

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

Journée-atelier “IA et images en SHS”

15 mai 2024

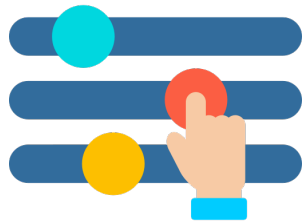
Christopher Kermorvant

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

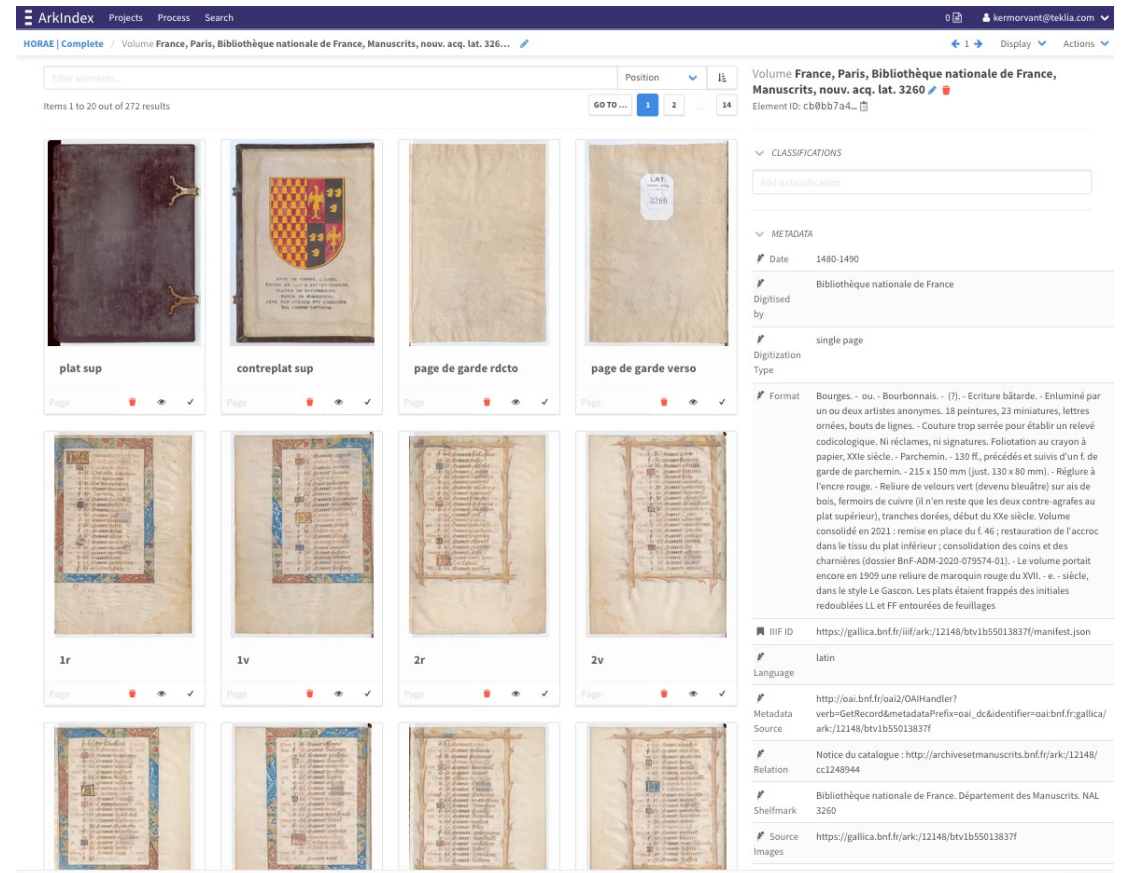


Diffusion et participation de la communauté

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

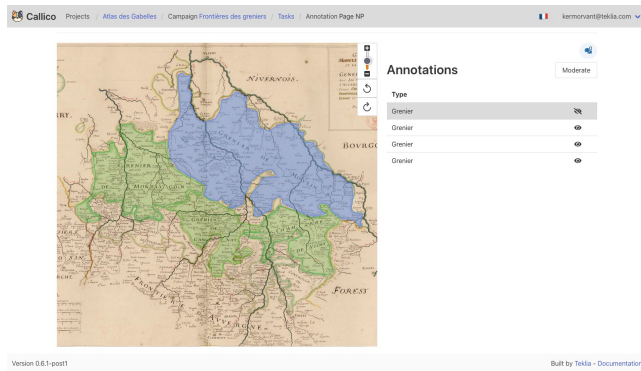


The screenshot shows the ArkindeX web interface. At the top, there's a navigation bar with 'ArkindeX', 'Projects', 'Process', and 'Search'. Below that, the breadcrumb trail reads 'HORAE | Complete / Volume France, Paris, Bibliothèque nationale de France, Manuscrits, nouv. acq. lat. 326...'. The main 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>

Callico : Annotation et validation

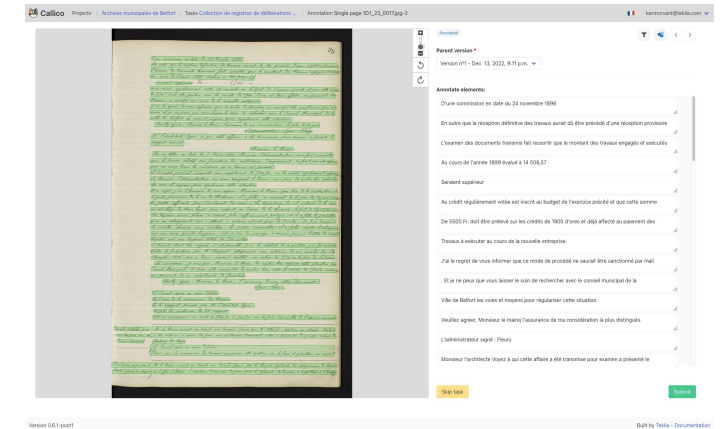
Zonage



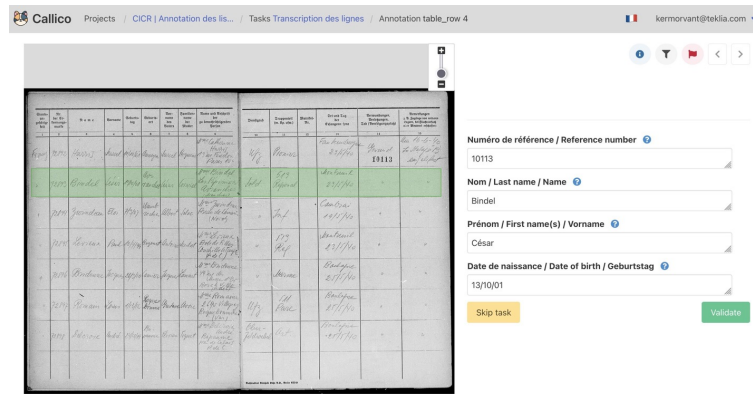
Classification



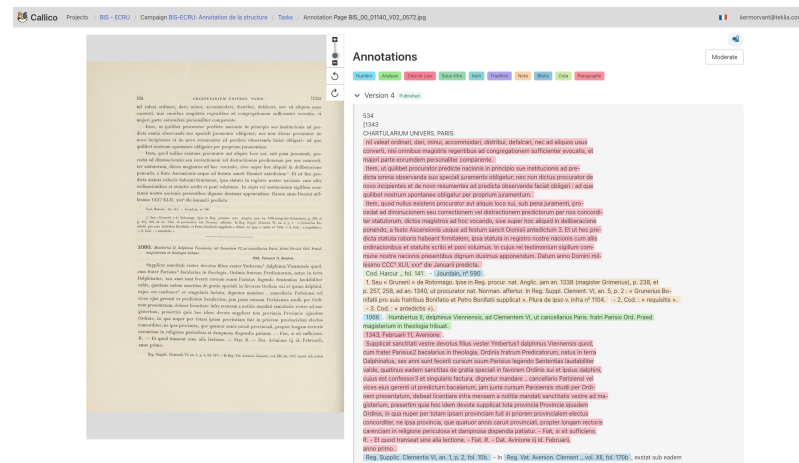
Transcription



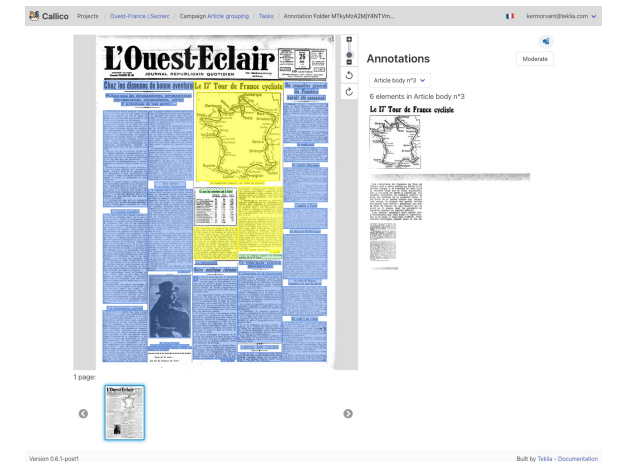
Clé-valeur



Entités nommées

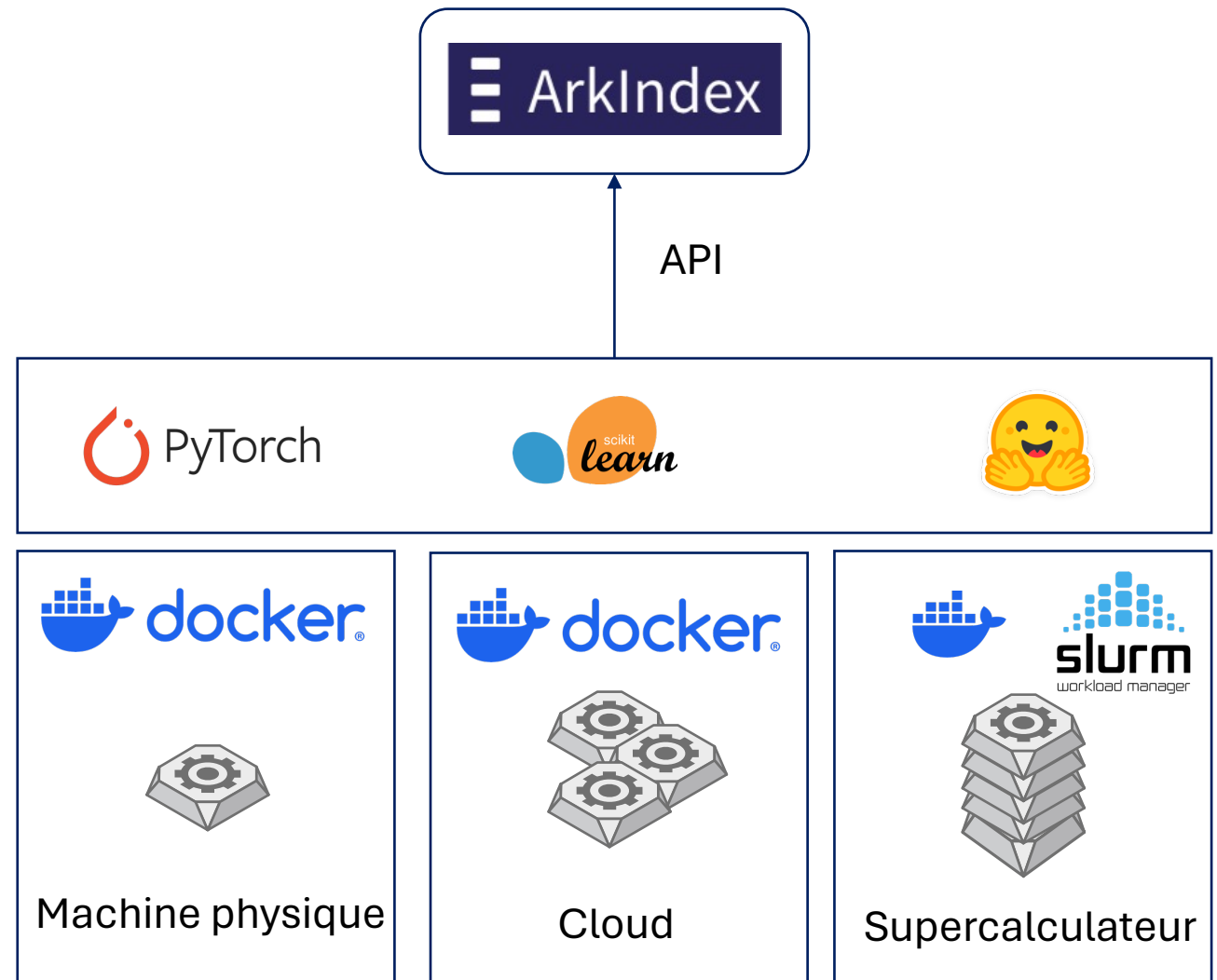


Groupement



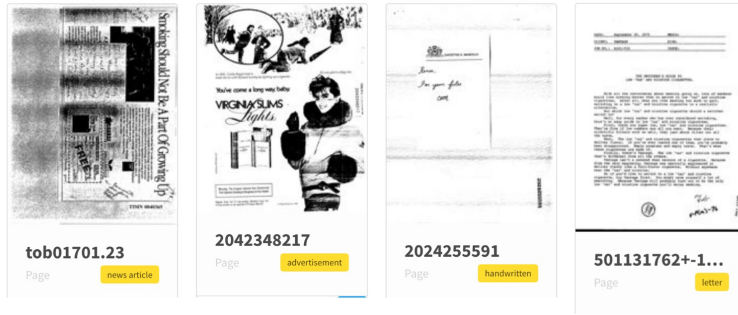
Arkindex : Intégration de modèles/algorithmes

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

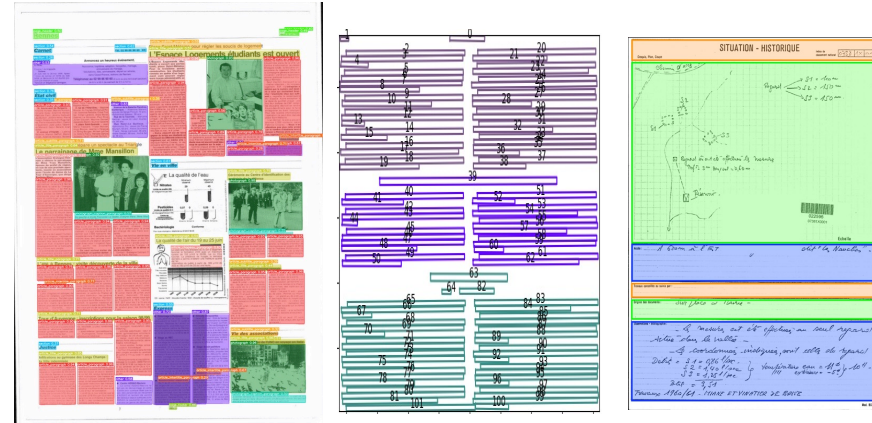


Arkindex : applications

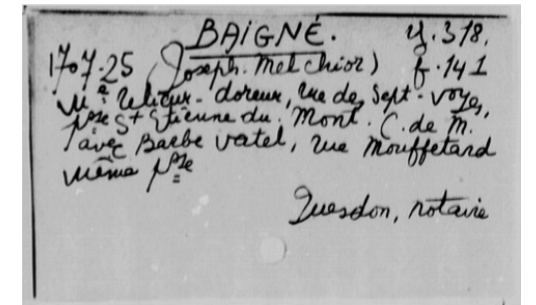
Classification



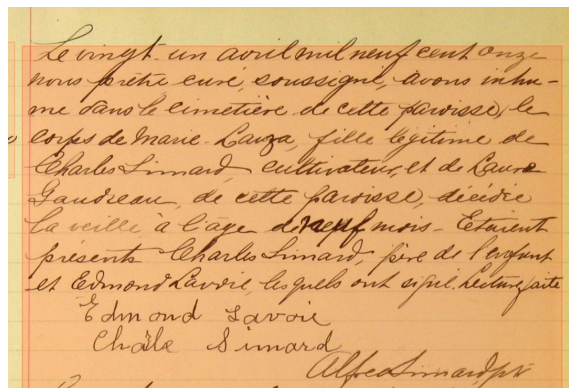
Structuration



OCR/HTR



Extraction d'entités



Analyse de photos



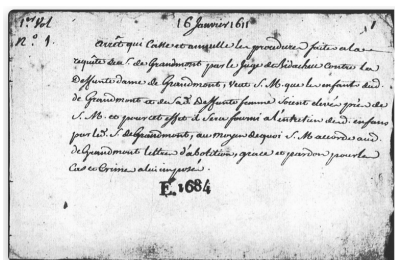
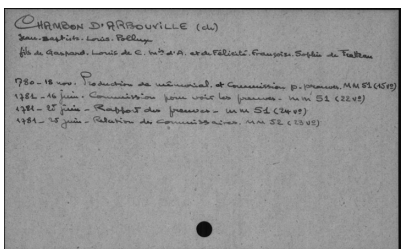
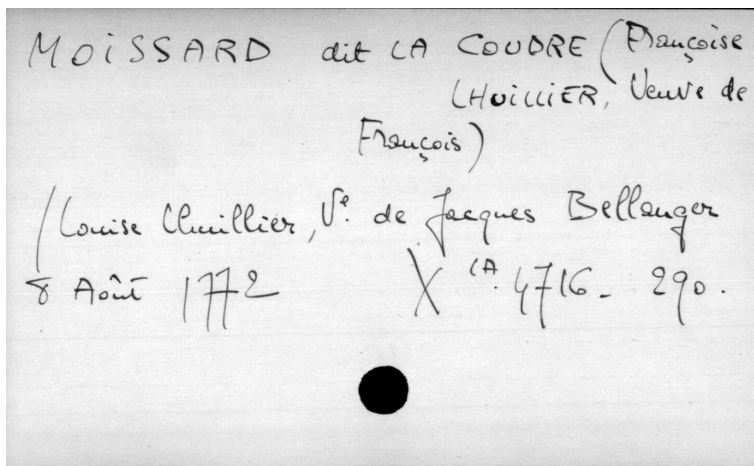
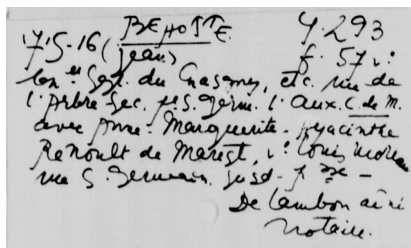
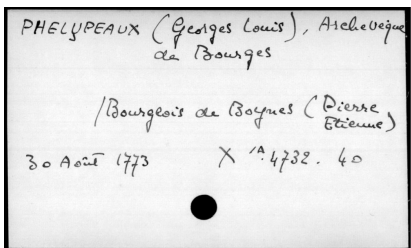
- tag group of people
- tag samurai
- tag map
- tag swords
- tag kimono
- tag potted plant

Archives Nationales

SIMARA, 2021-2023

Traitement de 400 000 fiches d'index

Reconnaissance d'écriture manuscrite, extraction d'entités



▼ TRANSCRIPTIONS

Created by **Simara (HTR + NER)** ⚙️

Simara (HTR + NER)...

99%



INTITULE MOISSARD dit LA COUDRE (Françoise LHUILLIER, veuve de François)

ANALYSE_COMPL Contre Louise Chuillier, veuve de Jacques Bellanger **DATE** 8 août 1772

COTE_SERIE X1A **COTE_ARTICLE** 4716

PRECISIONS_SUR_COTE 290

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	SEX	AGE	NOMS	ANNÉE de naissance	LIEU de naissance	NATIONALITÉ	SITUATION par rapport au chef de ménage	PROFESSION
144	M	45	Chévenet Charles	1813	Montrigny Corrèze	F	chef	Néant
145	F	46	Théveuet Marguerite	1815	Nouluner P Loire	F	épouse	Néant
146	F	47	Legouet Marie Antré	1866	Laloure de G	F	épouse	Néant
147	F	48	Jolivet Marie	1910	Venas	F	chef	couturière
148	F	49	Jolivet Simon	1908	Venas	F	frère	Mécanicien
149	F	50	V.Jaulines Marie	1852	Marigoy	F	grand mère	Néant
150	F	51	Lamarque Virginie	1870	Douvrv ord	F	chef	Néant
151	F	52	Sabathie Joseph E desnoud	1883	Toulonne	F	chef	Receveur Enregistrement
152	F	53	Sabathie Marguerite	1883	Gremoble	F	épouse	Néant
153	F	54	Ballaire Albert	1884	Vemennes	F	chef	profice de Pait
154	F	55	Ballaire Adélaïde	1876	Maillet	F	épouse	Néant
155	F	56	Rouy Louis	1895	Bronde	F	chef	Platrice peintre
156	F	57	Rouy Emilienne	1896	Hérison	F	épouse	Néant
157	F	58	Rouy Jean Robert	1920	Herisses	F	frères	apprenti platre
158	F	59	Lepée Louis	1893	Bizenemille	F	chef	Agricalten
159	F	60	Pepie Marie	1903	Gaumond	F	épouse	Néant
160	F	61	Pepce Jean	1925	Henison	F	frères	Néant
161	F	62	Lepée Edmond	1928	Heinon	F	frères	Néant
162	F	63	Guillonnet Henri	1910	Herison	F	chef	Camier
163	F	64	Guillomet Marguerite	1913	Chaubeux	F	épouse	Niant
164	F	65	Guillonnet Philippe	1882	Estesancies	F	père	Entreprenent frances frobles
165	F	66	V.Veuat Adélaïde	1882	Venisa P	F	chef	Néant
166	F	67	Riotte Jean	1911	Heinon	F	dat. civil	Courreur
167	F	68	Ve Martinet Marie	1860	Venise	F	chef	ménagère
168	F	69	V.Lachasagne Clarice	1870	Heinon	F	chef	Néant
169	F	70	Lachanagne Georges	1892	Vevas	F	frères	Electricien
170	F	71	V.Périnet Hélène	1862	Charbens	F	chef	Néant
171	F	72	Bergerat Jean	1864	Heinon	F	chef	jardinier
172	F	73	Bergerat Jean	1902	Heinon	F	frères	Camis
173	F	74	V.Malochet Amilie	1865	Maillet	F	chef	Néant

nom Chévenet prénom Charles date_naissance 1813 lieux_naissance Montrigny Corrèze relation P profession chef profession Néant

nom Théveuet prénom Marguerite date_naissance 1815 lieux_naissance Nouluner P Loire relation F profession épouse profession Néant

nom Legouet prénom Marie Antré date_naissance 1866 lieux_naissance Laloure de G état_civil V relation chef profession Néant

nom Jolivet prénom Marie date_naissance 1910 lieux_naissance Venas nationalité F 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 Sabathie 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 Bronde 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 Henison nationalité F relation frères profession Néant

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

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

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

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

nom V.Veuat prénom Adélaïde date_naissance 1882 lieux_naissance Venisa P relation chef profession Néant

nom Riotte prénom Jean date_naissance 1911 lieux_naissance Heinon 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 Heinon 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 Heinon nationalité F relation chef profession jardinier employeur Dr Benon

nom Bergerat prénom Jean date_naissance 1902 lieux_naissance Heinon 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



IIF / **Cantaloupe IIF Docker image** 
Docker Image from Teklia for the Cantaloupe IIF image server



IIF / **Cantaloupe Multiple S3 Buckets**   GNU General I
Delegate script to support multiple S3/Minio buckets

Arkindex : pour qui ?

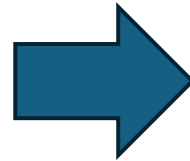


Personnalisation

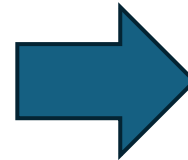
Traiter tout type de documents



Passage à l'échelle



Puissance
Complexité



Expert



Documentation
Formation
Développement
UI/UX



Pictoria ?