

Arkindex, an open-source platform for document image recognition applied to archeology

Christopher Kermorvant, TEKLIÀ, Paris

Mélodie Boillet, TEKLIÀ, Paris

Pablo Ciezar, INRAP

Anne-Violaine Szabados, CNRS - ArScAn

Julien Schuh, UPN - Université Paris Nanterre



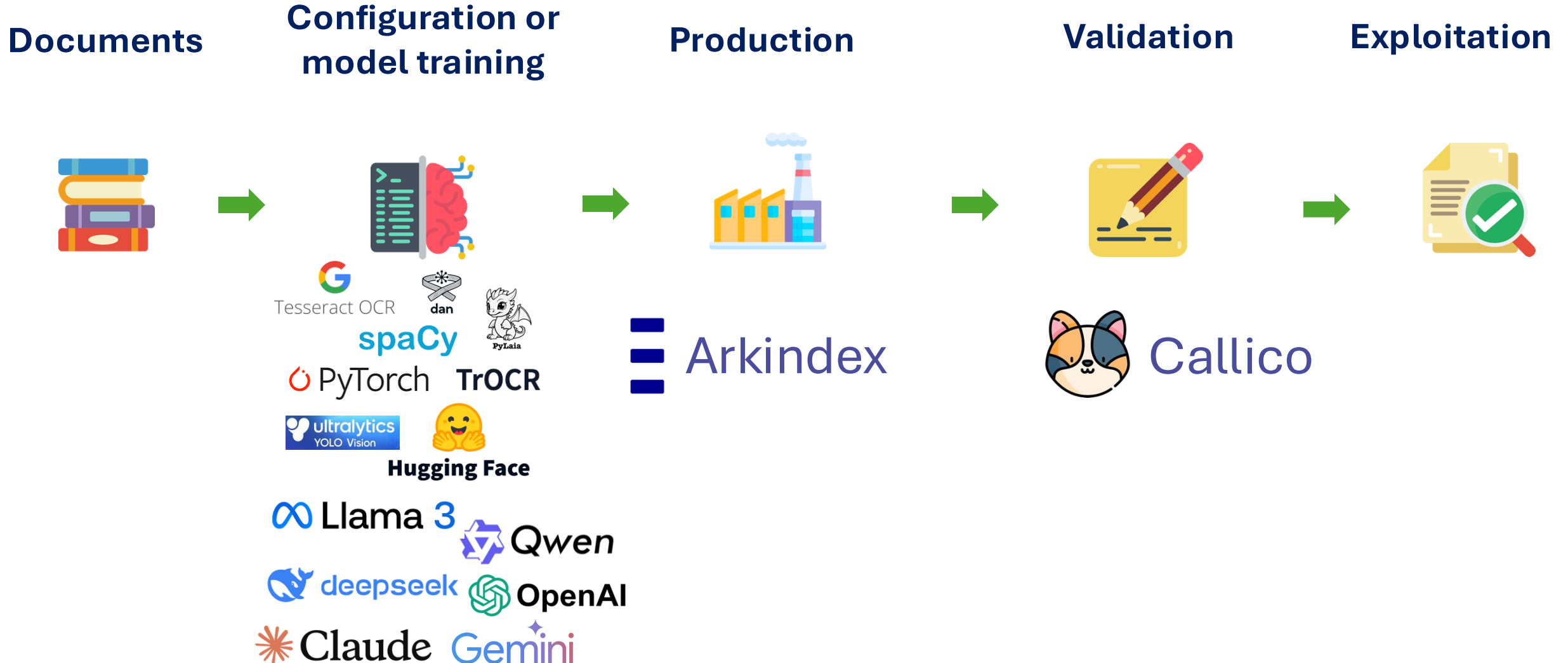
Automatic document processing with AI

AI models **can** recognize, understand and extract data from historical documents

The challenges are now:

- **How to choose** the right model: accurate, easy to run and cheap
- **How to model** data coming from different sources (meta-data, human annotation, predictions from AI)
- **How to process** documents at large scale: reliability, completeness, quality control

Document processing workflow

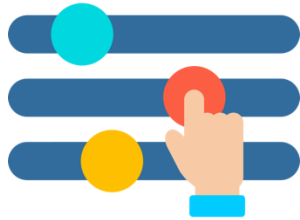


Arkindex: a generic document processing platform

Developed by TEKLIA since 2019

Already used for more than 60 projects

Customization



Process all types of documents

Scaling up



Process 1000 or 10 million pages

Open-source



Dissemination and community participation

<https://gitlab.teklia.com/arkindex>

Arkindex: a generic data model

The screenshot displays the Arkindex web interface for document analysis. The interface is divided into several sections:

- Header:** Includes the Arkindex logo, navigation tabs (Projects, Process, Search), and user information (kermorvant@tekla.com).
- Breadcrumb:** Shows the current path: CAA / Folder Argonnes / Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg.
- Left Sidebar (Red Box):** A hierarchical tree structure for document analysis, including filters and a list of elements with visibility toggles.
 - Filter by type... (dropdown)
 - Filter by worker run... (dropdown)
 - Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg A+
 - Fragment 1
 - Motif 1
 - Motif 2
 - Motif 3
 - Motif 4
 - Motif 5
 - Motif 6
 - Label 1
 - Text line 2
 - Text line 3
 - Text line 4
 - Text line 8
 - Text line 7
 - Text line 6
 - Text line 10
 - Text line 9
 - Text line 11
 - Text line 12
 - Label 2
 - Text line 1
 - Text line 5
- Main View:** A large image of a document page with handwritten text and a ruler. The text includes "UC 28" and "Gem. Maastricht". A red box highlights the text "Custom hierarchical layout analysis".
- Right Panel:** Contains metadata and analysis results for the selected page.
 - Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg
 - Element ID: 96e67ff8...
 - Worker File Import @ version 3
 - ORIENTATION
 - CLASSIFICATIONS
 - TRANSCRIPTIONS
 - Worker TMOCR @ main (123a5a)
 - 94%
 - METADATA
 - Date: 1982
 - Location: Maastricht

Arkindex: a generic data model

The screenshot displays the Arkindex web interface. The main area shows a document image with handwritten text, including "UC 28" and "Gem. Maastricht". A red box highlights the "CLASSIFICATIONS" section on the right, which shows a manual classification for "UC_28" with a 100% confidence score and a green checkmark. The interface also includes a sidebar with a list of elements (Motif 1-6, Label 1, Text line 1-12) and a top navigation bar with "Projects", "Process", and "Search" tabs.

Arkindex Projects Process Search

CAA / Folder Argonnes / Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg

Filter by type... Hide

Filter by worker run...

Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg A+

Fragment 1

- Motif 1
- Motif 2
- Motif 3
- Motif 4
- Motif 5
- Motif 6

Label 1

- Text line 2
- Text line 4
- Text line 8
- Text line 7
- Text line 6
- Text line 10
- Text line 9
- Text line 11
- Text line 12

Label 2

- Text line 1
- Text line 5

Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg

Element ID: 96e67ff8...

Worker File Import @ version 3

ORIENTATION

CLASSIFICATIONS

Manual

UC_28 100% ✓

TRANSCRIPTIONS

Worker TMOCR @ main (123a5a)

94%

Gem. Maastricht,

MANAGE

METADATA

Date 1982

Location Maastricht

View source image List elements on this image

Arkindex: a generic data model

The screenshot displays the Arkindex web interface. The top navigation bar includes 'Arkindex', 'Projects', 'Process', and 'Search'. The main content area shows a document page titled 'Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg'. On the left, a sidebar lists 'Fragment 1' with 'Motif 1' through 'Motif 6', and 'Label 1' with 'Text line 3' through 'Text line 12', and 'Label 2' with 'Text line 1' and 'Text line 5'. The central image shows a document with handwritten text and a ruler. A red box highlights the 'TRANSCRIPTIONS' section on the right, which shows a transcription of 'Gem. Maastricht,' with a 94% confidence score and a 'MANAGE' button. The bottom right section shows 'METADATA' with 'Date' 1982 and 'Location' Maastricht.

Manual and automatic text transcription
(printed and handwritten)

Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg

Element ID: 96e67ff8...

Worker File Import @ version 3

ORIENTATION

CLASSIFICATIONS

Manual

UC_28

100%

TRANSCRIPTIONS

Worker TMOCR @ main (123a5a)

94%

Gem. Maastricht,

MANAGE

METADATA

Date 1982

Location Maastricht

Arkindex: a generic data model

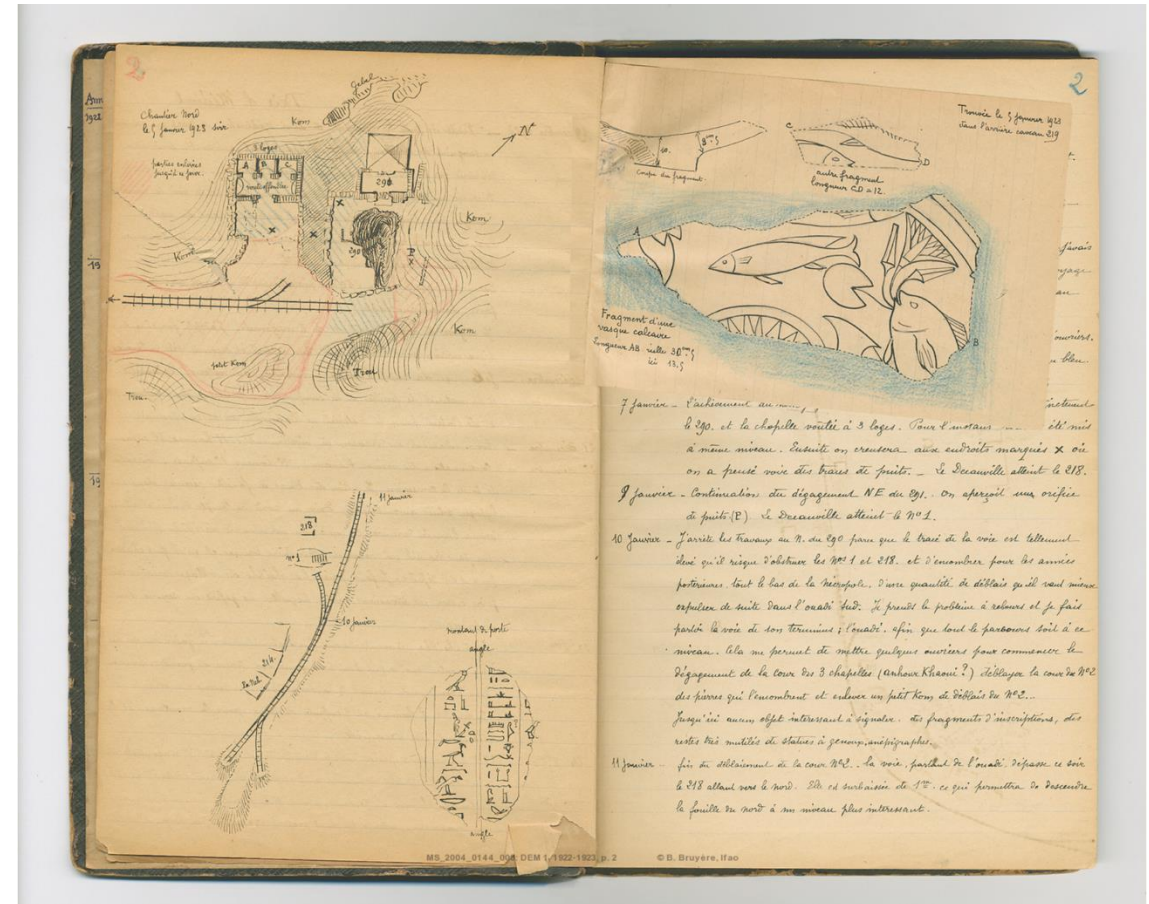
The screenshot displays the Arkindex web application interface. The top navigation bar includes 'Arkindex', 'Projects', 'Process', and 'Search'. The breadcrumb trail shows 'CAA / Folder Argonnes / Page 76cec7a4-bf95-4ea1-8a39-0eeef7cfa2c5.jpg'. The left sidebar contains a tree view of the document's structure, including 'Fragment 1' with 'Motif 1' through 'Motif 6', 'Label 1' with 'Text line 2' through 'Text line 12', and 'Label 2' with 'Text line 1' and 'Text line 5'. The central image viewer shows a document page with handwritten text and a ruler. A red box highlights the 'METADATA' section at the bottom right, which contains the following information:

| ▼ METADATA | |
|------------|------------|
| Date | 1982 |
| Location | Maastricht |

B. Bruyère's handwritten excavation journals

Journals of excavations by Egyptologist Bernard Bruyère from 1922 to 1951 at Deir el-Medina, on the left bank of Thebes

A lot of information, drawings of objects and surveys of monuments were not included in the published reports and are only available in the handwritten journals.



B. Bruyère's handwritten excavation journals

Automatic transcription (HTR)
of 1000
handwritten pages

Manual illustration detection and annotation (meta-data)

Indexing and search

IFAEO | Journaux de fouilles de... / Folder Journal de fouilles 1931-1938 / Page ms_2019_03566

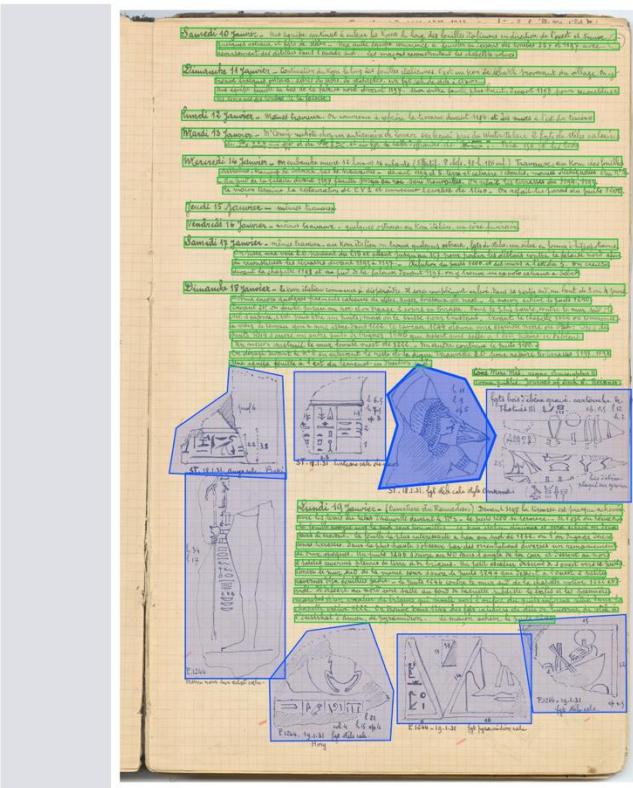
← 18 → Elements Processes Display

Filter by type... Hide

Filter by worker run...

Page ms_2019_03566

- Illustration ms_2019_03566...
- Text line 2
- Illustration ms_2019_03566...
- Text line 3
- Illustration ms_20...
- Text line 4
- Illustration ms_2019_03566...
- Text line 5
- Illustration ms_2019_03566...
- Text line 6
- Illustration ms_2019_03566...
- Text line 7
- Illustration ms_2019_03566...
- Text line 8
- Illustration ms_2019_03566...
- Text line 9
- Text line 10
- Text line 11
- Text line 12
- Text line 13
- Text line 14
- Text line 15
- Text line 16
- Text line 17
- Text line 18
- Text line 19
- Text line 20
- Text line 21
- Text line 22



View source image List elements on this image

Illustration ms_2019_03566_03

Element ID: f73c387b...

Created by Ifao_Bruyère

ORIENTATION

CLASSIFICATIONS

Manual

Stèle privée 100% ✓

Add a classification

TRANSCRIPTIONS

100% ✓

Dimanche 18 Janvier . Le Kom italien commence à disparaître . Il sera complètement enlevé . dans sa partie sud , au bout de 3 ou 4 jours .

METADATA

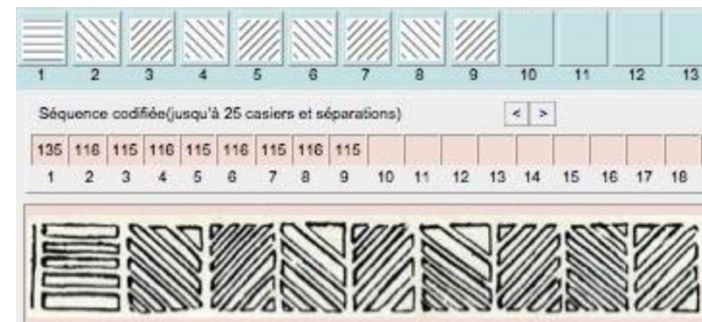
| | |
|----------------|-------------------------------|
| date of find | 1931, January 18 |
| found_location | Temple ptolémaïque, au sud du |



Identification of Argonne roller-stamped motifs

Late Antique Argonne ware was produced on a massive scale in Northwestern Europe between the 4th and 6th centuries.

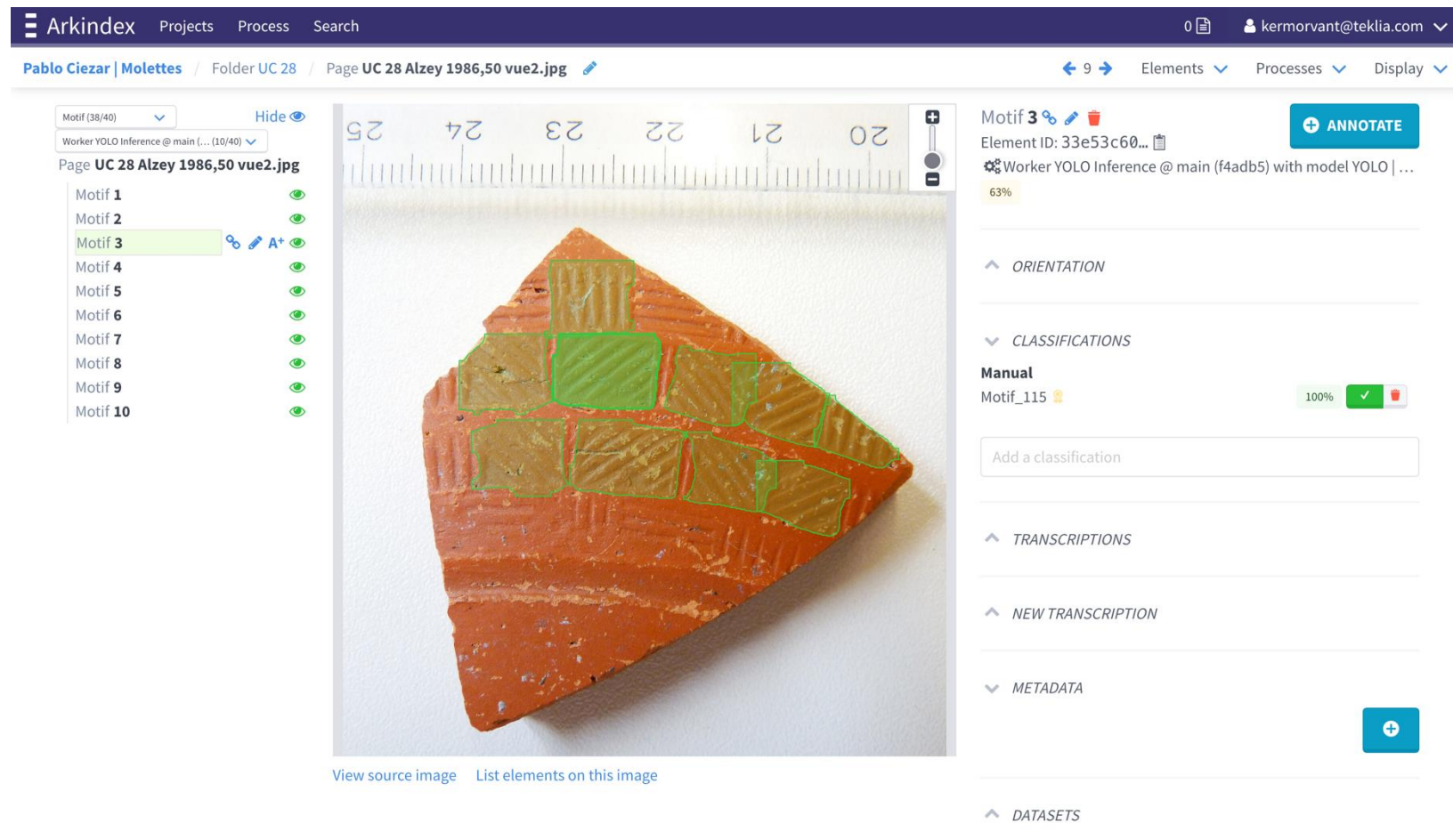
The roller-embossed decoration allows for the identification of production sites and workshops, and provides elements of dating.



Identification of Argonne roller-stamped motifs

Detection of the motifs: image segmentation algorithm (YOLO), stored as image element

Type of motif: classification on the detected element



The screenshot displays the Arkindex web application interface. At the top, the navigation bar includes 'Arkindex', 'Projects', 'Process', and 'Search'. The user is logged in as 'kermorvant@teklia.com'. The breadcrumb trail shows the current location: 'Pablo Ciezar | Molettes / Folder UC 28 / Page UC 28 Alzey 1986,50 vue2.jpg'.

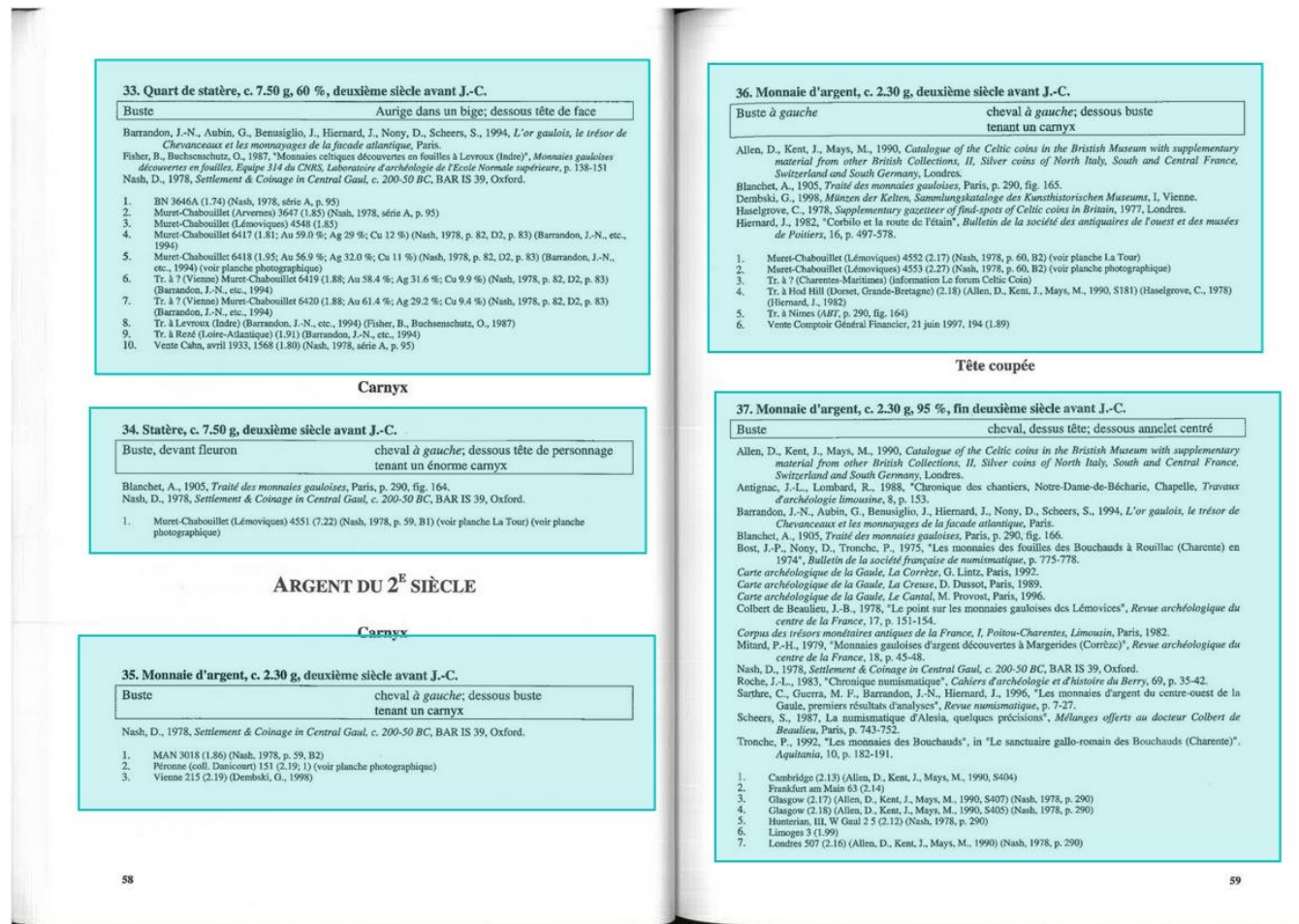
On the left, a list of motifs (Motif 1 to Motif 10) is shown, with 'Motif 3' selected. The main image area displays a photograph of a reddish-brown pottery fragment with several green rectangular bounding boxes highlighting detected motifs. A ruler is visible at the top of the image for scale.

On the right, the details for 'Motif 3' are shown. It includes the 'Element ID: 33e53c60...', the processing information 'Worker YOLO Inference @ main (f4adb5) with model YOLO | ...', and a '63%' confidence score. Below this, there are sections for 'ORIENTATION', 'CLASSIFICATIONS' (showing a 'Manual' classification for 'Motif_115' with 100% confidence), 'TRANSCRIPTIONS', 'NEW TRANSCRIPTION', 'METADATA', and 'DATASETS'. An 'ANNOTATE' button is located at the top right of the motif details.

Information extraction from Celtic coin catalogues

A vast amount of information on coins is available in paper publications.

These publications contain descriptive, geographical and bibliographical information in semi-structured form.



Information extraction from Celtic coin catalogues

OCR and Structured information extraction using multimodal Large Language Models (vLLM)

Arkindex
Projects
Process
Search
8
0
kermorvant@teklia.com

ENS-PSL | AOROC
Folder 3 - De l'Atlantique aux Arvernes.pdf
Page 31
Text zone 1

1
Elements
Processes
Display

Text zone 1
Element ID: 146a7c9b...
Created by CKE
ANNOTATE

33. Quart de statère, c. 7.50 g, 60 %, deuxième siècle avant J.-C.

| Buste | Aurige dans un bige; dessous tête de face |
|--|--|
| Barrandon, J.-N., Aubin, G., Benusiglio, J., Hiernard, J., Nony, D., Scheers, S., 1994, <i>L'or gaulois, le trésor de Chevaux et les monnayages de la façade atlantique</i> , Paris. Fisher, B., Buchsenschutz, O., 1987, "Monnaies celtiques découvertes en fouilles à Levroux (Indre)", <i>Monnaies gauloises découvertes en fouilles, Equipe 314 du CNRS, Laboratoire d'archéologie de l'Ecole Normale supérieure</i> , p. 138-151 Nash, D., 1978, <i>Settlement & Coinage in Central Gaul, c. 200-50 BC</i> , BAR IS 39, Oxford. | |
| 1. | BN 3646A (1.74) (Nash, 1978, série A, p. 95) |
| 2. | Muret-Chabouillet (Arvernes) 3647 (1.85) (Nash, 1978, série A, p. 95) |
| 3. | Muret-Chabouillet (Lémoviques) 4548 (1.85) |
| 4. | Muret-Chabouillet 6417 (1.81; Au 59.0 %; Ag 29 %; Cu 12 %) (Nash, 1978, p. 82, D2, p. 83) (Barrandon, J.-N., etc., 1994) |
| 5. | Muret-Chabouillet 6418 (1.95; Au 56.9 %; Ag 32.0 %; Cu 11 %) (Nash, 1978, p. 82, D2, p. 83) (Barrandon, J.-N., etc., 1994) (voir planche photographique) |
| 6. | Tr. à ? (Vienne) Muret-Chabouillet 6419 (1.88; Au 58.4 %; Ag 31.6 %; Cu 9.9 %) (Nash, 1978, p. 82, D2, p. 83) (Barrandon, J.-N., etc., 1994) |
| 7. | Tr. à ? (Vienne) Muret-Chabouillet 6420 (1.88; Au 61.4 %; Ag 29.2 %; Cu 9.4 %) (Nash, 1978, p. 82, D2, p. 83) (Barrandon, J.-N., etc., 1994) |
| 8. | Tr. à Levroux (Indre) (Barrandon, J.-N., etc., 1994) (Fisher, B., Buchsenschutz, O., 1987) |
| 9. | Tr. à Rezé (Loire-Atlantique) (1.91) (Barrandon, J.-N., etc., 1994) |
| 10. | Vente Cahn, avril 1933, 1568 (1.80) (Nash, 1978, série A, p. 95) |

View source image
List elements on this image

^ ORIENTATION

^ TRANSCRIPTIONS

Worker LLM OpenAI @ master (60d241) using configuration

```

{
  "denomination": "Quart de statère",
  "dating": "Deuxième siècle avant J.-C.",
  "obverse_description": "Buste",
  "reverse_description": "Aurige dans un bige ; dessous tête de face",
  "specimens": [
    {
      "coin_number": 1,
      "commune": "—",
      "department": "—",
      "museum": "BNF",
      "museum_number": "3646A",
      "main_metal": "—",
      "weight_g": 1.74,
      "references": ["Nash, 1978, série A, p. 95"]
    },
    {
      "coin_number": 2
    }
  ]
}

```



Scaling-up processing

Arkindex Projects Process Search 0 kermorvant@tekli.com

Configure process

Configure and run a process

[SELECT WORKERS](#) [CREATE TEMPLATE](#) [SELECT TEMPLATE](#) [CLEAR PROCESS](#)

| Worker Version | Model Version | Use GPU | Configuration | Dependencies | Actions |
|--|--|--------------------------|-------------------|--|--------------------|
| OBJECT-DETECTOR Doc-UCFN Inference https://gitlab.tekli.com/workers/ufcn/commit/eb227b1def08254fb5de4ac1cad94ef5bb... | gitlab-import Doc-UCFN Line Modern&Historica... 37676ff9... 📄 🔗 | <input type="checkbox"/> | + | + | 🗑️ |
| LINE-ATR PyLaia Inference https://gitlab.tekli.com/workers/pylaia/commit/d0cc88ad51fc4422b6e884e2d0d8856... | PyLaia French/Latin Medieval... 156b5c9d... 📄 🔗 | <input type="checkbox"/> | + | Doc-UCFN Inference - + | 🗑️ |
| IMAGE-SEGMENTER YOLO Inference https://gitlab.tekli.com/workers/yolo/commit/603287bbab3ea31d3e70723b79ab4789c8... | CATMuS Yolo medium 6123b5be... 📄 🔗 | <input type="checkbox"/> | + | + | 🗑️ |

Start

```
graph LR; Start --> YOLO_Inference[YOLO Inference]; Start --> Doc-UCFN_Inference[Doc-UCFN Inference]; Doc-UCFN_Inference --> PyLaia_Inference[PyLaia Inference];
```

[GO BACK TO ELEMENTS FILTERING](#) [Advanced settings](#) [RUN PROCESS](#)

Version 1.8.1 Enterprise Edition Built by Teklia [Support](#) [Documentation for Arkindex & CLI](#)

Processing workflow definition

Arkindex Projects Process Search 5 0 kermorvant@tekli.com

Process status

Process ID: 9c701bd9...

| Name | Project | Mode | Farm | Worker activities | Status | ACTIONS |
|--------------------------|-------------------|---------|-----------------|-------------------|-----------|-------------------|
| Socface AD044 Classify | Socface FRAD044 | Workers | Socface Workers | Ready | Completed | ⌵ |

initialisation [Completed](#)

yolo-classifier_a3f2d3_1 [Completed](#)

yolo-classifier_a3f2d3_10 [Completed](#)

yolo-classifier_a3f2d3_2 [Completed](#)

yolo-classifier_a3f2d3_3 [Completed](#)

yolo-classifier_a3f2d3_4 [Completed](#)

yolo-classifier_a3f2d3_5 [Completed](#)

yolo-classifier_a3f2d3_6 [Completed](#)

yolo-classifier_a3f2d3_7 [Completed](#)

yolo-classifier_a3f2d3_8 [Completed](#)

yolo-classifier_a3f2d3_9 [Completed](#)

initialisation

yolo-classifier_a3f2d3_2

yolo-classifier_a3f2d3_7

yolo-classifier_a3f2d3_9

yolo-classifier_a3f2d3_3

yolo-classifier_a3f2d3_8

yolo-classifier_a3f2d3_6

yolo-classifier_a3f2d3_4

yolo-classifier_a3f2d3_5

yolo-classifier_a3f2d3_1

yolo-classifier_a3f2d3_10

Hold down [Ctrl](#) to select multiple tasks

Version 1.8.2 Enterprise Edition Built by Teklia [Support](#) [Documentation for Arkindex & CLI](#)

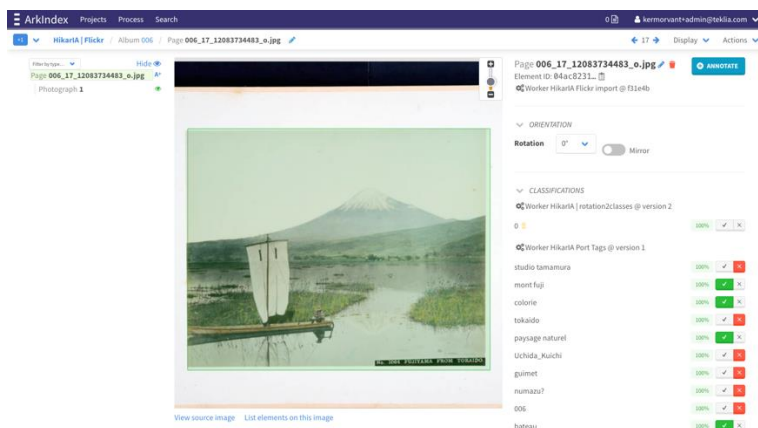
Distributed processing
(multiple servers, HPC)



TEKLIA's open-source software suite for document processing

Arkindex

Document processing



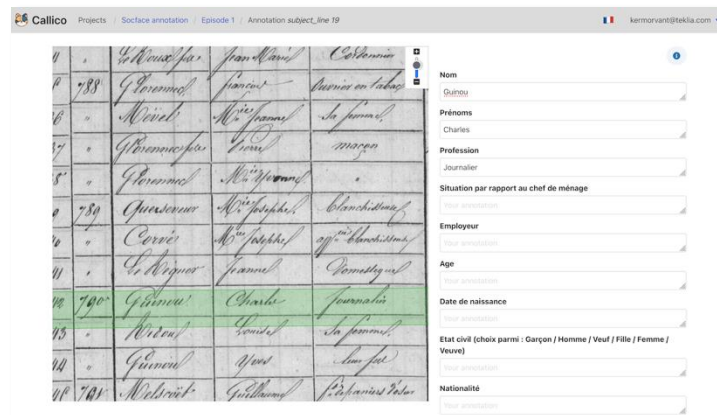
Open-source and Enterprise
licences

<https://gitlab.teklia.com>

<https://support.teklia.com>

Callico

Collaborative annotation campaigns

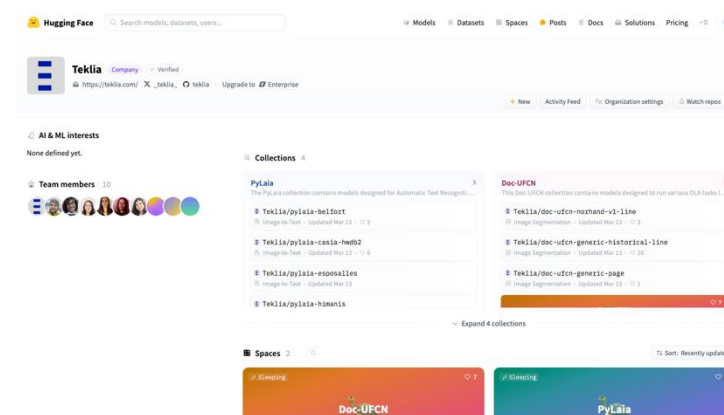


Open-source

kermorvant@teklia.com

HuggingFace

Models and datasets



Open-source, open-weights

<https://huggingface.co/Teklia>