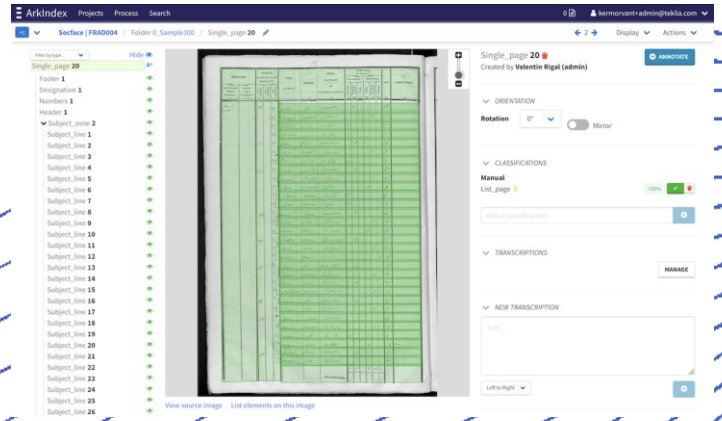
Document processing with Arkindex

Solène Tarride
Mélodie Boillet
Christopher Kermorvant

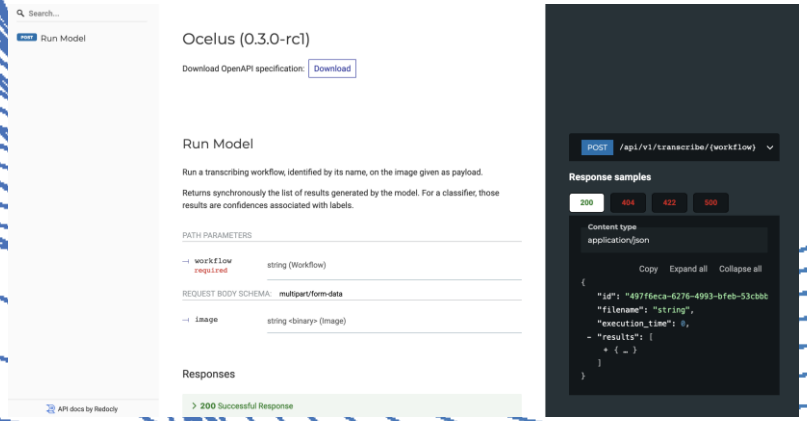
TEKLIA’s software suite for document processing

Arkindex



Callico

Ocelus

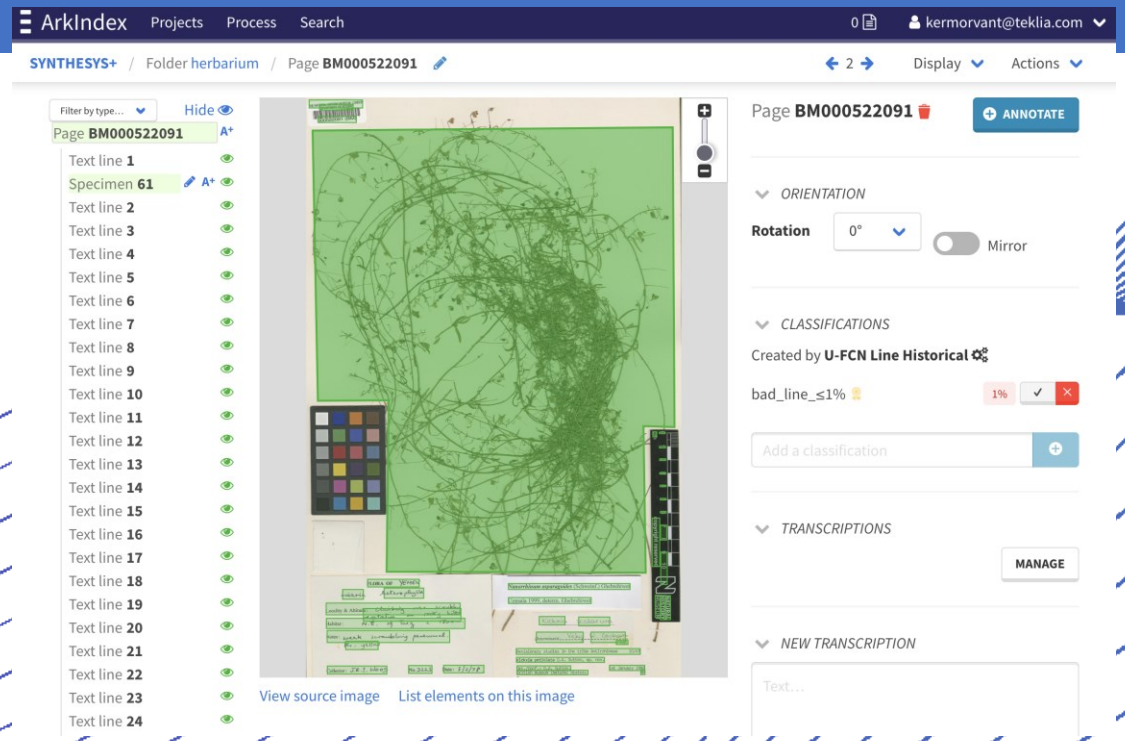


Yet another document processing platform?

Yes, but

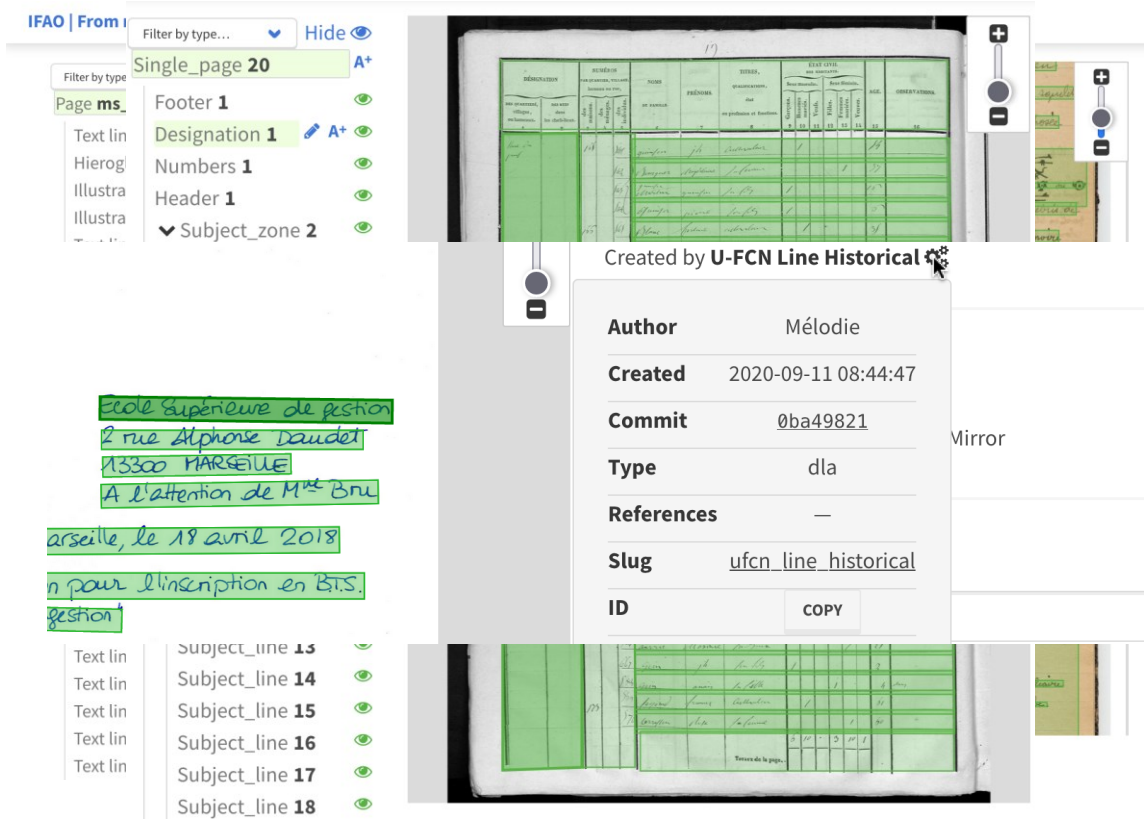
- Active development since 2019 (33 releases since June 2020)

Arkindex : document processing platform

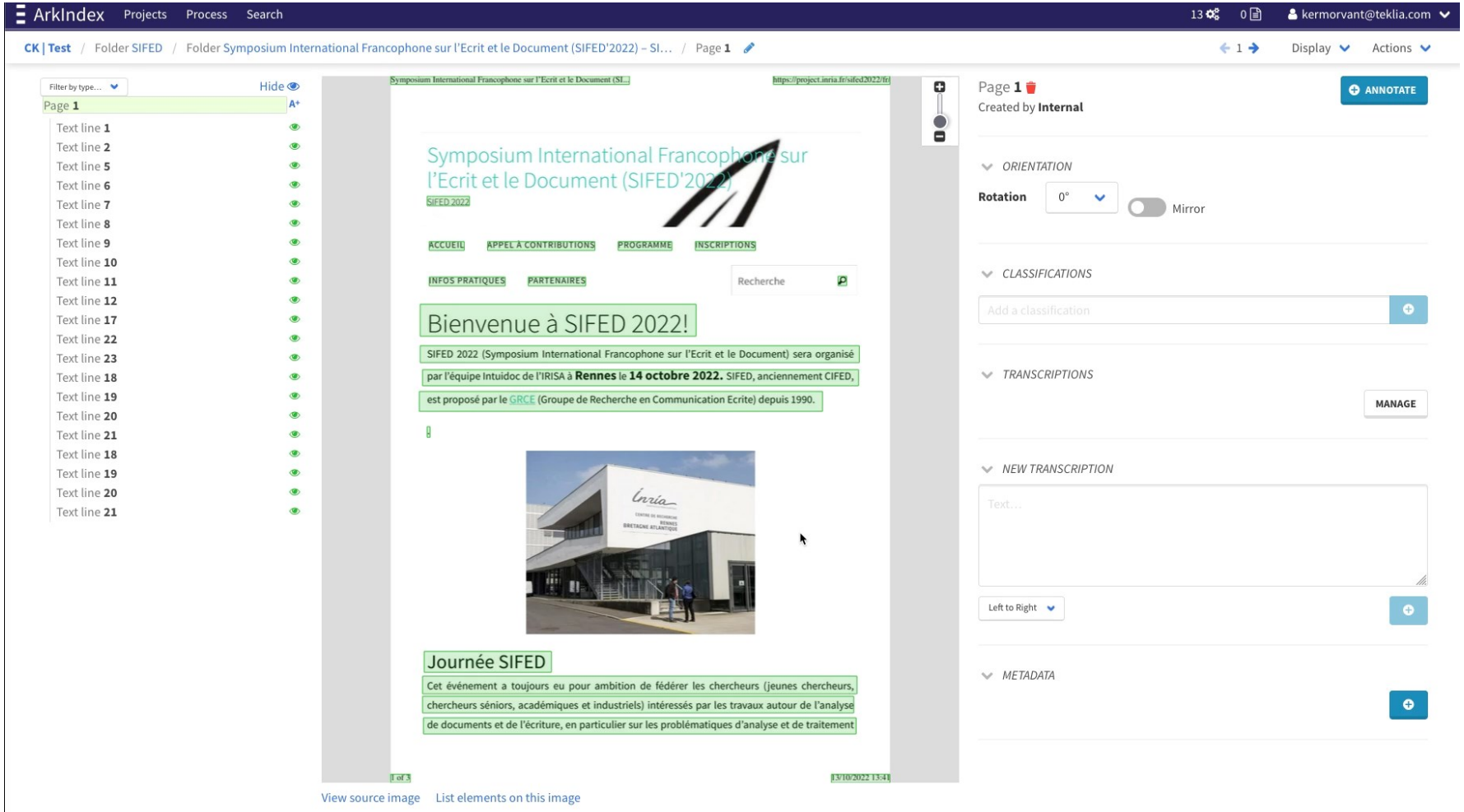


Document structure description

- 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°)

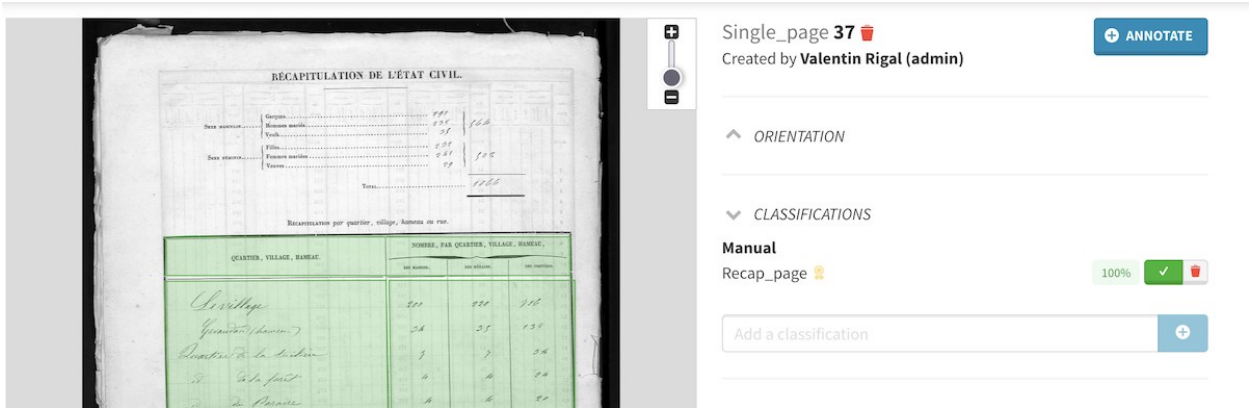


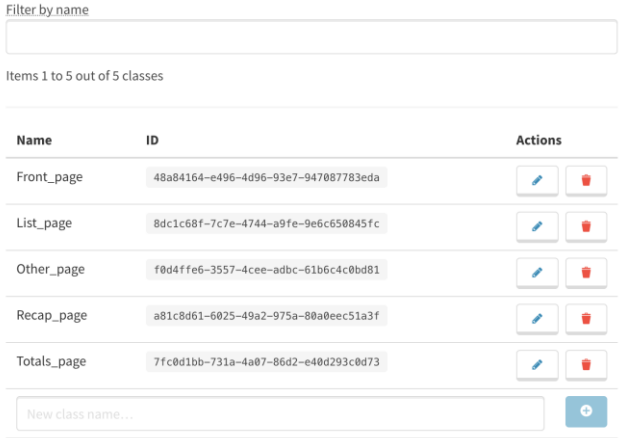
Document structure annotation

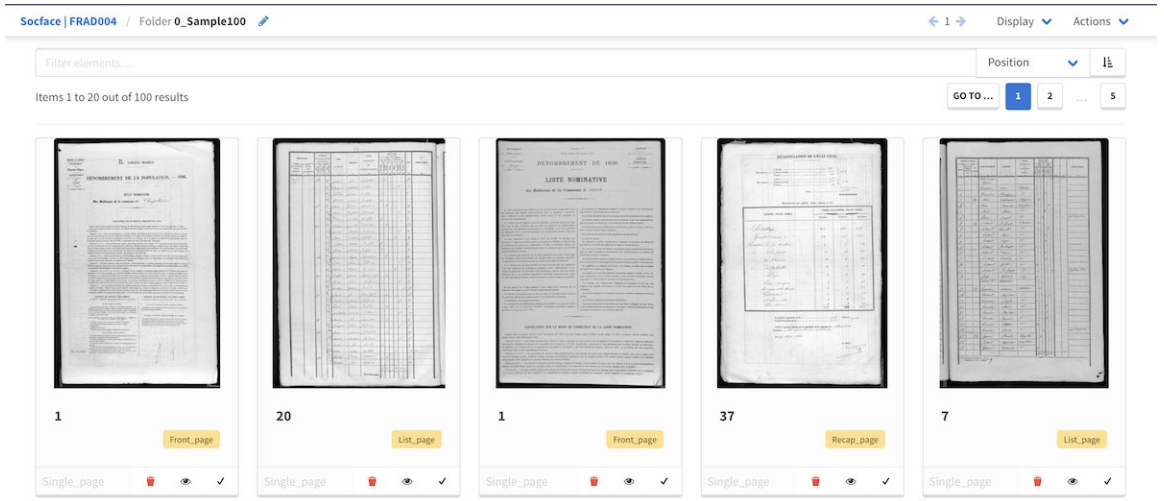


Element classification

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

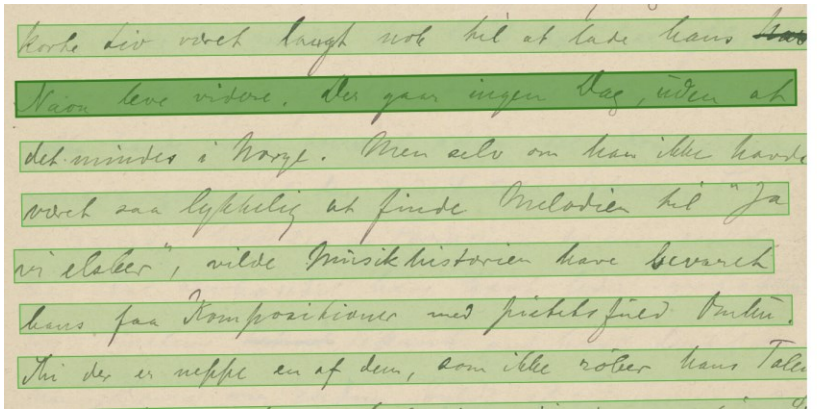


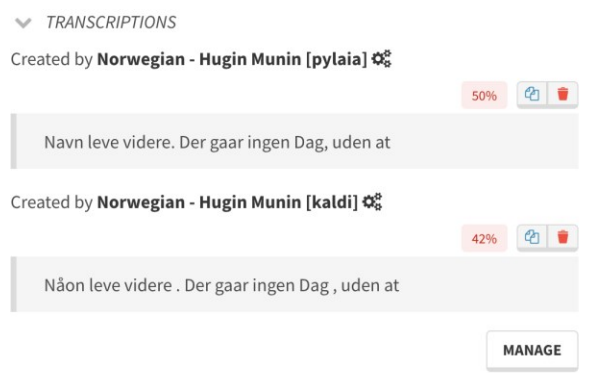




Element transcription

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





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

Filter by name

Francorum

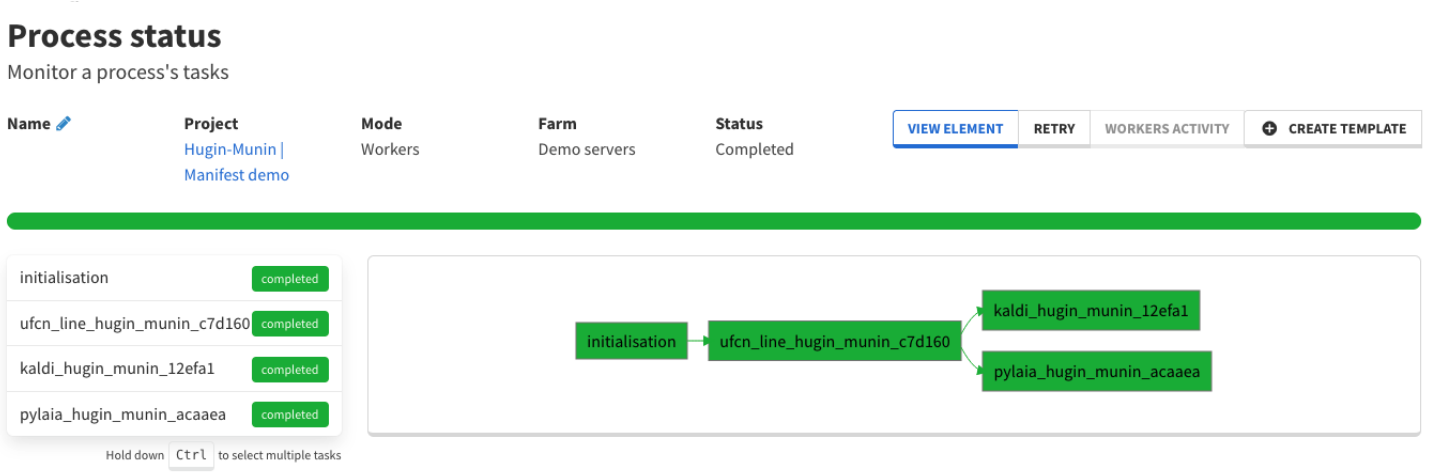
Items 1 to 20 out of 36 results

1 2

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 illuarii 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 columpna 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 fevroillustris 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

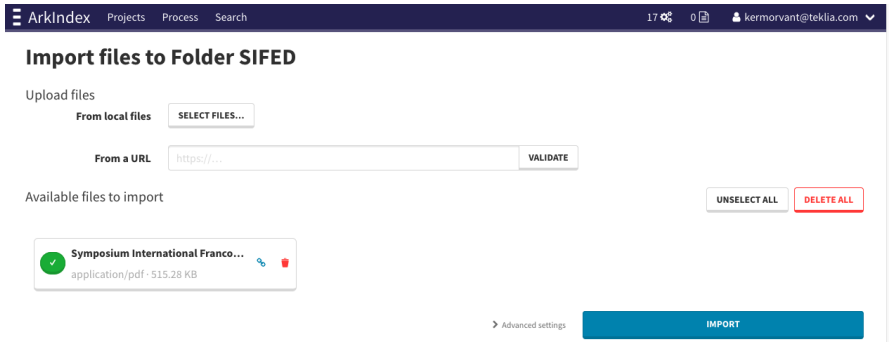


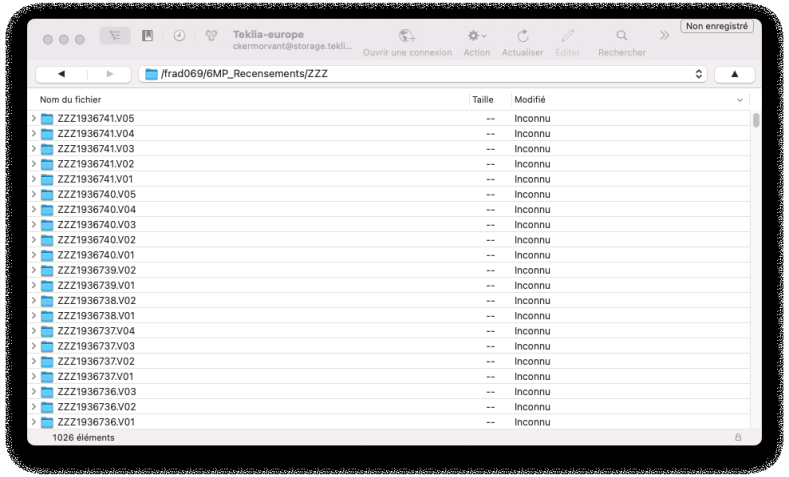
Import/Export

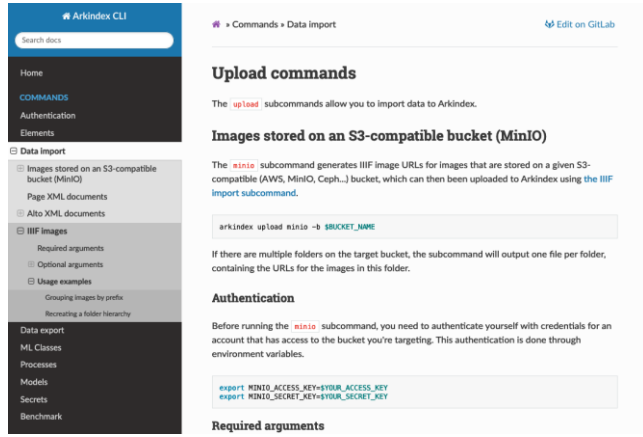
Import Web
- Local images, PDF, PDF-text
- Directories of images (S3)
- IIIF Manifest

Import CLI `https://cli.arkindex.org/upload/`
- Transkribus collections (wip)
- PageXML
- AltoXML

Export `https://cli.arkindex.org/export/`
- PDF-text
- Alto XML
- CSV
- Fully custom (sqlite database)







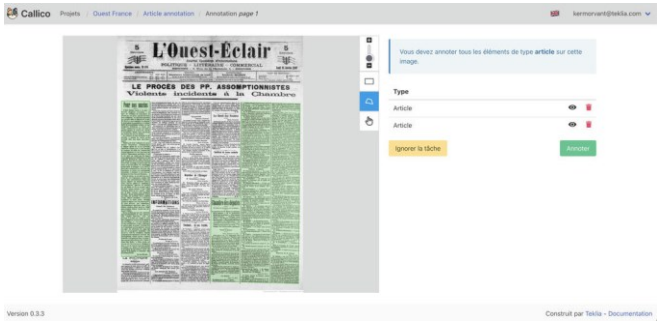
Usage/Profile matrix

	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 *Knows how to click*	✓	✓	✓					
Power user *Reads the doc*				✓	✓	✓		
Developer *Knows python/API*							✓	✓

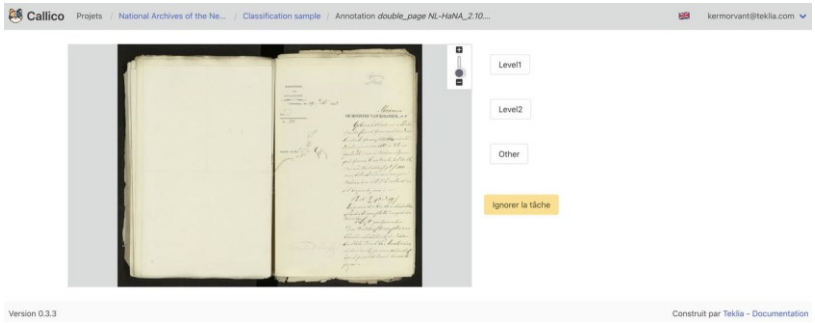
CALLICO

(open-source) Annotation and validation platform

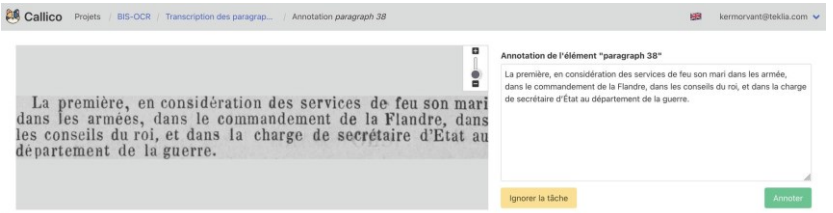
Layout



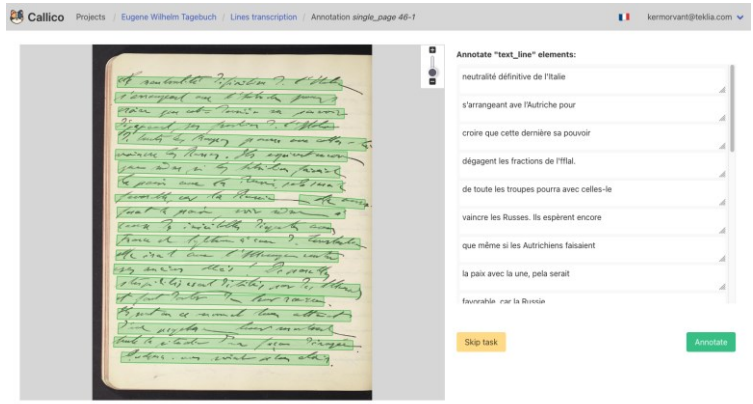
Classification



Transcription

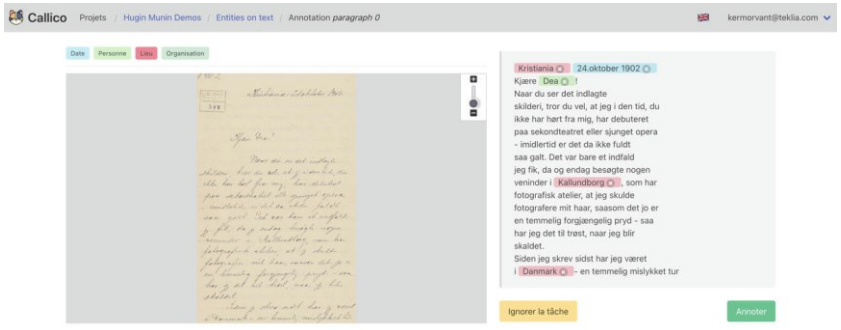


Multi-line transcription



Meta-data

Named-entity



+ annotator management, crowd-sourcing, campaign management

Coming soon

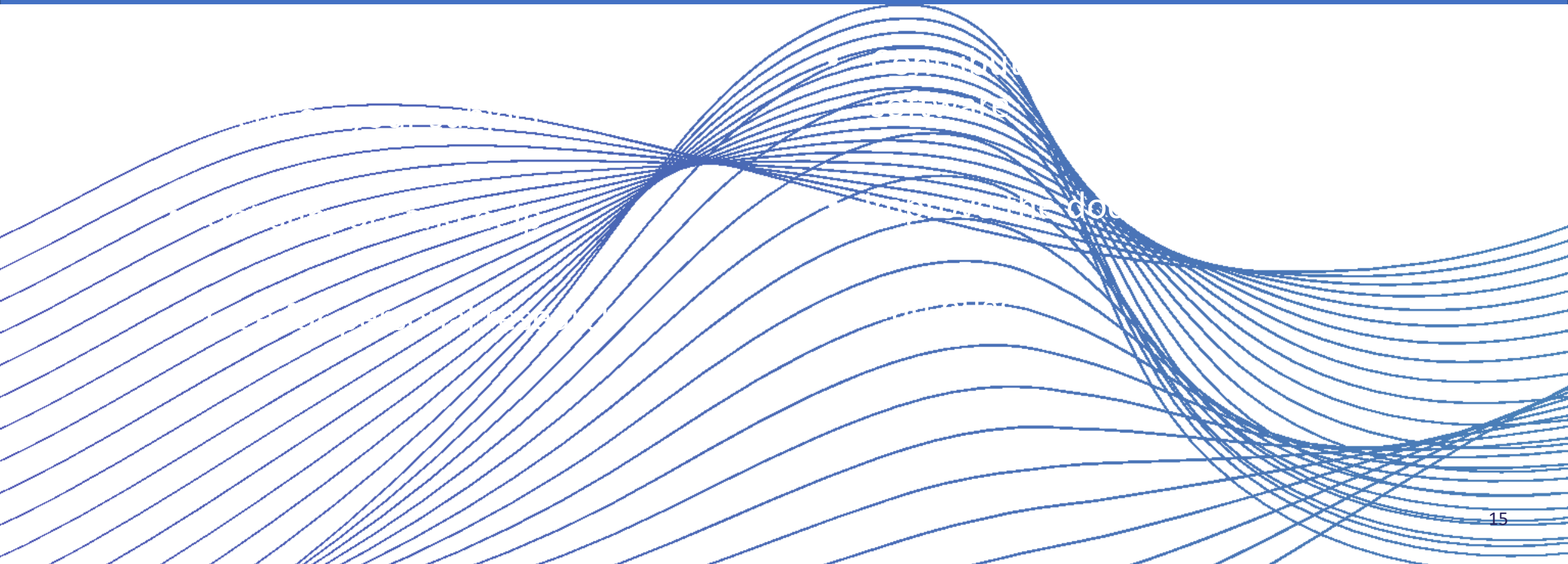
End of 2022

- 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@teklia.com

Mélodie Boillet
boillet@teklia.com

Christopher Kermorvant
kermorvant@teklia.com