Traitement automatique de documents numérisés avec Arkindex et Callico

Séminaire "Plateformes collaboratives et IIIF"

BNF – 23 janvier2025

Christopher Kermorvant - TEKLIA

Arkindex/Callico : Principes

Plateforme développée par TEKLIA depuis 2017
Utilisée en interne pour plus de 80 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

https://doc.arkindex.org/

Arkindex : fonctionnalités

Importer

Stocker

Organiser

Enrichir (humain/machine)

Visualiser

Exporter

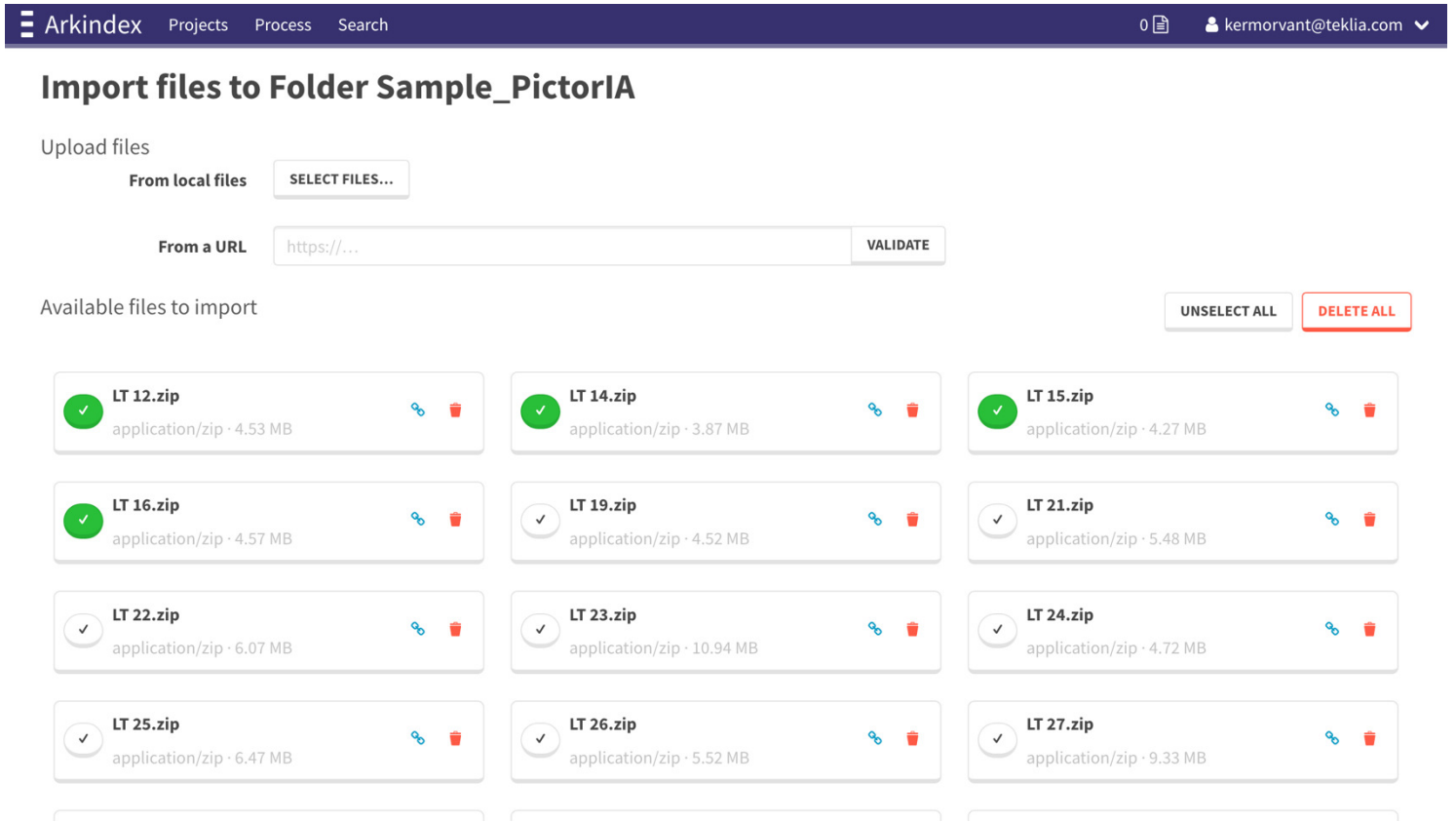
n’importe quel type de documents

Arkindex : Importer

Import de fichiers
- Images
- PDF image et texte
- ZIP
- Manifeste IIIF
- URL de manifeste IIIF

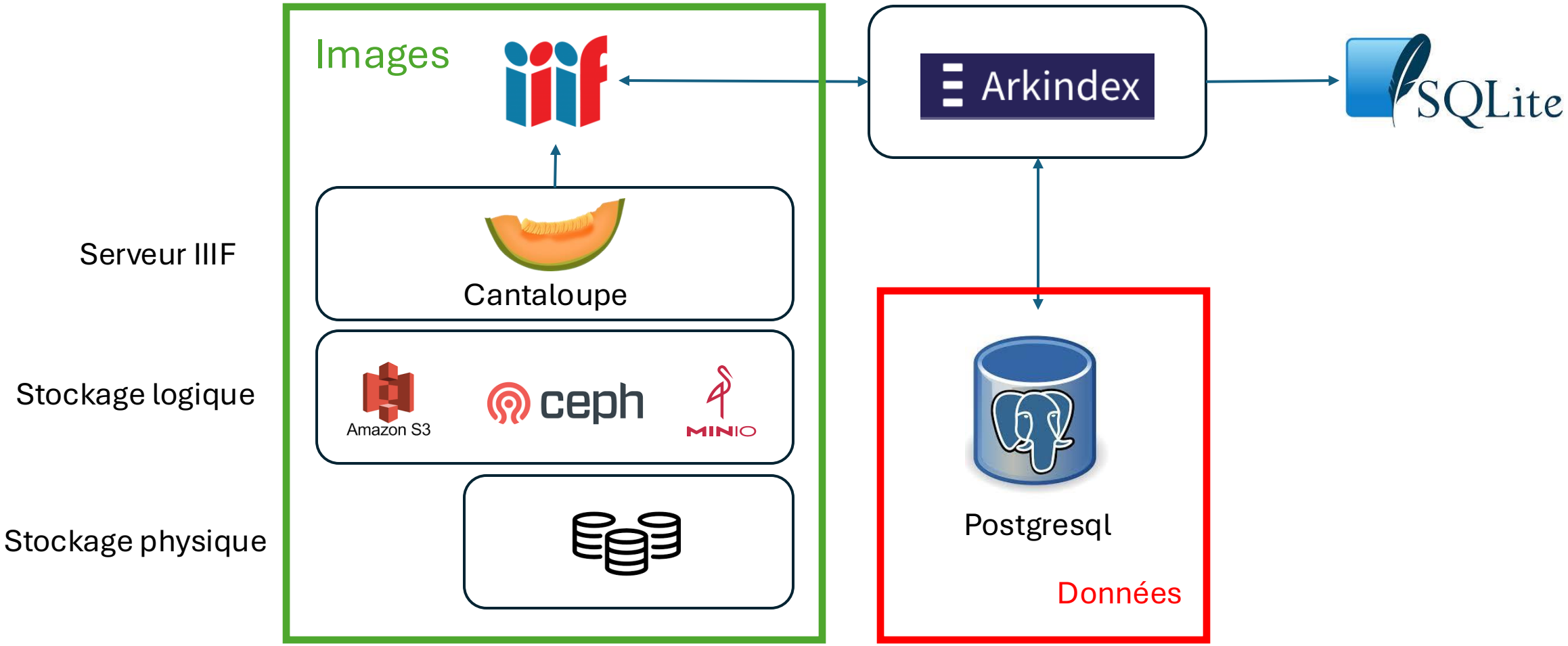
Import en masse
- À partir d'un serveur de stockage S3

Import en CLI
- pageXML
- ALTO & METS aLTO



https://cli.arkindex.org

Arkindex : Stocker

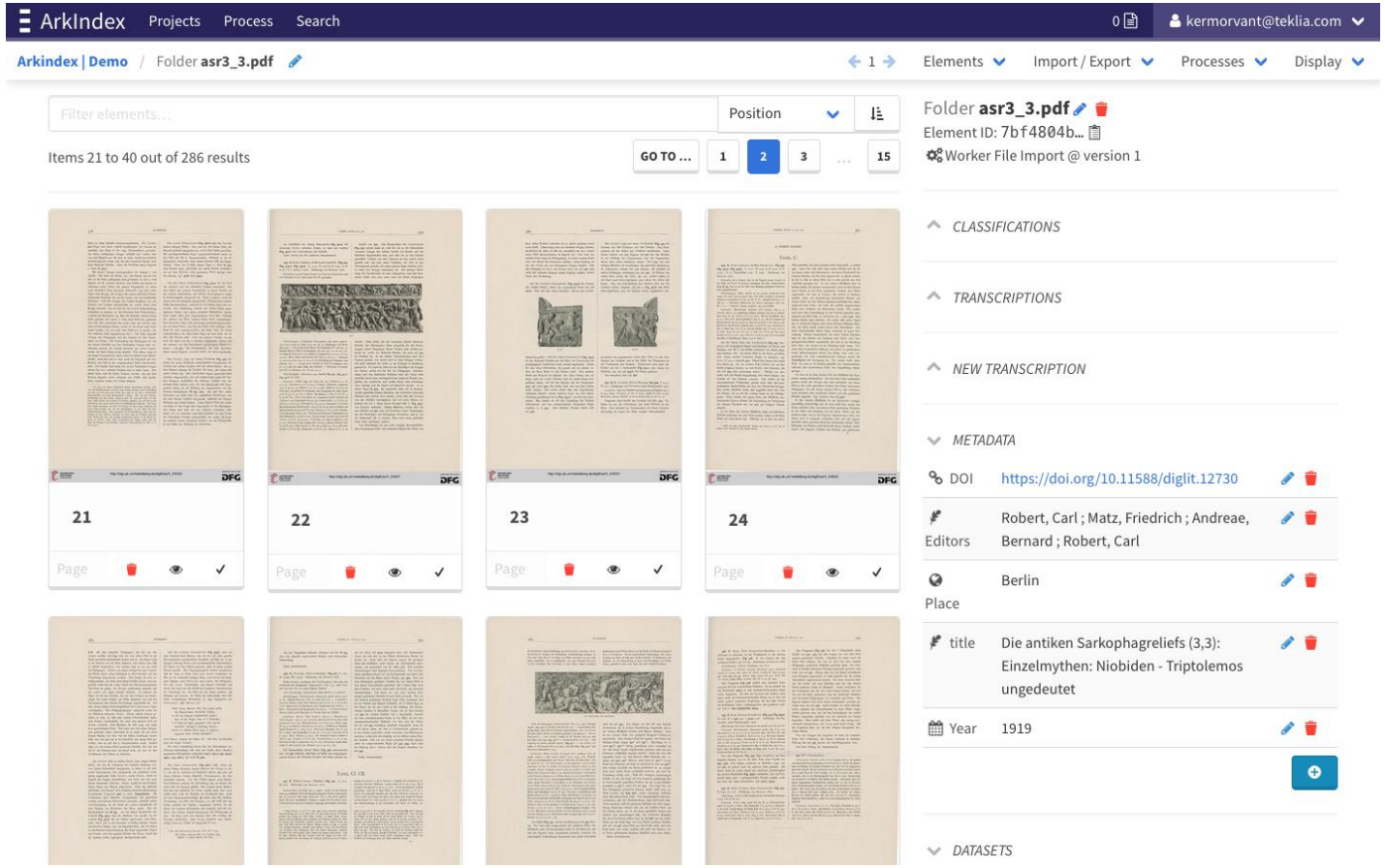


Arkindex : Organiser

Projet : gestion des droits, définitions des types d'éléments, etc.

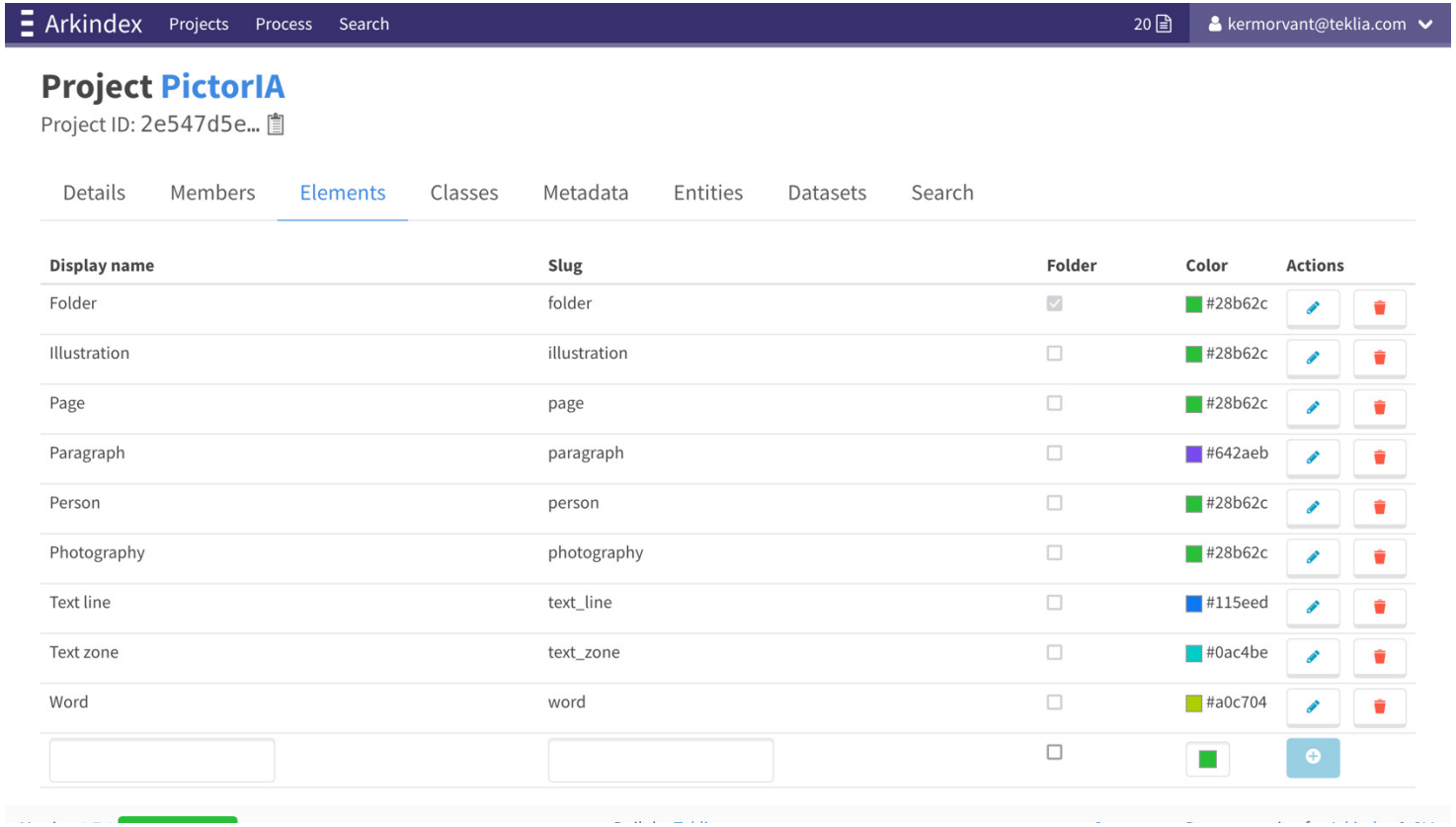
Page : type par défaut des images importées

Folder : type par défaut pour regrouper les images importées



Arkindex : Organiser

***Element* :** tout niveau de structure d'un document (physique/logique)



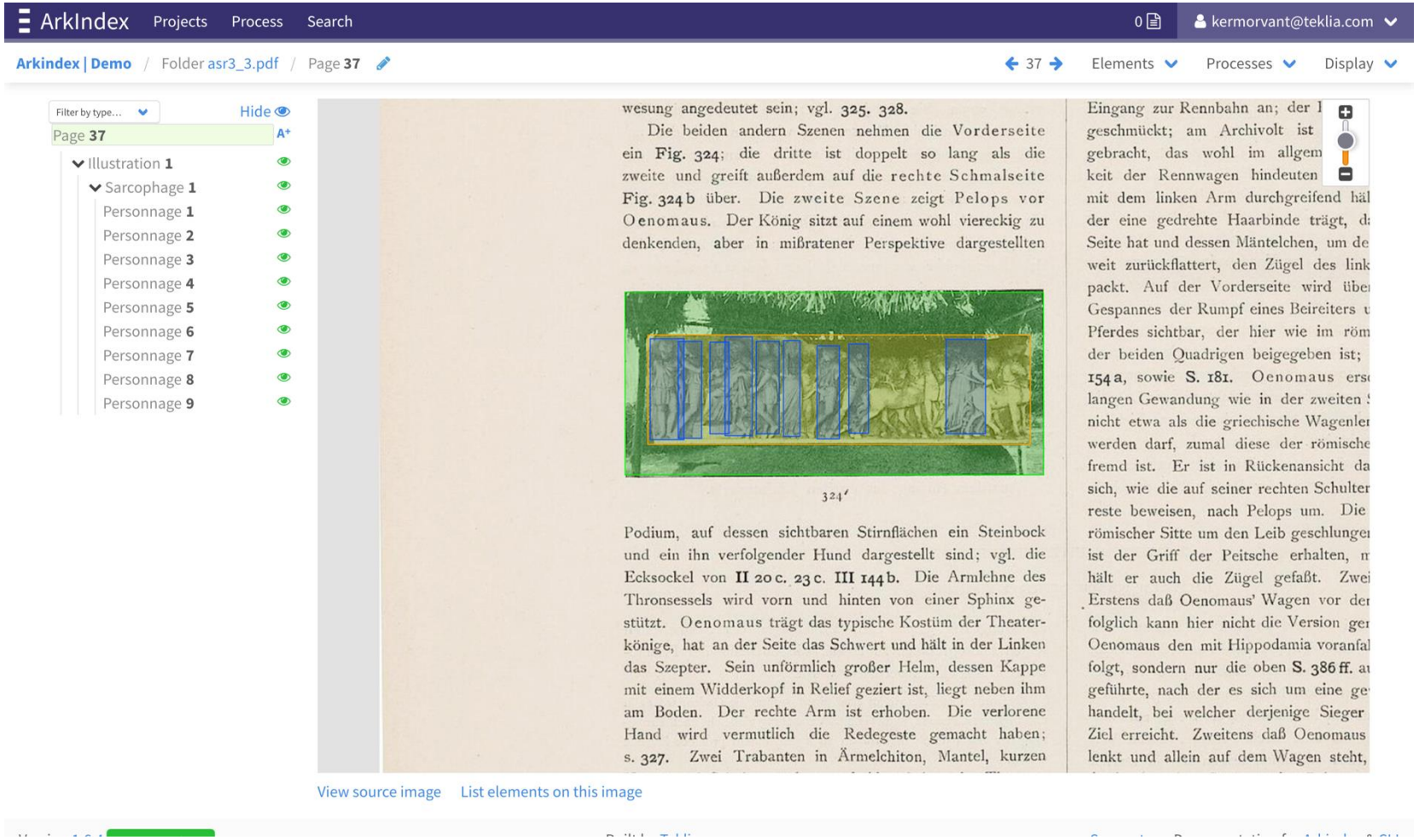
Nom affiché

Nom interne = slug

Arkindex : Enrichir - Hiérarchie

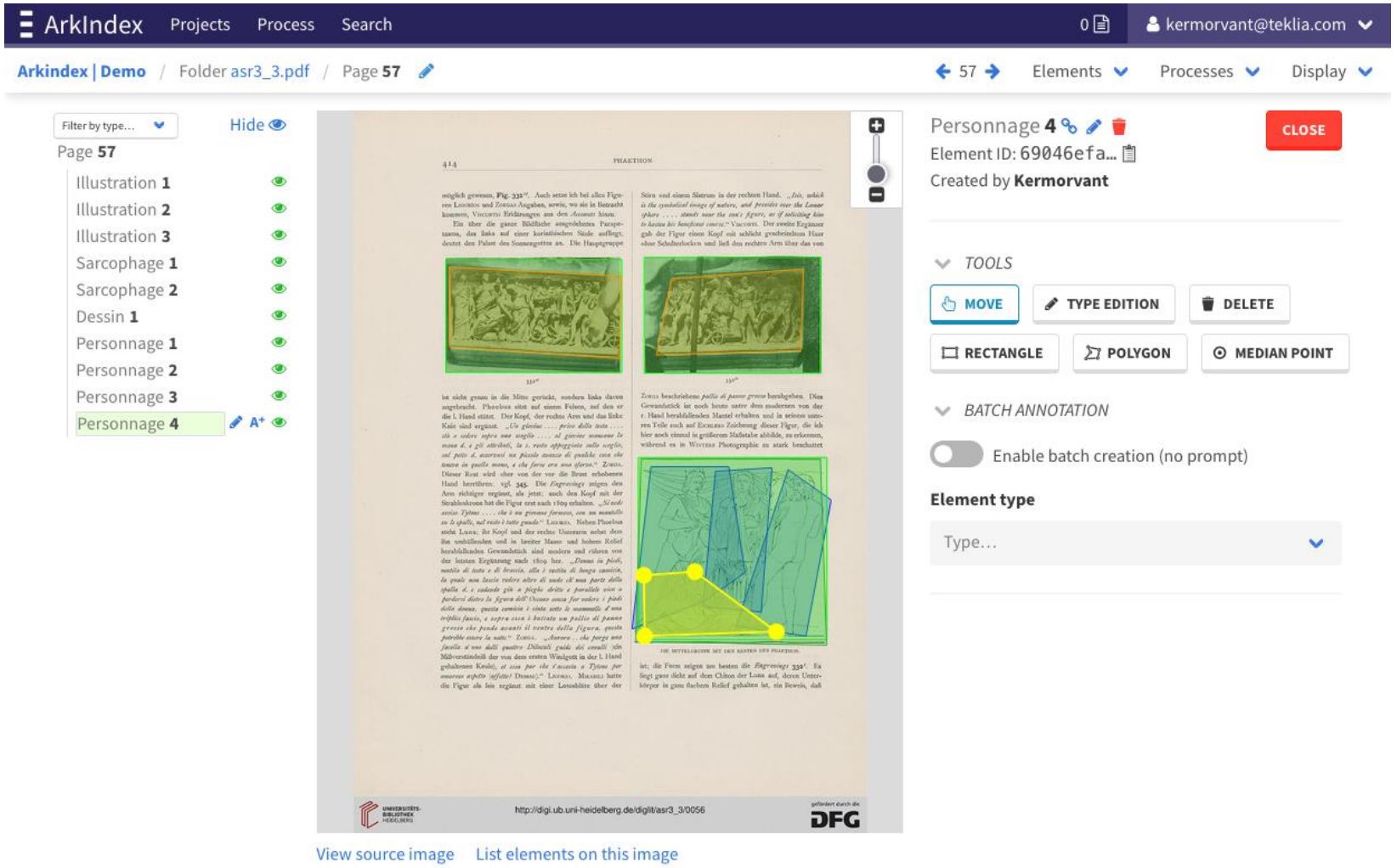
Annotation hiérarchique des éléments :

Folder
  Page
    Illustration
      Sarcophage
        Personnage



Arkindex : Enrichir

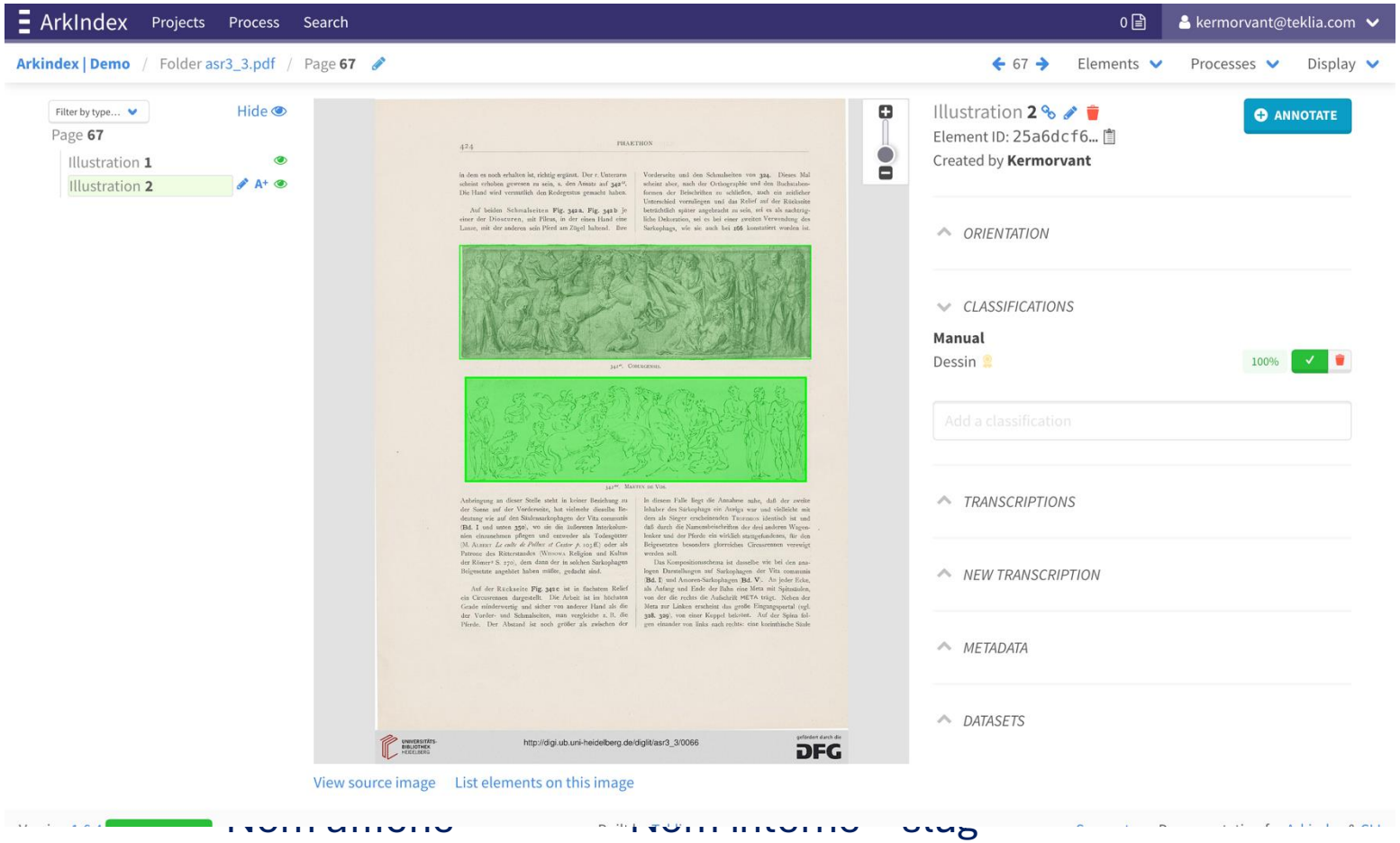
Création et Annotation des éléments



Arkindex : Enrichir - Classe

Les **classes** permettent de qualifier plus précisément un *element*

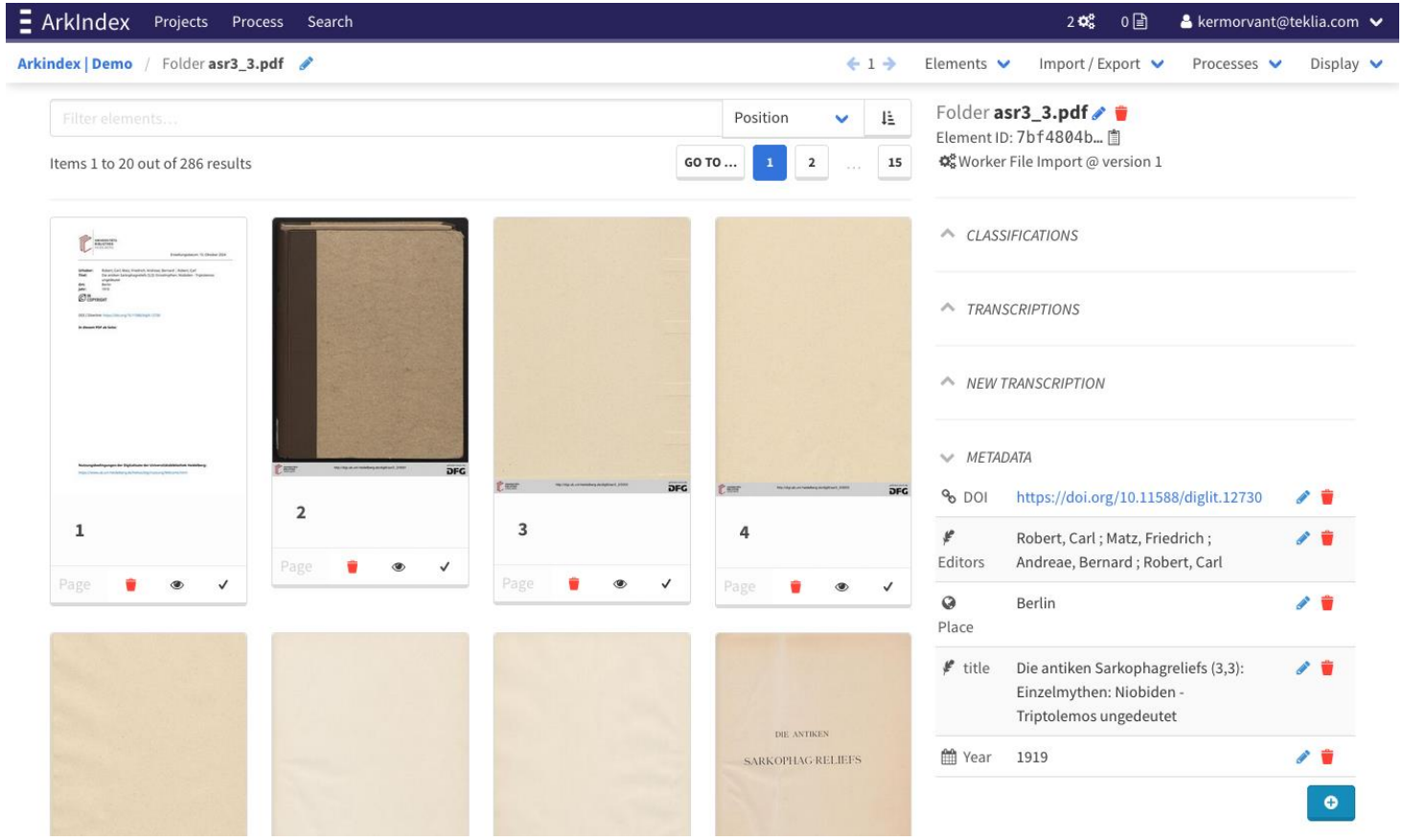
Element : illustration
Class : dessin ou photo



Arkindex : Enrichir - Métadata

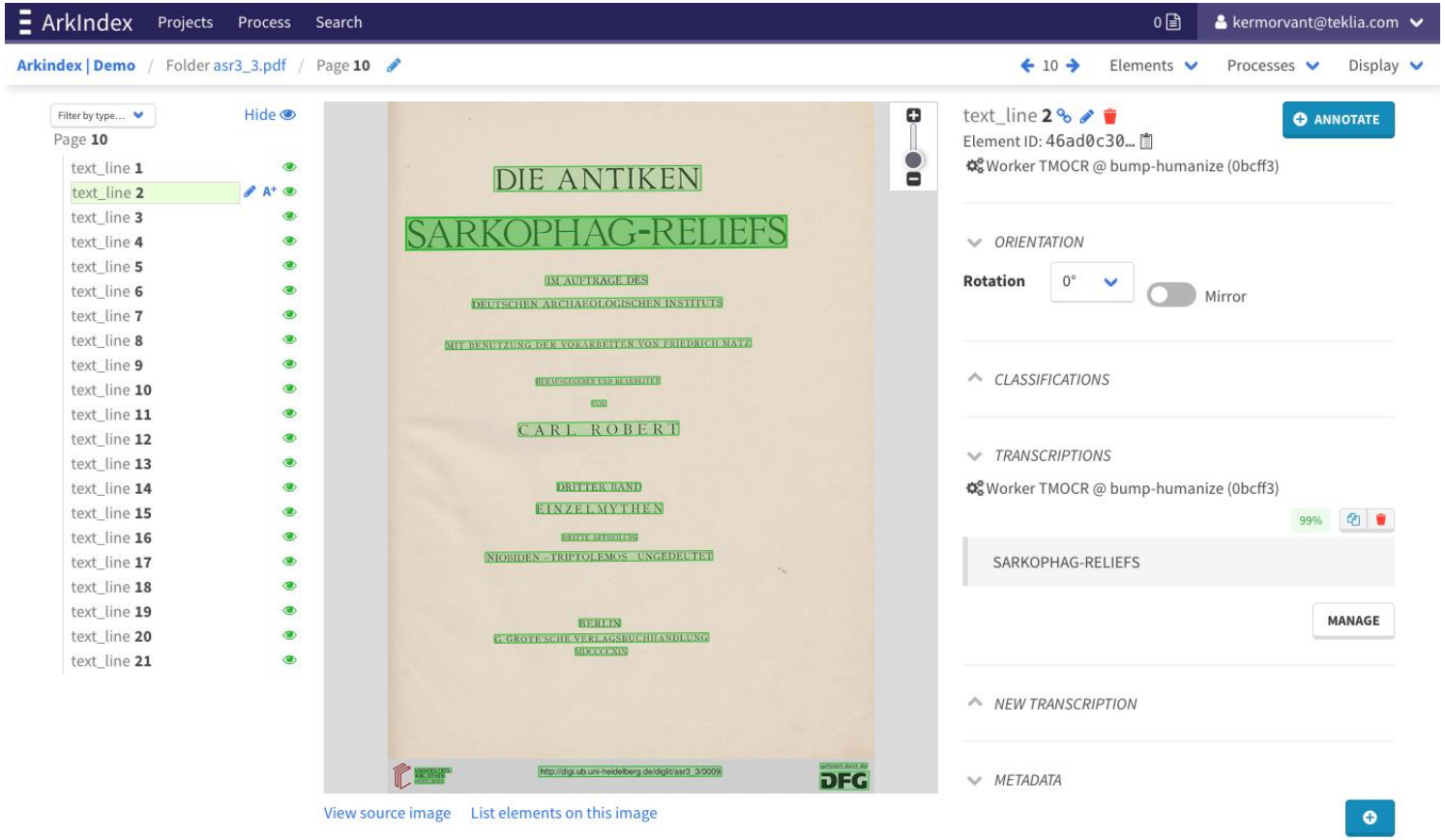
Méta-data : toute information supplémentaire

Types de metadata :
Text, Location, Date, URL, Reference, Numeric, Markdown



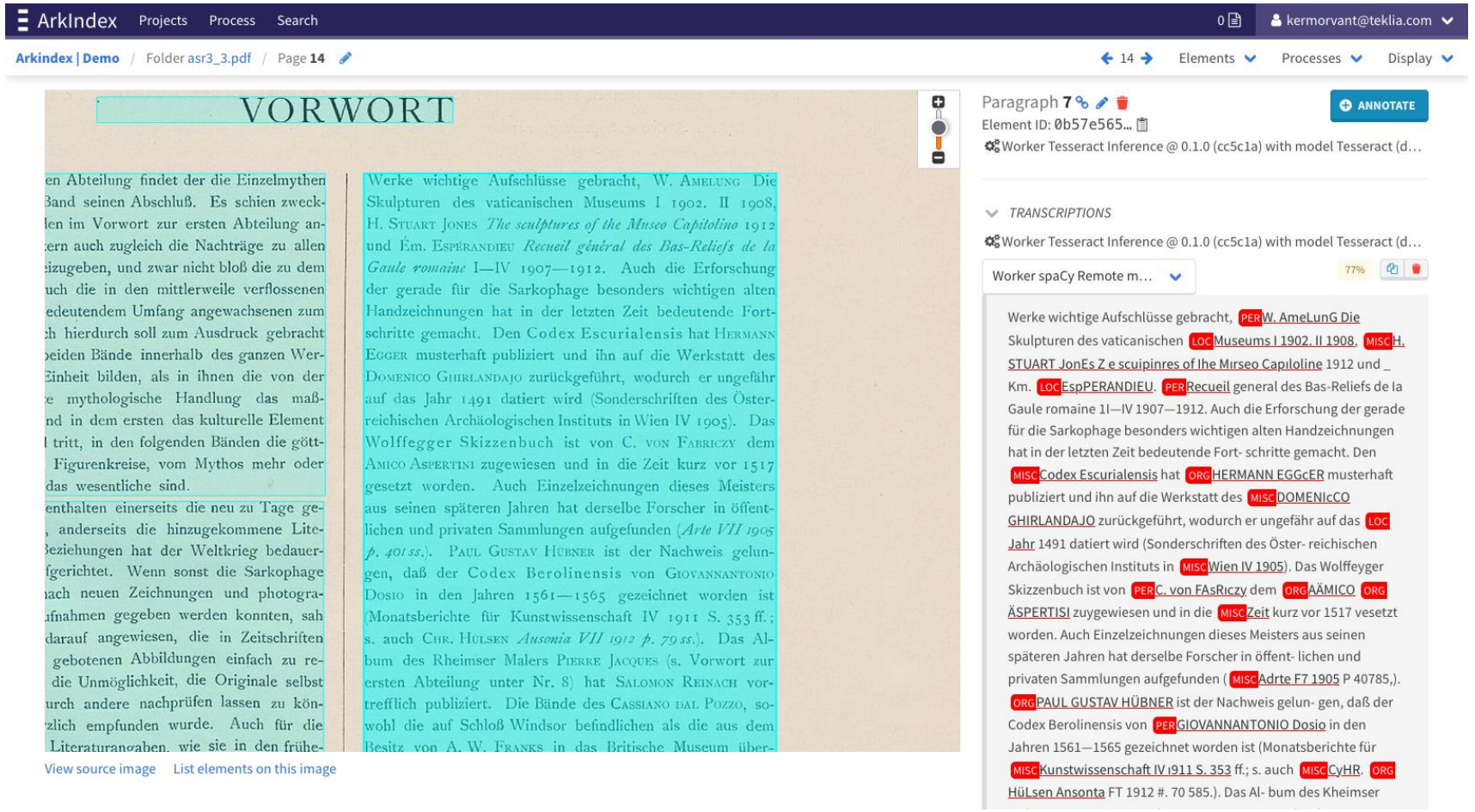
Arkindex : Enrichir - Transcription

Transcription du texte présent sur un élément



Arkindex : Enrichir – Entités nommées

Identification des **entités nommées** sur une transcription



Arkindex : Enrichir automatiquement

Processing type	Modules	Type d'algorithme
Text Recognition	Tesseract OCR	open-source
	Kraken HTR	open-source
	PyLaia OCR/HTR	open-source by TEKLIA
	PyLaia Training	open-source by TEKLIA
	DAN OCR/HTR	open-source by TEKLIA
	DAN Training	open-source by TEKLIA
	Google VisionOCR	proprietary API
	Microsoft OCR	proprietary API
Document Object Detection	Doc-UFCN	open-source by TEKLIA
	Doc-UFCN Training	open-source by TEKLIA
	Yolo segmenter	open-source
	Yolo segmenter training	open-source
	Mask-RCNN	open-source
	Grounding DINO	open-source
	Layout Parser (table detector)	open-source
Named-Entity Recognition	Spacy	open-source
	spaCy training	open-source
	Flair	open-source
	Flair training	open-source
	Stanza	open-source

Processing type	Modules	Type d'algorithme
Information Extraction	DAN	open-source by TEKLIA
	DAN IE training	open-source by TEKLIA
	Microsoft Document AI (forms processing)	proprietary API
	Microsoft Document AI (receipt processing)	proprietary API
Document Classifier	XGBoost text classifier/image	open-source
	XGBoost training	open-source
	Yolo classifier	open-source
	Yolo classifier training	open-source
	CLIP Text/image classifier	open-source
	Resnet classifier	open-source
Visual LLMs	OpenAI ChatGPT	proprietary API
	Anthropic Claude	proprietary API
Text Translation	MarianMT	open-source
	T5	open-source

+ integration de n'importe quel service ou bibliothèque dans un container docker

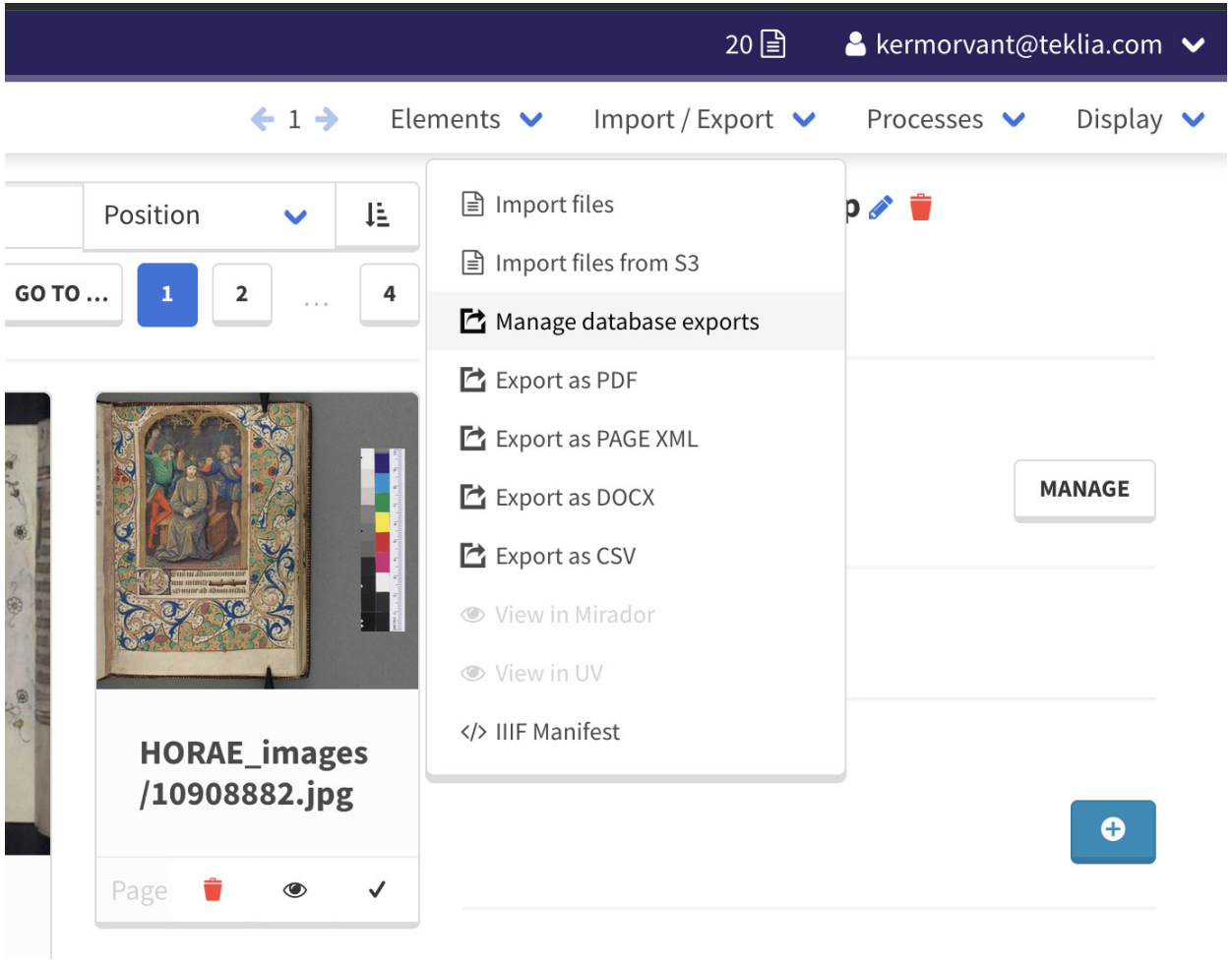
https://gitlab.teklia.com/workers/base-worker

Arkindex : Exporter

Export de fichiers
- PDF image et texte
- pageXML
- docx
- csv

Import de DB
- sqlite

Import en CLI
+ ALTO



Arkindex / Callico

ArkIndex

Plateforme de traitement automatique de documents

- Passage à l'échelle
- Tout type de traitements
- Tout type d'algorithmes

→ À destination d'utilisateurs avertis

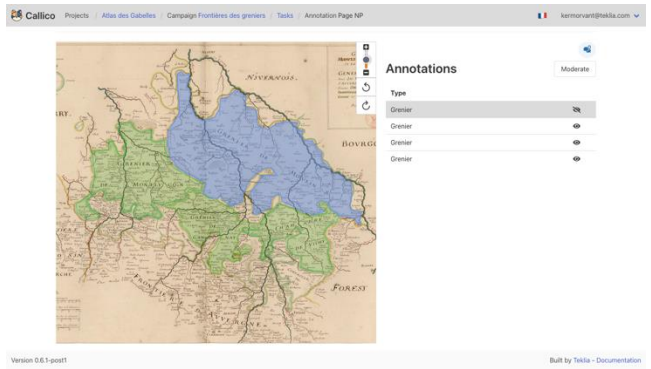
Application web d'annotation et validation de documents

- Multiples contributeurs
- Tâches simples et unitaires
- Gestion de campagnes d'annotation ou validation

→ Tout public

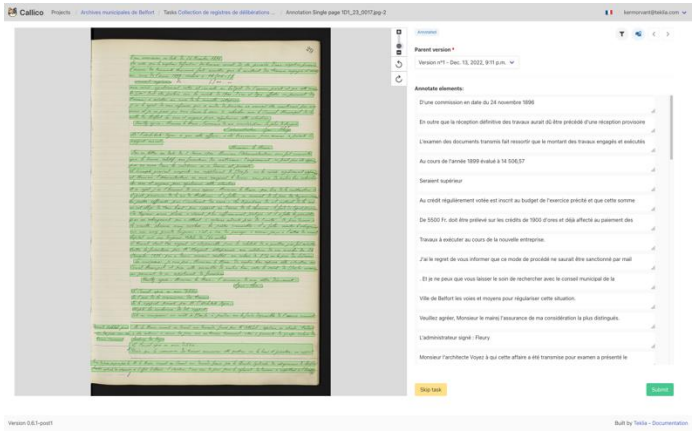
Callico : Annotation et validation

Zonage

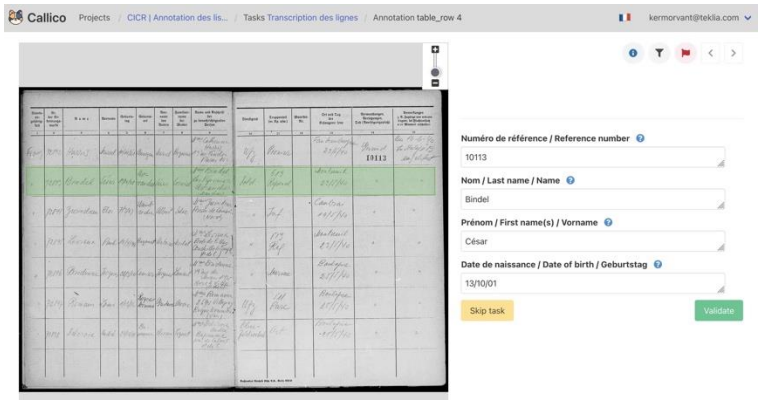


Classification

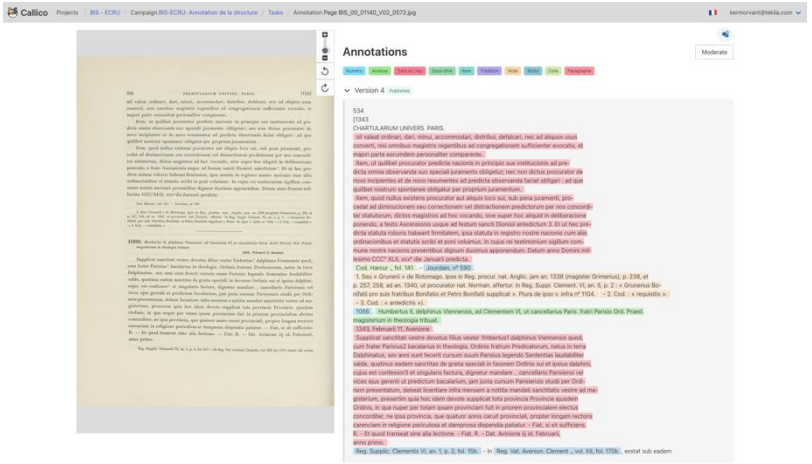
Transcription



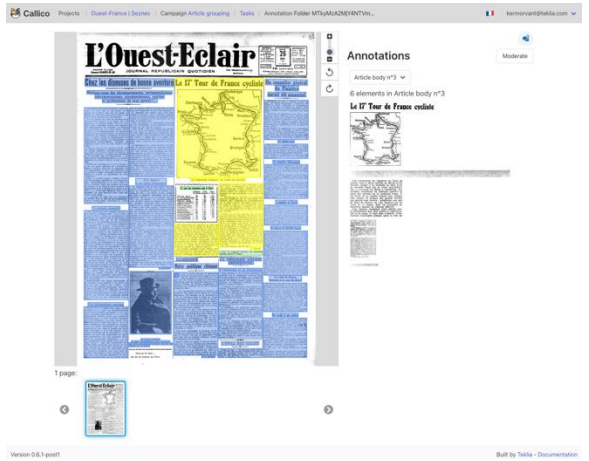
Clé-valeur



Entités nommées



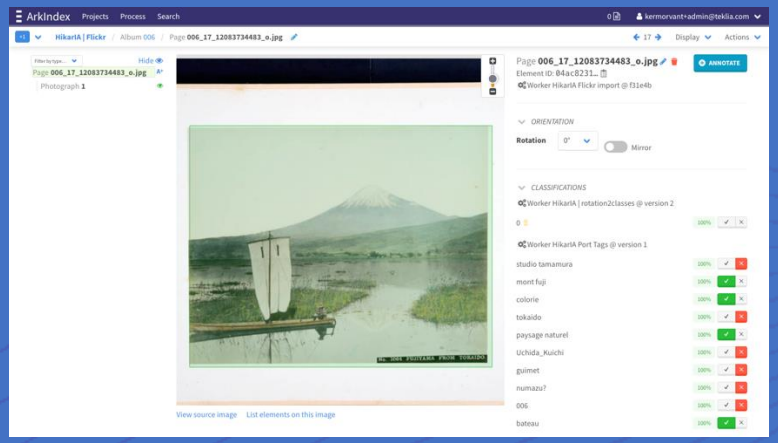
Groupement



TEKLIA's open-source software suite for document processing

Arkindex

Document processing

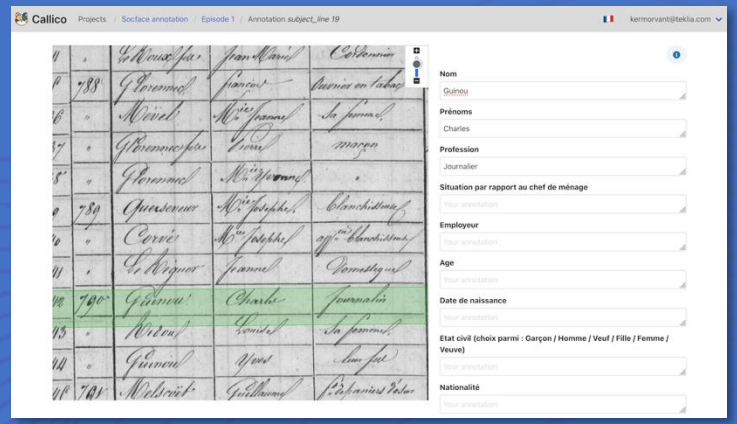


Open-source and Enterprise licences

https://gitlab.teklia.com

Callico

Collaborative annotation campaigns

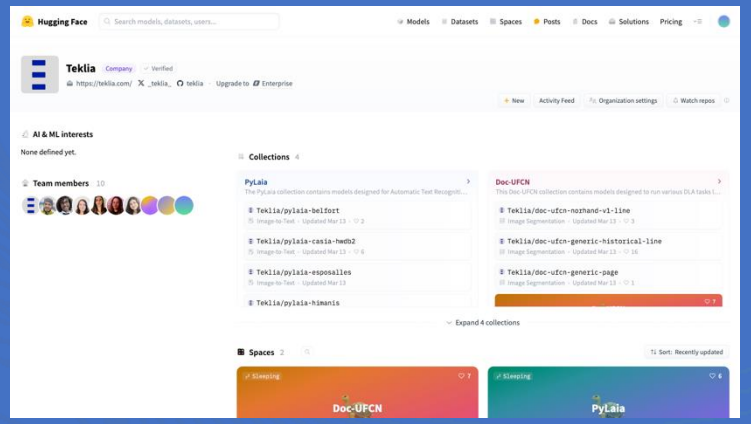


Open-source

https://support.teklia.com

HuggingFace

Models and datasets



Open-source, open-weights

https://huggingface.co/Teklia