



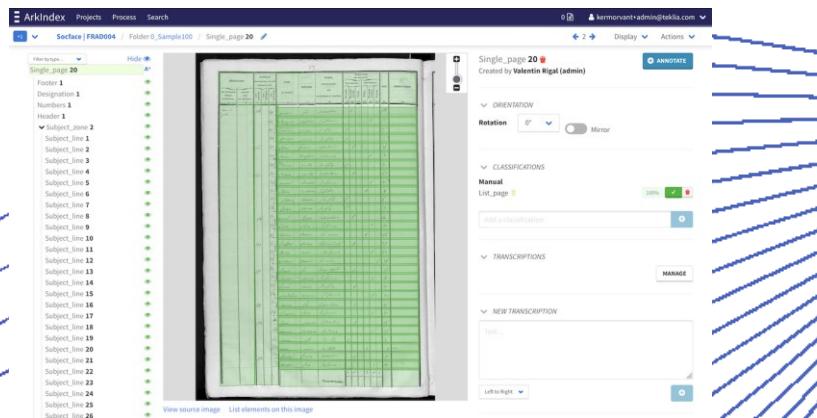
Document processing with Arkindex

Solène Tarride
Mélodie Boillet
Christopher Kermorvant

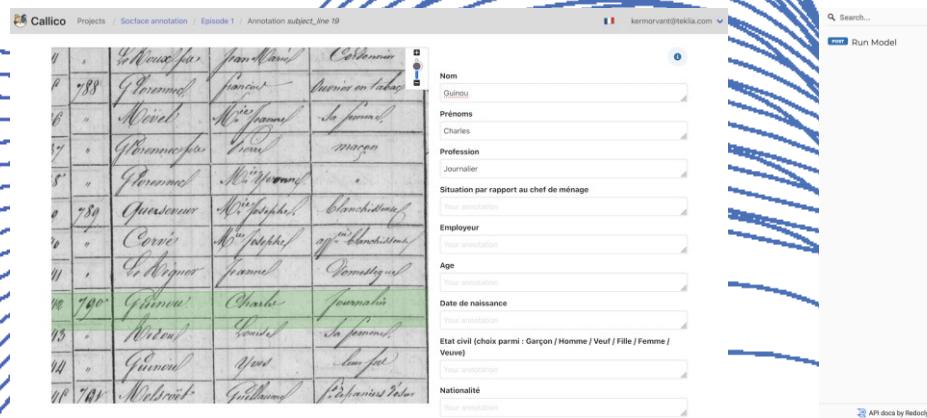


TEKLIA's software suite for document processing

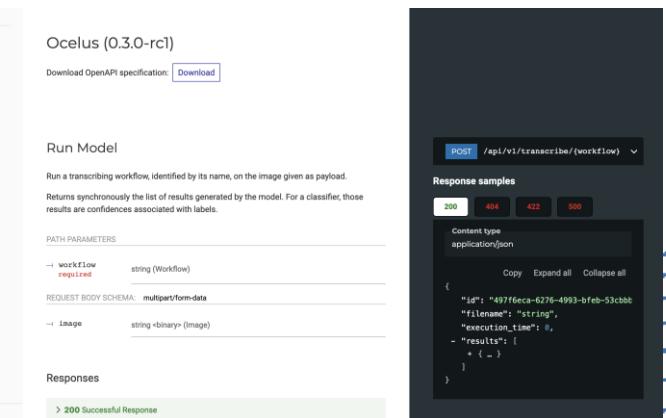
Arkindex



Callico



Ocelus



Yet another document processing platform?

Yes, but

- Active development since 2010 (23 releases since June 2020)



THE IKEA EFFECT
WE LOVE IT MORE IF WE MADE IT

Arkindex : document processing platform

Arkindex Projects Process Search

SYNTHESYS+ / Folder herbarium / Page BM000522091

Page BM000522091 A+

Filter by type... Hide

Text line 1 Specimen 61

Text line 2

Text line 3

Text line 4

Text line 5

Text line 6

Text line 7

Text line 8

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

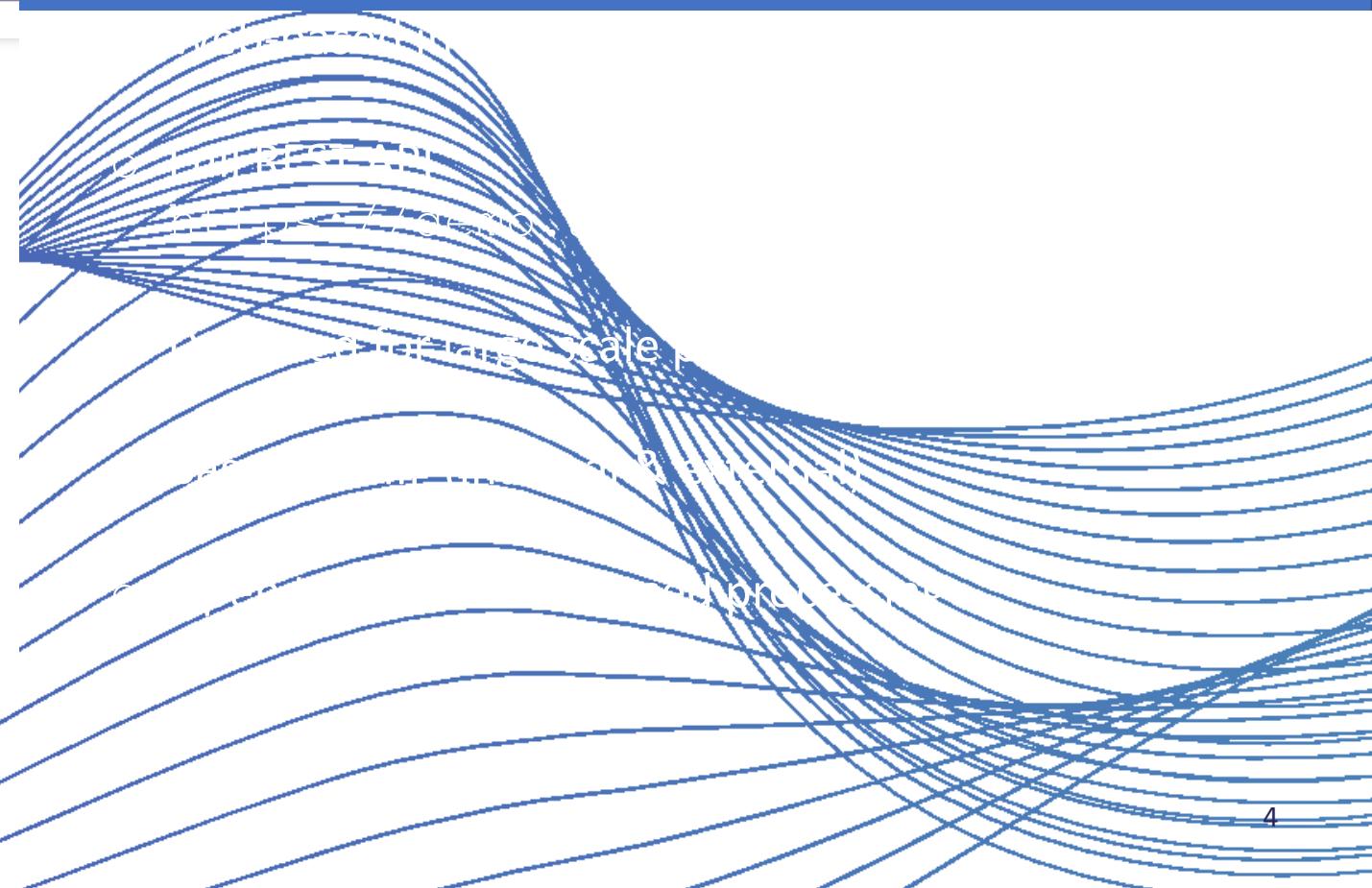
Text line 23

Text line 24

View source image List elements on this image

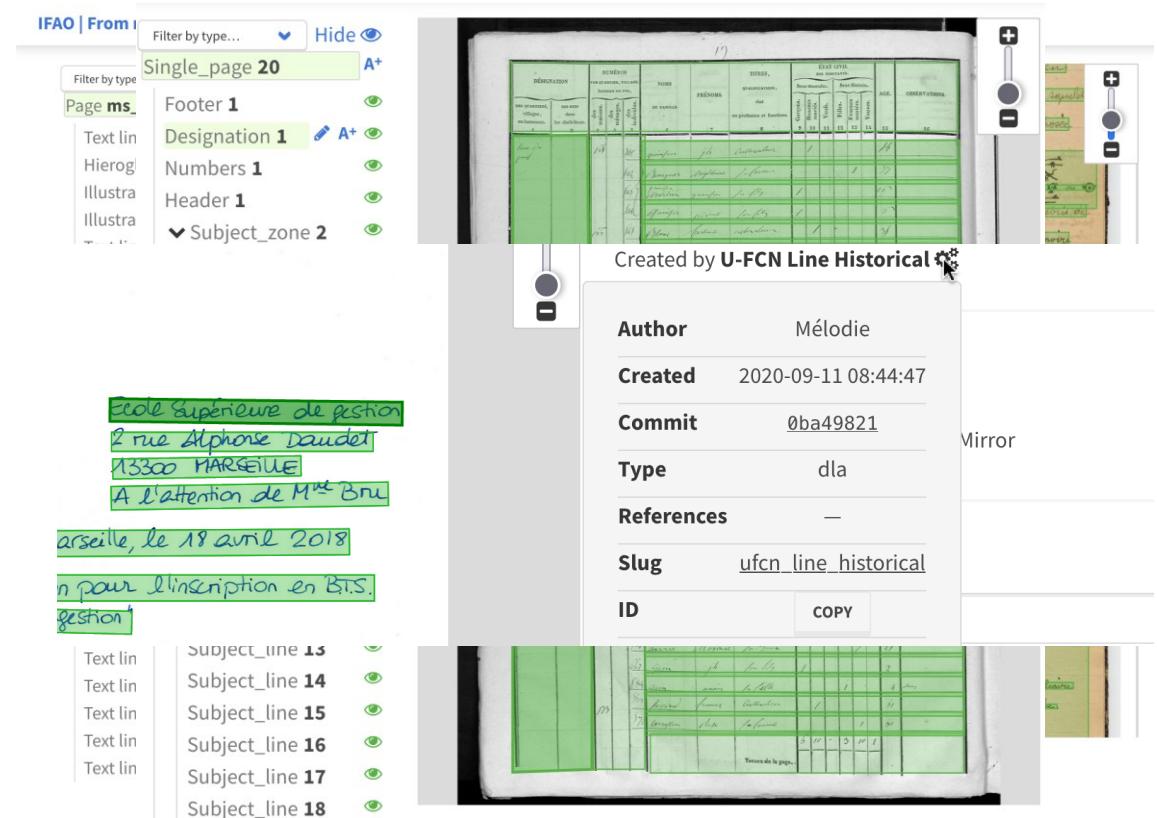
The screenshot shows the Arkindex interface for a herbarium specimen. On the left, a thumbnail of the scanned page is displayed with various green highlights and annotations. A sidebar on the left lists 24 text lines and one specimen entry, each with a visibility toggle. The main panel on the right contains the following sections:

- Page BM000522091** (with a red error icon)
- ANNOTATE** button
- ORIENTATION** section with **Rotation** set to **0°** and a **Mirror** toggle switch.
- CLASSIFICATIONS** section showing a classification by **U-FCN Line Historical** with a confidence of **0%**. It includes a **bad_line_≤1%** entry with a 1% confidence level and a delete button.
- TRANSCRIPTIONS** section with a **MANAGE** button.
- NEW TRANSCRIPTION** section with a text input field labeled **Text...**.



Document structure description

- Fully custom element definition
 - for collection description
 - for document description
- Fully custom element hierarchy
- Identification of the source (manual, commit number)
- Confidence score
- Orientation (0°, 90°, 180°, 270°)



Document structure annotation

ArkIndex Projects Process Search

CK | Test / Folder SIFED / Folder Symposium International Francophone sur l'Ecrit et le Document (SIFED'2022) - Si... / Page 1

Filter by type... Hide

Symposium International Francophone sur l'Ecrit et le Document (SIFED'2022) <https://project.inria.fr/sifed2022/fr>

Page 1 Created by Internal

ORIENTATION Rotation 0° Mirror

CLASSIFICATIONS Add a classification

TRANSCRIPTIONS

NEW TRANSCRIPTION Text...

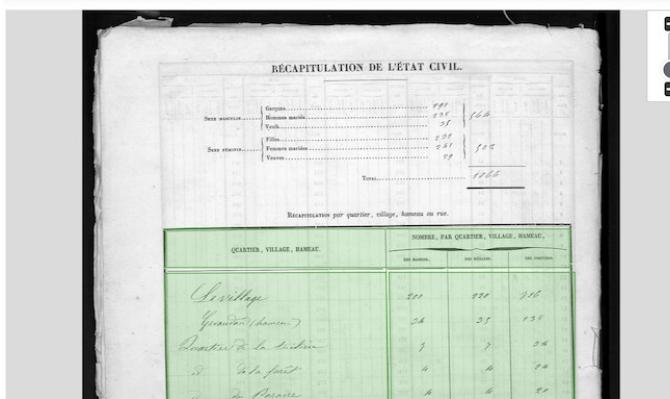
METADATA

Page 1 of 3 View source image List elements on this image 13/10/2022 13:41

The screenshot shows the ArkIndex platform interface for document structure annotation. The left sidebar lists 'Text line 1' through 'Text line 21' with corresponding icons. The main content area displays the 'Symposium International Francophone sur l'Ecrit et le Document (SIFED'2022)' website. The right panel contains annotation tools for 'Page 1', including orientation controls (Rotation 0°, Mirror), classification fields, transcription fields (with a 'Left to Right' dropdown), and metadata fields. A large blue 'ANNOTATE' button is prominent.

Element classification

- A class can be added to any element
- Custom class definition, auto-complete
- Multilabel classification, confidence level
- Identification of the source (manual, commit number)



RECAPITULATION DE L'ÉTAT CIVIL

Single_page 37 🗑️

Created by Valentin Rigal (admin)

+ ANNOTATE

ORIENTATION

CLASSIFICATIONS

Manual

Recap_page 📈

Add a classification

Filter by name

Items 1 to 5 out of 5 classes

Name	ID	Actions
Front_page	48a84164-e496-4d96-93e7-947087783eda	
List_page	8dc1c68f-7c7e-4744-a9fe-9ec650845fc	
Other_page	f0d4ffe6-3557-4cee-adbc-61b6c4c0bd81	
Recap_page	a81c8d61-6025-49a2-975a-80a0eec51a3f	
Totals_page	7fc0d1bb-731a-4a07-86d2-e40d293c0d73	

New class name...

Socface | FRAD004 / Folder 0_Sample100

Filter elements...

Items 1 to 20 out of 100 results

Page	File Type	Annotations
1	Front_page	
20	List_page	
1	Front_page	
37	Recap_page	
7	List_page	

Element transcription

- Transcription on any element
- Multiple versions supported
- Identification of the source (manual, commit number)
- Confidence score
- Right-to-left support (Arabic)

The screenshot shows a handwritten note in Norwegian. The text reads:

Korle lev vech laug vob til at hæve hæns ~~taas~~
Nåon leve videre. Der gaar ingen Dag, uden at
det mindes i Norge. Men selv om han ikke hænde
væch saa lykkelig at finde Melodien til "Ja
vi elsker", vilde Musikhistorien have bevare
hæns jaa Kompositionen med pæbels findt Punkten.
Så der er næppe en af dem, som ikke rober hæns Tale.

Below the text, there are two transcription examples:

TRANSCRIPTIONS
Created by Norwegian - Hugin Munin [pylaia] 50%
Nåon leve videre. Der gaar ingen Dag, uden at
Created by Norwegian - Hugin Munin [kaldi] 42%
Nåon leve videre . Der gaar ingen Dag , uden at

MANAGE

Named-entity

- Custom entities with types and subtypes
- Position of entities on transcription
- Link to entity reference list (entity linking)
- Right-to-left support (Arabic)

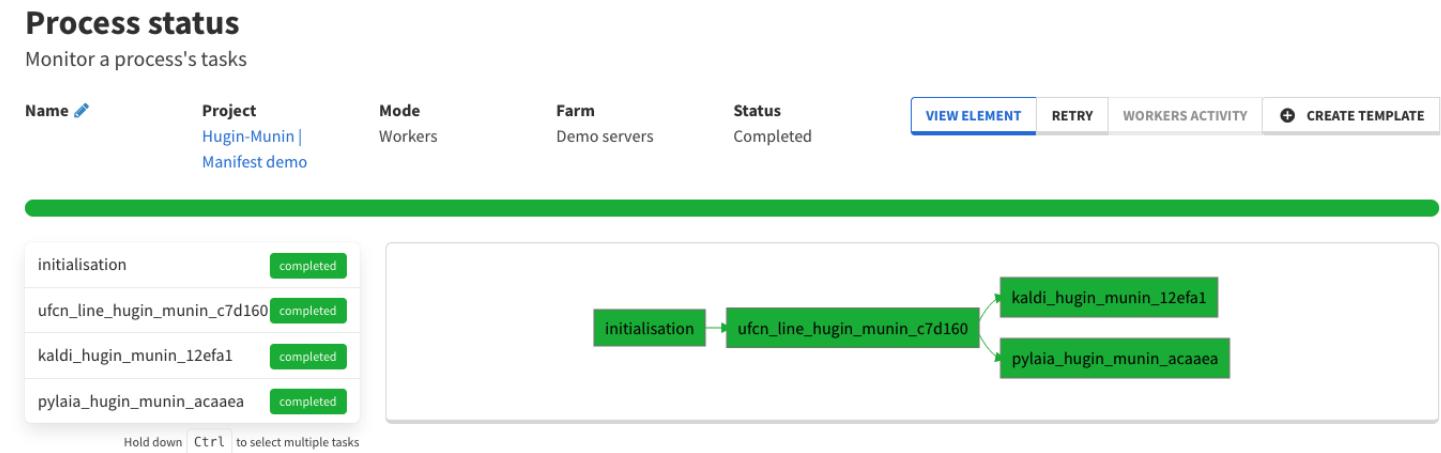


Entities in Himanis | TEKLIA processing

Name	Dates	Metadata	Worker
beati dyonisii in francorum		—	subtype: LOC spaCy Multilingual Home C3PO4 [All data]
cardinal dei gratia francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
francorum		—	subtype: LOC spaCy Multilingual Home C3PO4 [All data]
girardi de bosco philippus dei gracia francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
guillelmo deiemovicus clerico nuncium illarui domini nostri apudovici dei gracia regis francorum		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
hugo de ainsiato clerici domini nostri francorum regis		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
hugonem de cella militer ch dei gratia francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
hugonem ilz par la grace de dieu roys de francorum		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
ludovicus dei gratia francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
ludovicus quondam francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petro dominum dei gratia francorum et navarre rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de broto miles domini nostri francorum		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de calardo miles domini regis francorum dominus		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de columpa militis domini bien francorum regis		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de ferrarisius miles domini nostri regis francorum		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de fevoillustris francorum rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus dei gracia francorum registrorumiversis		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de mornayo miles dominus de feritate nobilberti senesc lugarcassent impendent domini nostri francorum regis		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus de pugnibus domini nostri francorum regum		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]
petrus opus autissiodorum karolus dei gratia francorum et navarre rex		—	subtype: PERS spaCy Multilingual Home C3PO4 [All data]

Processing workflow

- Select the elements to process
- Select the “workers”
- Configure the workflow
- Select the resources (cluster, GPU, splits)
- Monitor the progress



Import/Export

Import Web

- Local images, PDF, PDF-text
- Directories of images (S3)
- IIIF Manifest

Import CLI <https://cli.arkindex.org/upload/>

- Transkribus collections (wip)
- PageXML
- AltoXML

ArkIndex Projects Process Search

Import files to Folder SIFED

Upload files

From local files

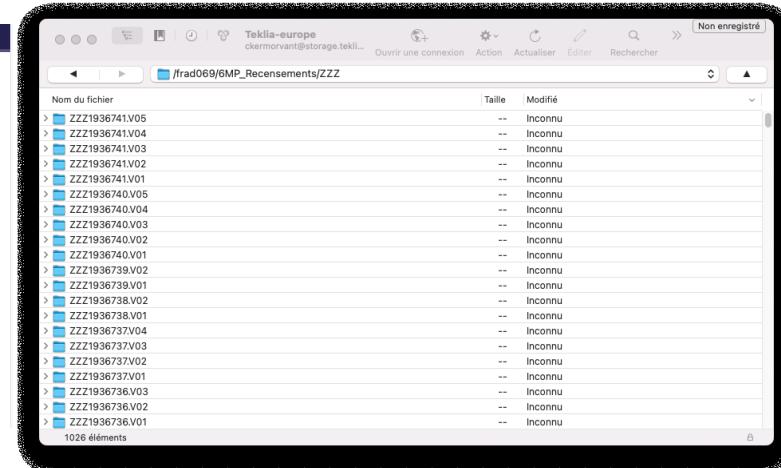
From a URL

Available files to import

Symposium International Franco... application/pdf - 515.28 KB

UNSELECT ALL

Advanced settings



Export <https://cli.arkindex.org/export/>

- PDF-text
- Alto XML
- CSV
- Fully custom (sqlite database)

ArkIndex CLI

Commands Authentication Elements

Data import

- Images stored on an S3-compatible bucket (MinIO)
- Page XML documents
- Alto XML documents

IIF Images

- Required arguments
- Optional arguments
- Usage examples
- Grouping images by prefix
- Recreating a folder hierarchy

Data export

- ML Classes
- Processes
- Models
- Secrets
- Benchmark

Upload commands

The `upload` subcommands allow you to import data to Arkindex.

Images stored on an S3-compatible bucket (MinIO)

The `minio` subcommand generates IIF URLS for images that are stored on a given S3-compatible (AWS, MinIO, Ceph...) bucket, which can then be uploaded to Arkindex using the `IIF import` subcommand.

```
arkindex upload minio -b $BUCKET_NAME
```

If there are multiple folders on the target bucket, the subcommand will output one file per folder, containing the URLs for the images in this folder.

Authentication

Before running the `minio` subcommand, you need to authenticate yourself with credentials for an account that has access to the bucket you're targeting. This authentication is done through environment variables.

```
export MINIO_ACCESS_KEY=$YOUR_ACCESS_KEY  
export MINIO_SECRET_KEY=$YOUR_SECRET_KEY
```

Required arguments

Usage/Profile matrix

	Web interface					Command Line Interface		API	
Profile	Visualization /annotation	Few documents	OCR	Complex workflow	Standard Import / Export	Many documents	Custom Import / Export	Custom workers	
Simple web user <i>Knows how to click</i>	✓	✓	✓						
Power user <i>Reads the doc</i>				✓	✓	✓			
Developer <i>Knows python/API</i>							✓	✓	

CALLICO

(open-source) Annotation and validation platform

Layout

This screenshot shows the CALLICO interface for annotation. On the left is a thumbnail of a newspaper page titled 'L'Ouest-Eclair'. A sidebar on the right lists 'Type' options: Article, Article, and Article. Buttons for 'Ignorer la tâche' (Ignore task) and 'Annoter' (Annotate) are present. The footer indicates 'Version 0.3.3'.

Classification

This screenshot shows the classification interface. It displays a double-page spread of a historical manuscript. To the right, classification levels are listed: Level1, Level2, and Other. A 'Fill in the tâche' button is also visible. The footer indicates 'Version 0.3.3'.

Transcription

This screenshot shows the transcription interface. It displays a paragraph of French text: "La première, en considération des services de feu son mari dans les armées, dans le commandement de la Flandre, dans les conseils du roi, et dans la charge de secrétaire d'Etat au département de la guerre." A text input field for transcription is shown, along with 'Ignorer la tâche' and 'Annoter' buttons. The footer indicates 'Annotation de l'élément "paragraph 38"'.

Multi-line transcription

This screenshot shows the multi-line transcription interface. It displays a handwritten letter with several lines of text underlined for annotation. A sidebar on the right lists 'Annotate "text_line" elements': neutralité définitive de l'Italie, s'arrangeant avec l'Autriche pour, croire que cette dernière sa pourvoir, dégagé les fractions de l'Italie, de toute les troupes pourra avec celles-là vaincre les Russes. Ils espèrent encore que même si les Autrichiens faisaient la paix avec la une, cela serait favorable pour la Russie. A sidebar at the bottom right includes 'Skip task' and 'Annotate' buttons. The footer indicates 'Version 0.3.3'.

Meta-data

This screenshot shows the meta-data interface. It displays a stamp with fields for 'Ville envoiée' (Address of the receiver) and 'Date'. A sidebar on the right lists 'Fill in the sending date from the document' and 'Address of the receiver (street, town, ...)' with a 'Fill in' button. A 'Fill in the tâche' button is also present. The footer indicates 'Version 0.3.3'.

Named-entity

This screenshot shows the named-entity interface. It displays a handwritten note with several entities annotated with colored boxes. A sidebar on the right lists 'Date', 'Personne', 'Lieu', and 'Organisation'. A 'Fill in the tâche' button is also present. The footer indicates 'Kristiania 24.oktober 1902'.

+ annotator management, crowd-sourcing, campaign management

Coming soon

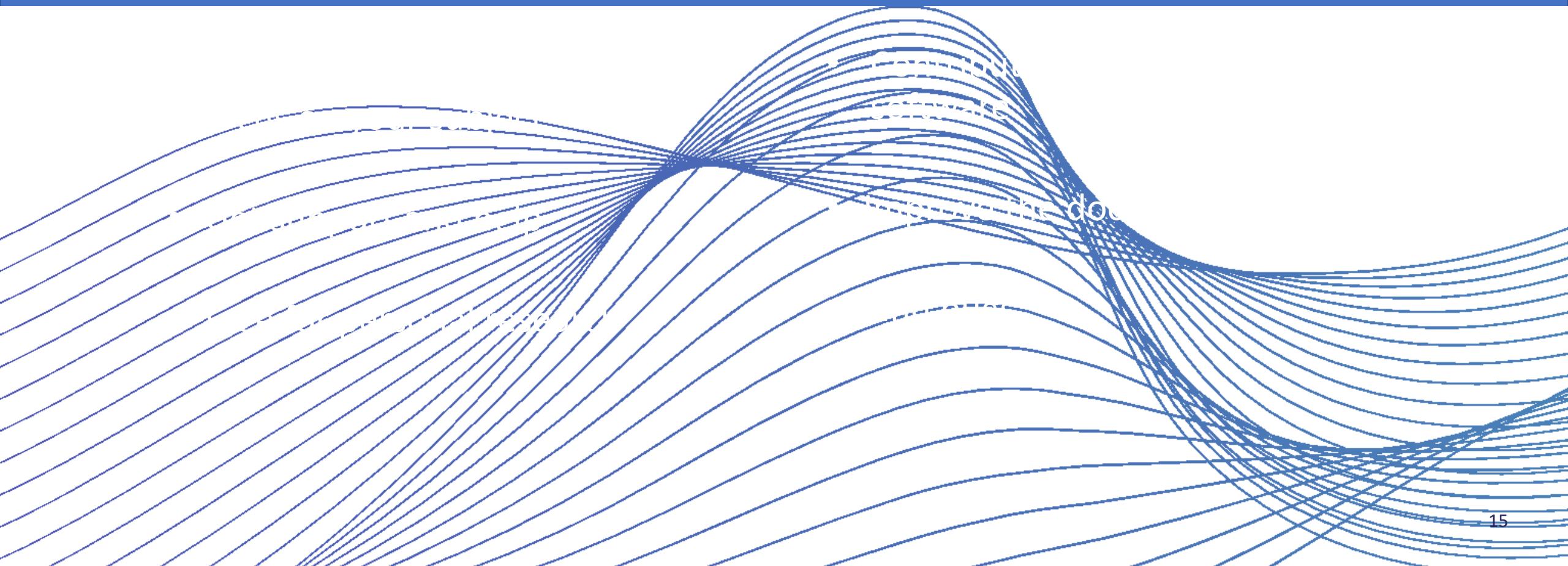
End of 2022

- Training models in Arkindex
- Open-sourcing Callico
- Connexion Arkindex - Callico

Why should you use Arkindex?

For you

For us



See you at the demo !

Solène Tarride
starride@teklia.com

Mélodie Boillet
boillet@teklia.com

Christopher Kermorvant
kermorvant@teklia.com