

Traitement automatique de documents numérisés  
avec Arkindex et Callico

Séminaire "Plateformes collaboratives et IIF"

BNF – 23 janvier 2025

Christopher Kermorvant - TEKLIÀ

**T E K L I A**

# 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

# 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

Arkindex Projects Process Search 0 kermorvant@teklia.com

### Import files to Folder Sample\_Pictoria

Upload files

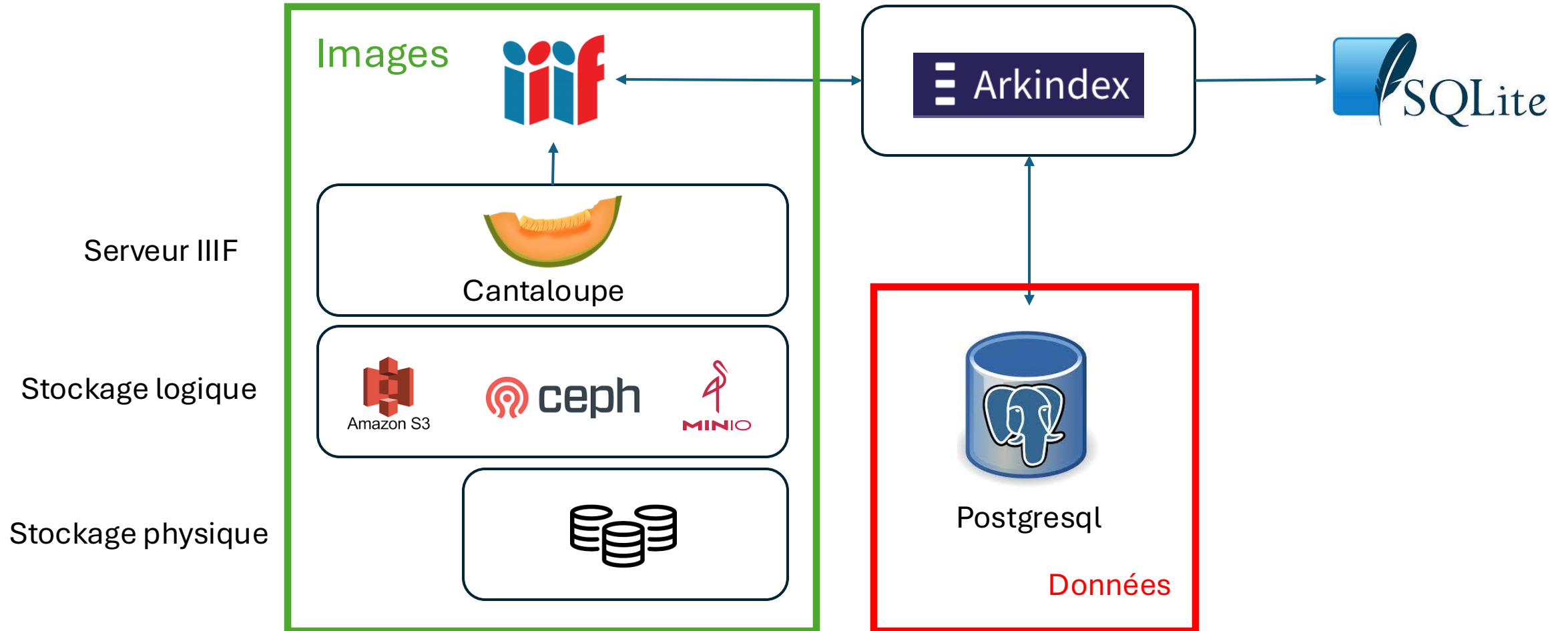
From local files

From a URL

Available files to import

<input checked="" type="checkbox"/> LT 12.zip application/zip · 4.53 MB	<input checked="" type="checkbox"/> LT 14.zip application/zip · 3.87 MB	<input checked="" type="checkbox"/> LT 15.zip application/zip · 4.27 MB
<input checked="" type="checkbox"/> LT 16.zip application/zip · 4.57 MB	<input type="checkbox"/> LT 19.zip application/zip · 4.52 MB	<input type="checkbox"/> LT 21.zip application/zip · 5.48 MB
<input type="checkbox"/> LT 22.zip application/zip · 6.07 MB	<input type="checkbox"/> LT 23.zip application/zip · 10.94 MB	<input type="checkbox"/> LT 24.zip application/zip · 4.72 MB
<input type="checkbox"/> LT 25.zip application/zip · 6.47 MB	<input type="checkbox"/> LT 26.zip application/zip · 5.52 MB	<input type="checkbox"/> LT 27.zip application/zip · 9.33 MB

# Arkindex : Stocker

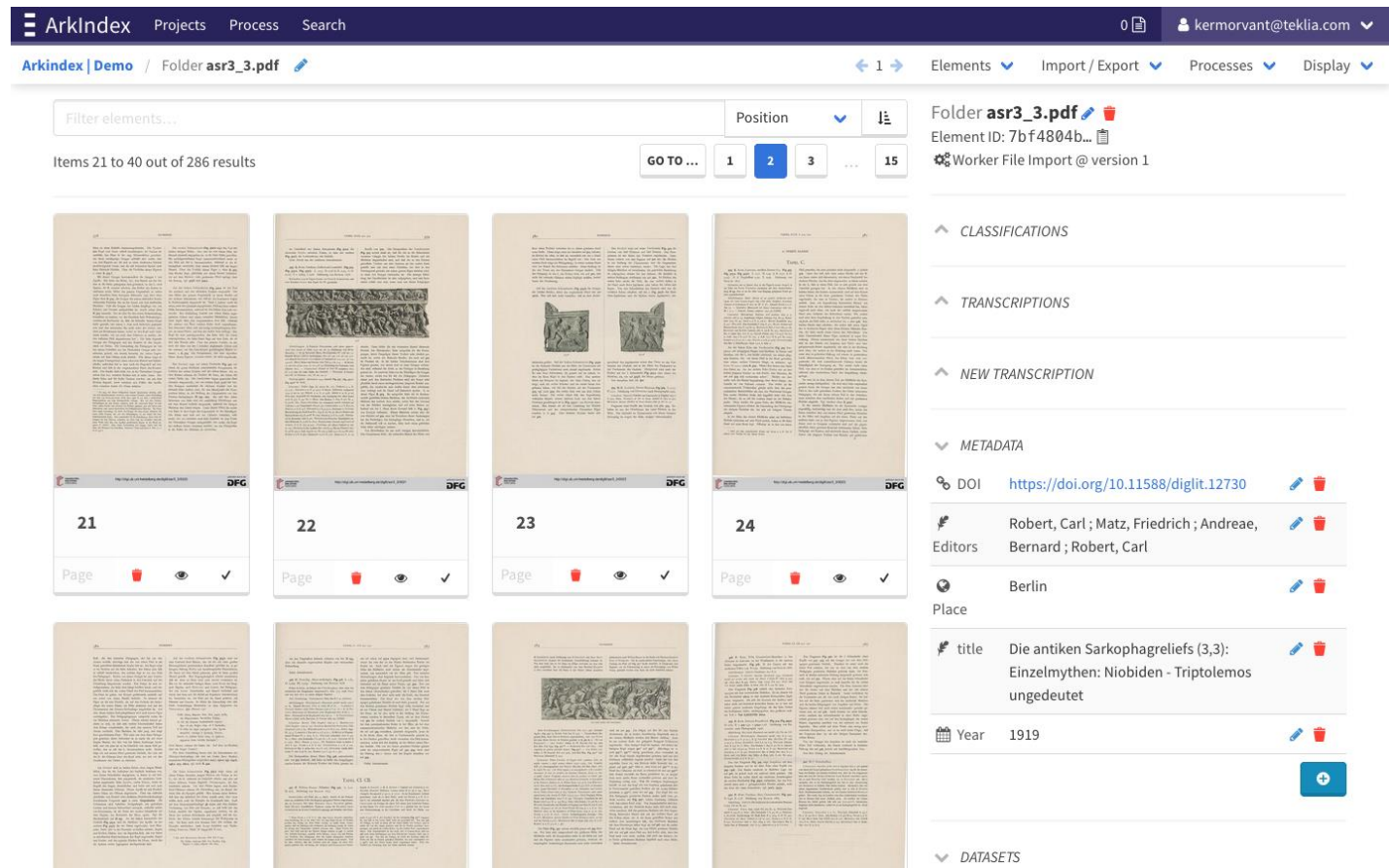


# Arkindex : Organiser

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

**Page** : type par défaut des images importées

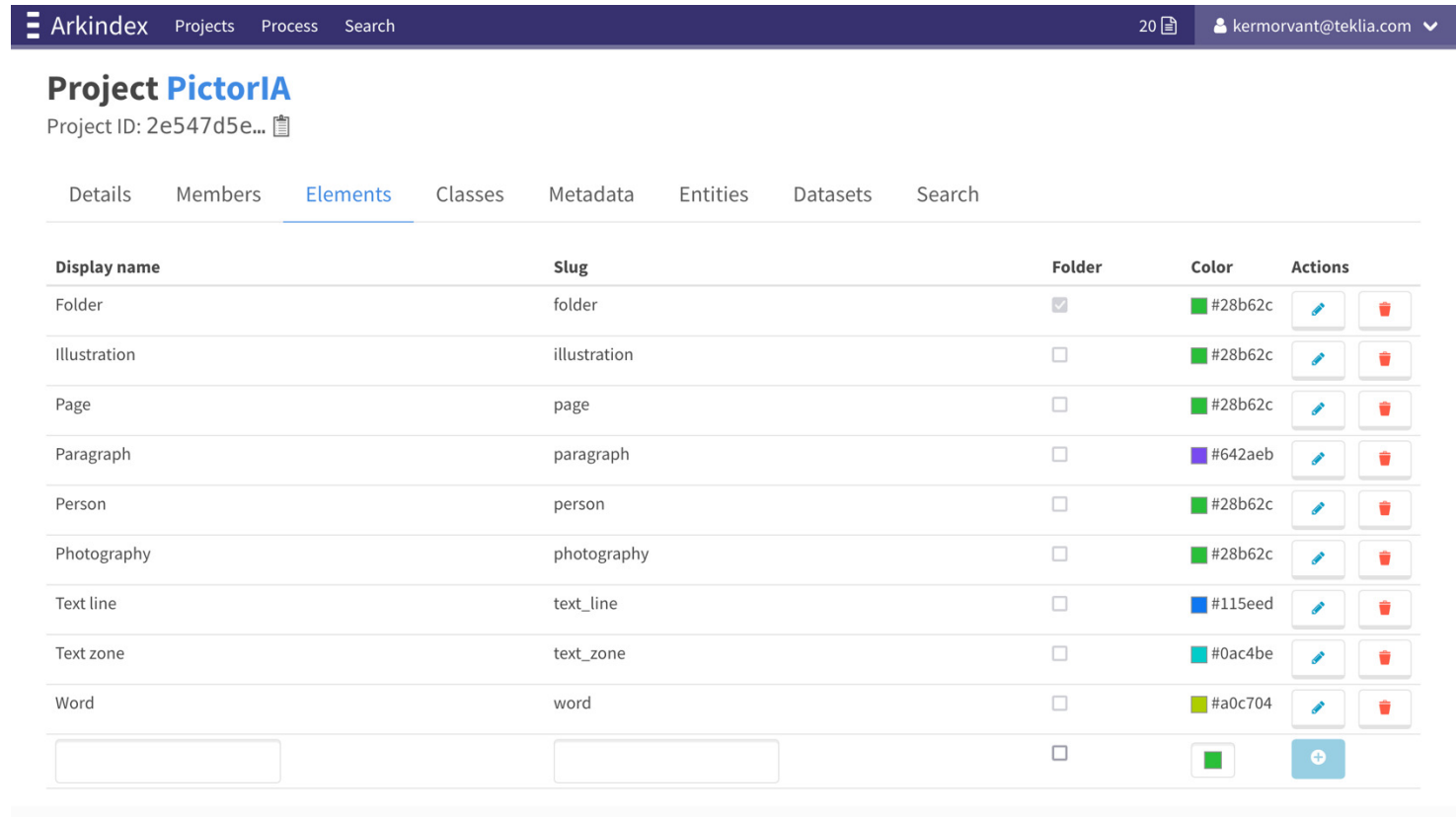
**Folder** : type par défaut pour regrouper les images importées



The screenshot displays the ArkIndex web interface. At the top, there is a navigation bar with 'ArkIndex' and menu items 'Projects', 'Process', and 'Search'. The user is logged in as 'kermorvant@teklia.com'. Below the navigation bar, the current view is 'Folder asr3\_3.pdf'. A search bar and a 'Filter elements...' input are visible. The main content area shows a grid of eight scanned pages, numbered 21 to 24. Each page thumbnail includes a 'Page' label and icons for deletion, viewing, and confirmation. The right sidebar provides metadata for the folder, including 'CLASSIFICATIONS', 'TRANSCRIPTIONS', 'NEW TRANSCRIPTION', and 'METADATA'. The metadata section lists the DOI (<https://doi.org/10.11588/digit.12730>), Editors (Robert, Carl ; Matz, Friedrich ; Andreae, Bernard ; Robert, Carl), Place (Berlin), title (Die antiken Sarkophagreliefs (3,3): Einzelmythen: Niobiden - Triptolemos ungedeutet), and Year (1919). A 'DATASETS' section is also visible at the bottom of the sidebar.

# Arkindex : Organiser

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






















Arkindex Projects Process Search 20 kermorvant@tekla.com

### Project Pictoria

Project ID: 2e547d5e...

Details Members **Elements** Classes Metadata Entities Datasets Search

Display name	Slug	Folder	Color	Actions
Folder	folder	<input checked="" type="checkbox"/>	#28b62c	 
Illustration	illustration	<input type="checkbox"/>	#28b62c	 
Page	page	<input type="checkbox"/>	#28b62c	 
Paragraph	paragraph	<input type="checkbox"/>	#642aeb	 
Person	person	<input type="checkbox"/>	#28b62c	 
Photography	photography	<input type="checkbox"/>	#28b62c	 
Text_line	text_line	<input type="checkbox"/>	#115eed	 
Text_zone	text_zone	<input type="checkbox"/>	#0ac4be	 
Word	word	<input type="checkbox"/>	#a0c704	 
<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input type="color"/>	

Nom affiché

Nom interne = slug

# Arkindex : Enrichir - Hiérarchie

Annotation hiérarchique  
des éléments :

Folder

Page

Illustration

Sarcophage

Personnage

Arkindex Projects Process Search 0 kermorvant@tekli.com

Arkindex | Demo / Folder asr3\_3.pdf / Page 37

Filter by type... Hide Page 37 A+

- Illustration 1
  - Sarcophage 1
    - Personnage 1
    - Personnage 2
    - Personnage 3
    - Personnage 4
    - Personnage 5
    - Personnage 6
    - Personnage 7
    - Personnage 8
    - Personnage 9

wesung angedeutet sein; vgl. 325. 328.

Die beiden andern Szenen nehmen die Vorderseite ein Fig. 324; die dritte ist doppelt so lang als die zweite und greift außerdem auf die rechte Schmalseite Fig. 324 b über. Die zweite Szene zeigt Pelops vor Oenomaus. Der König sitzt auf einem wohl viereckig zu denkenden, aber in mißratener Perspektive dargestellten



324'

Podium, auf dessen sichtbaren Stirnflächen ein Steinbock und ein ihn verfolgender Hund dargestellt sind; vgl. die Ecksockel von II 20 c. 23 c. III 144 b. Die Armlehne des Thronsessels wird vorn und hinten von einer Sphinx gestützt. Oenomaus trägt das typische Kostüm der Theaterkönige, hat an der Seite das Schwert und hält in der Linken das Szepter. Sein unförmlich großer Helm, dessen Kappe mit einem Widderkopf in Relief geziert ist, liegt neben ihm am Boden. Der rechte Arm ist erhoben. Die verlorene Hand wird vermutlich die Redegeste gemacht haben; s. 327. Zwei Trabanten in Ärmelchiton, Mantel, kurzen

Eingang zur Rennbahn an; der I geschmückt; am Archivolte ist gebracht, das wohl im allgem keit der Rennwagen hindeuten mit dem linken Arm durchgreifend hält der eine gedrehte Haarbinde trägt, d Seite hat und dessen Mäntelchen, um de weit zurückflattert, den Zügel des link packt. Auf der Vorderseite wird über Gespannes der Rumpf eines Beireiters u Pferdes sichtbar, der hier wie im röm der beiden Quadrigen beigegeben ist; 154 a, sowie S. 181. Oenomaus erso langen Gewandung wie in der zweiten ! nicht etwa als die griechische Wagenler werden darf, zumal diese der römische fremd ist. Er ist in Rückenansicht da sich, wie die auf seiner rechten Schulter reste beweisen, nach Pelops um. Die römischer Sitte um den Leib geschlungen ist der Griff der Peitsche erhalten, n hält er auch die Zügel gefaßt. Zwei Erstens daß Oenomaus' Wagen vor der folglich kann hier nicht die Version ger Oenomaus den mit Hippodamia voranfalfolgt, sondern nur die oben S. 386 ff. au geführte, nach der es sich um eine ge handelt, bei welcher derjenige Sieger Ziel erreicht. Zweitens daß Oenomaus lenkt und allein auf dem Wagen steht,

View source image List elements on this image



# Arkindex : Enrichir

## Création et Annotation des éléments

ArkIndex Projects Process Search 0 kermorvant@teklia.com

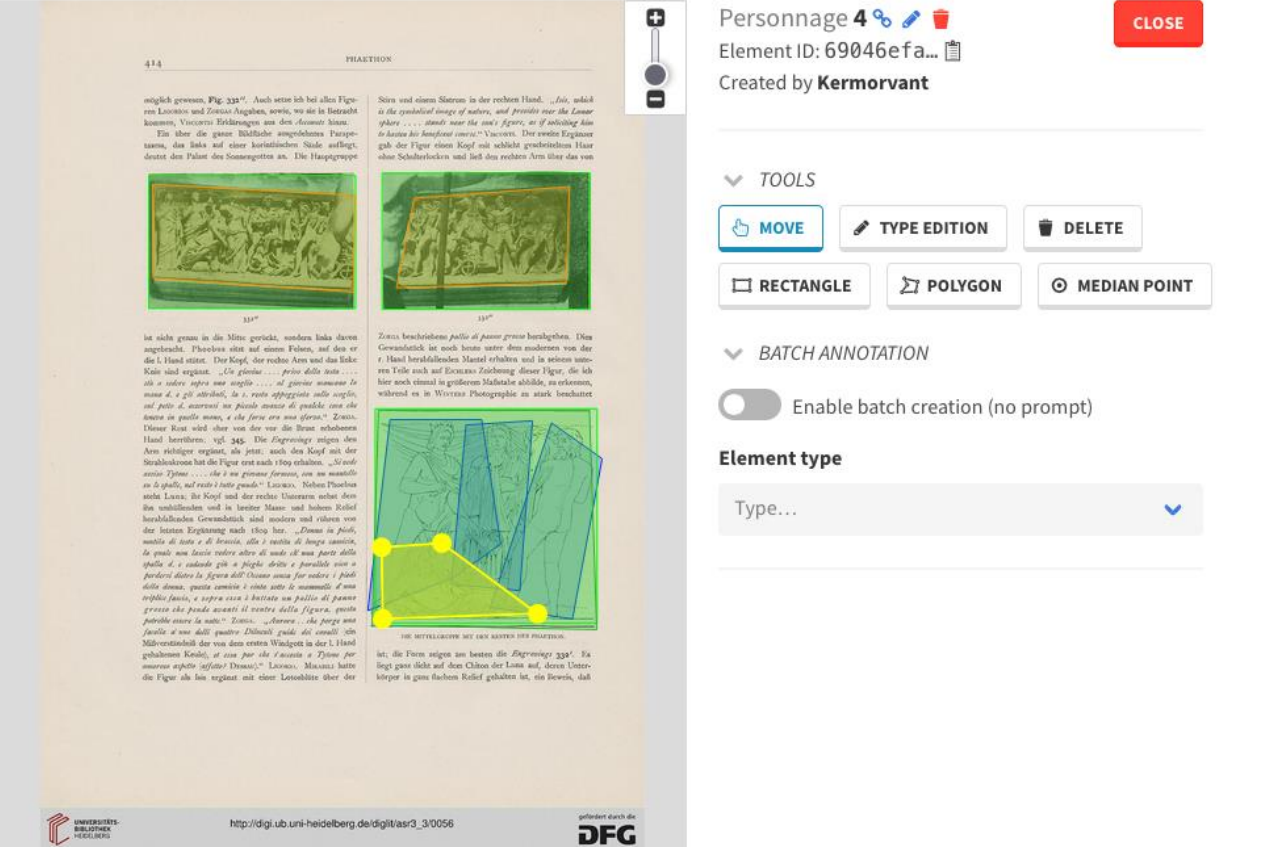
Arkindex | Demo / Folder asr3\_3.pdf / Page 57

← 57 → Elements Processes Display

Filter by type... Hide

Page 57

- Illustration 1
- Illustration 2
- Illustration 3
- Sarcophage 1
- Sarcophage 2
- Dessin 1
- Personnage 1
- Personnage 2
- Personnage 3
- Personnage 4



Personnage 4  
Element ID: 69046efa...  
Created by Kermorvant

TOOLS

MOVE TYPE EDITION DELETE

RECTANGLE POLYGON MEDIAN POINT

BATCH ANNOTATION

Enable batch creation (no prompt)

Element type

Type...

# Arkindex : Enrichir - Classe

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

*Element* : illustration  
*Class* : dessin ou photo

The screenshot displays the ArkIndex web application interface. At the top, there is a navigation bar with 'ArkIndex', 'Projects', 'Process', and 'Search' menus. The user is logged in as 'kermorvant@teklia.com'. The current view is 'Arkindex | Demo / Folder asr3\_3.pdf / Page 67'. The main content area shows a document page with two illustrations highlighted in green. The first illustration is labeled '341<sup>re</sup> OBERGANG' and the second is '341<sup>re</sup> MATRIE IN VIT'. The right sidebar contains a metadata panel for 'Illustration 2' with an ID of '25a6dcf6...'. It includes an 'ANNOTATE' button and a list of classification categories: 'ORIENTATION', 'CLASSIFICATIONS', 'Manual' (with a 'Dessin' tag and a 100% completion indicator), 'TRANSCRIPTIONS', 'NEW TRANSCRIPTION', 'METADATA', and 'DATASETS'. The bottom of the page features logos for 'UNIVERSITÄT HEIDELBERG DIGITAL RESEARCH' and 'DFG'.

[View source image](#) [List elements on this image](#)

# Arkindex : Enrichir - Métadatas

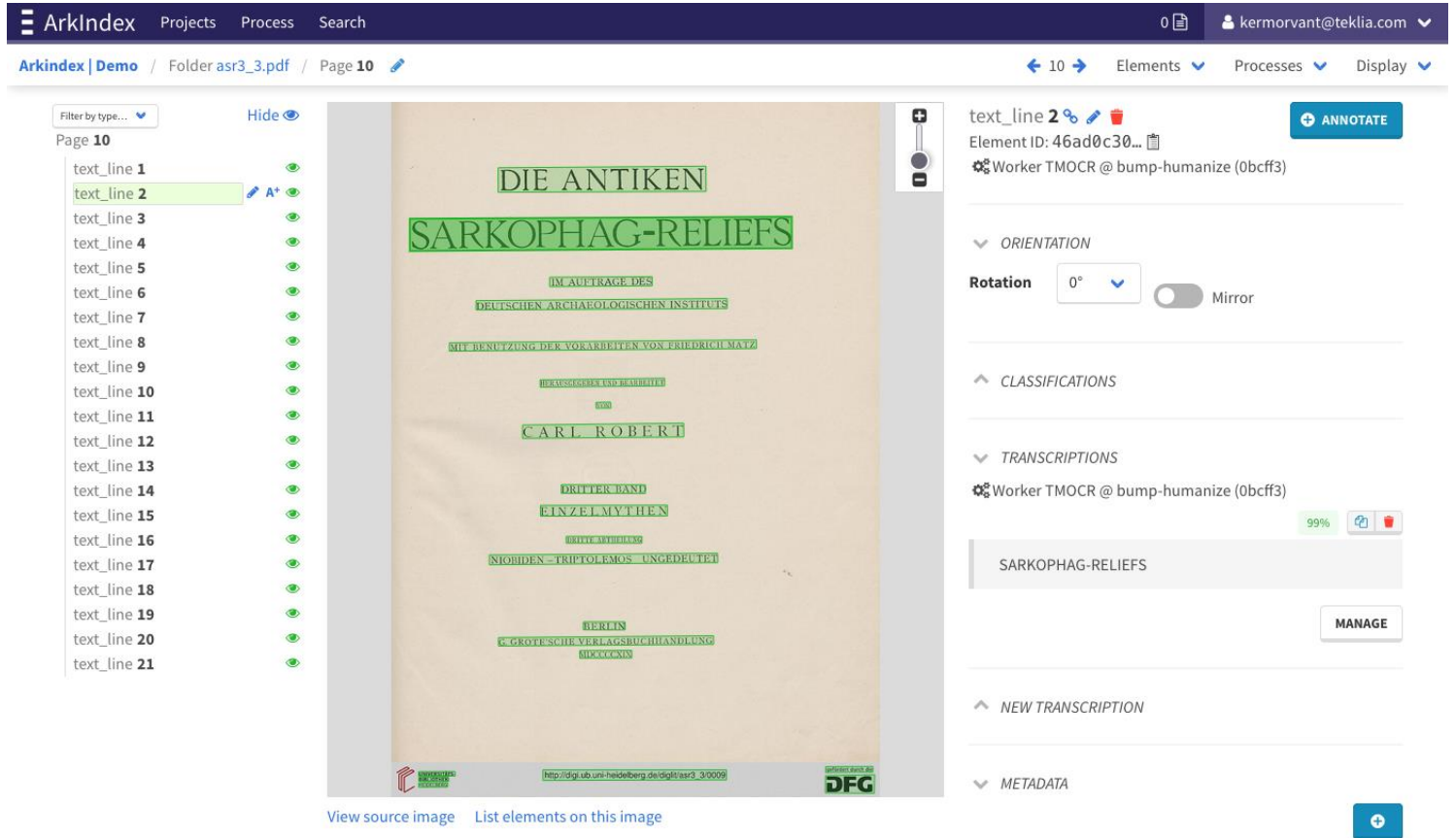
**Méta-data** : toute information supplémentaire

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

The screenshot shows the Arkindex web interface. At the top, there is a navigation bar with 'Arkindex', 'Projects', 'Process', and 'Search'. The user is logged in as 'kermorvant@tekli.com'. The current view is 'Folder asr3\_3.pdf' with 286 results. A 'Filter elements...' search bar is present. Below the search bar, there are navigation controls for 'GO TO ...' with page numbers 1, 2, and 15. The main content area displays a grid of document thumbnails. The first thumbnail (page 1) shows a document with text and a logo. The second thumbnail (page 2) shows a brown cover. The third and fourth thumbnails (pages 3 and 4) show plain brown covers. The right sidebar contains metadata for the folder 'asr3\_3.pdf'. The metadata includes: DOI: <https://doi.org/10.11588/diglit.12730>; Editors: Robert, Carl ; Matz, Friedrich ; Andraea, Bernard ; Robert, Carl; Place: Berlin; title: Die antiken Sarkophagreliefs (3,3): Einzelmythen: Niobiden - Triptolemos ungedeutet; Year: 1919. There are icons for editing and deleting each metadata field.

# Arkindex : Enrichir - Transcription

## Transcription du texte présent sur un élément



The screenshot displays the ArkIndex web interface. At the top, the navigation bar includes 'ArkIndex', 'Projects', 'Process', and 'Search'. The user is logged in as 'kermorvant@tekli.com'. The current view is 'Folder asr3\_3.pdf / Page 10'. The main content area shows a document page with the following text:

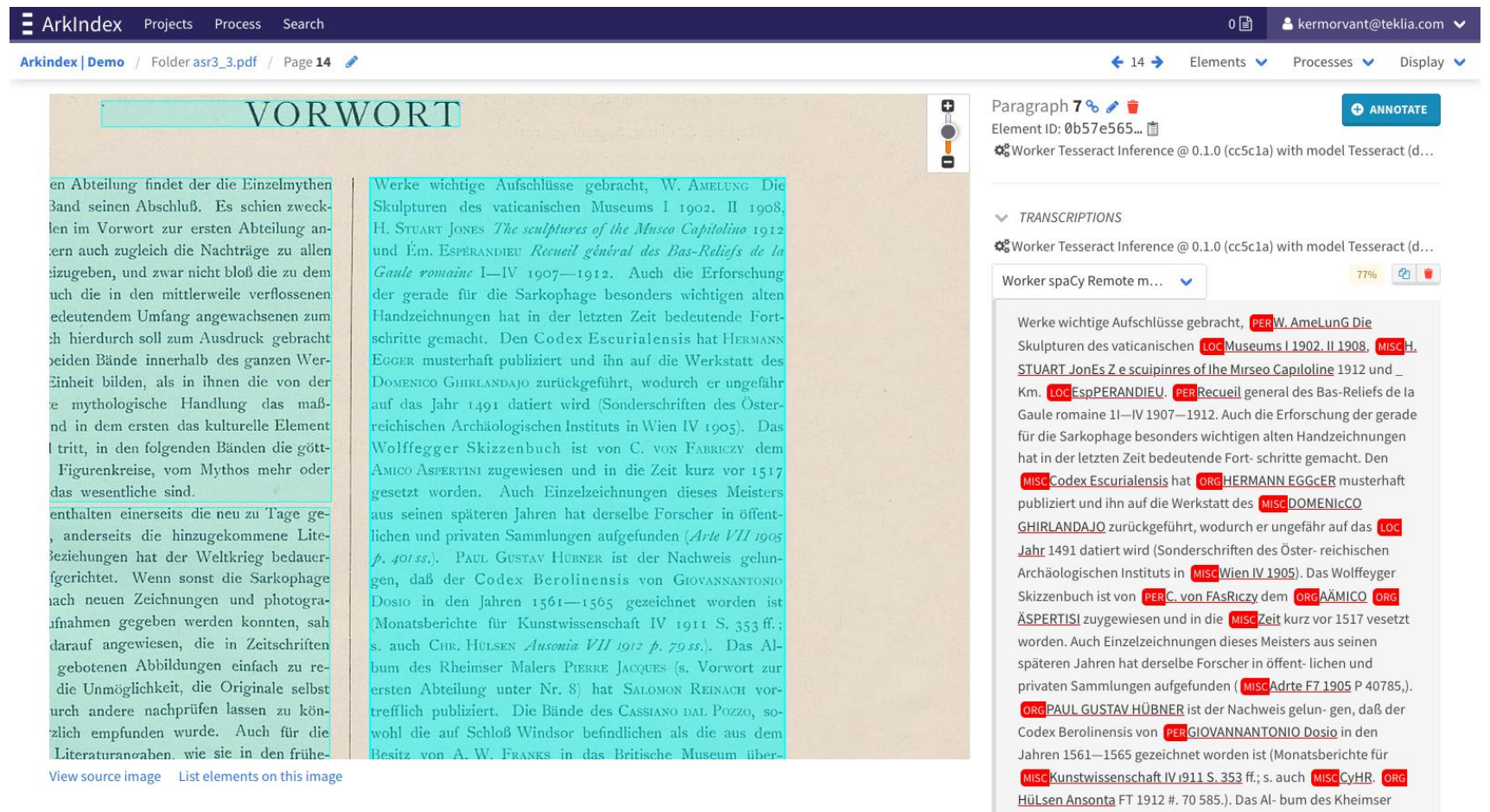
DIE ANTIKEN  
SARKOPHAG-RELIEFS  
IN AUFRAGE DES  
DEUTSCHEN ARCHÄOLOGISCHEN INSTITUTS  
MIT BENUTZUNG DER VORARBEITEN VON FRIEDRICH NAU  
ÜBERSETZT UND RAUTHE  
VON  
CARL ROBERT  
DRITTER BAND  
EINZELMYTHEN  
DIE ANTIKEN  
NORIDEN = TRIPTOLEMOS - UNGEDRUCKT  
BERLIN  
C. GROTESCHE VERLAGSBUCHHANDLUNG  
MEEUWEN

The left sidebar shows a list of text lines from 'text\_line 1' to 'text\_line 21'. 'text\_line 2' is highlighted. The right sidebar shows the details for 'text\_line 2', including the element ID '46ad0c30...' and the worker 'Worker TMOCR @ bump-humanize (0bcff3)'. The 'ORIENTATION' section shows 'Rotation 0°' and a 'Mirror' toggle. The 'CLASSIFICATIONS' section is empty. The 'TRANSCRIPTIONS' section shows a transcription of 'SARKOPHAG-RELIEFS' with a 99% confidence score and a 'MANAGE' button. The 'NEW TRANSCRIPTION' and 'METADATA' sections are also visible.



# Arkindex : Enrichir – Entités nommées

## Identification des entités nommées sur une transcription



The screenshot displays the ArkIndex web interface. At the top, there is a navigation bar with 'ArkIndex', 'Projects', 'Process', and 'Search' menus. Below this, a breadcrumb trail shows 'Arkindex | Demo / Folder asr3\_3.pdf / Page 14'. The main content area shows a document page with the title 'VORWORT' highlighted in a light blue box. The text on the page is transcribed and contains several named entities highlighted in red boxes. The right-hand sidebar contains a 'Paragraph 7' section with an 'ANNOTATE' button and a 'TRANSCRIPTIONS' section with a 'Worker spaCy Remote m...' dropdown. The sidebar also shows a progress indicator at 77% and a 'View source image' link at the bottom.

ArkIndex Projects Process Search

Arkindex | Demo / Folder asr3\_3.pdf / Page 14

← 14 → Elements Processes Display

Paragraph 7  
Element ID: 0b57e565...  
Worker Tesseract Inference @ 0.1.0 (cc5c1a) with model Tesseract (d...)

TRANSCRIPTIONS  
Worker Tesseract Inference @ 0.1.0 (cc5c1a) with model Tesseract (d...)  
Worker spaCy Remote m... 77%

VORWORT

en Abteilung findet der die Einzelmythen Band seinen Abschluß. Es schien zwecklen im Vorwort zur ersten Abteilung anern auch zugleich die Nachträge zu allen zugeben, und zwar nicht bloß die zu dem uch die in den mittlerweile verlossenen edeutendem Umfang angewachsenen zum ch hierdurch soll zum Ausdruck gebracht beiden Bände innerhalb des ganzen Wer Einheit bilden, als in ihnen die von der e mythologische Handlung das maßnd in dem ersten das kulturelle Element tritt, in den folgenden Bänden die gött Figurenkreise, vom Mythos mehr oder das wesentliche sind.

Werke wichtige Aufschlüsse gebracht, W. AMELUNG Die Skulpturen des vaticanischen Museums I 1902. II 1908, H. STUART JONES *The sculptures of the Museo Capitolino* 1912 und Ém. ESPERANDIEU *Recueil général des Bas-Reliefs de la Gaule romaine* I—IV 1907—1912. Auch die Erforschung der gerade für die Sarkophage besonders wichtigen alten Handzeichnungen hat in der letzten Zeit bedeutende Fortschritte gemacht. Den Codex Escorialensis hat HERMANN EGGER musterhaft publiziert und ihn auf die Werkstatt des DOMENICO GHIRLANDAJO zurückgeführt, wodurch er ungefähr auf das Jahr 1491 datiert wird (Sonderschriften des Österreichischen Archäologischen Instituts in Wien IV 1905). Das Wolffegger Skizzenbuch ist von C. VON FABRICZY dem AMICO ASPERTINI zugewiesen und in die Zeit kurz vor 1517 gesetzt worden. Auch Einzelzeichnungen dieses Meisters aus seinen späteren Jahren hat derselbe Forscher in öffentlichen und privaten Sammlungen aufgefunden (*Arte VII* 1905 p. 401 ss.). PAUL GUSTAV HÜBNER ist der Nachweis gelungen, daß der Codex Berolinensis von GIOVANNANTONIO DOSIO in den Jahren 1561—1565 gezeichnet worden ist (Monatsberichte für Kunstwissenschaft IV 1911 S. 353 ff.; s. auch CHR. HULSEN *Ausonia VII* 1912 p. 79 ss.). Das Album des Rheinischer Malers PIERRE JACQUES (s. Vorwort zur ersten Abteilung unter Nr. 8) hat SALOMON REINACH vortrefflich publiziert. Die Bände des CASSIANO DAL POZZO, sowohl die auf Schloß Windsor befindlichen als die aus dem Besitz von A. W. FRANKS in das Britische Museum über-

View source image List elements on this image

# 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
	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
Visual LLMs	Resnet classifier	open-source
	OpenAI ChatGPT	proprietary API
Text Translation	Anthropic Claude	proprietary API
	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

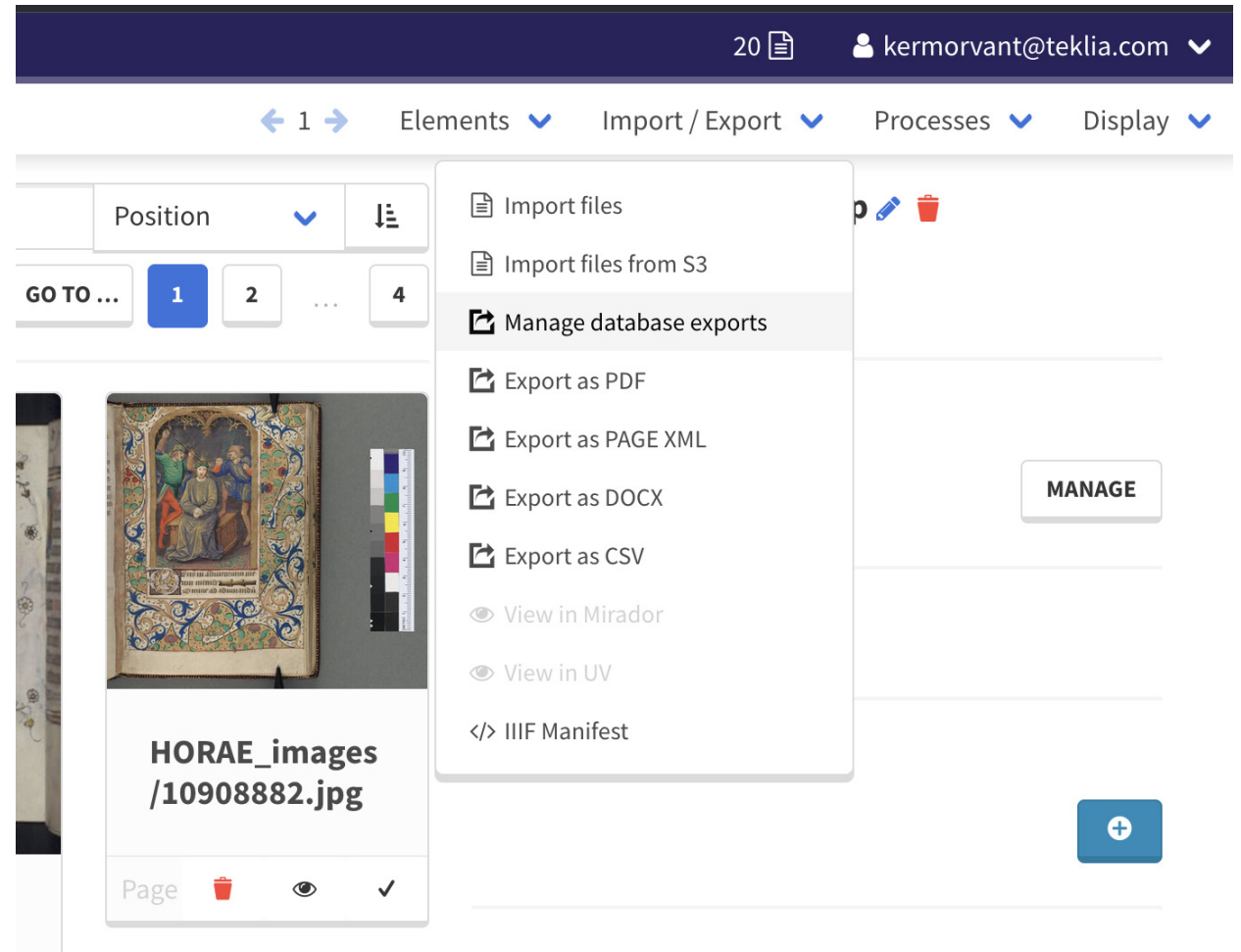
- PDF image et texte
- pageXML
- docx
- csv

## Import de DB

- sqlite

## Import en CLI

+ ALTO



The screenshot displays the Arkindex web interface. At the top, there is a dark blue header with the user's email 'kermorvant@tekli.com' and a dropdown arrow. Below the header, a navigation bar contains 'Elements', 'Import / Export', 'Processes', and 'Display', each with a dropdown arrow. The 'Import / Export' menu is open, showing options: 'Import files', 'Import files from S3', 'Manage database exports', 'Export as PDF', 'Export as PAGE XML', 'Export as DOCX', 'Export as CSV', 'View in Mirador', 'View in UV', and '</> IIIF Manifest'. The main content area shows a list of items. The first item is a manuscript page image with the caption 'HORAE\_images /10908882.jpg'. Below the image, there are icons for 'Page', a trash can, an eye, and a checkmark. A 'MANAGE' button is visible to the right of the image. At the bottom right, there is a blue button with a white plus sign.

# Arkindex / Callico



## 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

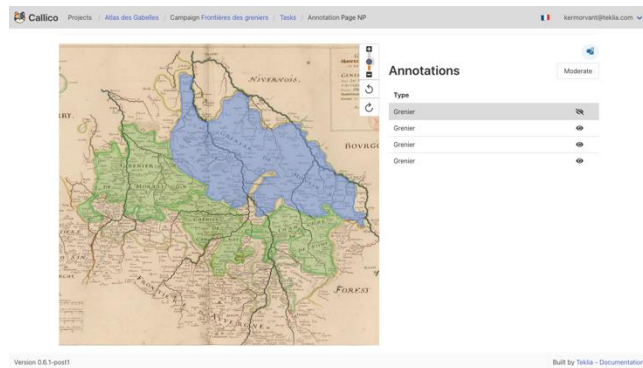
- Multiples contributeurs
- Tâches simples et unitaires
- Gestion de campagnes d'annotation ou validation

➔ Tout public



# Callico : Annotation et validation

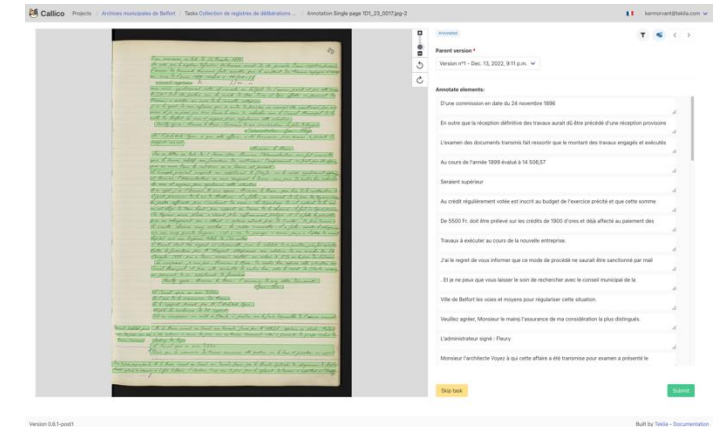
## Zonage



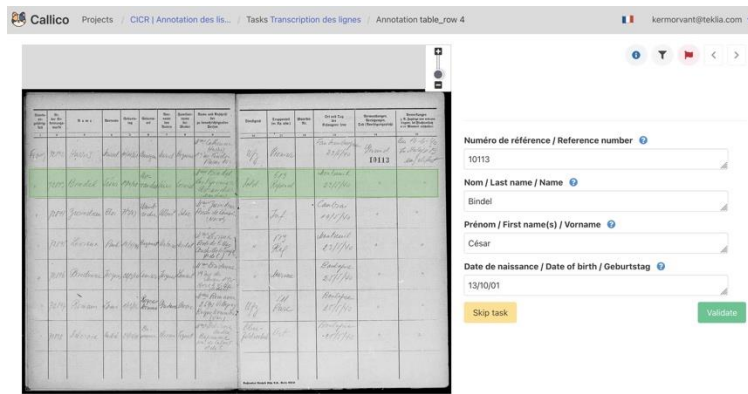
## Classification



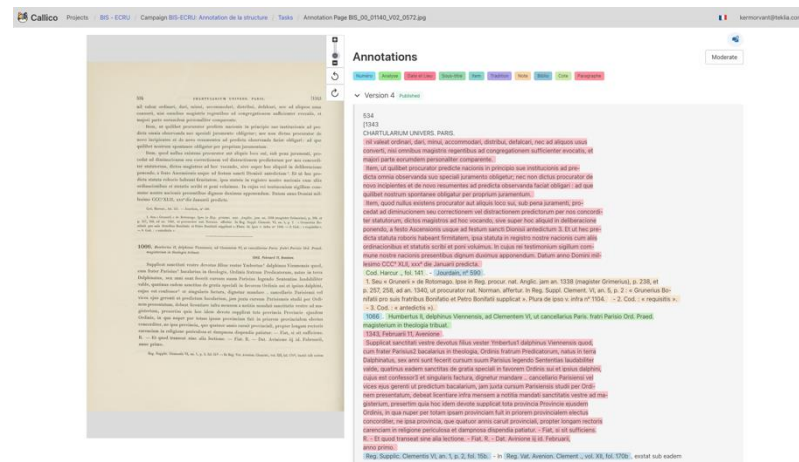
## Transcription



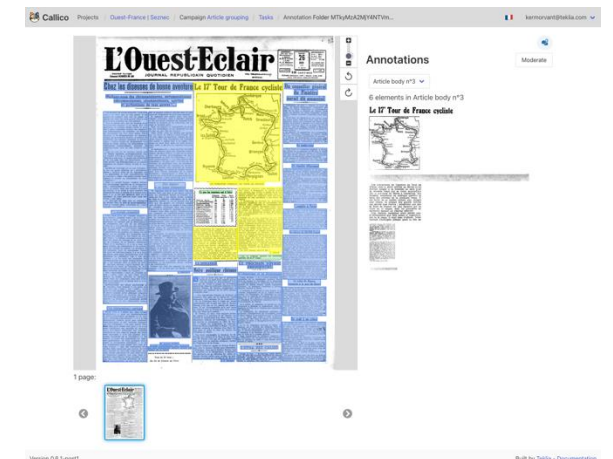
## Clé-valeur



## Entités nommées



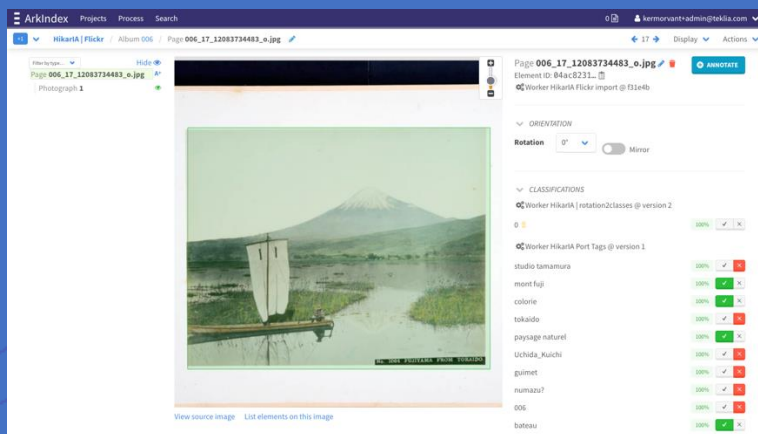
## Groupement



# TEKLIA's open-source software suite for document processing

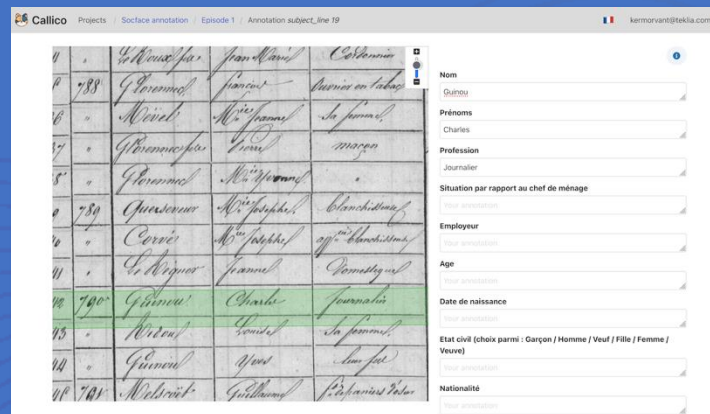
## Arkindex

Document processing



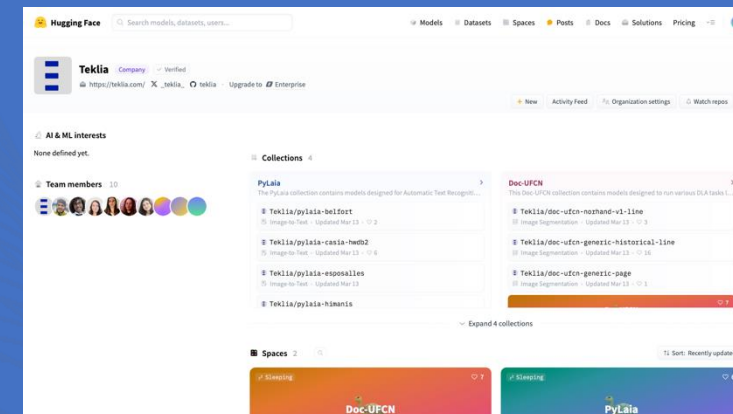
## Callico

Collaborative annotation campaigns



## HuggingFace

Models and datasets



Open-source and Enterprise licences

<https://gitlab.teklia.com>

Open-source

<https://support.teklia.com>

Open-source, open-weights

<https://huggingface.co/Teklia>