

TEKLIA

Document processing with Arkindex

Solène Tarride

Mélodie Boillet

Christopher Kermorvant

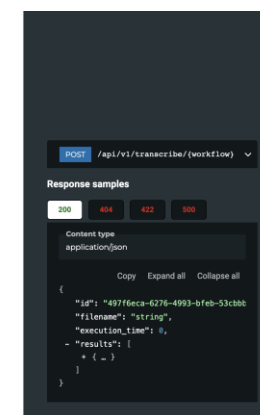
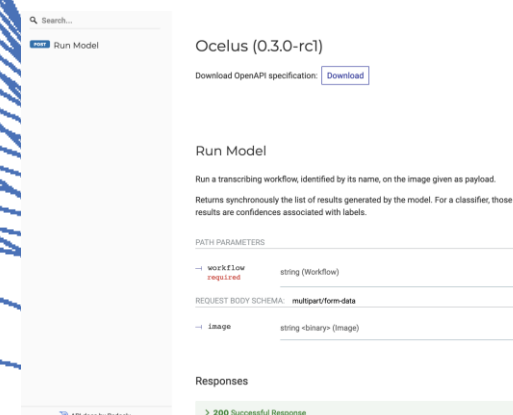
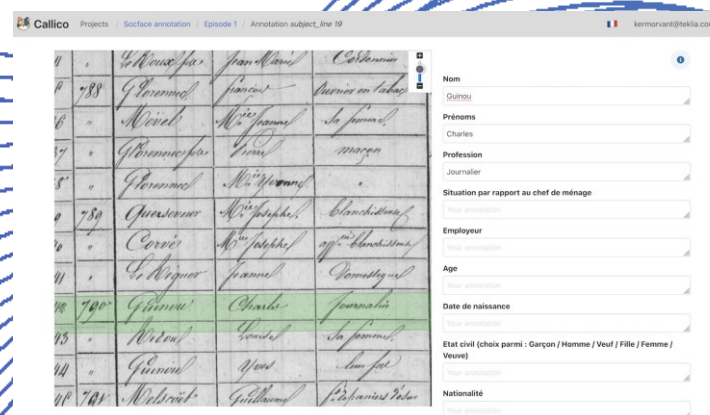
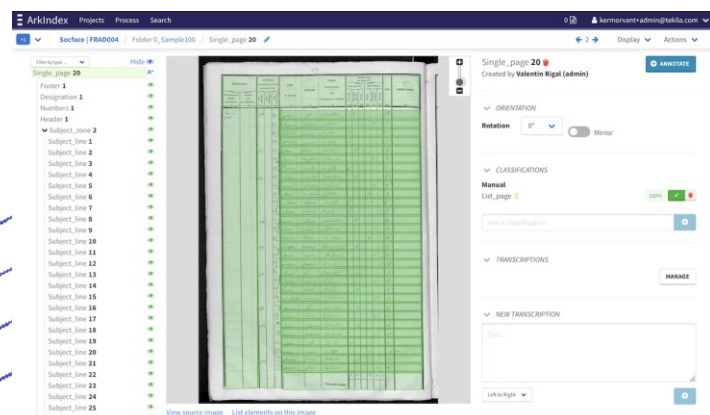


TEKLIA's software suite for document processing

Arkindex

Callico

Ocelus



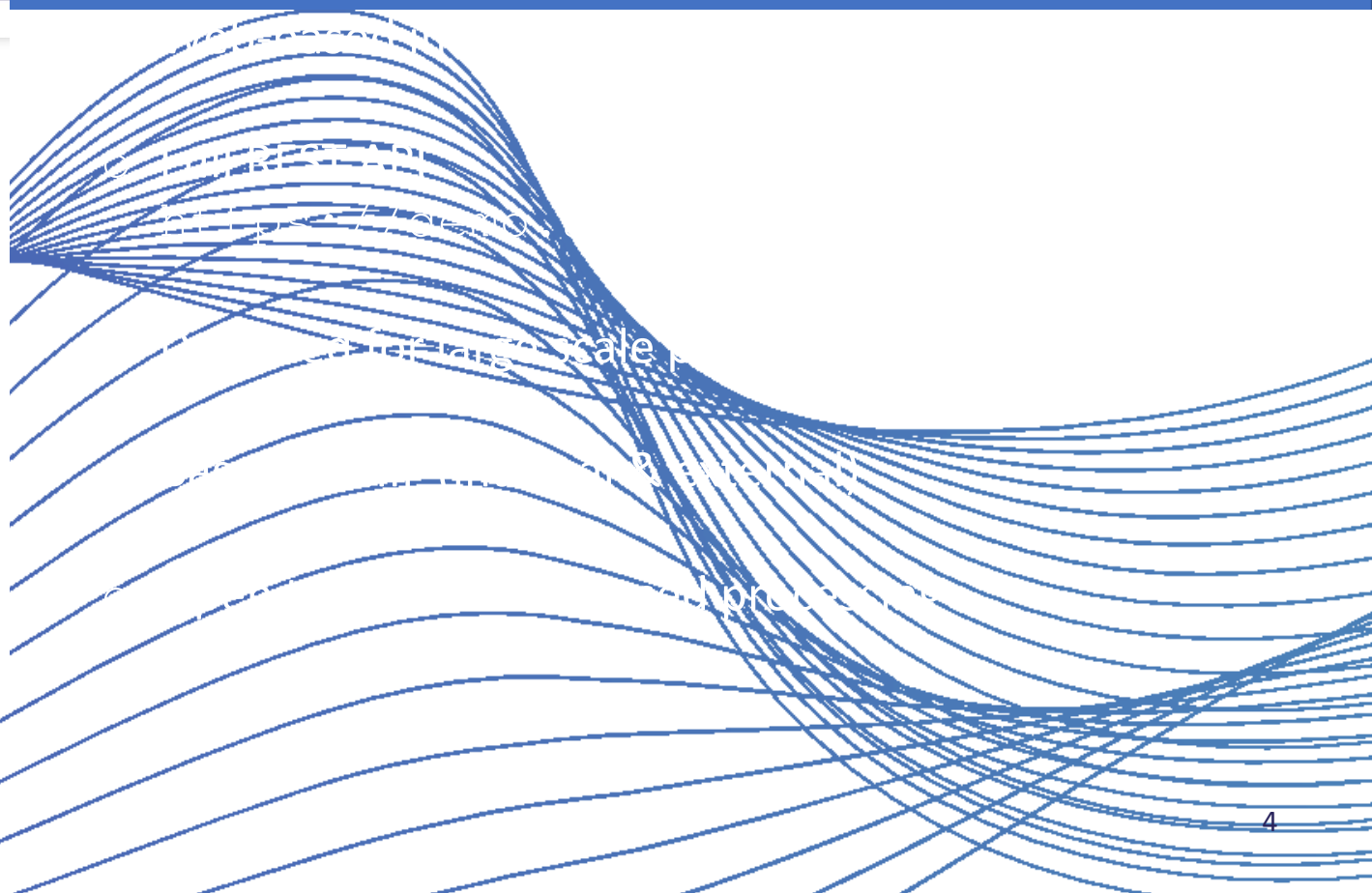
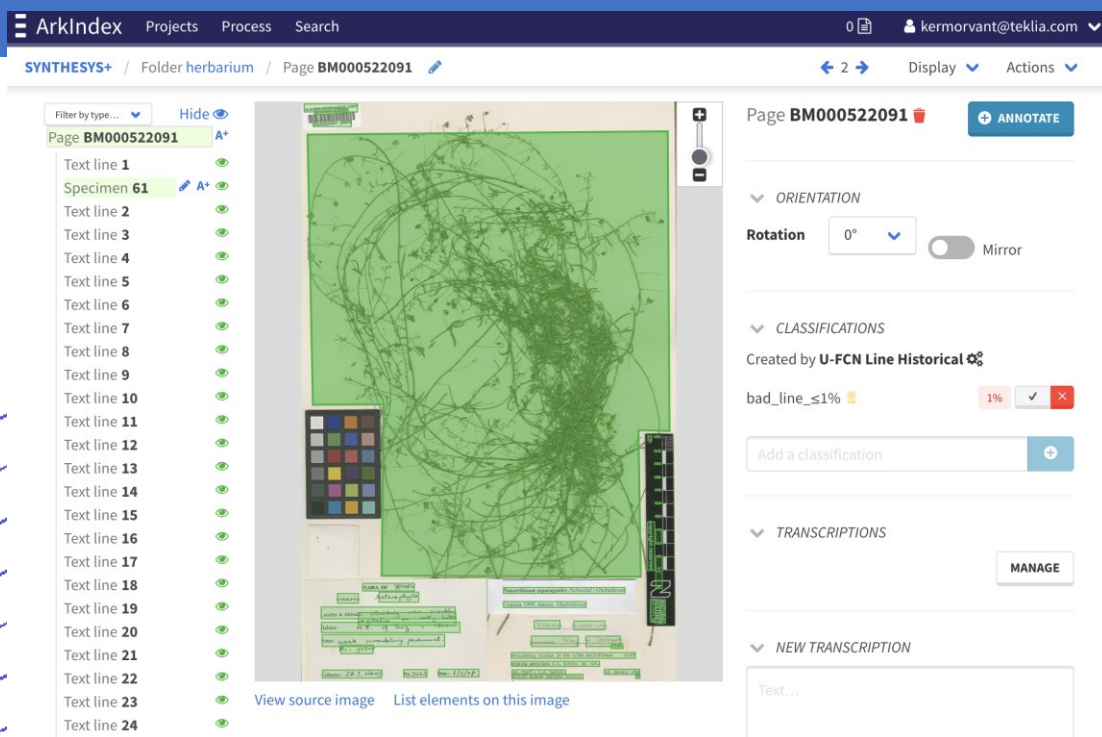
Yet another document processing platform?

Yes, but

- Active development since 2010 (22 releases since June 2020)



Arkindex : document processing platform



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°)

The screenshot displays the IFAO | From IFAO interface. On the left, a sidebar shows a list of elements: Page ms, Text lin, Hierog, Illustra, and Illustra. The main area shows a document page with a table and text. A metadata panel on the right contains the following information:

- Created by: U-FCN Line Historical
- Author: Mélodie
- Created: 2020-09-11 08:44:47
- Commit: 0ba49821
- Type: dla
- References: —
- Slug: ufcn_line_historical
- ID: COPY

The document page shows a table with columns: RÉSUMÉ, DÉSIGNATION, DÉTAILS, and DÉTAILS. The table contains handwritten entries. Below the table, there is a text area with the following text:

École Supérieure de Gestion
2 rue Alphonse Daudet
13300 MARSEILLE
A l'attention de M^{me} Bru
Marseille, le 18 avril 2018
pour l'inscription en BTS.
Gestion

Document structure annotation

ArkIndex Projects Process Search 13 0 kermorvant@tekli.com

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

Filter by type... Hide Page 1

- Text line 1
- Text line 2
- 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 17
- Text line 22
- Text line 23
- Text line 18
- Text line 19
- Text line 20
- Text line 21


Symposium International Francophone sur l'Ecrit et le Document (SIFED'2022)

ACCUEIL APPEL A CONTRIBUTIONS PROGRAMME INSCRIPTIONS

INFOS PRATIQUES PARTENAIRES Recherche

Bienvenue à SIFED 2022!

SIFED 2022 (Symposium International Francophone sur l'Ecrit et le Document) sera organisé par l'équipe Intuidoc de l'IRISA à **Rennes le 14 octobre 2022**. SIFED, anciennement CIFED, est proposé par le **GRCE** (Groupe de Recherche en Communication Ecrite) depuis 1990.



Journée SIFED

Cet événement a toujours eu pour ambition de fédérer les chercheurs (jeunes chercheurs, chercheurs seniors, académiques et industriels) intéressés par les travaux autour de l'analyse de documents et de l'écriture, en particulier sur les problématiques d'analyse et de traitement

Page 1 Created by Internal

ORIENTATION

Rotation 0° Mirror

CLASSIFICATIONS

Add a classification

TRANSCRIPTIONS

MANAGE

NEW TRANSCRIPTION

Text...

Left to Right

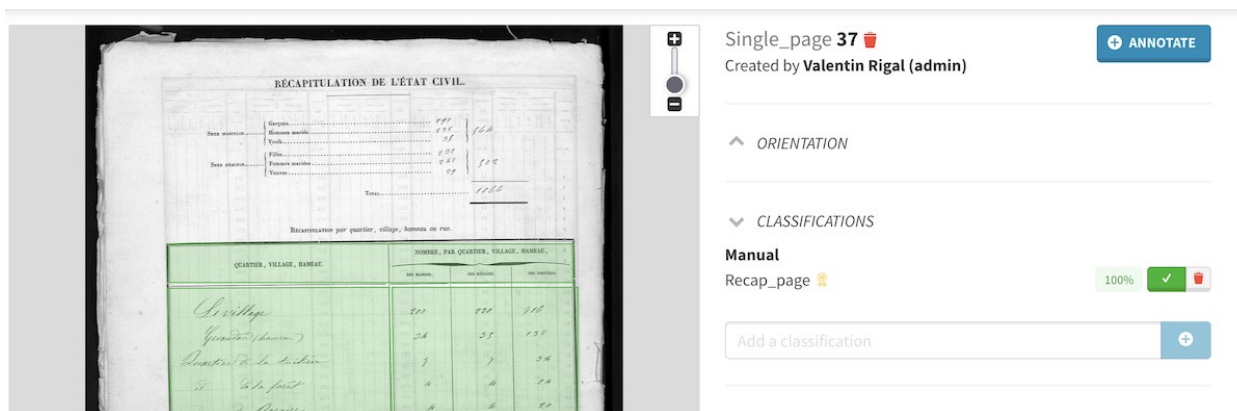
METADATA

1 of 3 8/10/2022 13:41

View source image List elements on this image

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)



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-9e6c650845fc	
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...

Position 1 2 ... 5

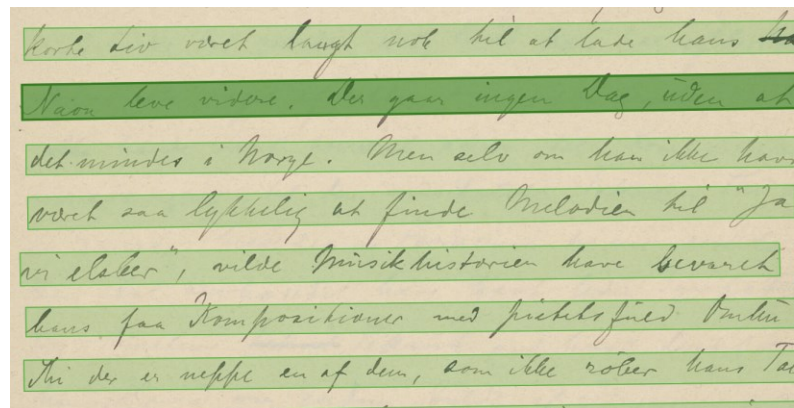
GO TO ... 1 2 ... 5

Items 1 to 20 out of 100 results

1	20	1	37	7
Front_page	List_page	Front_page	Recap_page	List_page
Single_page	Single_page	Single_page	Single_page	Single_page

Element transcription

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



▼ TRANSCRIPTIONS

Created by **Norwegian - Hugin Munin [pylaia]**  

50%  

Navn 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 | TEKLI processing

Filter by name

Francorum

Items 1 to 20 out of 36 results

Name	Dates	Metadata	Worker
 beati dyonisii in francorum	—	subtype: LOC	spaCy Multilingual Home C3P04 [All data]
 cardinal dei gratia francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 francorum	—	subtype: LOC	spaCy Multilingual Home C3P04 [All data]
 girardi de bosco philippus dei gracia francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 guillelmo deiemovicus clericus nuncium illuarii domini nostri apudovici dei gracia regis francorum	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 hugo de ainsiato clericus domini nostri francorum regis	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 hugonem de cella militis ch dei gratia francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 hugonem ilz par la grace de dieu roys de francorum	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 ludovicus dei gratia francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 ludovicus quondam francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petro dominum dei gratia francorum et navarre rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de broto miles domini nostri francorum	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de calardo miles domini regis francorum dominus	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de columpna militis domini bien francorum regis	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de ferrarius miles domini nostri regis francorum	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de fevroillustris francorum rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus dei gracia francorum registorumversis	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de mornayo miles dominus de feritate nobilberti senesc lugarcassent impendent domini nostri francorum regis	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus de pugibus domini nostri francorum regum	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]
 petrus opus autissiodorum karolus dei gratia francorum et navarre rex	—	subtype: PERS	spaCy Multilingual Home C3P04 [All data]

Processing workflow

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

Process status

Monitor a process's tasks

Name 

Project

[Hugin-Munin](#) |
[Manifest demo](#)

Mode

Workers

Farm

Demo servers

Status

Completed

[VIEW ELEMENT](#)

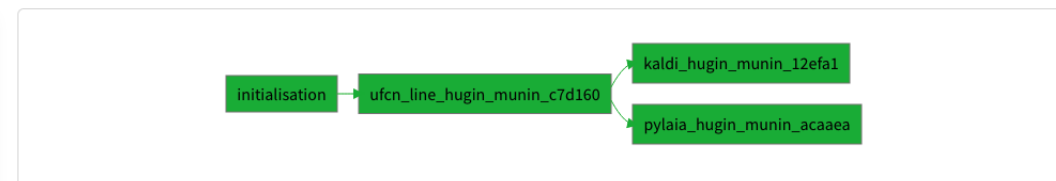
[RETRY](#)

[WORKERS ACTIVITY](#)

[+ CREATE TEMPLATE](#)

initialisation	completed
ufcn_line_hugin_munin_c7d160	completed
kaldi_hugin_munin_12efa1	completed
pylaia_hugin_munin_acaaea	completed

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



Import/Export

Import Web

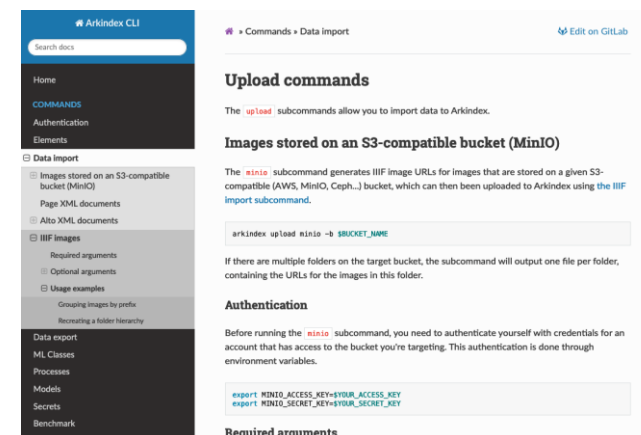
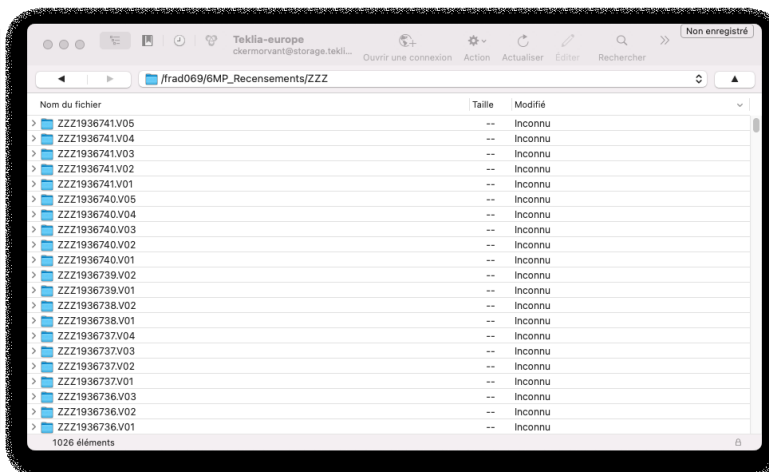
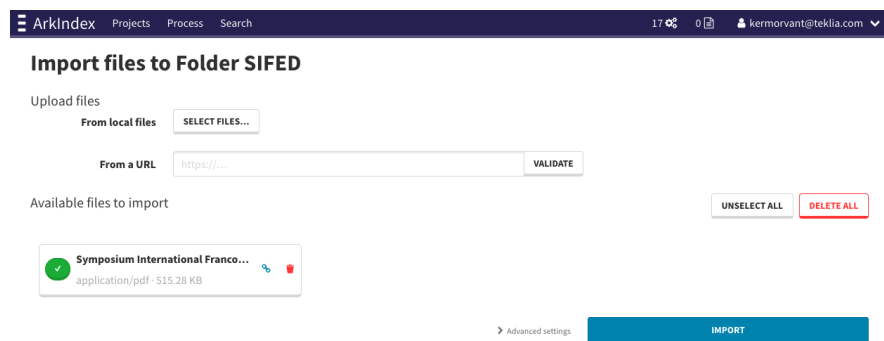
- Local images, PDF, PDF-text
- Directories of images (S3)
- IIIF Manifest

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

- Transkribus collections (wip)
- PageXML
- AltoXML

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

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



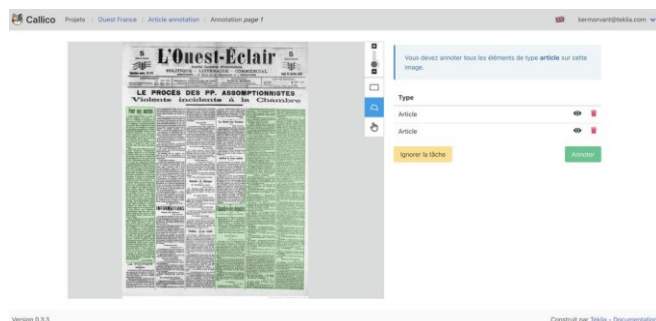
Usage/Profile matrix

	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>							✓	✓

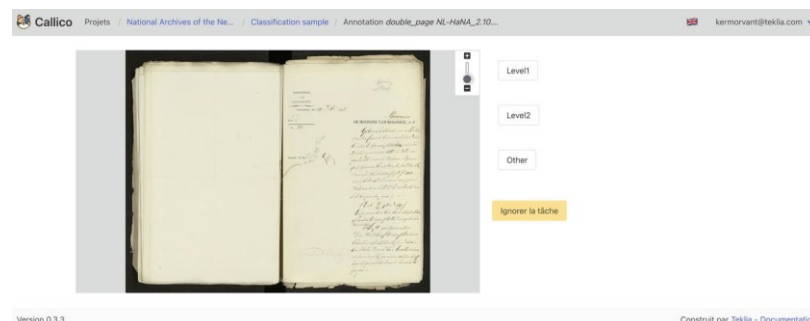
CALICO

(open-source) Annotation and validation platform

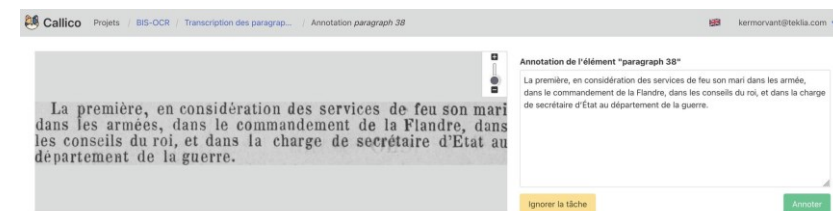
Layout



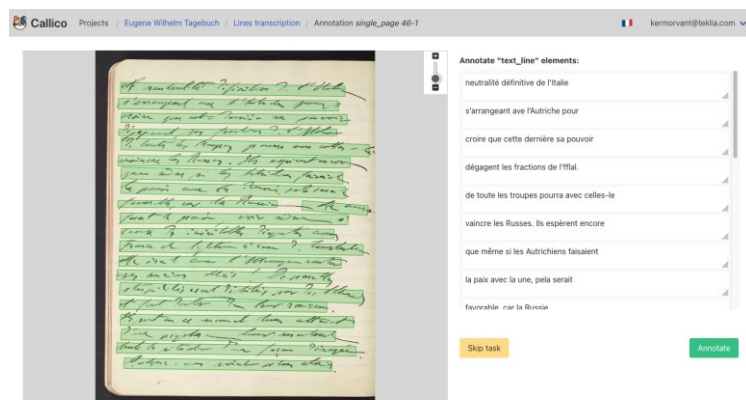
Classification



Transcription



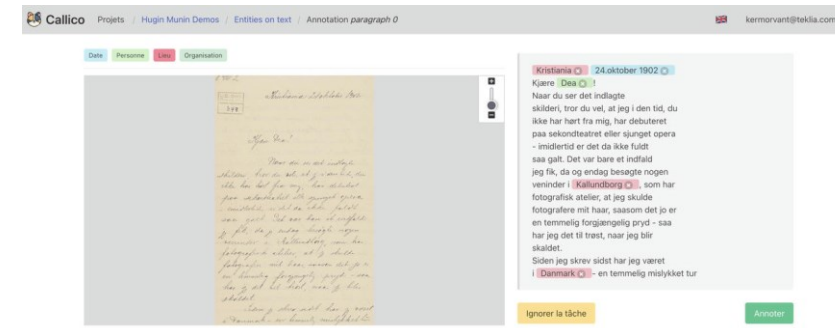
Multi-line transcription



Meta-data



Named-entity



+ annotator management, crowd-sourcing, campaign management

Coming soon

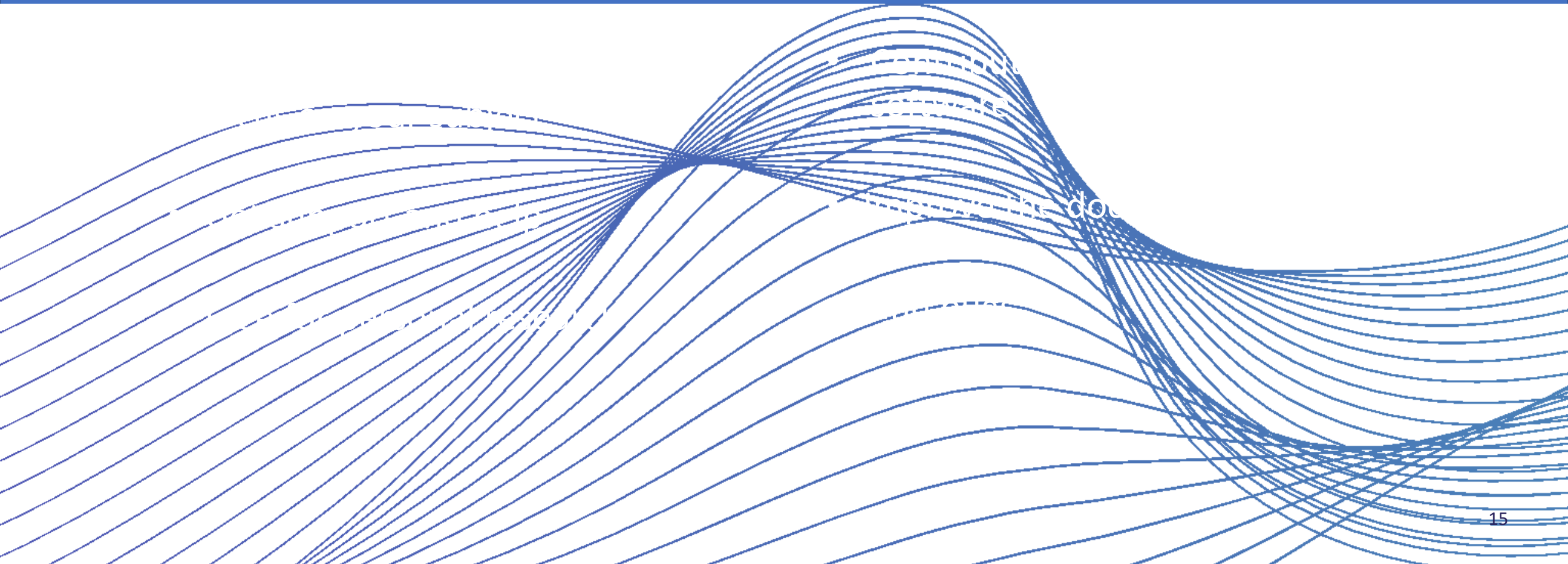
End of 2022

- 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@tekliia.com

Mélodie Boillet
boillet@tekliia.com

Christopher Kermorvant
kermorvant@tekliia.com



| www.tekliia.com