORE.
démonstration
Référentiels | API | À propos | Mentions légales | Contact


isidore.science
Fr
Thème
Documents
Rechercher dans les 6 274 865 documents de ISIDORE...
Recherche av
Mes auteurs ont récemment publié
Scholarly conversation on Twitter
Frédéric Clavert
25 mars 2021
Billets de blog
The Swiss research galaxy
Martin Grandjean
2016
Articles
Sciences de l'information et de la communication
The Swiss research galaxy
Martin Grandjean
Articles
Sciences de l'information et de la commu
Mes derniers documents à lire
Les expulsions des étrangers dans le monde romain (IIe siècle av. J.-C...
Michael Lionel Mihindou
22 nov. 2023
Missale ad usum Fratrum Praedicatorum
Groupe dominicain. Enlumineur
Manuscrits
COLLOQUE NATIONALE : REPENSER LA GESTION DE LA VILLE AFRICAINE A L'ERE...
Abdelouahab Bouchareb
9 déc. 2021
SP
Mon profil
Back-office
Se déconnecter
Mes bibliothèques
Mon historique
Mes requêtes
Mes auteurs
+7
Mes alertes

isidore.science

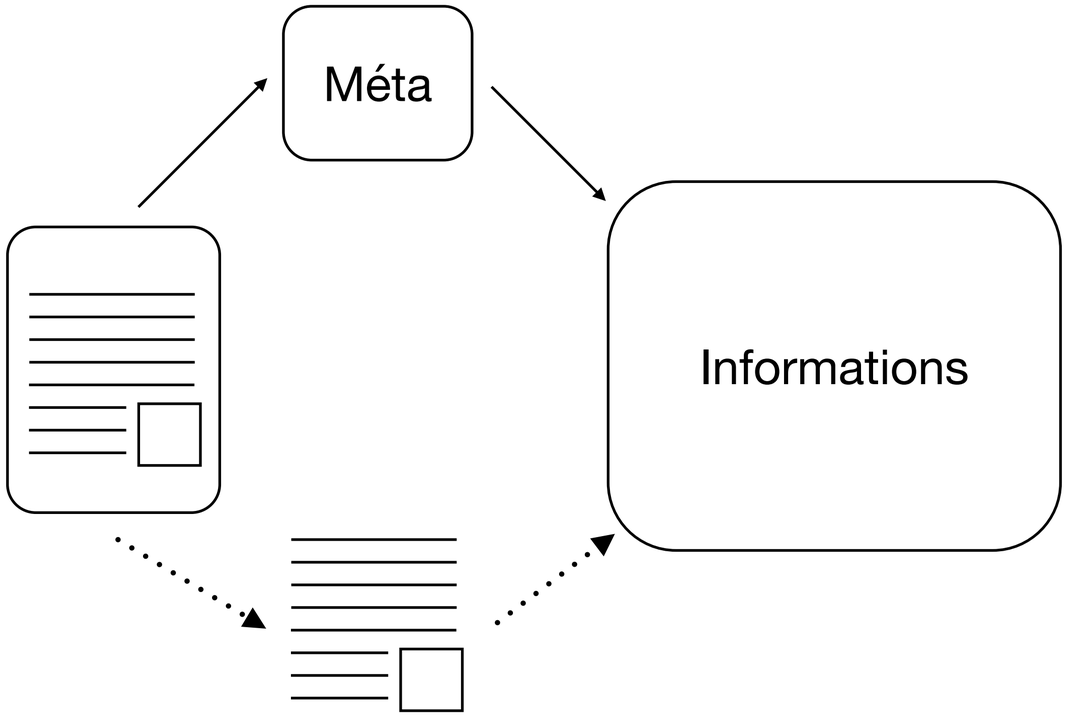
En production depuis le 8 déc. 2010

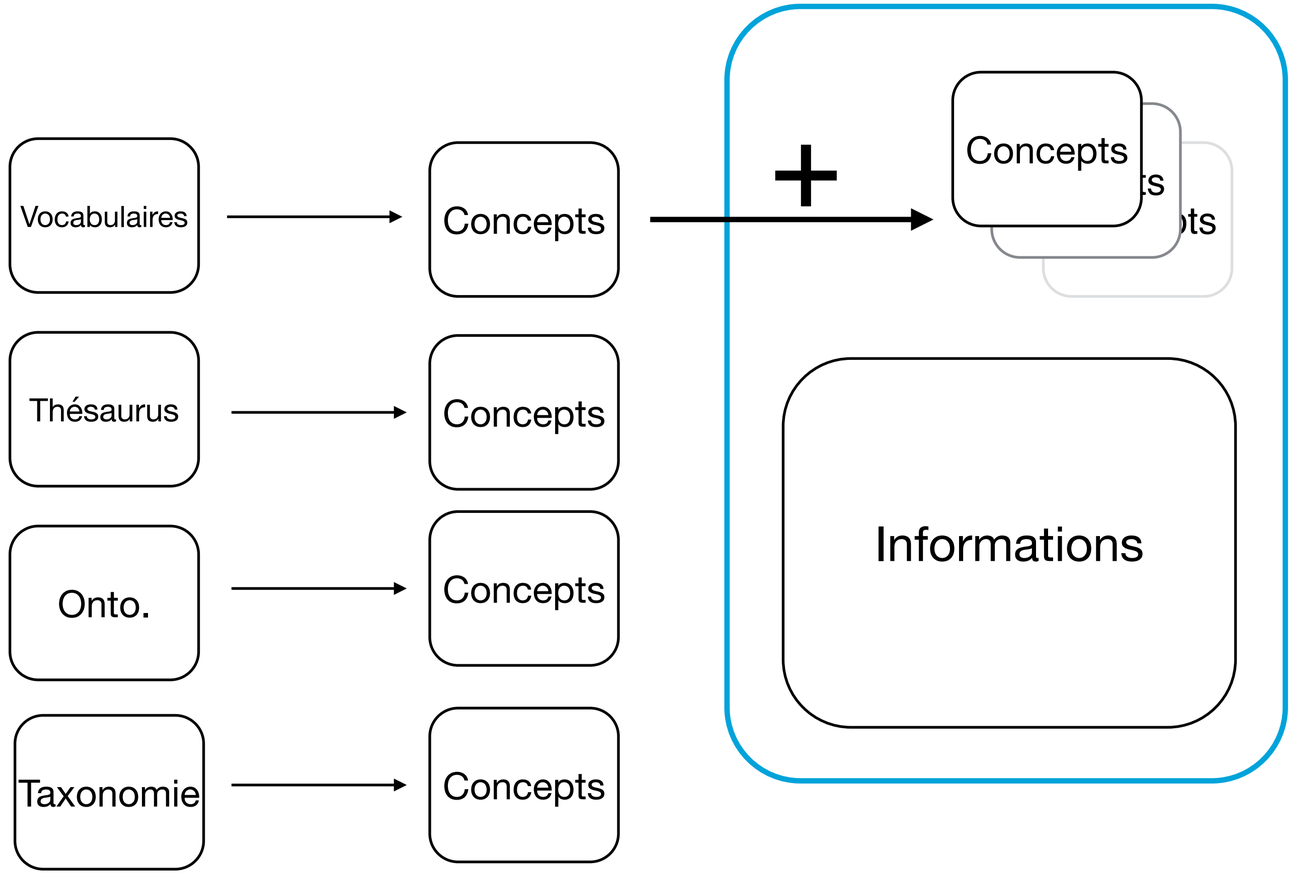
6 millions de documents*
et jeux de données de sciences humaines et sociales (SHS)
signalés, enrichis, « reliés » entre eux

* articles, jeux de données, actualités, thèses, mémoires, fonds d’archives, audiovisuel, etc.

Venant de 10000 « collections » du monde entier

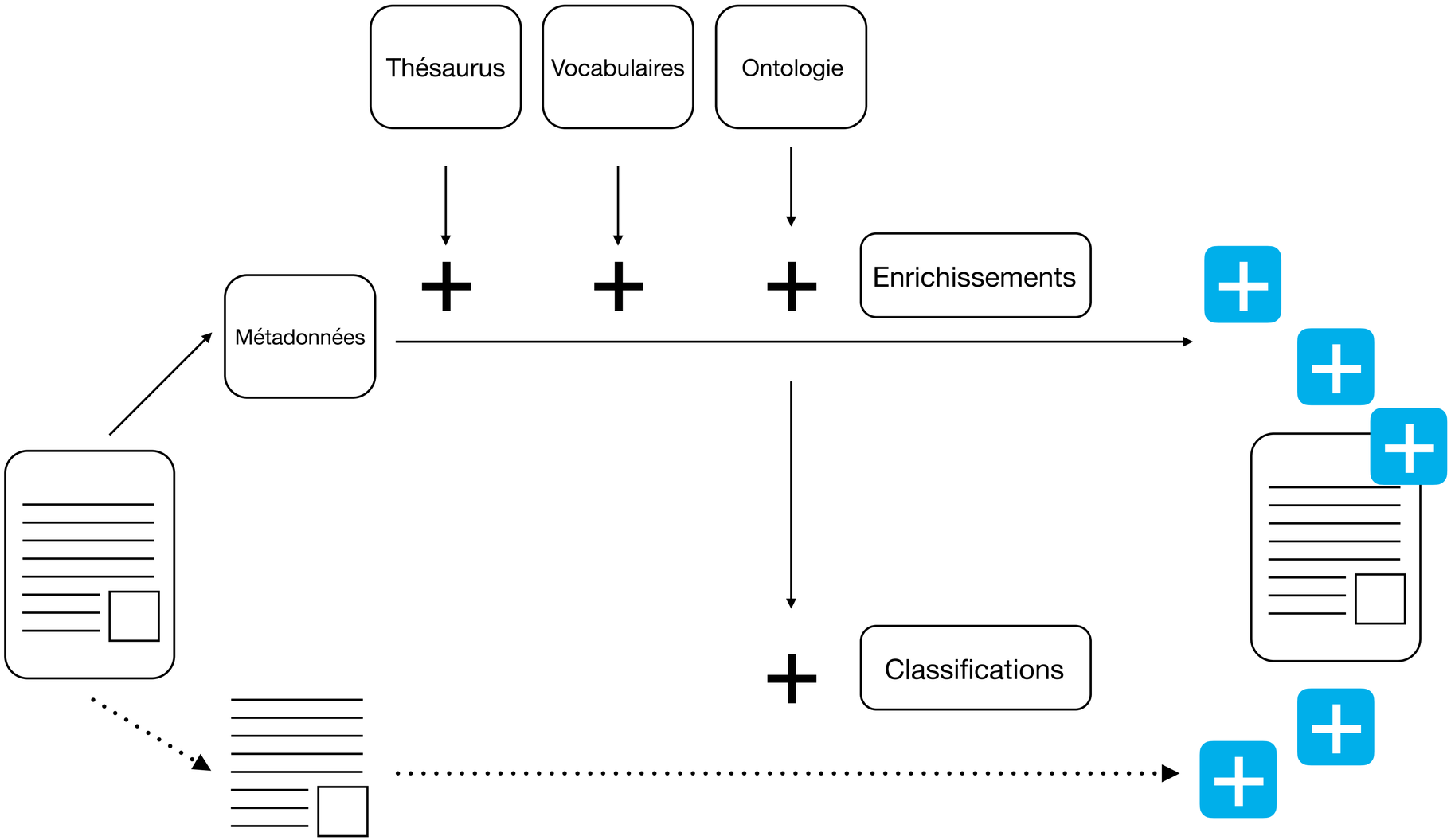
Moyenne de 400.000 utilisateurs / an
Moyenne de 75.000.000 requêtes / an


Méta
Informations



Concepts : *subjects*

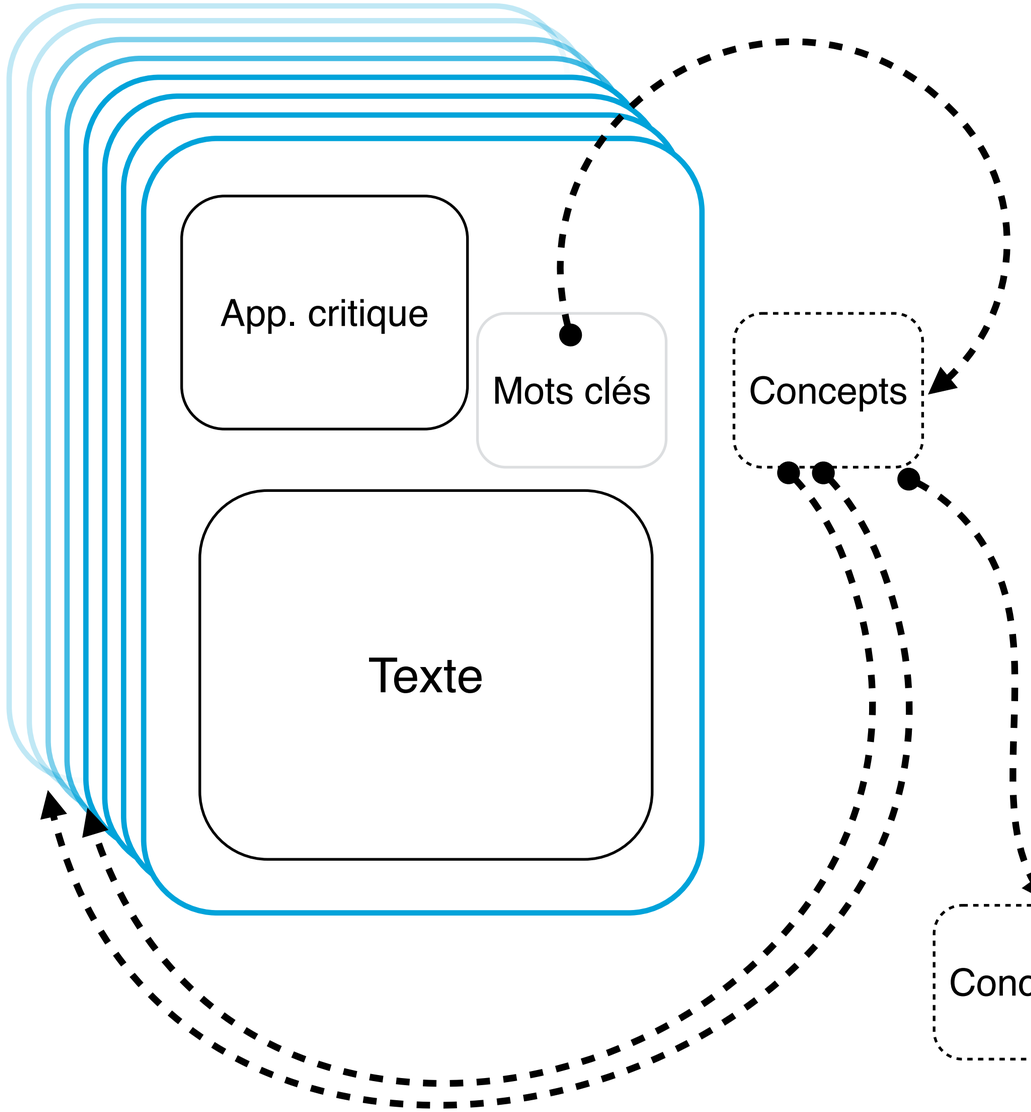
Concept : une URI + des labels en plusieurs langues


Thésaurus
Vocabulaires
Ontologie
Enrichissements
Métadonnées
Classifications

Relier des parties d'informations (et/ou des documents) avec des concepts scientifiques, donc d'autres documents et par ex. des auteurs entre eux en utilisant :

- la création d'enrichissements « sémantiques » à partir de concepts (*subjects*) issus des publications/données SHS

- La catégorisation des documents dans 27 disciplines SHS pour pondérer les résultats du moteur de recherche



Semantic Web

Word Wide Web Consortium

WIKIDATA

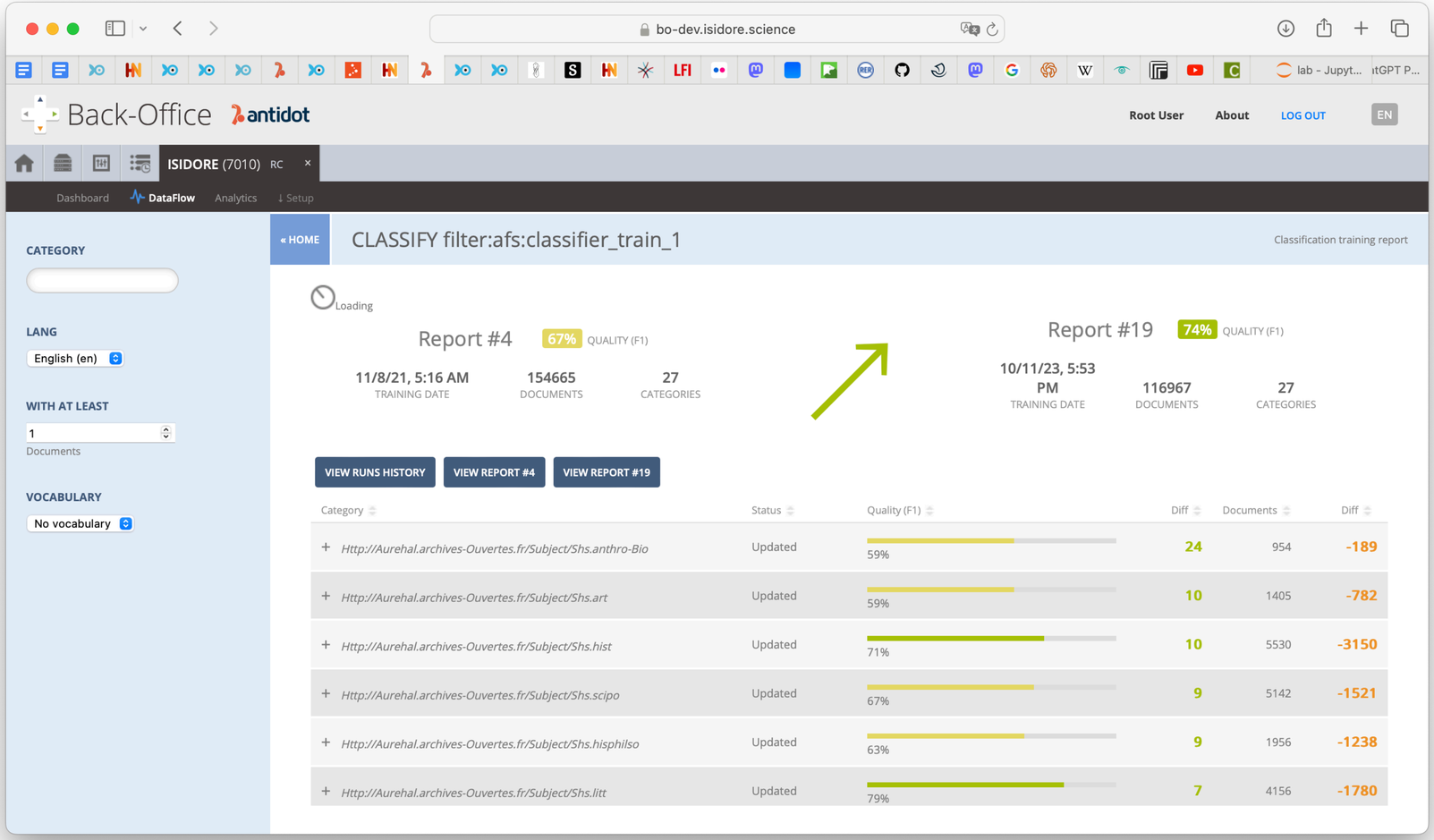
ArchiRès

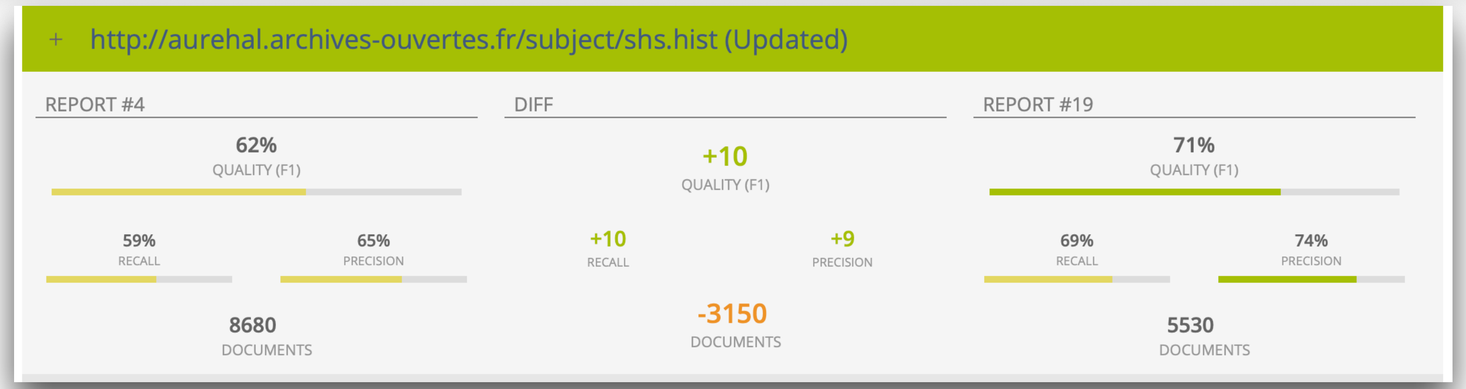
{BnF data.bnf.fr

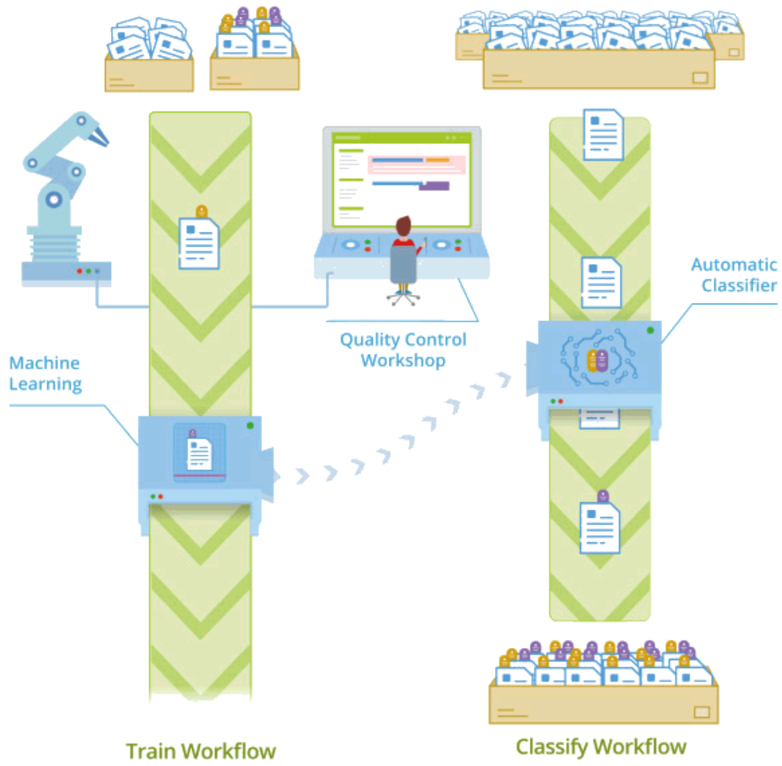
ORCID
Connecting Research and Researchers

Enrichissements : algorithme fondé sur une analyse morphologique des termes (regex + *machine learning* supervisé)

Classification
utilisant un moteur d’entrainement
(corpus de référence + analyse de
proximité + *machine learning* supervisé)


bo-dev.isidore.science
Back-Office
antidot
Root User
About
LOG OUT
EN
ISIDORE (7010) RC
Dashboard
DataFlow
Analytics
Setup
CATEGORY
LANG
English (en)
WITH AT LEAST
1
Documents
VOCABULARY
No vocabulary
« HOME
CLASSIFY filter:afs:classifier_train_1
Classification training report
Loading
Report #4
67% QUALITY (F1)
11/8/21, 5:16 AM
TRAINING DATE
154665
DOCUMENTS
27
CATEGORIES
Report #19
74% QUALITY (F1)
10/11/23, 5:53 PM
TRAINING DATE
116967
DOCUMENTS
27
CATEGORIES
VIEW RUNS HISTORY
VIEW REPORT #4
VIEW REPORT #19
Category
Status
Quality (F1)
Diff
Documents
Diff
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.anthro-Bio Updated 59% 24 954 -189
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.art Updated 59% 10 1405 -782
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.hist Updated 71% 10 5530 -3150
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.scipo Updated 67% 9 5142 -1521
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.hisphilso Updated 63% 9 1956 -1238
+ Http://Aurehal.archives-Ouvertes.fr/Subject/Shs.litt Updated 79% 7 4156 -1780


+ http://aurehal.archives-ouvertes.fr/subject/shs.hist (Updated)
REPORT #4
62%
QUALITY (F1)
59%
RECALL
65%
PRECISION
8680
DOCUMENTS
DIFF
+10
QUALITY (F1)
+10
RECALL
+9
PRECISION
-3150
DOCUMENTS
REPORT #19
71%
QUALITY (F1)
69%
RECALL
74%
PRECISION
5530
DOCUMENTS


Machine Learning
Quality Control Workshop
Automatic Classifier
Train Workflow
Classify Workflow

Retour d’expérience après 13 ans de fonctionnement

- Dispositif assez rigide : uniquement 27 disciplines, peu agile pour accueillir de nouvelles communautés émergeantes
- Dispositif assez ancien (*machine learning* supervisé)
- Dispositif éprouvé mais peu agile en terme d’évolution technologique (adhérence entre savoir-faire d’innovation et d’exploitation)
- Adhérence complexe à maintenir entre les besoins des chercheur·e·s dans l’évolution de la recherche SHS et les savoir-faire documentaires, en IST, en informatique (développement, exploitation)


2030

Définition du programme

- Repenser les chaines de traitement (impliquant la refonte des IA déjà présentes dans ISIDORE)

- Mettre en œuvre les nouvelles fonctionnalités en incluant une réflexion sur l’apport et les limites des IA dans le traitement des données en SHS

- Développer de nouvelles interfaces pour expliquer le fonctionnement mais surtout les limites des IA que nous utilisons

- Proposer un ISIDORE sans doute plus « modulaire » : respectant mieux les besoins des communautés multiples des SHS

”We” ecosystem : a proposal for the futur of search engine and discovery tools

Stéphane POUYLLAU Jean-Luc MINEL
Nicolas SAURET

“We” ecosystem : a proposal for the futur of search engine and discovery tools

Abstract

This keynote is a proposal to define a new ecosystem and plateform, called “We”, that places the practices of researchers in SSH at the heart of search engine and discovery tools features. The keynote will first situate the “We” proposal in the ecosystem of and for academic tools. Secondly, it will propose a conceptual representation for an innovative platform based on the discovery of experts, topics and trends for scientific information using classification, linking and enrichment of data tools. It will finally describe its notions (expert, Topic box) and functionalities such as integrated tools (Python notebooks) and technological advances (Deep Learning). We want to point out that this proposal provides a big picture, which ideas are yet to be discussed and enriched. In particular, the different diagrams do not pretend to present an exact design of the HMI but instead aim at illustrating its forseen functionalities. As a proposal, “We” might be implemented for any academic search engines, such as ISIDORE as a primary proof of concept, then eventually as a seamless service. This presentation will rely on recent research works devloped by HN-Lab team of IR* Huma-Num through projects like “Revue 2.0”, “Huma-num Open Science” and some proofs of concept like Callisto, ISIDORE Jupyter Notebooks, ISIDORE connectors for Zotero.

A classifier using ISIDORE, the social and humanities search engine and Keras API for Deep Learning

Stéphane Pouyllau (0000-0002-9619-1002)[1]

[1]CNRS Research Engineer, Huma-Num, Paris.

August 21, 2020

Abstract

This document presents the creation of a text classifier to predict the disciplinary affiliation of titles (from books or articles) in the humanities and social sciences. It implements ISIDORE and Keras API for Deep Learning.

1 Introduction

The purpose of this document is to present the implementation of a title classifier using the data from ISIDORE[1] and Keras API[2] for Deep Learning. It describes the construction of the classifier, its training and use to predict whether titles (from books or articles) can be identified as history documents.

The aim here is not to make an exhaustive and detailed example using all the finesse of Keras, but to demonstrate the potential and relative ease of Deep Learning in the social sciences and humanities (SSH) and to construct a simple example in the form of a demonstration for SSH audiences[3]. It is largely based and implemented the blog post *Practical Text Classification With Python and Keras*[1] and *How to Develop a Deep Convolutional Neural Network for Sentiment Analysis*[5]. We use the Jupyter notebooks to include code in the document and all data sets are available and executable directly with Binder. A non-exhaustive bibliography is available at the end of the document and will potentially guide the reader.

2 Implementation: from choosing training set to learning

2.1 Prepare data

We will use document titles, from ISIDORE itself, to train a neural network classifier using the possibilities of Keras API. This API allows you to create, train and use neural networks[2, p. 65]. The classification of text in neural networks requires vectorizing the titles to be able to process them afterwards. In our example we will use a simple representation of the data called "Bag-Of-Words" (BOW) using *CountVectorizer* from scikit-learn[3][1][4].

First of all, to train the classifier to recognize "history" data, we constitute a training set using the SPARQL query below[5]. ISIDORE categorizes more than 7 million documents every

[1]See https://isidore.science
[2]See https://keras.io
[3]I would like to thank Jean-Luc Minel, Professor Emeritus at the University of Paris Nanterre, for his advice in creating this classifier.
[4]See https://scikit-learn.org/stable
[5]ISIDORE proposes an API and an endpoint SPARQL: https://isidore.science/sparql and https://isidore.science/sqe

- POUYLLAU S., MINEL J-L, SAURET N. (2023, January 14). “We” ecosystem : a proposal for the futur of search engine and discovery tools. https://doi.org/10.5281/zenodo.7536441

- POUYLLAU, Stéphane. (2020). A classifier using ISIDORE, the social and humanities search engine and Keras API for Deep Learning (Version V1). Zenodo. https://doi.org/10.5281/zenodo.3994126

10 ans d'ISIDORE

Pouyllau Stéphane, Minel Jean-Luc, Capelli Laurent, Bunel Mélanie, Sauret Nicolas, Capelli Laurent, Busonera Pauline, Desseigne Adrien, Baude Olivier, Jouguet Hélène

2021/10/19

Mot-clés : isidore, moteur de recherche, outil de découverte, réseau social académique, topic modelling, latent dirichlet analysis, réseaux de neurones
Keywords: search engine, discovery tool, academic social network, isidore, topic modelling, latent dirichlet analysis, neural networks

ISIDORE a 10 ans !

En 2011 paraissait la 1ère version d'ISIDORE. Son 10ème anniversaire est l'occasion de rappeller l'historique du projet et de présenter ses futures grandes lignes en cours de définition dans le cadre du programme "Huma-Num Science Ouverte" [1].

Cette année ISIDORE a franchi les 10 millions de documents et de données indexés, enrichis et catégorisés venant de plus de 9000 bases et entrepôts de données du monde entier. Déjà plus de 2000 chercheur · e · s y possèdent un compte, profitant ainsi des nombreuses fonctionnalités de veille scientifique et de découverte personnalisables.

Au fait, qu'est-ce qu'ISIDORE ?

Initialement, ISIDORE est un moteur de recherche permettant de découvrir et de trouver des publications, des données numériques et profils de chercheur · e · s en sciences humaines et sociales (SHS) venant du monde entier.

Il permet de rechercher dans le texte intégral de plusieurs millions de documents (articles, thèses et mémoires, rapports, jeux de données, pages Web, notices

1. Programme financé par le fond national pour la science ouverte du ministère de la recherche, de l'enseignement-supérieur et de l'innovation.

« Refonder » ISIDORE : feuille de route pour redécouvrir les informations et les données des SHS

POUYLLAU Stéphane

Note liminaire

Document issue du travail collectif pour les 10 ans d'ISIDORE : POUYLLAU, Stéphane & al. (2021). ISIDORE a 10 ans. Zenodo. https ://doi.org/10.5281/zenodo.5699997

Cette note a bénéficié des échanges (entre mars 2023 et juin 2023) entre l'auteur et Fabrice Lacroix (PDG d'Antidot SA), Jean Delahousse (consultant en Web sémantique et graphe de données), Susan Brown (Guelph University), Marcello Vitalli-Rosati (Université de Montréal). Elle compile ainsi un nombre de réflexions, avis, lectures élaborées durant cette période.

La multiplication des moteurs de recherche et de découverte depuis le début des années 2000 a complexifié, le paysage de la recherche d'information pour les chercheurs. ISIDORE, par son inclusion en 2013 dans l'écosystème d'Huma-Num, profite cependant, autour de lui, de la mise en œuvre d'une intégration de services numériques (authentification centralisée, entrepôts de données créés avec NAKALA] et moissonnables, etc.). Mais si les interfaces Web ont fleuri, les fonctionnalités de recherche, qui s'appuient sur la qualité des données indexées, sur la capacité du dispositif à les analyser, à les regrouper, à les trier et à les relier, sont restées le plus souvent dans leurs états initiaux. La création de chaînes de traitement en *Machine Learning* et surtout leurs exploitations, utilisant comme ISIDORE un apprentissage entrainé par référentiels, nécessitent des cycles de vie rapides et souvent complexes sur le plan des choix et de la validation scientifique des référentiels d'enri-

- POUYLLAU, Stéphane, CAPELLI, Laurent, MINEL, Jean-Luc, BUNEL, Mélanie, SAURET, Nicolas, BAUDE, Olivier, JOUGUET, Hélène, BUSONERA, Pauline, & DESSEIGNE, Adrien. (2021). ISIDORE a 10 ans. Zenodo. https://doi.org/10.5281/zenodo.5699997

- POUYLLAU Stéphane. (2023). « Refonder » ISIDORE : feuille de route pour redécouvrir les informations et les données des SHS. Zenodo. https://doi.org/10.5281/zenodo.8089406

Pourquoi ?

Traiter et mieux enrichir des métadonnées et le texte intégral

Analyser du texte et des données complexes (données sérielles, etc.) dans le but de rapprocher des informations

Suivant les besoins des disciplines proposer des fonctionnalités d’écritures assistées

Répondre aux différents besoins

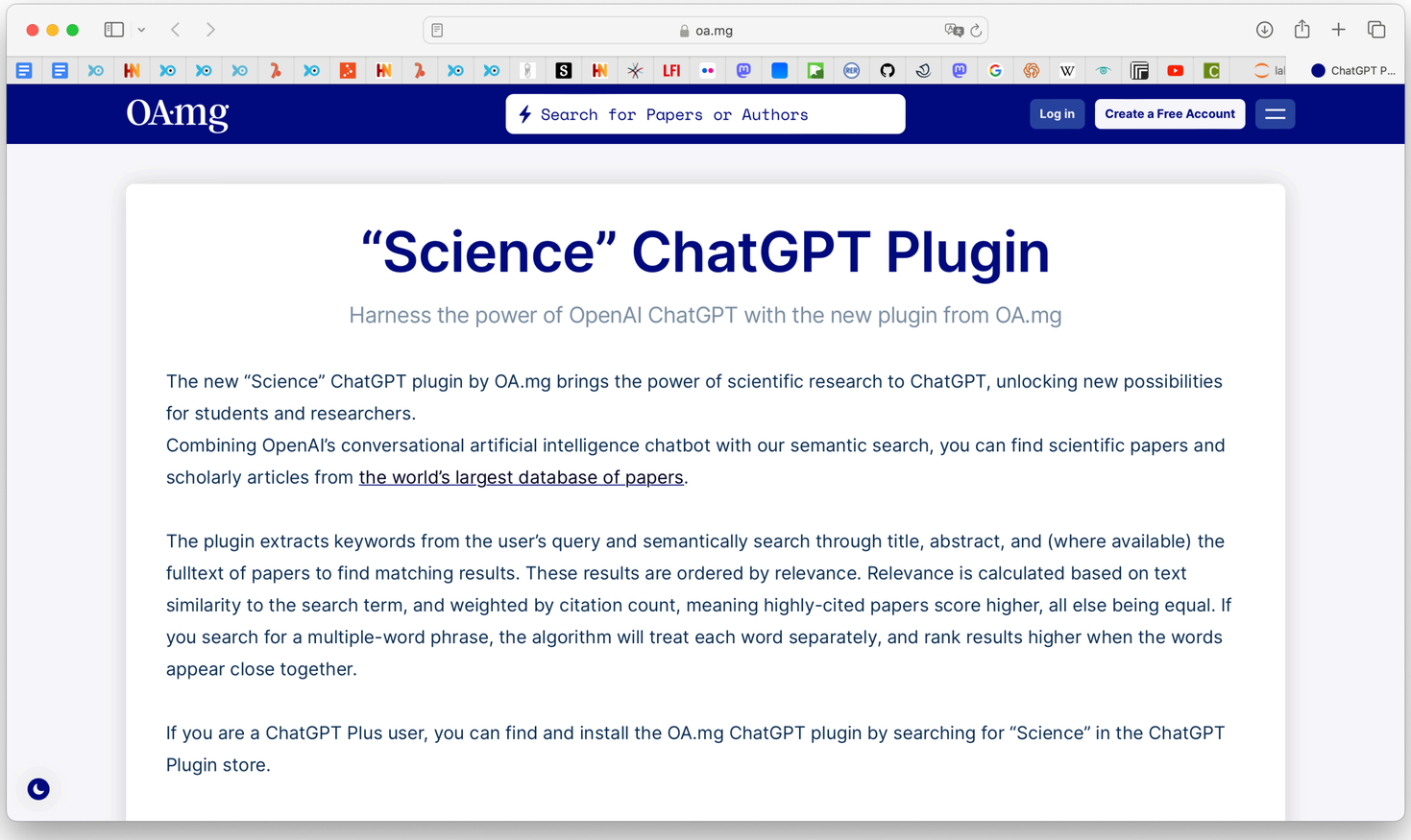
Communautés très différentes
dans leurs pratiques
de recherche et d'utilisation
des informations
(sources, document, code, …)

Documents et des informations
hétérogènes

Des pratiques de
« mises en données »
différentes

Des pratiques de diffusion
différentes

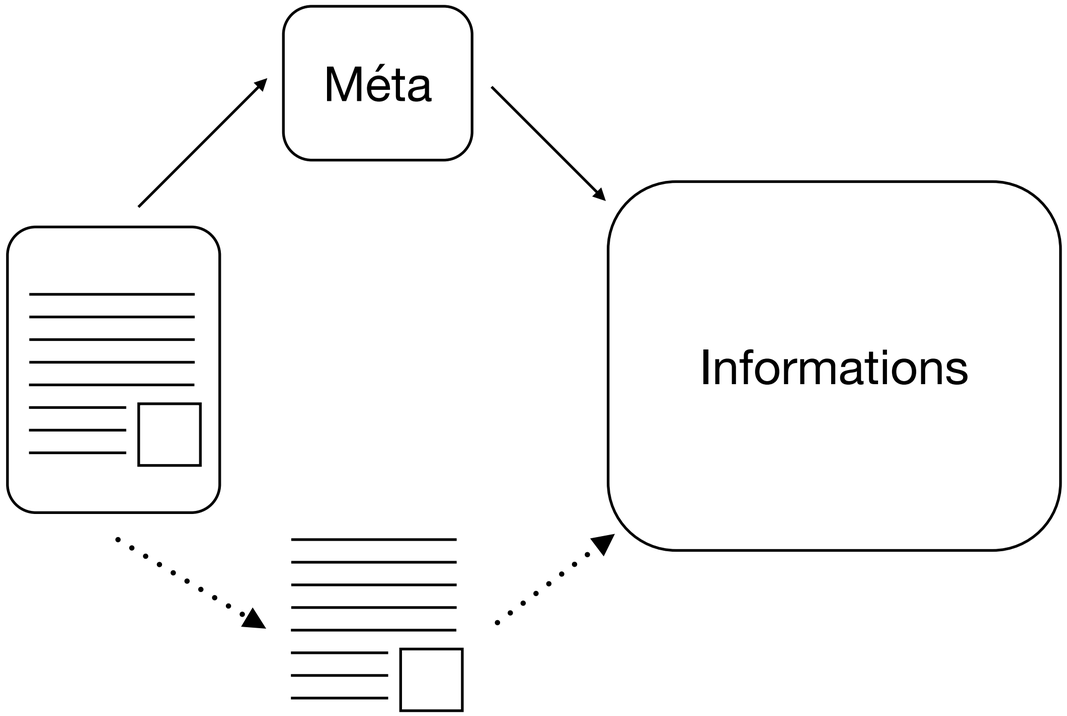
Ce que l’on ne veut pas faire



https://oa.mg/chatgpt

Ce que l’on veut faire

Re-questionner le principe initial d’ISIDORE



vulcain.cnrs.fr

File Edit View Run Kernel Git Tabs Settings Help

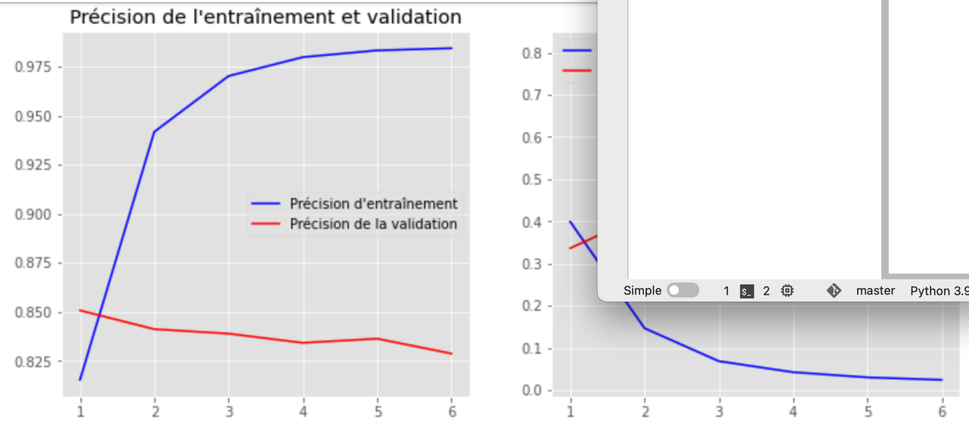
Filter files by name

/

Name	Last Modified
isidore-jupyter	8 hours ago
R	6 months ago

Classifieur Keras utilisant IS

Markdown git



Utilisation du classifieur pour prédire si des titres d'articles sont de l'histoire

On définit des titres d'articles à classifier et on les vectorise pour les classifier avec le classifieur :

[7]:
```
sentences_apredire = ['Stratégies éditoriales des musées. Une approche de la médiation par l'accès ouvert aux données nu
                      'Le monde karstique',
                      'Au plus près des âmes et des corps. Une histoire intime des catholiques au xixe siècle',
                      'Cuba: pour une géographie du socialisme',
                      'Migrations en Turquie',
                      'Les autoroutes et informations',
                      'Les seigneuries et baronies au Moyen-age',
                      'La guerre civile espagnole',
                      'De la Terre à la Lune',
                      'Hommes et structures du Moyen Âge',
                      'Guerriers et Paysans, VIIe – XIIe siècles : premier essor de l'économie européenne']
```

Simple 1 0 No Kernel | Initializing Mode: Command Ln 1, Col 1 Classifieur Keras utilisant ISIDORE.ipynb 1

vulcain.cnrs.fr

File Edit View Run Kernel Git Tabs Settings Help

Launcher Classifieur Keras utilisant IS

Markdown git Python 3.9 DL Keras

Filter files by name

/ public / ClassifieurTitres_ISIDORE /

Name	Last Modified
Classifieur Ker...	8 hours ago
Isidore-extrac...	8 hours ago
MotsISIDORE-...	8 hours ago
MotsISIDORE-...	8 hours ago
MotsISIDORE-...	8 hours ago
MotsISIDORE-...	8 hours ago
README.md	8 hours ago

[5]:
```
history = model.fit(X_train, y_train,
                    epochs=6,
                    verbose=1,
                    validation_data=(X_test, y_test),
                    batch_size=10)

loss, accuracy = model.evaluate(X_train, y_train, verbose=False)
print("Précision de l'entraînement: {:.4f}".format(accuracy))
loss, accuracy = model.evaluate(X_test, y_test, verbose=False)
print("Exactitude des tests: {:.4f}".format(accuracy))
```

```
Epoch 1/6
2022-02-11 08:56:03.258648: I tensorflow/compiler/mlir/mlir_graph_optimization_pass.cc:116] None of the MLIR optimizatio
n passes are enabled (registered 2)
2022-02-11 08:56:03.274063: I tensorflow/core/platform/profile_utils/cpu_utils.cc:112] CPU Frequency: 2399995000 Hz
1340/1340 [==============================] - 8s 5ms/step - loss: 0.4715 - accuracy: 0.7697 - val_loss: 0.3364 - val_accu
racy: 0.8508
Epoch 2/6
1340/1340 [==============================] - 6s 5ms/step - loss: 0.1413 - accuracy: 0.9451 - val_loss: 0.4162 - val_accu
racy: 0.8412
Epoch 3/6
1340/1340 [==============================] - 7s 5ms/step - loss: 0.0619 - accuracy: 0.9732 - val_loss: 0.5191 - val_accu
racy: 0.8389
Epoch 4/6
1340/1340 [==============================] - 6s 5ms/step - loss: 0.0372 - accuracy: 0.9831 - val_loss: 0.6352 - val_accu
racy: 0.8342
Epoch 5/6
1340/1340 [==============================] - 6s 5ms/step - loss: 0.0289 - accuracy: 0.9845 - val_loss: 0.7064 - val_accu
racy: 0.8364
Epoch 6/6
1340/1340 [==============================] - 6s 5ms/step - loss: 0.0213 - accuracy: 0.9856 - val_loss: 0.8104 - val_accu
racy: 0.8288
Précision de l'entraînement: 0.9872
Exactitude des tests: 0.8288
```

Simple 1 2 master Python 3.9 DL Keras | Idle Mode: Command Ln 1, Col 1 Classifieur Keras utilisant ISIDORE.ipynb 1

Keras

Simple. Flexible. Powerful.


LIBRARY OF
CONGRESS


{BnF data.bnf.fr


BIBLIOTECA
NACIONAL
DE ESPAÑA
BNE


Isla ö eiland otok
ejland hsù nisi oy
GeoEthno
Île Insula Ihla ey
Isola Insel Island


Semantic
Web


ORCiD
Connecting Research
and Researchers


ArchiRès

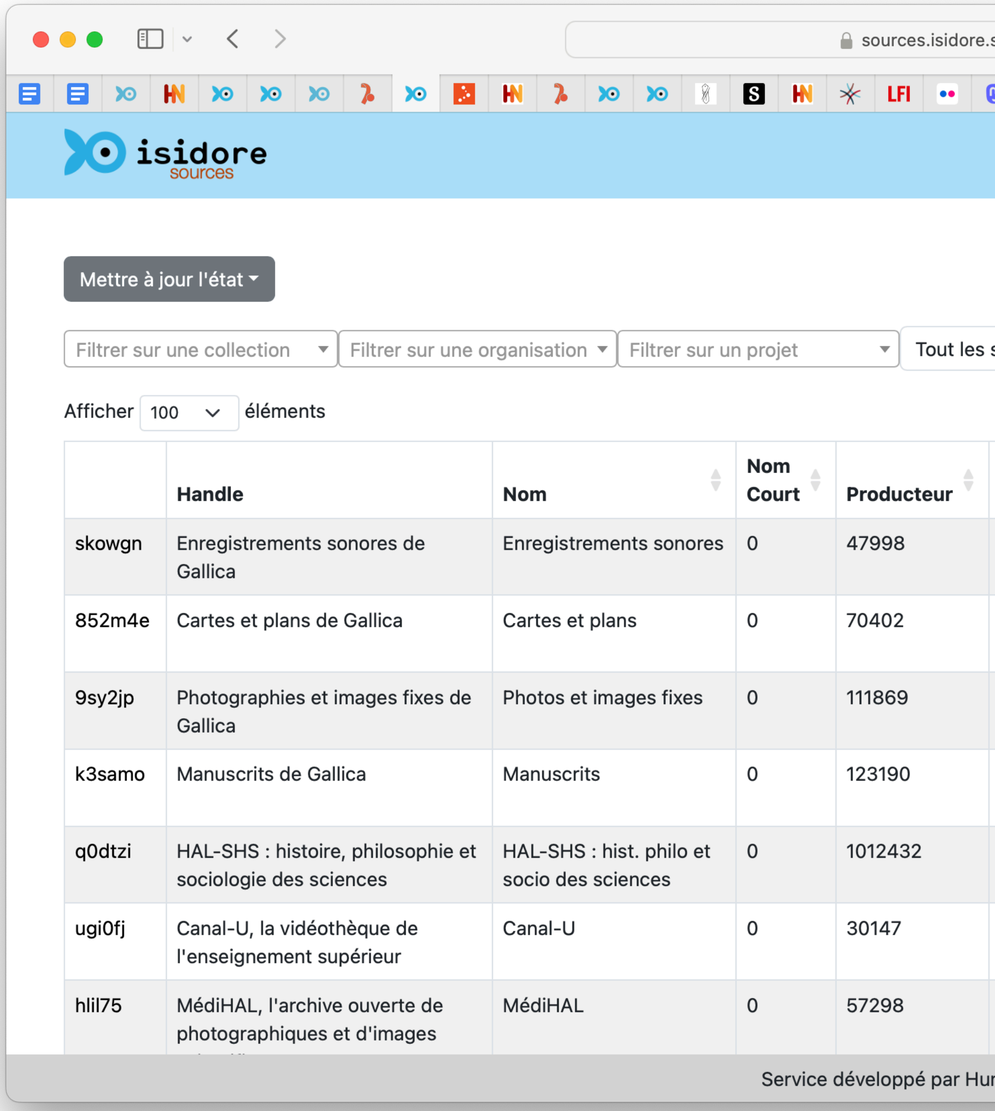

Gemet
European Environment Agency

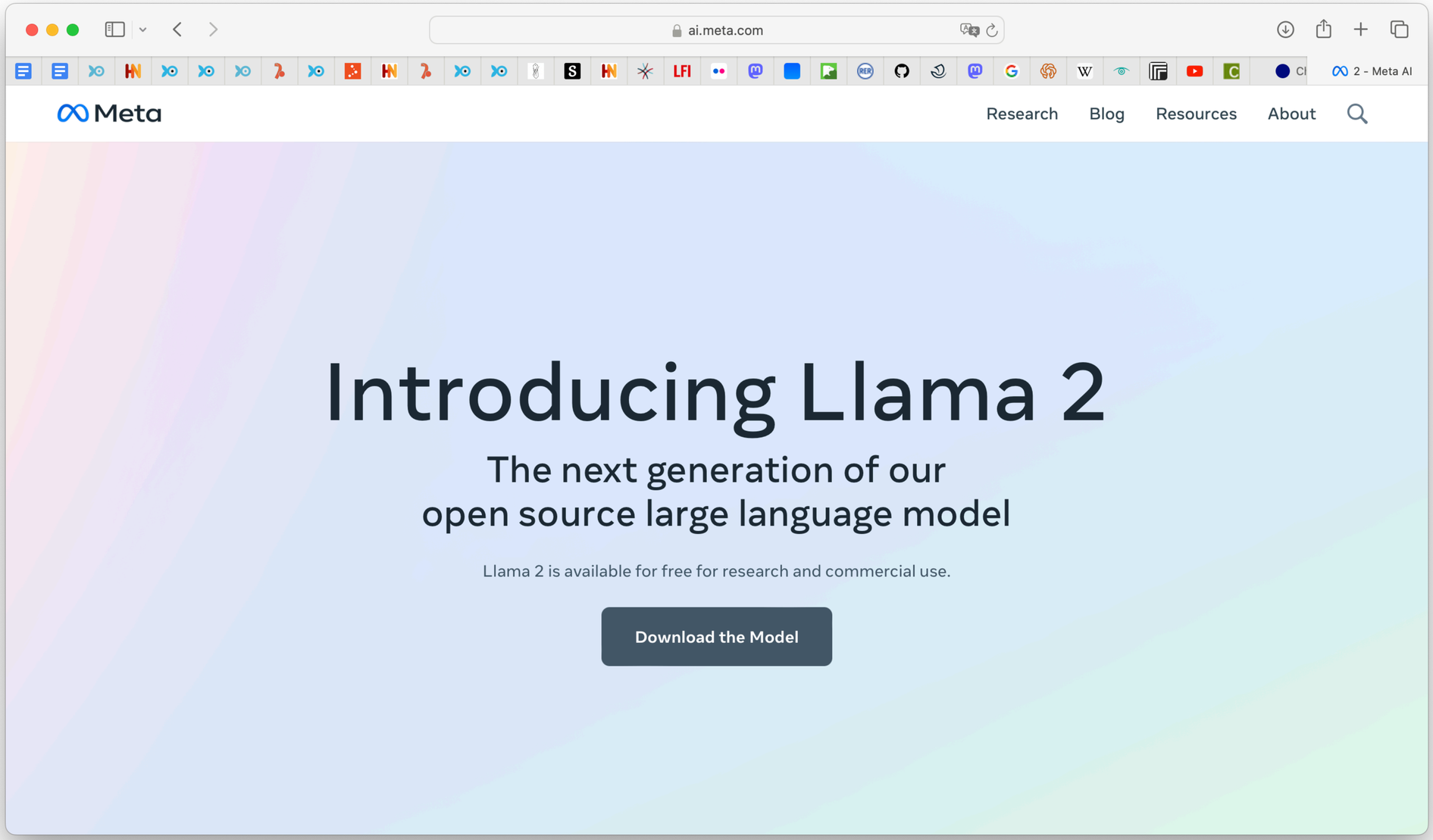

HAL
archives-ouvertes.fr


GeoNames


FRANTIQ


OpenEdition
HAL
SCIENCES DE L'HOMME ET DE LA SOCIÉTÉ
Gallica
BIBLIOTHÈQUE NUMÉRIQUE
L'ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
CAIRN.INFO
CHERCHER, REPÉRER, AVANCER
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
BIBLIOTECA NACIONAL DE ESPAÑA
BNE
cnrs
persée
open
Open Access
Publishing in European Networks
SOAS
University of London
CPU
CONFÉRENCE DES PRÉSIDENTS D'UNIVERSITÉ
fondation maison des sciences de l'homme
érudit
Promouvoir et diffuser la recherche et la création
Université de Montréal
UQÀM
ULB
redalyc
LA HEMEROTECA CIENTÍFICA EN LÍNEA
INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN


sources.isidore.s
isidore
sources
Mettre à jour l'état
Filtrer sur une collection
Filtrer sur une organisation
Filtrer sur un projet
Tout les
Afficher 100 éléments
Handle
Nom
Nom Court
Producteur
skowgn
Enregistrements sonores de Gallica
Enregistrements sonores
0
47998
852m4e
Cartes et plans de Gallica
Cartes et plans
0
70402
9sy2jp
Photographies et images fixes de Gallica
Photos et images fixes
0
111869
k3samo
Manuscrits de Gallica
Manuscrits
0
123190
q0dtzi
HAL-SHS : histoire, philosophie et sociologie des sciences
HAL-SHS : hist. philo et socio des sciences
0
1012432
ugi0fj
Canal-U, la vidéothèque de l'enseignement supérieur
Canal-U
0
30147
hlil75
MédiHAL, l'archive ouverte de photographiques et d'images
MédiHAL
0
57298
Service développé par Hu


ai.meta.com
2 - Meta AI
Meta
Research
Blog
Resources
About
Introducing Llama 2
The next generation of our
open source large language model
Llama 2 is available for free for research and commercial use.
Download the Model

bigscience.huggingface.co

BigScience Blog

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

🌸Introducing The World's Largest Open Multilingual Language Model: BLOOM🌸

Large language models (LLMs) have made a significant impact on AI research. These powerful, general mode
can take on a wide variety of new language tasks from a user's instructions. However, academia, nonprofits ar
smaller companies' research labs find it difficult to create, study, or even use LLMs as only a few industrial labs
with the necessary resources and exclusive rights can fully access them. Today, we release BLOOM, the first
multilingual LLM trained in complete transparency, to change this status quo — the result of the largest
collaboration of AI researchers ever involved in a single research project.

Who is organizing BigScience

BigScience is not a consortium nor an officially incorporated entity. It's an open collaboration boot-strapped by HuggingFace, GENCI and IDRIS, and organised as a research workshop. This research workshop gathers academic, industrial and independent researchers from many affiliations and whose research interests span many fields of research across AI, NLP, social sciences, legal, ethics and public policy.

While there is no formal relationship between any of the affiliation entities of the participants to the workshop and working group, the BigScience initiative is thankful for the freedom to participate to the workshop that the academic and industrial institutions behind all the participants have been providing. In particular, we would like to acknowledge and thank the support provided by:

keras.io

Google Cloud Présentation Solutions Produits Tarifs Ressources Docs Assistance Language Console

Nous contacter Commencer l'essai gratuit

- Cloud TPU (Tensor Processing Units)
- Présentation
- Fonctionnement
- Utilisations courantes
  - Exécutez des charges de travail d'entraînement d'IA à grande échelle
  - Ajustez les modèles d'IA de base
  - Diffusez des charges de travail d'inférence d'IA à grande échelle
  - Cloud TPU dans GKE
  - Cloud TPU dans Vertex AI
- Tarifs

Tutoriels, documentation et

Lancement de **Cloud TPU v5e**, notre Cloud TPU le plus rentable, polyvalent et évolutif à ce jour.

Cloud TPU (Tensor Processing Units)

Accélérez le développement de l'IA avec les TPU Google Cloud

Les Cloud TPU optimisent les performances et les coûts de toutes les charges de travail d'IA, de l'entraînement à l'inférence. Grâce à une infrastructure de centre de données de pointe, les TPU offrent une fiabilité élevée, une haute disponibilité et une sécurité optimale.

Profiter d'un essai gratuit | Contacter le service commercial

Vous ne savez pas si les TPU sont la réponse à vos besoins ? Découvrez quand utiliser des GPU ou des CPU sur des instances Compute Engine pour exécuter vos charges de travail de machine learning.

Infos clés

- Exécutez des charges de travail d'entraînement d'IA à grande échelle
- Ajustez les modèles d'IA de base
- Diffusez des charges de travail d'inférence d'IA à grande échelle

Cloud TPU v5e est maintenant disponible en version Preview publique.

APERCU

consensus.app

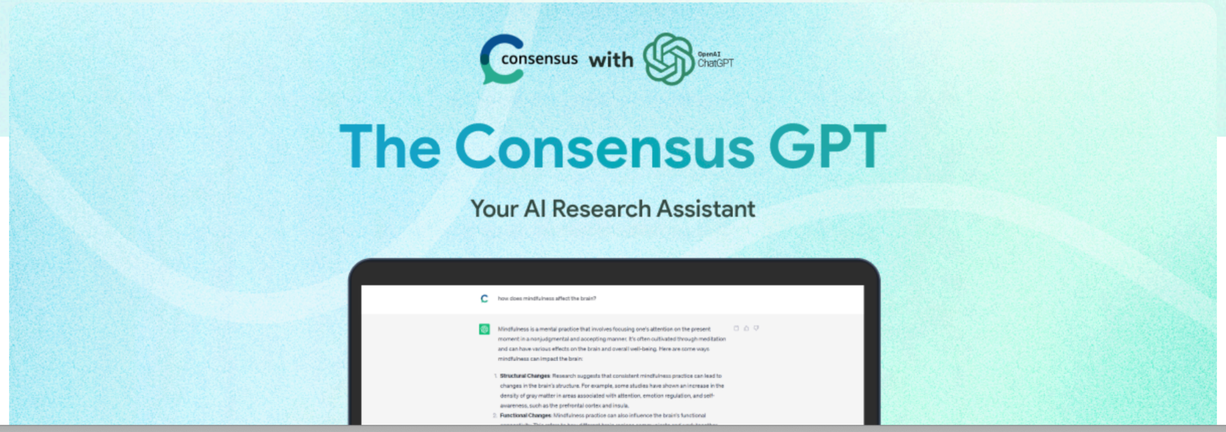
consensus

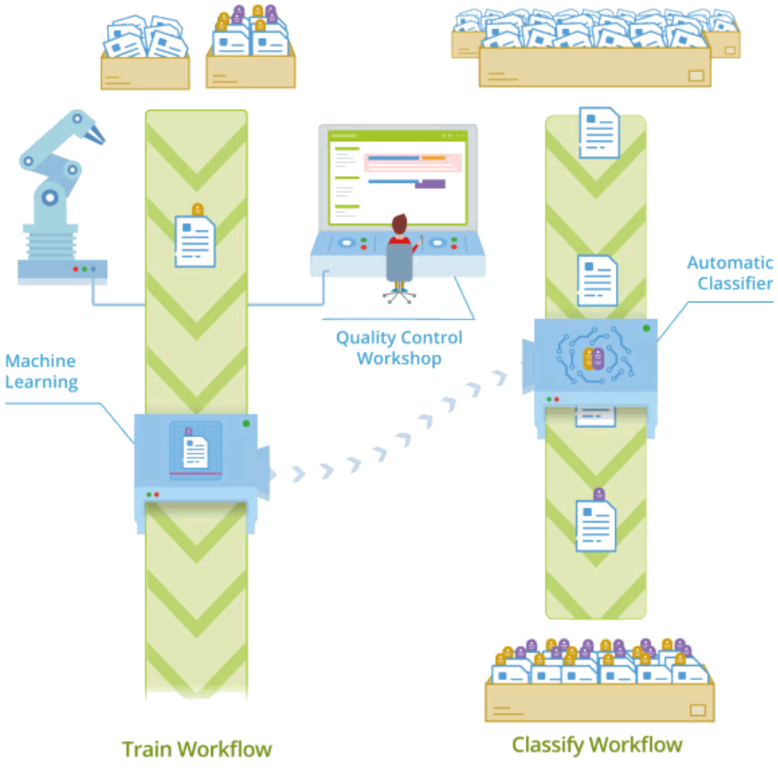
Home About Blog Contact News Pricing Login Sign Up

Introducing: Consensus GPT, Your AI Research Assistant

January 10, 2024 • By Consensus

Share on:





github.com

Product Solutions Open Source Pricing Search Sign in Sign up

opinionscience / FabriqueLLM Public

Notifications Fork 0 Star 1

Code Issues 1 Pull requests Actions Projects Security Insights

main 1 branch 0 tags

Pclanglais Update README.md e9b77e

falcon	Add files via upload
LICENSE	Initial commit
README.md	Update README.md

README.md

FabriqueLLM

FabriqueLLM is a compilation of code notebook and educational resources to fine-tune lea
(Falcon, Llama, MPT...). The project was originally conceived for fine-tuning in French but c
contexts.

The lightest model used (7B parameters in 4-bit) can be used directly on the free version c
1h30. They provide a good introduction to the fine-tuning LLM. Even a simple fine-tuning c
satisfies a variety of uses that GPT-4 would not fulfill (historical variants, dialects, sociolec

Te code notebooks of FabriqueLLM can be used in combination with the datasets made av
InstructionFr repository: https://github.com/opinionscience/InstructionFr

Falcon

FabriqueLLM include a documented code notebook in French and English to fine-tuning Fa
quantized at 4bit. This is currently one of the most cost effective model for fine-tuning. Th
(one epoch) should run in 1h15-1h30 on the free version of Google Colab and in around 15
Colab Pro+ (with a A100 with 40 Go of Vram).

The code notebook comes under two variants:

- **Fine_tune_avec_falcon_4bit.ipynb**: Generic example with a French translation of Alp
Probably the best place to start ([https://colab.research.google.com/drive/1wu56PqmF
zJIlxk_ypbdiuYwrkVJH9?usp=sharing](Google Colab))
- **Fine_tune_avec_falcon_4bit_novel17.ipynb**: More advanced use with a retraining on
novels with the original script. Settings have been slightly tweaked as a result (with a H
([https://colab.research.google.com/drive/1grBHTUuYdodLZFnV0ehUFSUoa4DJkgA2?
Colab))

Generation is slow but should be significantly enhanced in future versions. The sample out
notebooks provide a good illustration of the different capacity of fine-tuning.

With a standard approach (fine-tuning on Vicogne) you get chatGPT-like results:

> **Instruction**: Peux-tu m'expliquer en détail comment je pourrais me rendre sur la Lune ?
>
> **Output (Falcon-Vicogne)**: Pour se rendre sur la Lune, une personne devrait embarquer sur un vaisseau spatial en orbite autour de la Terre, puis parcourir environ 384 400 miles (643 200 kilomètres) pour se rendre à la surface de la Lune. Le vaisseau spatial doit être conçu pour être capable de voyager dans l'espace et devrait être propulsé par un moteur à combustion pour le faire fonctionner. Le vaisseau spatial doit également être conçu pour protéger les passagers contre les effets de l'environnement extérieur, tels que la gravité, les

opsci.ai

opsci

À ceux qui veulent composer avec la société,
nous offrons la juste mesure des données.

EXPERTISES

COLLECTE AUGMENTÉE

- Data-mining multiplateforme : Twitter, Facebook, Instagram, YouTube, Tiktok, Reddit, Quora, Pinterest, Twitch...
- Presse internationale, nationale, locale
- Open data, statistiques publiques, sondages
- OSINT

TRAITEMENT SUPERVISÉ

- Topic-mining et audience-mining
- Deep learning, NLP, semantic mapping
- Analyse dynamique de réseaux
- Détection d'émergence et d'accélération
- Détection de comportement inauthentique coordonné

ANALYSE COMPARÉE

- Qualification humaine et rétroaction de modélisation
- Mise en perspective de l'affordance des plateformes
- Étude sémiotique, sémiologique et socio-politique
- Cartographies des communautés, controverses et conflits
- Réseau d'analystes internationaux

EXPERTISES

MONDE ACADÉMIQUE | SPHÈRE PUBLIQUE | MONDE ÉCONOMIQUE

Des grandes entreprise commes des startups s'appuient sur Opsci pour l'étude de leurs audiences, de leur marché d'opinion, pour l'analyse des tendances, pour l'aide au profilage de leurs cibles et de leurs campagnes, pour le monitoring de réputation, la veille de crise, et pour la mesure de la performance.

Pour optimiser votre expérience, nous collectons des données de navigation en utilisant les cookies (lire les mentions légales). Êtes-vous d'accord ? oui / non

Ce que l’on veut faire

Repérer des documents
et des données (moteur de recherche)

Proposer un écosystème pour la
veille scientifique (réseau social)

Communautés très différentes
dans leurs pratiques
de recherche et d’utilisation
des informations
(sources, documents, codes,
…)

Documents et des informations
hétérogènes

Des pratiques de
« mises en données »
différentes

Des pratiques de diffusion
différentes

?

Proposer un écosystème pour
favoriser les pistes de collaboration
(entre collègues, entre communautés
autour de revues, etc.)

Proposer un écosystème
pour favoriser
l’écriture collaborative avec des
données

Proposer un écosystème pour
favoriser l’écriture collaborative avec
des données

Permettre d’assister les chercheur·e·s
dans leurs productions d’écrits
(plans, états de l’art, etc.)

Construction : des processus itératifs

Type d’IA dans ISIDORE 2030

machine leaning	deep learning (transformer)	*Fine tuning* de LLM et MLM (Middle Language Model) utilisant des données SHS

isidore

- Doctorant·e·s en commun
- Post-doc
- Temps de calcul au Genci
- Programme de recherche

2030

A discuter : la proposition d'un laboratoire commun antidot/CNRS

cnrs


cnrs


Université
de Montréal


HN Lab


CRIHN

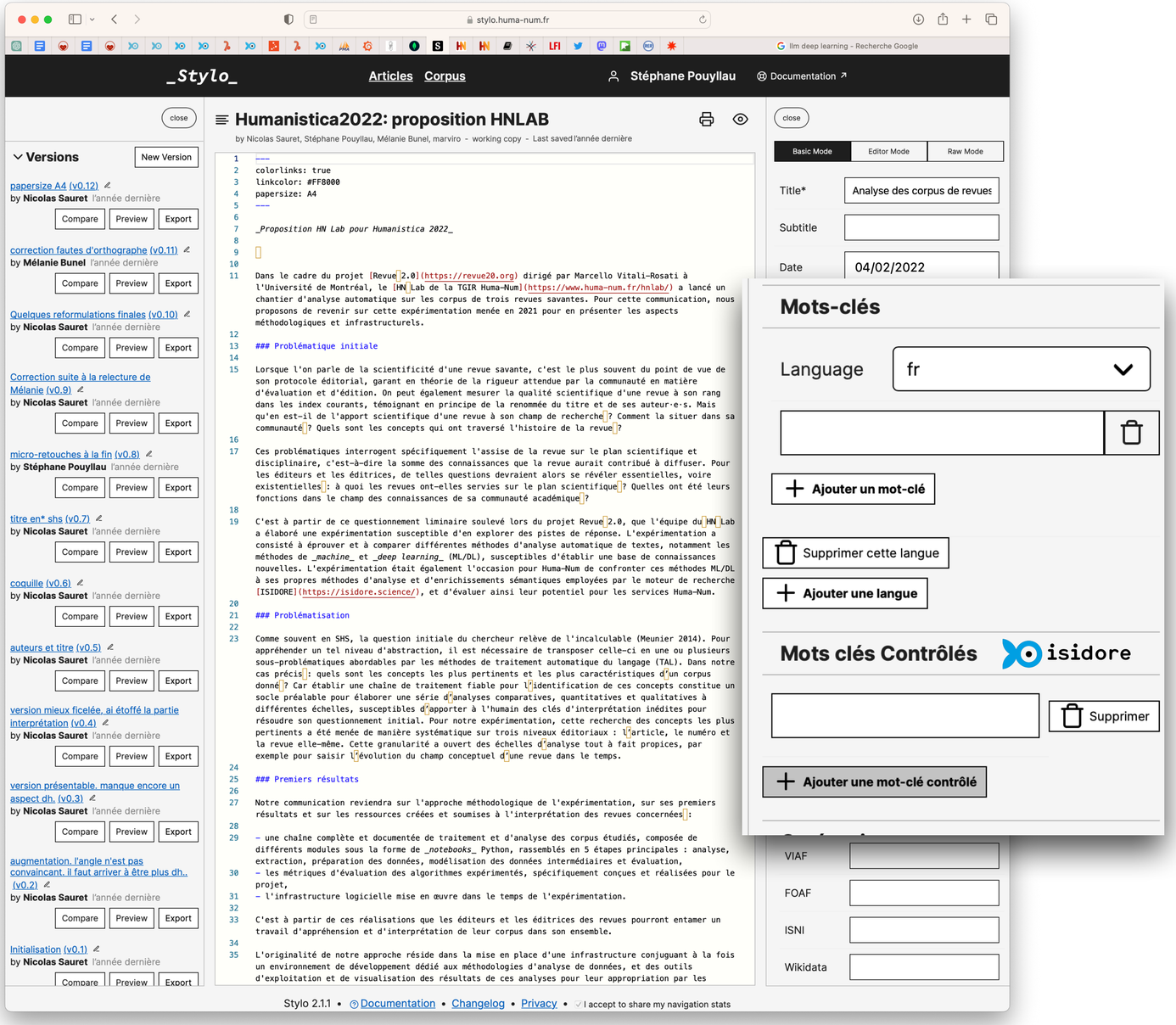
ÉCRITURES NUMÉRIQUES
Chaire de recherche du Canada

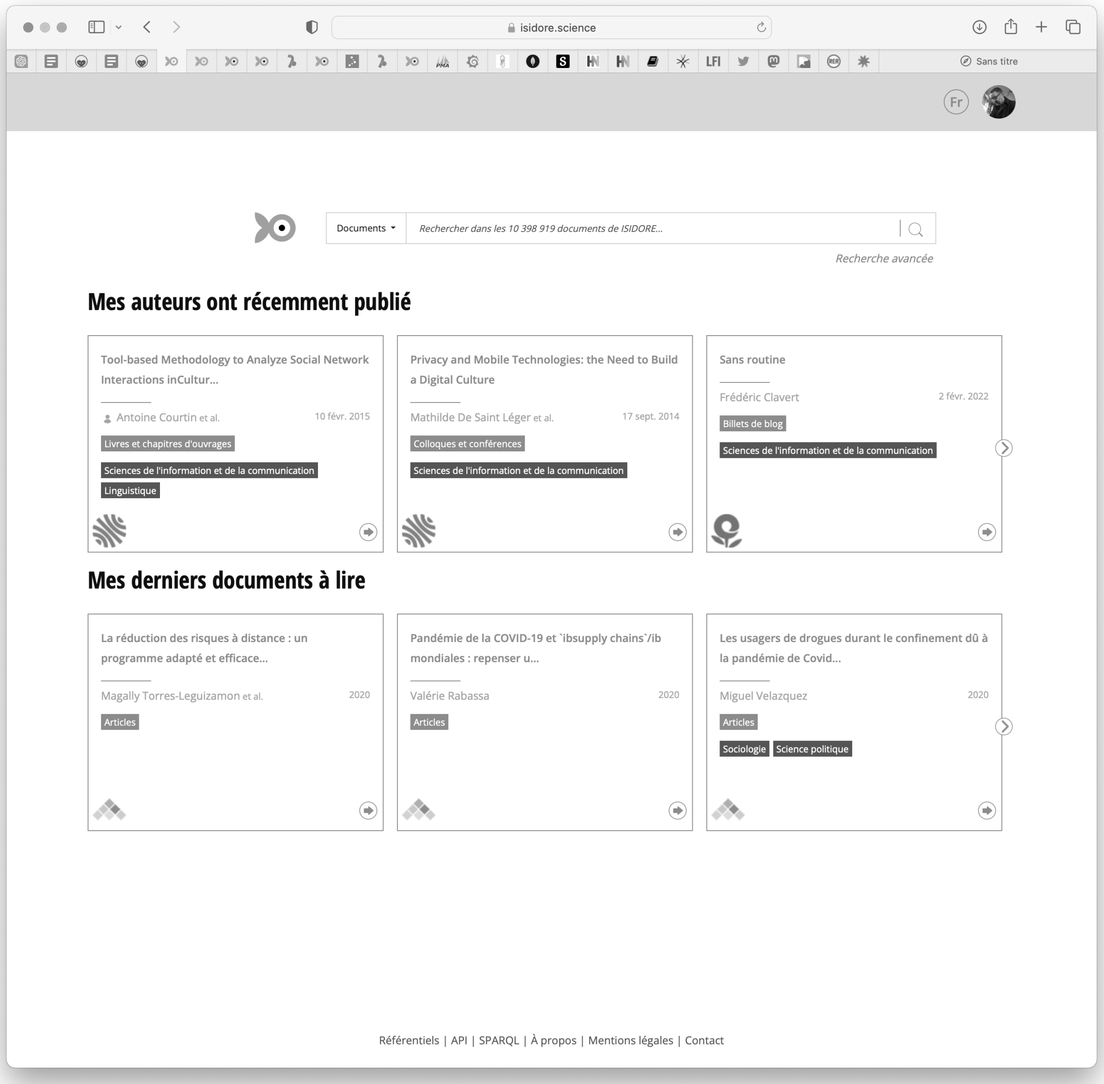
+

HN Lab

+

métopes
méthodes et outils
pour l'édition structurée




isidore.science
Sans titre
Documents
Rechercher dans les 10 398 919 documents de ISIDORE...
Recherche avancée
Mes auteurs ont récemment publié
Tool-based Methodology to Analyze Social Network Interactions inCultur...
Antoine Courtin et al.
10 févr. 2015
Livres et chapitres d'ouvrages
Sciences de l'information et de la communication
Linguistique
Privacy and Mobile Technologies: the Need to Build a Digital Culture
Mathilde De Saint Léger et al.
17 sept. 2014
Colloques et conférences
Sciences de l'information et de la communication
Sans routine
Frédéric Clavert
2 févr. 2022
Billets de blog
Sciences de l'information et de la communication
Mes derniers documents à lire
La réduction des risques à distance : un programme adapté et efficace...
Magally Torres-Leguizamon et al.
2020
Articles
Pandémie de la COVID-19 et `ibsupply chains`/ib mondiales : repenser u...
Valérie Rabassa
2020
Articles
Les usagers de drogues durant le confinement dû à la pandémie de Covid...
Miguel Velazquez
2020
Articles
Sociologie
Science politique
Référentiels | API | SPARQL | À propos | Mentions légales | Contact


stylo.huma-num.fr
Stylo
Articles
Corpus
Stéphane Pouyllau
Documentation
close
Versions
New Version
papersize A4 (v0.12)
by Nicolas Sauret l'année dernière
Compare
Preview
Export
correction fautes d'orthographe (v0.11)
by Mélanie Bunel l'année dernière
Quelques reformulations finales (v0.10)
by Nicolas Sauret l'année dernière
Correction suite à la relecture de Mélanie (v0.9)
by Nicolas Sauret l'année dernière
micro-retouches à la fin (v0.8)
by Stéphane Pouyllau l'année dernière
titre en* shs (v0.7)
by Nicolas Sauret l'année dernière
coquille (v0.6)
by Nicolas Sauret l'année dernière
auteurs et titre (v0.5)
by Nicolas Sauret l'année dernière
version mieux ficelée, ai étoffé la partie interprétation (v0.4)
by Nicolas Sauret l'année dernière
version présentable, manque encore un aspect dh. (v0.3)
by Nicolas Sauret l'année dernière
augmentation. l'angle n'est pas convaincant. il faut arriver à être plus dh.. (v0.2)
by Nicolas Sauret l'année dernière
Initialisation (v0.1)
by Nicolas Sauret l'année dernière
Humanistica2022: proposition HNLAB
by Nicolas Sauret, Stéphane Pouyllau, Mélanie Bunel, marviro - working copy - Last saved l'année dernière
Basic Mode
Editor Mode
Raw Mode
Title*
Analyse des corpus de revues
Subtitle
Date
04/02/2022
Language
fr
License
Creative Commons Attribution
Journal issue
Journal issue
Acknowledgements
Acknowledgements
Bibliography
Display
All citations
Authors
Last name
Sauret
First name
Nicolas
Affiliations
Biography
Email
ORCID
0000-0001-7516-3427
VIAF
FOAF
ISNI
Wikidata
Stylo 2.1.1
Documentation
Changelog
Privacy
I accept to share my navigation stats

Forces

Maturité du dispositif ISIDORE

Briques technologiques disponibles

Positionnement institutionnel

Partenaires industriels (mise en production)

Faiblesses

Complexité dans la maitrise des ≠ IA

Frontière entre possibilité et besoins

Enjeux éco-numériques

Opportunités

Demandes des chercheur·e·s complexes et divergentes

Permettre d'associer + les communautés SHS au développement des IA (et surtout de leurs usages)

Passer de la critique décentré à la critique d'usage de l'IA

Menaces

« Ringardiser » un instrument de travail existant

Concurrences d'autres dispositifs moins dans les mains des chercheur·e·s

Se perdre dans dans de l'innovation sans fin

ISIDORE 2030 : questions ?

Stéphane Pouyllau, ingénieur de recherche au CNRS
co-fondateur d'Huma-Num.
Responsable du HN Lab et en charge d'ISIDORE