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

Journée-atelier “IA et images en SHS”

15 mai 2024

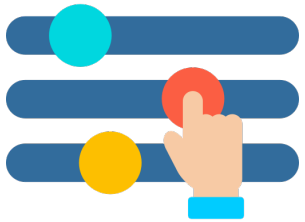
Christopher Kermorvant

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

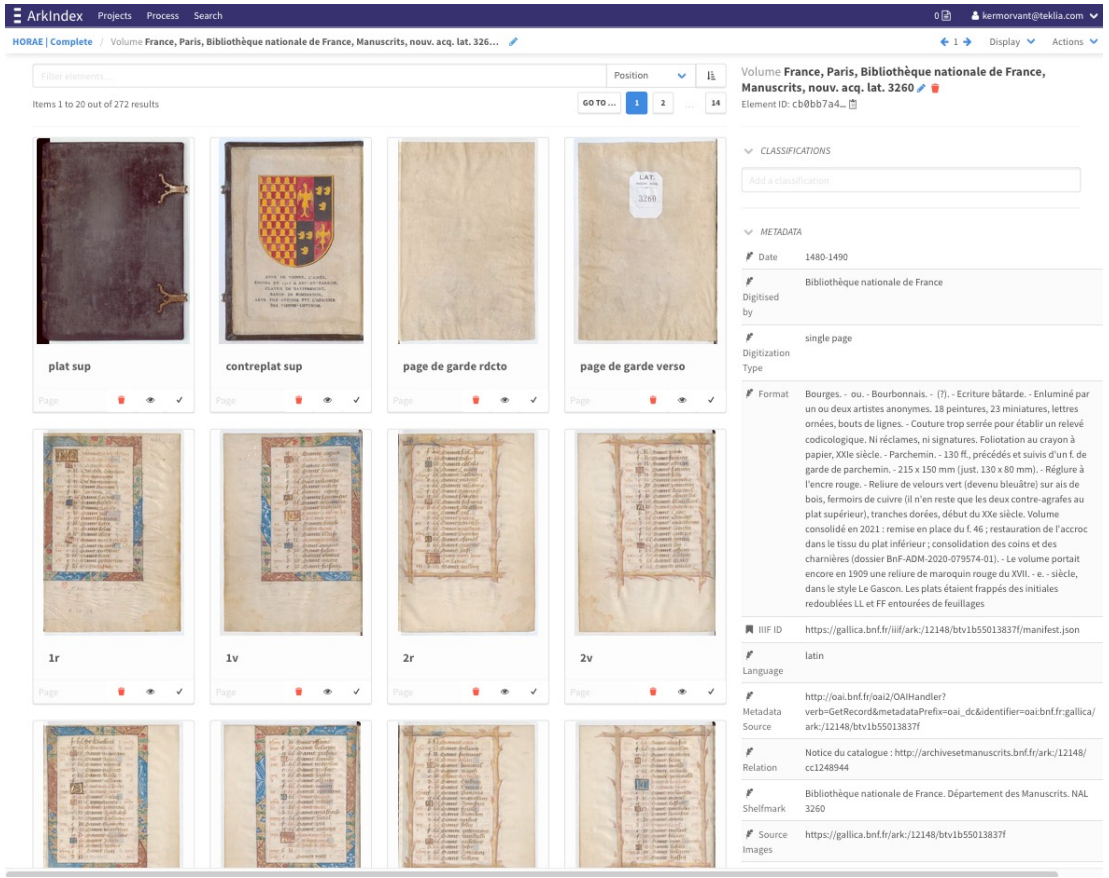
Open-source

Diffusion et participation de la communauté

PictorIA ?

Arkindex : Stockage et gestion des documents

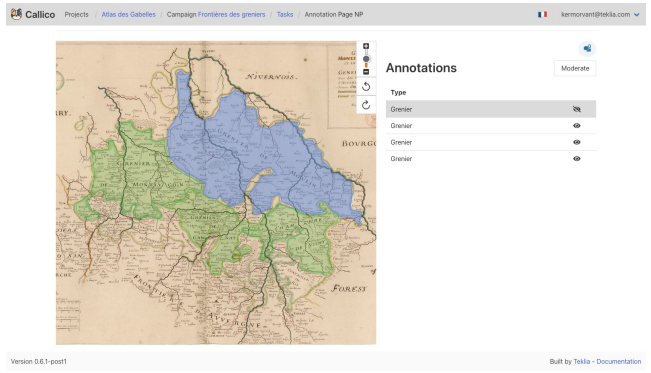
- Import web (petit corpus), S3 (gros corpus) ou par manifest IIIF
- Support des formats images et PDF
- Structuration hiérarchique des corpus complètement adaptable
- Gestion des métadonnées à tous les niveaux
- Visualisation, navigation



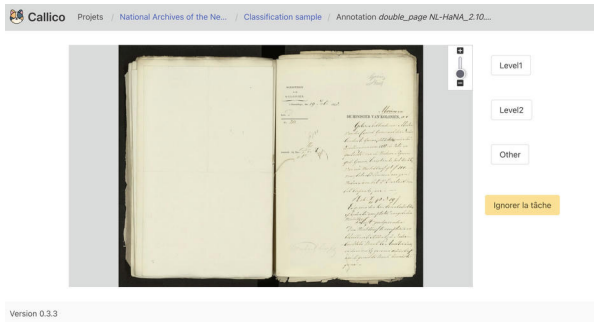
https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json

Callico : Annotation et validation

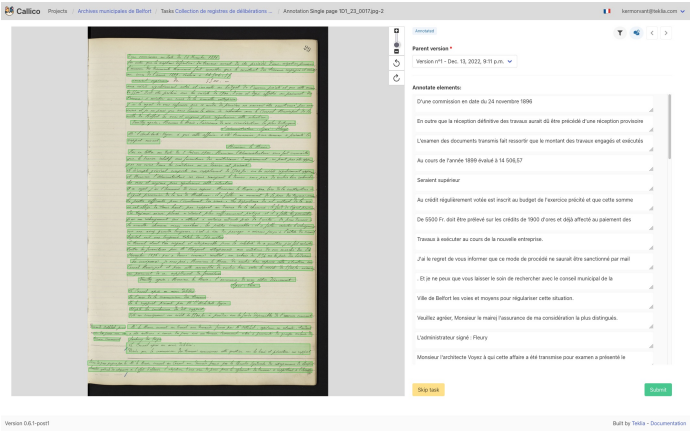
Zonage



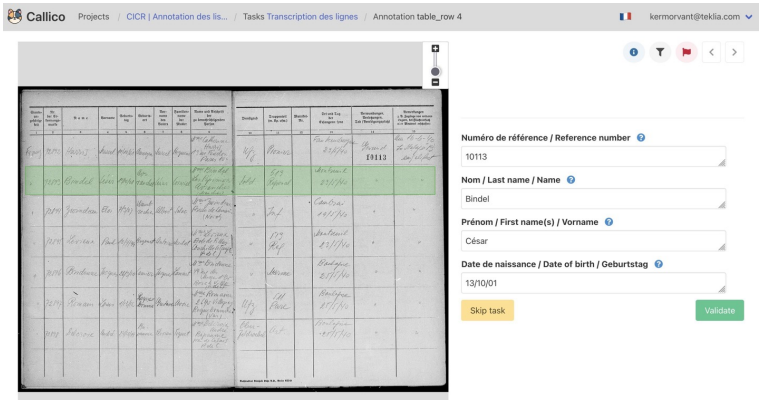
Classification



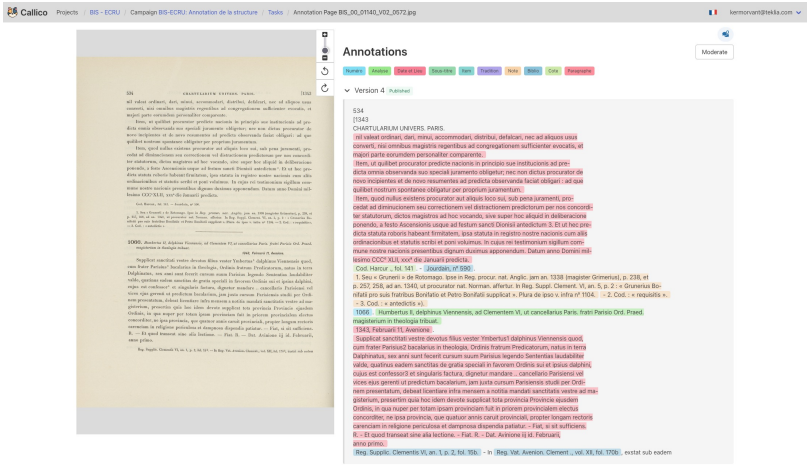
Transcription



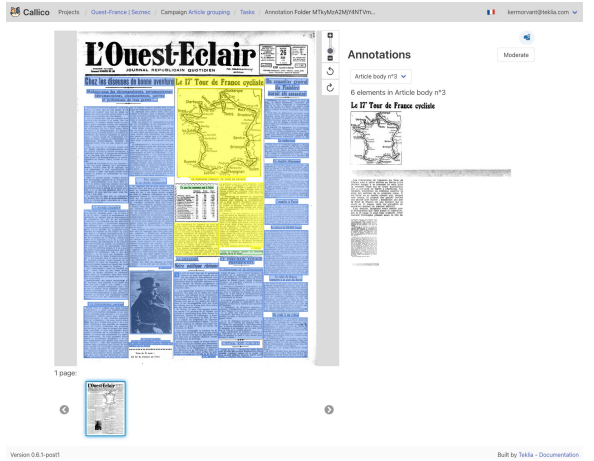
Clé-valeur



Entités nommées

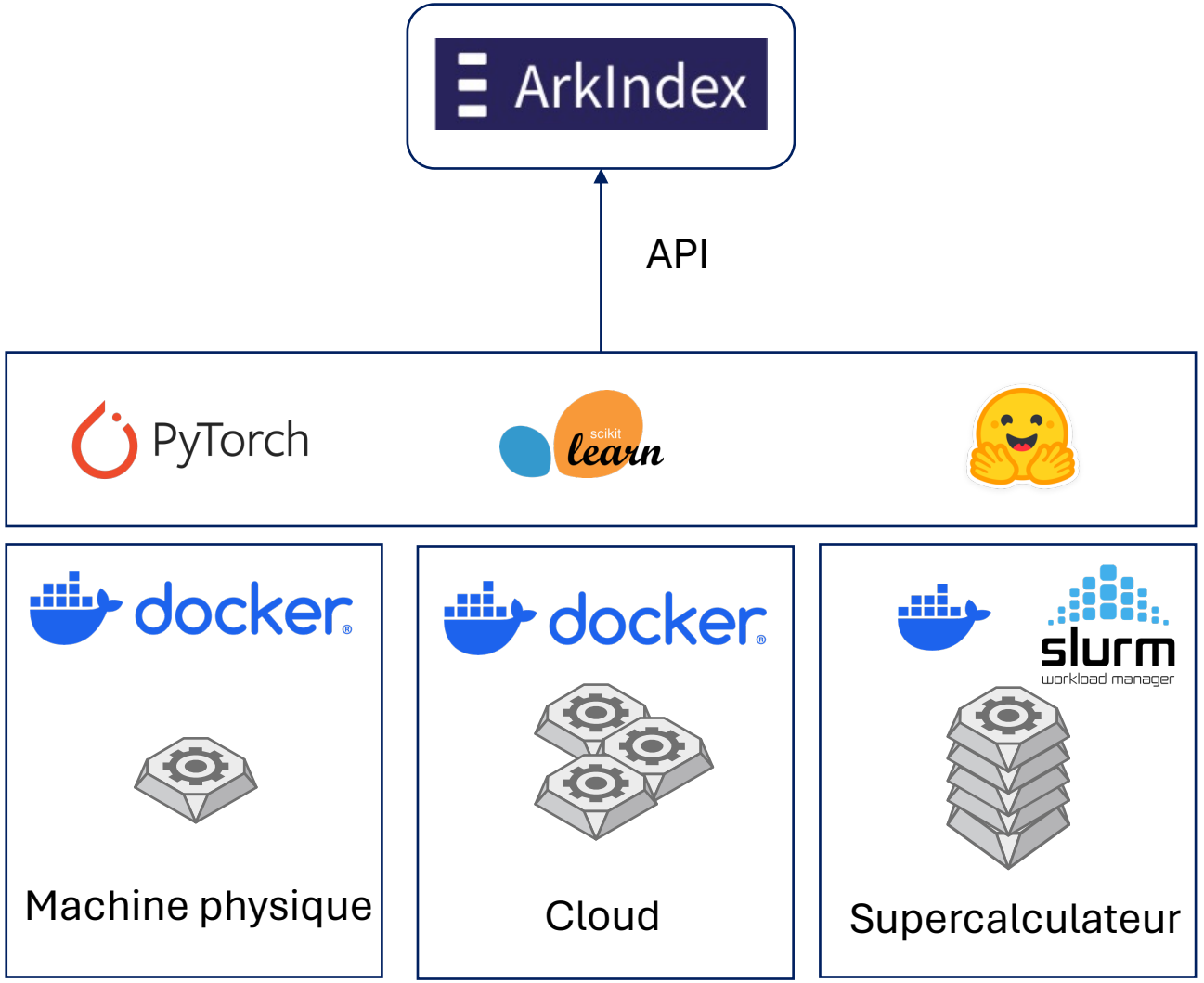


Groupement



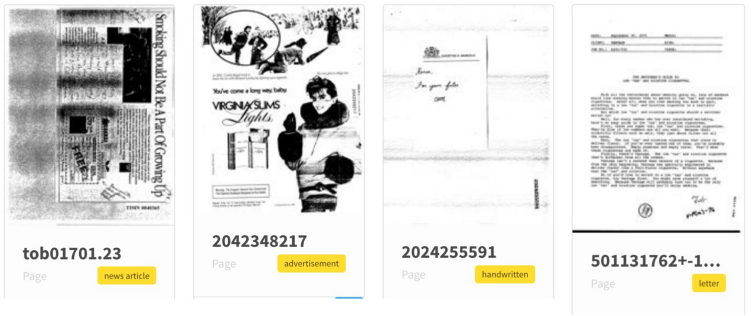
Arkindex : Intégration de modèles/algorithmes

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

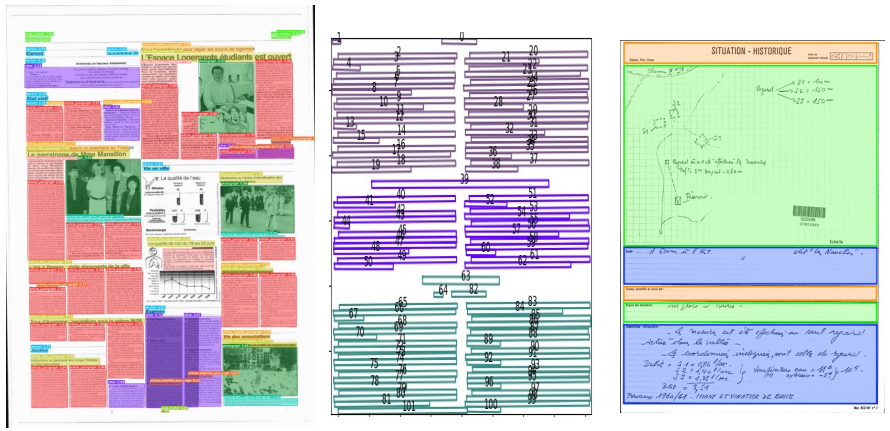


Arkindex : applications

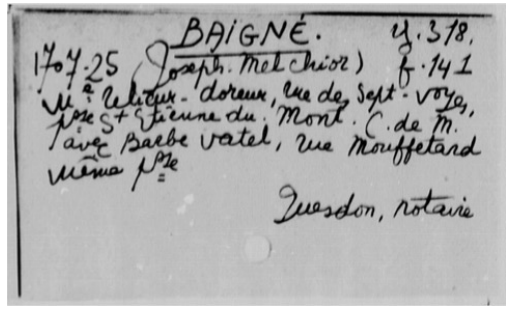
Classification



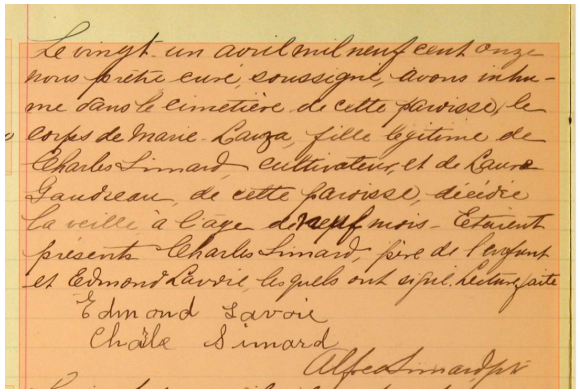
Structuration



OCR/HTR



Extraction d'entités



Le _DATE vingt - un avril mil neuf cent onze ✓ × nous prêtre curé
soussigné , avons inhu -
mé dans le cimetière de cette paroisse , le
corps de Marie - _PERSONNE Caronza , fill ✓ × e légitime de
Charles Simard _PROFESSION cultivateur ✓ × , et de Laura
_PERSONNE Gaudreau de cet ✓ × te paroisse , décédée
la veille a l ' âge de _AGE septf mois ✓ × . Etaient
présents _PRENOM Charles ✓ × Simard , père de l ' enfant
et Edmond _PERSONNE Davidré , les ✓ × quels ont signé .
Lecture faite .

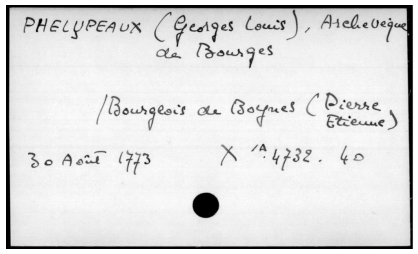
Analyse de photos

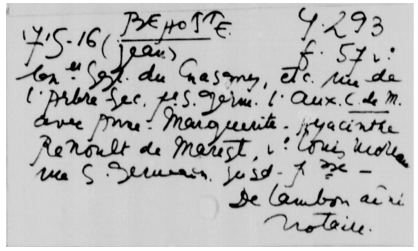
tag	group of people
tag	samourai
tag	map
tag	swords
tag	kimono
tag	potted plant

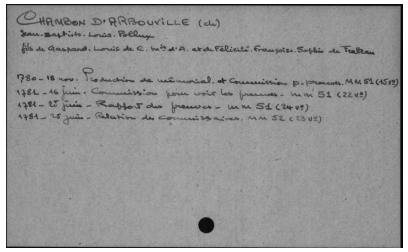
Archives Nationales
SIMARA, 2021-2023

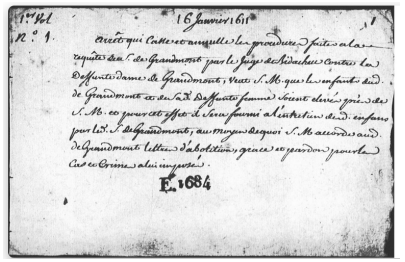
Traitement de 400 000 fiches d'index

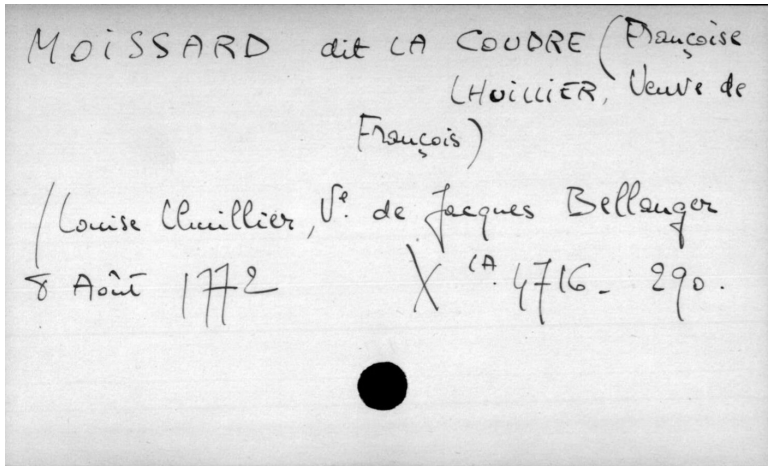
Reconnaissance d'écriture manuscrite, extraction d'entités











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

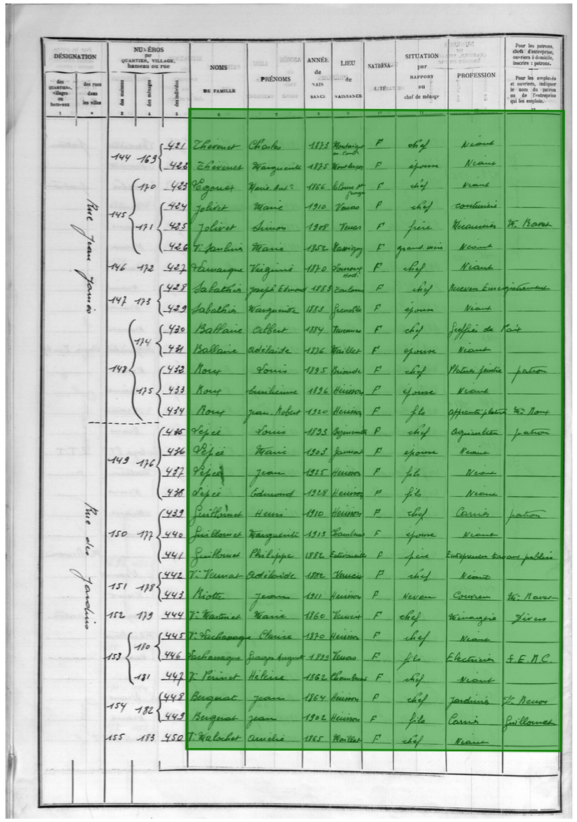
COTE_SERIE X1A COTE_ARTICLE 4716

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)

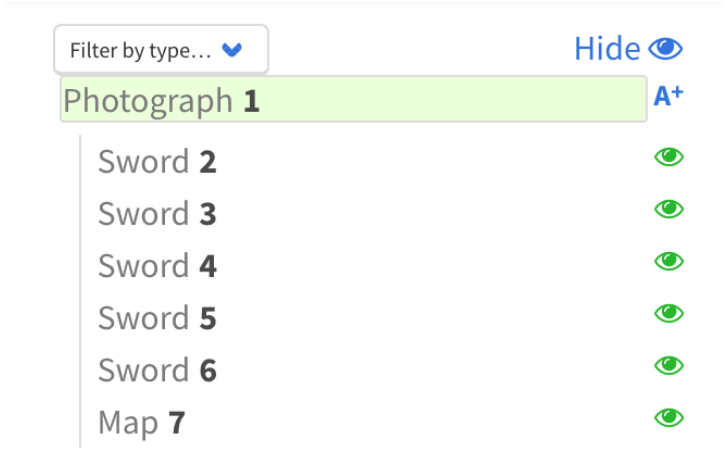


nom Chévenet prénom Charles date_naissance 1873 lieux_naissance Montrigny Corrèze relation P profession chef profession Néant
nom Théveuet prénom Marguerite date_naissance 1875 lieux_naissance Nouluner P Loire relation F profession épouse profession Neeant
nom Legouet prénom Marie Antré date_naissance 1866 lieux_naissance Laloure de G état_civil V relation chef profession néant
nom Jolivet prénom Marie date_naissance 1910 lieux_naissance Venas nationalité 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 Mariigoy nationalité F relation grand mère profession Néant
nom Lamarque prénom Virginie date_naissance 1870 lieux_naissance Douvrz 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 Sabalhier 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 proffice 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 Brionde nationalité F relation chef profession Platrice peintre employeur patron
nom Rouy prénom Emilienne date_naissance 1896 lieux_naissance Hérisson nationalité F relation épouse profession Néant
nom Roux prénom Jean Robert date_naissance 1920 lieux_naissance Herisses nationalité F relation fles 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 Neant
nom Pepce prénom Jean date_naissance 1925 lieux_naissance Henison nationalité F relation fils profession Néant
nom Lepée prénom Edmond date_naissance 1928 lieux_naissance Heinon nationalité P relation fils employeur 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 Guillomet prénom Philippe date_naissance 1882 lieux_naissance Estesancies nationalité F relation père profession Entrepresent frances froblies
nom V Veunat prénom Adèlaide date_naissance 1882 lieux_naissance Venisa P relation chef profession Néant
nom Riotte prénom Jean date_naissance 1911 lieux_naissance Heinon nationalité F état_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 Veuas nationalité F relation f es profession E lectricien 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 Heinoon nationalité F relation chef profession jardinier employeur Dr Benon
nom Bergerat prénom Jean date_naissance 1902 lieux_naissance Heinon nationalité F relation fils 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



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

https://teklia.com/research/open-source/

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

Arkindex / **Backend**

Python3/Django backend

Callico / **Callico**

Annnotation platform, companion of Arkindex

Arkindex / **Frontend**

Development & Tests frontend

C

IIIF / **Cantaloupe IIIF Docker image**

Docker Image from Teklia for the Cantaloupe IIIF image server

C

IIIF / **Cantaloupe Multiple S3 Buckets** GNU General

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 ?