

Les outils open-source de Teklia pour le traitement de documents numérisés

Journée-atelier “IA et images en SHS”

15 mai 2024

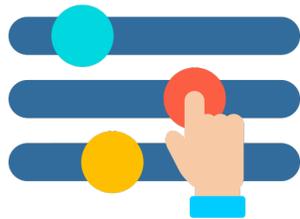
Christopher Kermorvant

kermorvant@teklia.com

Arkindex/Callico : Principes

Plateforme développée par TEKLIA depuis 2019
Utilisée en interne pour plus de 60 projets

Personnalisation



Traiter tout type de documents

Passage à l'échelle



Traiter 1000 ou 10 millions de pages

Open-source

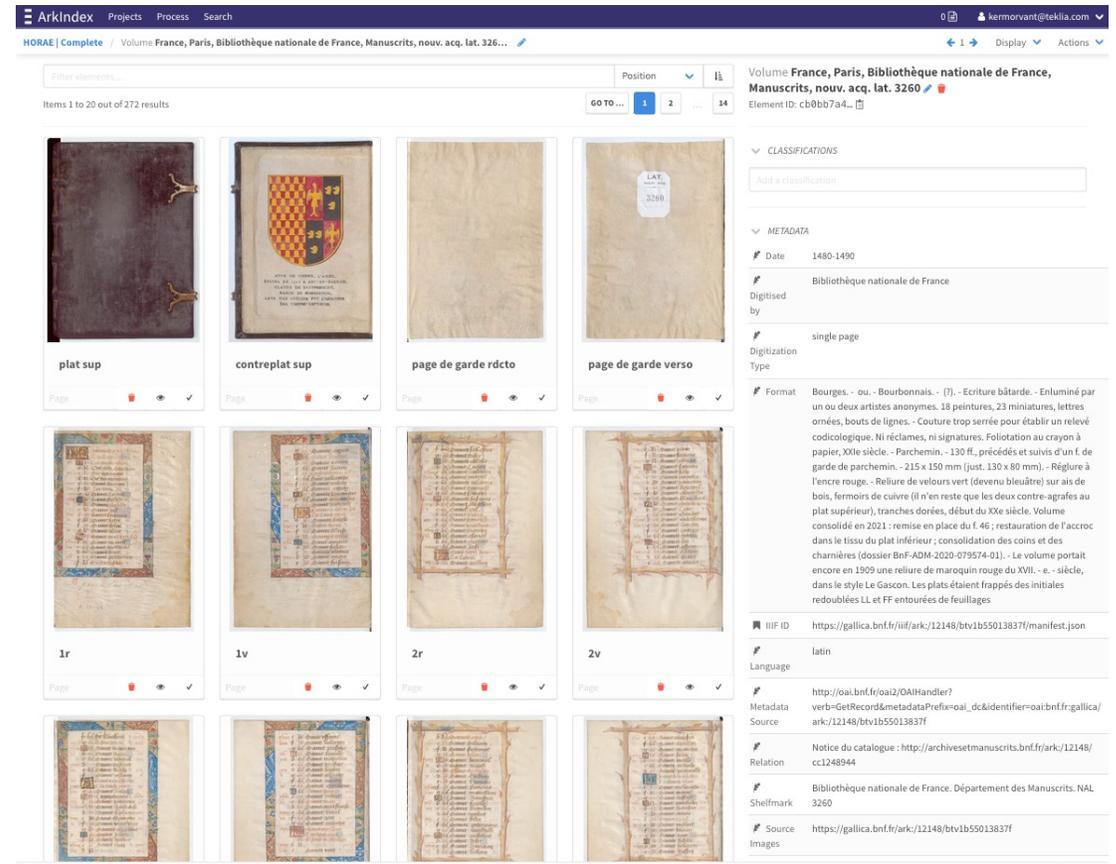


Diffusion et participation de la communauté

Pictoria ?

ArkindeX : Stockage et gestion des documents

- Import web (petit corpus), S3 (gros corpus) ou par manifest IIIF
- Support des formats images et PDF
- Structuration hiérarchique des corpus complètement adaptable
- Gestion des métadonnées à tous les niveaux
- Visualisation, navigation

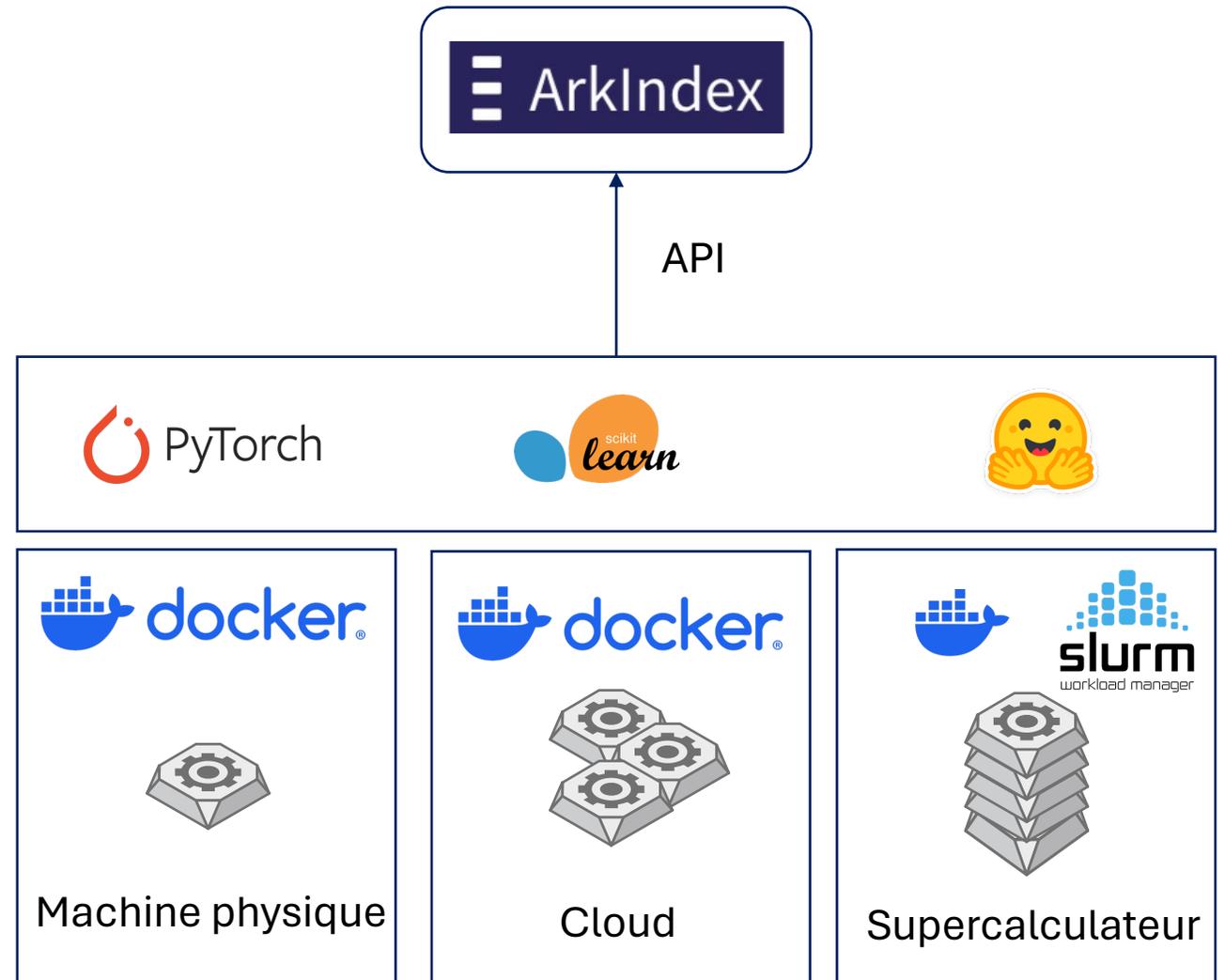


The screenshot shows the ArkindeX web interface. At the top, there's a navigation bar with 'ArkindeX', 'Projects', 'Process', and 'Search'. Below that, the breadcrumb trail reads 'HORAE | Complete / Volume France, Paris, Bibliothèque nationale de France, Manuscrits, nouv. acq. lat. 326...'. The main content area displays a grid of 12 thumbnail images of manuscript pages, labeled 'plat sup', 'contreplat sup', 'page de garde rdcto', 'page de garde verso', and folios '1r', '1v', '2r', '2v'. The right sidebar shows the 'Volume France, Paris, Bibliothèque nationale de France, Manuscrits, nouv. acq. lat. 3260' with element ID 'cb0bb7a4...'. It includes sections for 'CLASSIFICATIONS', 'METADATA', and 'IIIF ID'. The IIIF ID is <https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json>.

<https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json>

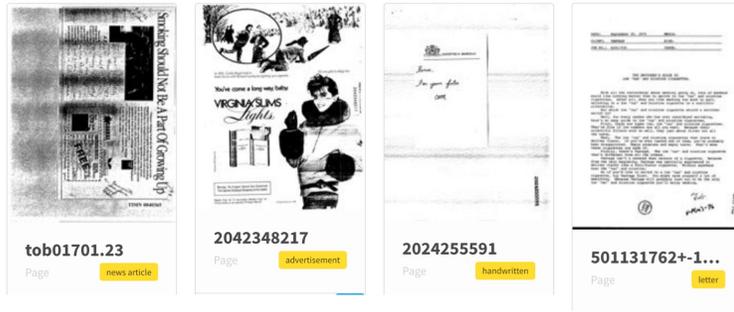
Arkindex : Intégration de modèles/algorithmes

- Intégration de n'importe quel langage/code/modèle
- Code de base python fourni
- Intégration par API
- Déploiement par Docker
- Entraînement et inférence

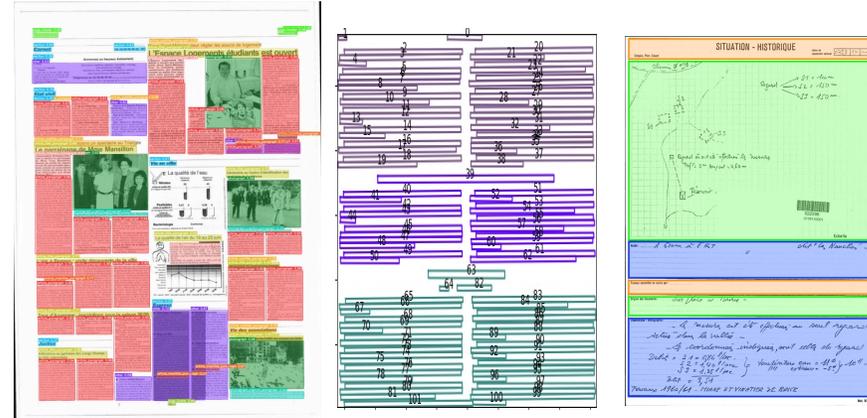


Arkindex : applications

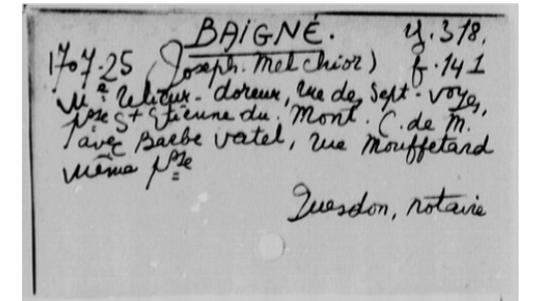
Classification



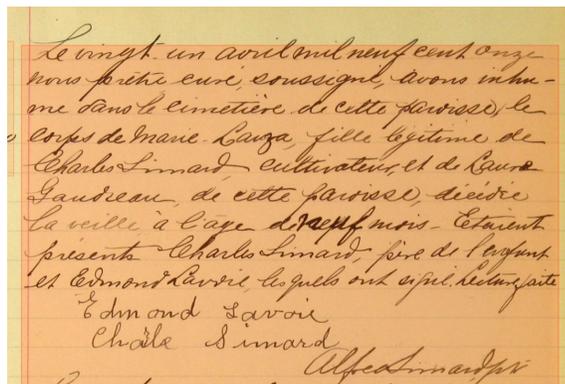
Structuration



OCR/HTR



Extraction d'entités



Analyse de photos



- tag group of people
- tag samurai
- tag map
- tag swords
- tag kimono
- tag potted plant

Projet SOCFACE

INED/PSE/SIAF/TEKLIA, 2022-2025

Classification, structuration, reconnaissance, linking de 100 ans de recensements français entre 1836-1936 (30M d'images)

DENOMINATION	NOI. ENREG.	NOI. ENREG. VILLAGES	NOI. ENREG. COMMUNES	ANNÉE de l'ENREG.	LIÉU de l'ENREG.	NAT. de l'ENREG.	SITUATION de l'ENREG.	PROF.
Chévenet Charles	1873	Montrigny	Corrèze	1873	Corrèze	P	chef	Néant
Théveuet Marguerite	1875	Noulun	P. Loire	1875	Noulun	F	épouse	Néant
Legouet Marie Antré	1866	Lalouère	G. G.	1866	Lalouère	F	épouse	Néant
Jolivet Marie	1910	Venas	nationalité P.	1910	Venas	F	chef	couturière
Jolivet Simon	1908	Venas	nationalité P.	1908	Venas	F	frère	Mécanicien
V. Jaulines Marie	1852	Marigoy	nationalité F.	1852	Marigoy	F	grand mère	Néant
Lamarque Virginie	1870	Douvrv	ord.	1870	Douvrv	F	chef	Néant
Sabathie Joseph E. desnoud	1883	Toulonne	nationalité P.	1883	Toulonne	F	chef	Receveur Enregistrement
Sabathier Marguerite	1883	Gremoble	nationalité F.	1883	Gremoble	F	épouse	Néant
Ballaire Albert	1884	Vemennes	nationalité F.	1884	Vemennes	F	chef	profice de Pait
Ballaire Adélaïde	1876	Maillet	nationalité F.	1876	Maillet	F	épouse	Néant
Rouy Louis	1895	Bionde	nationalité F.	1895	Bionde	F	chef	Platrice peintre
Rouy Emilienne	1896	Hérison	nationalité F.	1896	Hérison	F	épouse	Néant
Roux Jean Robert	1920	Herisses	nationalité F.	1920	Herisses	F	frères	apprenti platre
Lepée Louis	1893	Bizenemille	nationalité F.	1893	Bizenemille	F	chef	Agricalten
Pepie Marie	1903	Gaumond	nationalité F.	1903	Gaumond	F	épouse	Néant
Pepce Jean	1925	Heinson	nationalité F.	1925	Heinson	F	frères	Néant
Lepée Edmond	1928	Heinson	nationalité P.	1928	Heinson	F	frères	Néant
Guillonnet Henri	1910	Heinson	nationalité P.	1910	Heinson	F	chef	Camier
Guillomet Marguerite	1913	Chaubeux	nationalité P.	1913	Chaubeux	F	épouse	Néant
Guillomet Philippe	1882	Estesancies	nationalité F.	1882	Estesancies	F	père	Entreprenent frances frobles
V. Veunat Adélaïde	1882	Venas	P.	1882	Venas	F	chef	Néant
Riotte Jean	1911	Heinson	nationalité F.	1911	Heinson	F	dat. civil	Neveu
Ve Martinet Marie	1860	Venise	nationalité F.	1860	Venise	F	chef	ménagère
V. Lachasagne Clarice	1870	Heinson	nationalité F.	1870	Heinson	F	chef	Néant
Lachanagne Georges	1892	Vevas	nationalité F.	1892	Vevas	F	frères	Electricien
V. Périnet Hélène	1862	Charbens	nationalité F.	1862	Charbens	F	chef	Néant
Bergerat Jean	1864	Heinson	nationalité F.	1864	Heinson	F	chef	jardinier
Bergerat Jean	1902	Heinson	nationalité F.	1902	Heinson	F	frères	Camis
V. Malochet Amilie	1865	Maillet	nationalité F.	1865	Maillet	F	chef	Néant

nom Chévenet prénom Charles date_naissance 1873 lieux_naissance Montrigny Corrèze relation P profession chef profession Néant

nom Théveuet prénom Marguerite date_naissance 1875 lieux_naissance Noulun P. Loire relation F profession épouse profession Néant

nom Legouet prénom Marie Antré date_naissance 1866 lieux_naissance Lalouère de G. G. relation F profession épouse profession Néant

nom Jolivet prénom Marie date_naissance 1910 lieux_naissance Venas nationalité P relation chef profession couturière

nom Jolivet prénom Simon date_naissance 1908 lieux_naissance Venas nationalité P relation frère profession Mécanicien employeur M Ravet

nom V. Jaulines prénom Marie date_naissance 1852 lieux_naissance Marigoy nationalité F relation grand mère profession Néant

nom Lamarque prénom Virginie date_naissance 1870 lieux_naissance Douvrv ord relation chef profession Néant

nom Sabathie prénom Joseph E. desnoud date_naissance 1883 lieux_naissance Toulonne nationalité P relation chef profession Receveur Enregistrement

nom Sabathier prénom Marguerite date_naissance 1883 lieux_naissance Gremoble nationalité F relation épouse profession Néant

nom Ballaire prénom Albert date_naissance 1884 lieux_naissance Vemennes nationalité F relation chef profession profice de Pait

nom Ballaire prénom Adélaïde date_naissance 1876 lieux_naissance Maillet nationalité F relation épouse profession Néant

nom Rouy prénom Louis date_naissance 1895 lieux_naissance Bionde nationalité F relation chef profession Platrice peintre employeur patron

nom Rouy prénom Emilienne date_naissance 1896 lieux_naissance Hérison nationalité F relation épouse profession Néant

nom Roux prénom Jean Robert date_naissance 1920 lieux_naissance Herisses nationalité F relation frères profession apprenti platre employeur M Noot

nom Lepée prénom Louis date_naissance 1893 lieux_naissance Bizenemille nationalité F relation chef profession Agricalten employeur patron

nom Pepie prénom Marie date_naissance 1903 lieux_naissance Gaumond nationalité F relation épouse profession Néant

nom Pepce prénom Jean date_naissance 1925 lieux_naissance Heinson nationalité F relation frères profession Néant

nom Lepée prénom Edmond date_naissance 1928 lieux_naissance Heinson nationalité P relation frères profession Néant

nom Guillonnet prénom Henri date_naissance 1910 lieux_naissance Heinson nationalité P relation chef profession Camier employeur patron

nom Guillomet prénom Marguerite date_naissance 1913 lieux_naissance Chaubeux nationalité P relation épouse profession Néant

nom Guillomet prénom Philippe date_naissance 1882 lieux_naissance Estesancies nationalité F relation père profession Entreprenent frances frobles

nom V. Veunat prénom Adélaïde date_naissance 1882 lieux_naissance Venas P. relation chef profession Néant

nom Riotte prénom Jean date_naissance 1911 lieux_naissance Heinson nationalité F dat. civil Neveu profession Courreur employeur Min Naves

nom Ve Martinet prénom Marie date_naissance 1860 lieux_naissance Venise nationalité F relation chef profession ménagère employeur Diven

nom V. Lachasagne prénom Clarice date_naissance 1870 lieux_naissance Heinson nationalité F relation chef profession Néant

nom Lachanagne prénom Georges prénom Auguste date_naissance 1892 lieux_naissance Vevas nationalité F relation frères profession Electricien employeur S.T.M.C

nom V. Périnet prénom Hélène date_naissance 1862 lieux_naissance Charbens nationalité F relation chef profession Néant

nom Bergerat prénom Jean date_naissance 1864 lieux_naissance Heinson nationalité F relation chef profession jardinier employeur Dr Benon

nom Bergerat prénom Jean date_naissance 1902 lieux_naissance Heinson nationalité F relation frères profession Camis employeur Guillounet

nom V. Malochet prénom Amilie date_naissance 1865 lieux_naissance Maillet nationalité F relation chef profession Néant

Projet HikarIA

Musée Guimet/TEKLIA, 2023-2026

Labélisation, détection d'objets, recherche par similarité pour une collection d'albums de photographies japonaises du XIXème siècle



Filter by type... ▾ Hide 

Photograph 1 A+

- Sword 2 
- Sword 3 
- Sword 4 
- Sword 5 
- Sword 6 
- Map 7 

 tag	group of people
 tag	samurai
 tag	map
 tag	swords
 tag	kimono
 tag	potted plant

TEKLIA et l'Open-source

Librairies et outils de Deep Learning

- Segmentation / détection : Doc-UFCN
[PyPi](#) [GitLab](#)
- HTR : PyLaia [PyPi](#) [Gitlab](#)
- Evaluation NER : Nerval [GitLab](#)
- Evaluation segmentation : DISS [GitLab](#)

Données et modèles sur HuggingFace

- [13 modèles d'HTR avec PyLaia](#)
- [3 modèles de détection de lignes avec Doc-UFCN](#)
- [3 modèles de NER avec SpaCy](#)
- [12 jeux de données annotées](#)

<https://gitlab.teklia.com>

29 repositories



Arindex / **Backend**

Python3/Django backend



Callico / **Callico**

Annotation platform, companion of Arindex



Arindex / **Frontend**

Development & Tests frontend



IIIIF / **Cantaloupe IIIIF Docker image**

Docker Image from Teklia for the Cantaloupe IIIIF image server



IIIIF / **Cantaloupe Multiple S3 Buckets** GNU General I

Delegate script to support multiple S3/Minio buckets

Arkindex : pour qui ?

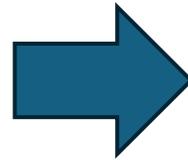


Personnalisation

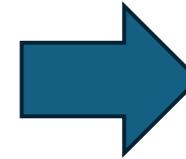
Traiter tout type de documents



Passage à l'échelle



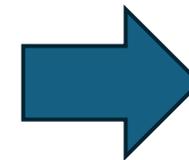
Puissance
Complexité



Expert



Documentation
Formation
Développement
UI/UX



PictorIA ?