



# Structuration, reconnaissance d'écriture et extraction d'information dans les documents avec Arkindex

Journée IA et Humanités Numériques - 2024

Christopher Kermorvant

TEKLIA, Paris, France

**T E K L I A**

# Les humanités numériques

Une discipline ou trans-discipline multiforme

- *Digitized humanities* : création et analyse ressources numérisées
- *Computational humanities* : mise au point de modèles computationnels
- *Humanities of the digital* : étude du phénomène digital
- *Public humanities* : communication numérique en humanités

Luhmann, J., & Burghardt, M. (2021). Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, 73(2), 148–171.

<https://doi.org/10.1002/asi.24533>

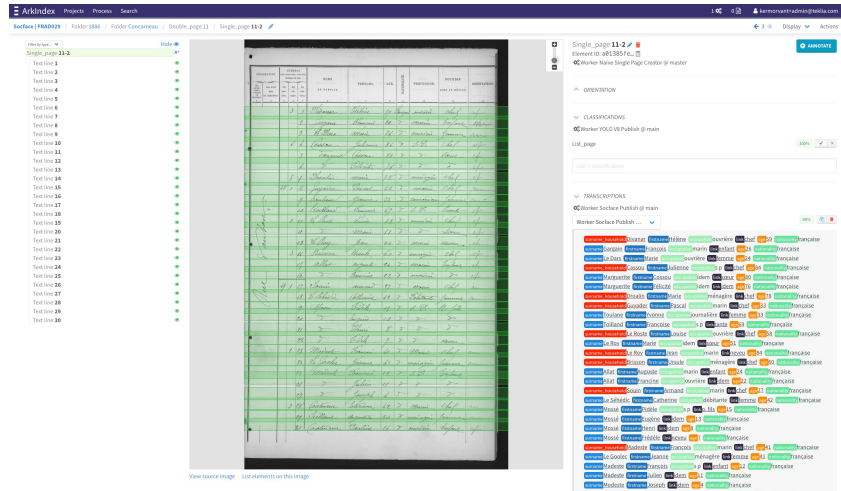
# Digitized humanities

*« développement d'outils informatiques pour numériser, stocker , traiter, collecter, connecter, organiser, diffuser, explorer et visualiser des corpus de textes, images, etc. »*

- *Ces outils sont communs à de nombreuses problématiques de recherche*
- *Ces outils sont soumis aux mêmes contraintes que les autres logiciels pour assurer leur pérennité (bonnes pratiques, maintenance, interopérabilité, documentation, open-source, financement, communauté, etc.)*

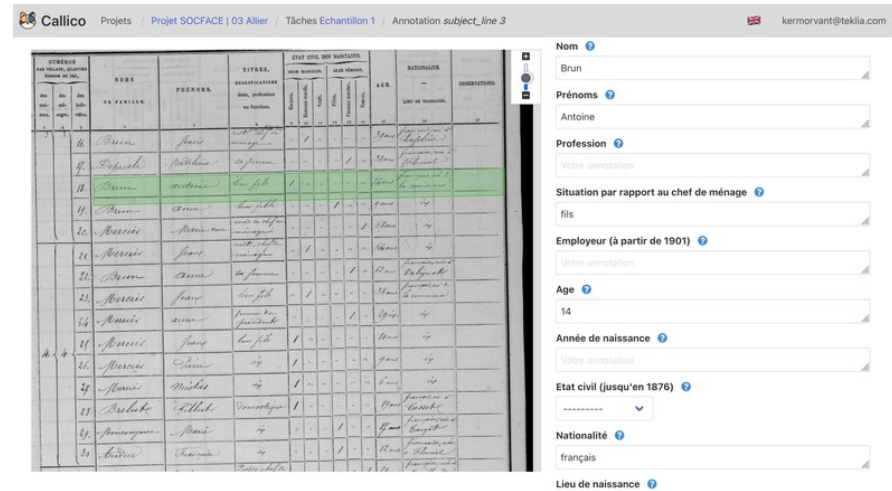
➤ *Création de la plateforme open-source Arkindex*

# Arkindex - Callico



Plateforme de traitement de documents numérisés

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



Application d'annotation et de validation de documents

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



Open-source : GNU AGPLv3  
Qualité de code : tests unitaires, CI/CD, revues  
Facilité de déploiement : Docker





CICR

# CICR : Listes de prisonniers français

Archives du Comité International de la Croix Rouge

36 M de noms dans les listes de prisonniers de la 2<sup>nd</sup> Guerre Mondiale

700 000 pages, principalement manuscrites



Stalay n°/a Liste n° 288 Seite 3

Staat- oder gebürtig- keit	Nr. der Ur- kennungs- marke	N a m e	Born- name	Geburts- tag	Geburts- ort	War- name bei Mutter	Familien- name bei Mutter	Name und Wohnort der zu benachteiligten Person	Dienstgrad	Truppenteil (in Nr. ufm.)	Matrikel- Nr.	Ort und Tag der Befangenahme	Verwundungen, Verletzungen, Tod (Beerdigungsort)	Bemerkungen (z. B. Angaben von anderen Quellen bei Unklarheit oder Unstimmigkeit)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frank- reich	Stalay n°/a 10260	Picardet	Joseph	20. 5. 1910	Flers- Lille	Desire	Neplon	M <sup>me</sup> Picardet Joseph Rue de Babelome Flers- Lille Nord	Ober Geführer	4 Motor. Dragoner 10 Escad.	2945	Flécles 29. 5. 40	6198	gestorben 8. 6. 1940 man des Fiskus
"	10261	Hubert	Roger	26. 8. 1907	Dijon	Emile	Theremin	M <sup>me</sup> Hubert Emile chez Bossinier au 5 <sup>e</sup> Sierrain Saint- Martin	S. 2 Kl. Pionnier	58. 101	1583	Bellefontaine Dijon 20. 11. 40	"	"
"	10263	Verinot	Robert	4. 7. 1913	Relichien	Jules	Bossard	M <sup>me</sup> Jules Verinot chez Bossard Quersing / Deule Nord	S. 2 Kl. Dragoner	4. Motor. 10 Escad.	5688	Castres Lille 29. 5. 40	"	"
"	10265	Batalin	Alexandre	23. 6. 1904	Petro- zovk Russic	Alexandre	Chernysch	M <sup>me</sup> Batalin 33 Rue Henri- Martin Vannes Seine	Offizier	54. 502	4238	Amras Seine 20. 5. 40	"	"
"	10264	Lemckayer	Francois	5. 7. 1917	Pierre	Niquel	M <sup>me</sup> Lemckayer Pierre Tremblay Mont- Caba du Nord	Geführer	131. R. stabil.	1866	1866	Vernouille 21. 5. 40	"	"
"	10265	Steur	Georges	4. 8. 1894	Palumet	Victor	Leccmic	M <sup>me</sup> Steur Gaston Rue du Pneu Bony Nord	S. 2 Kl.	19. R. de Train 2 Kl.	2576	Amiens Cambrai 14. 5. 40	"	"
"	10266	Bertremieux	Roger	26. 5. 1922	Avion	Louis	Derise	M <sup>me</sup> Bertremieux Denise Rue de Coite Fosseur Gonelle	S. 2 Kl.	151. R. 10 Kl.	211	C. Lamy Avion 25. 5. 40	"	"
"	10267	Tramaille	Raymond	20. 4. 1915	Sturima la Riviere	Clairide	Philippe	M <sup>me</sup> C. Tramaille Jeanne la Riviere Saine et Loire	S. 2 Kl.	22. 10 Kl.	848	Montreuil Macen 22. 5. 40	"	"

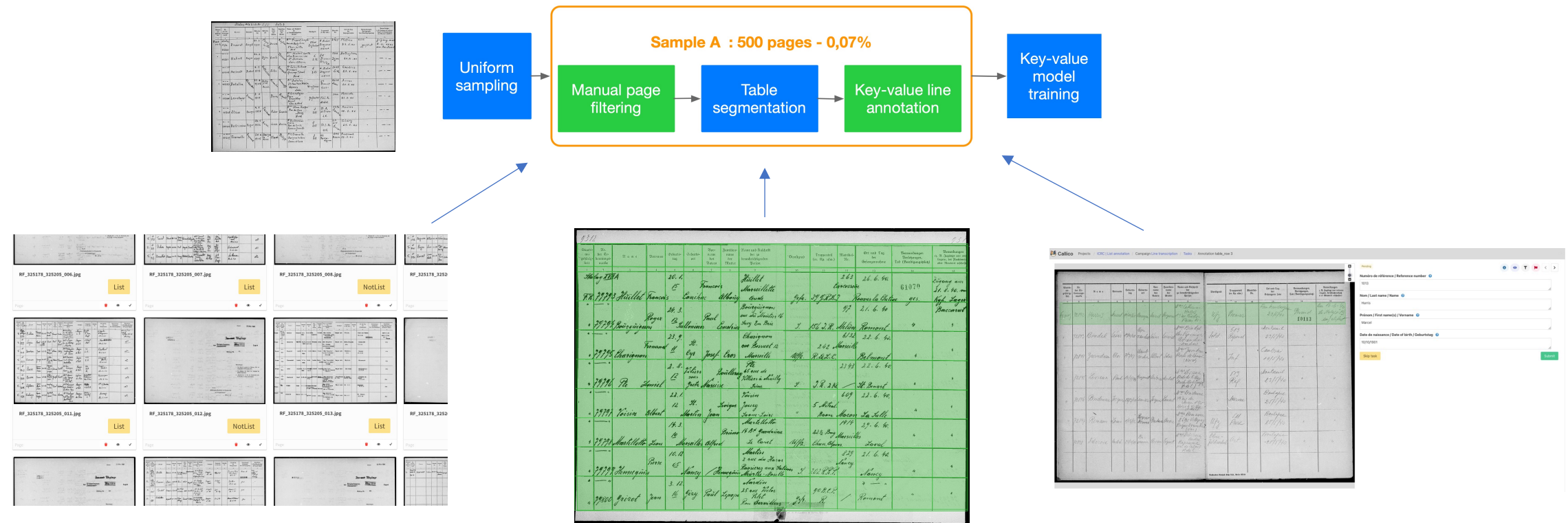




CICR

# CICR : Déroulement du projet

## Phase 1 : initialisation



Annotation des classes en batch dans Arkindex

Détection automatique des lignes de tableau dans Arkindex

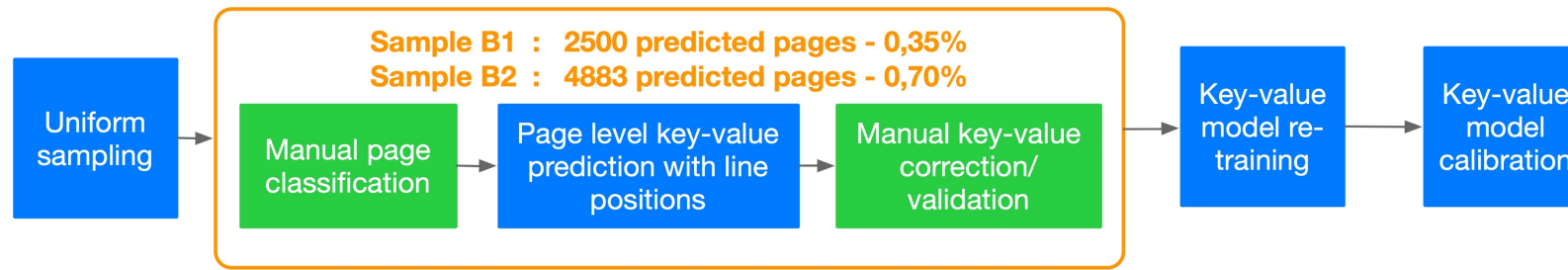
Transcription clé-valeur dans Callico



CICR

# CICR : Déroulement du projet

## Phase 2 : itérations



Prédiction clé-valeur + position dans Arkindex

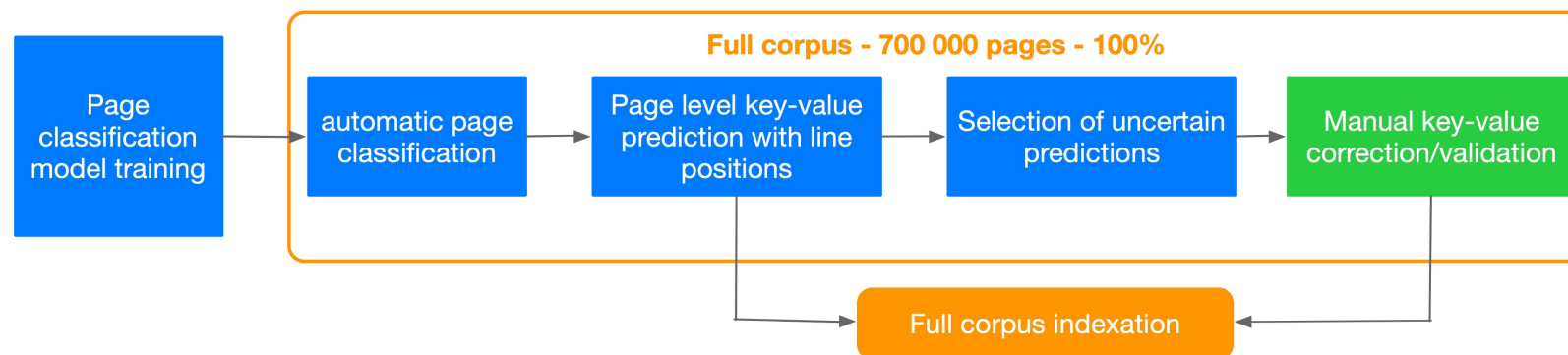
Correction clé-valeur dans Callico



CICR

# CICR : Déroulement du projet

## Phase 3 : Production

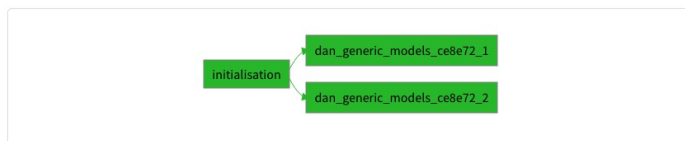


### Process status

Process ID: 54ab6083...

Name: CICR | 'Corpus' - DAN | Project: ICRC | Full dataset | Mode: Workers | Farm: Production | Status: Completed | ACTIONS

initialisation [Completed]  
dan\_generic\_models\_... [Completed]  
dan\_generic\_models\_... [Completed]



### Workers activity

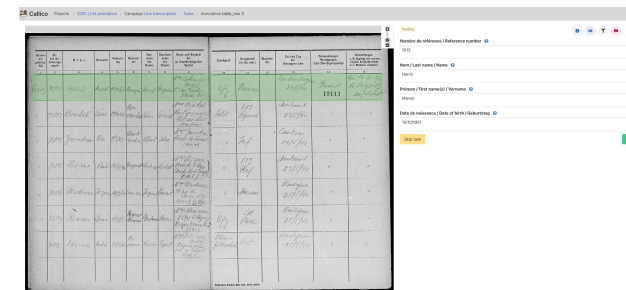
SELECT ALL FAILED ELEMENTS

State	Count	Percentage
processed	513866	100.00%
started	4	0.00%
total	513870	—

Element processing time	
Minimum	1.second 918 milliseconds
Maximum	3.minutes 8.seconds 956 milliseconds
Average	9.seconds 368 milliseconds
Median	8.seconds 928 milliseconds
Estimated time	17.seconds 856 milliseconds

Traitement distribué dans Arkindex  
Classification  
+  
Extraction Clé-Valeur



Validation/correction des incertains dans Calico





CICR

# CICR : Métriques

Performances des modèles :

- Classification (Yolo V8) : P=100%, R=100%
- Extraction Clé-Valeur (DAN) :

Itération	Taille Entrainement	Taille Test	CER %
Initialisation	400	50	8.85
Itération 1	2185	267	3.83
Itération 2	3912	267	3.06

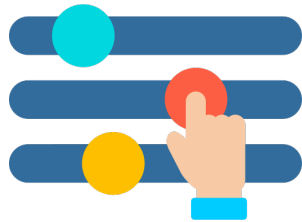
Temps d'annotation

- Initialisation : 70 heures pour 500 pages
- Itérations 1 et 2 : 140 heures pour 4883 pages

Temps d'ingénieur : 200 heures

# Arkindex : principes

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

# Stockage et gestion des documents

- Import web (petit corpus), S3 (gros corpus) ou par manifest IIIF
- Support des formats images et PDF
- Structuration hiérarchique des corpus complètement adaptable
- Gestion des méta-données à tous les niveaux

The screenshot shows the ArkIndex web interface. At the top, there are navigation tabs for 'Projects', 'Process', and 'Search'. The main content area displays a grid of 12 thumbnail images of manuscript pages, each with a label below it: 'plat sup', 'contreplat sup', 'page de garde rdcto', 'page de garde verso', '1r', '1v', '2r', '2v', and four more pages. To the right of the grid is a detailed metadata panel for the volume 'Manuscripts, nouv. acq. lat. 3260'. The panel includes sections for 'CLASSIFICATIONS', 'METADATA', and 'IIIF ID'. The 'METADATA' section lists 'Date: 1480-1490', 'Digitised by: Bibliothèque nationale de France', 'Digitization Type: single page', and 'Format: Bourges. - ou - Bourbonnais. - (?) - Ecriture bâtarde. - Enluminé par un ou deux artistes anonymes. 18 peintures, 23 miniatures, lettres ornées, bouts de lignes. - Couture trop serrée pour établir un relevé codicologique. Ni réclames, ni signatures. Foliotation au crayon à papier, XXe siècle. - Parchemin. - 130 ff., précédés et suivis d'un f. de garde de parchemin. - 215 x 150 mm (just. 130 x 80 mm). - Reliure à l'encre rouge. - Reliure de velours vert (devenu bleuâtre) sur ais de bois, fermoirs de cuivre (il n'en reste que les deux contre-agrafes au plat supérieur), tranches dorées, début du XXe siècle. Volume consolidé en 2021 : remise en place du f. 46 ; restauration de l'accroc dans le tissu du plat inférieur ; consolidation des coins et des charnières (dossier BnF-ADM-2020-079574-01). - Le volume portait encore en 1909 une reliure de maroquin rouge du XVII. - e. - siècle, dans le style Le Gascon. Les plats étaient frappés des initiales redoublées LL et FF entourées de feuillages.

IIIF ID: <https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json>

Language: latin

Metadata Source: [http://oai.bnf.fr/oai2/OAIHandler?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai:bnf:gallica/ark:/12148/btv1b55013837f](http://oai.bnf.fr/oai2/OAIHandler?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:bnf:gallica/ark:/12148/btv1b55013837f)

Relation: Notice du catalogue : <http://archivesetmanuscrits.bnf.fr/ark:/12148/cc1248944>

Shelfmark: Bibliothèque nationale de France. Département des Manuscrits. NAL 3260

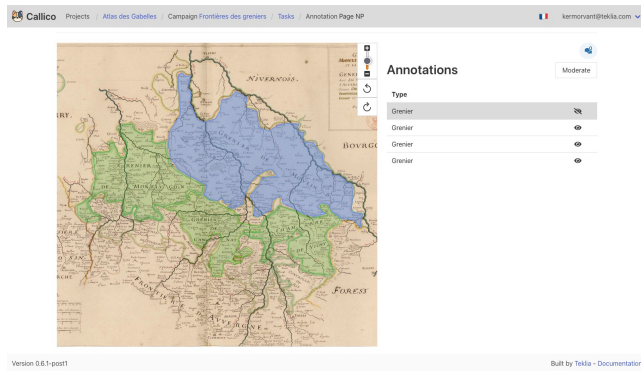
Source: <https://gallica.bnf.fr/ark:/12148/btv1b55013837f>

Images: <https://gallica.bnf.fr/ark:/12148/btv1b55013837f>


<https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json>

# Spécification par les annotations : Callico

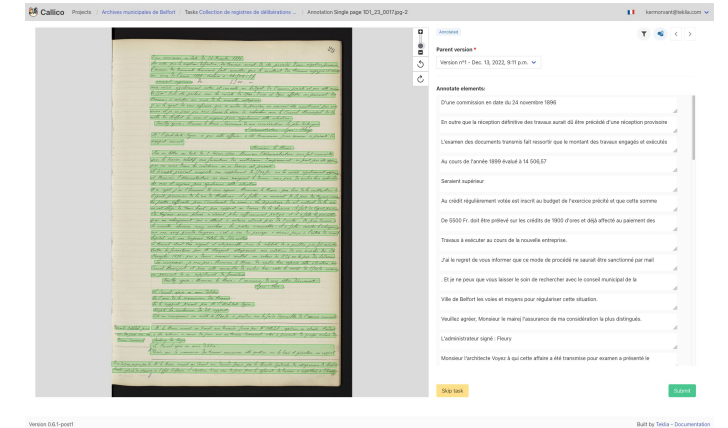
## Zonage



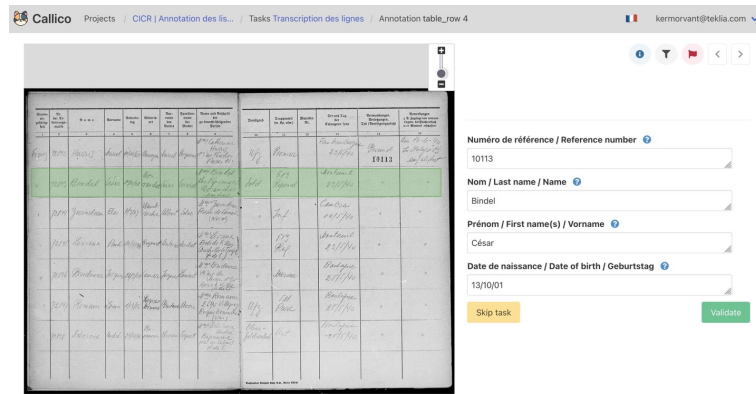
## Classification



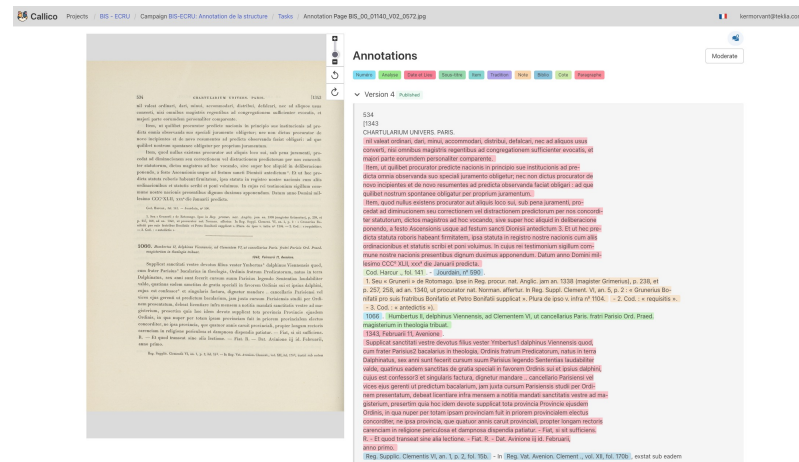
## Transcription



## Clé-valeur



## Entités nommées

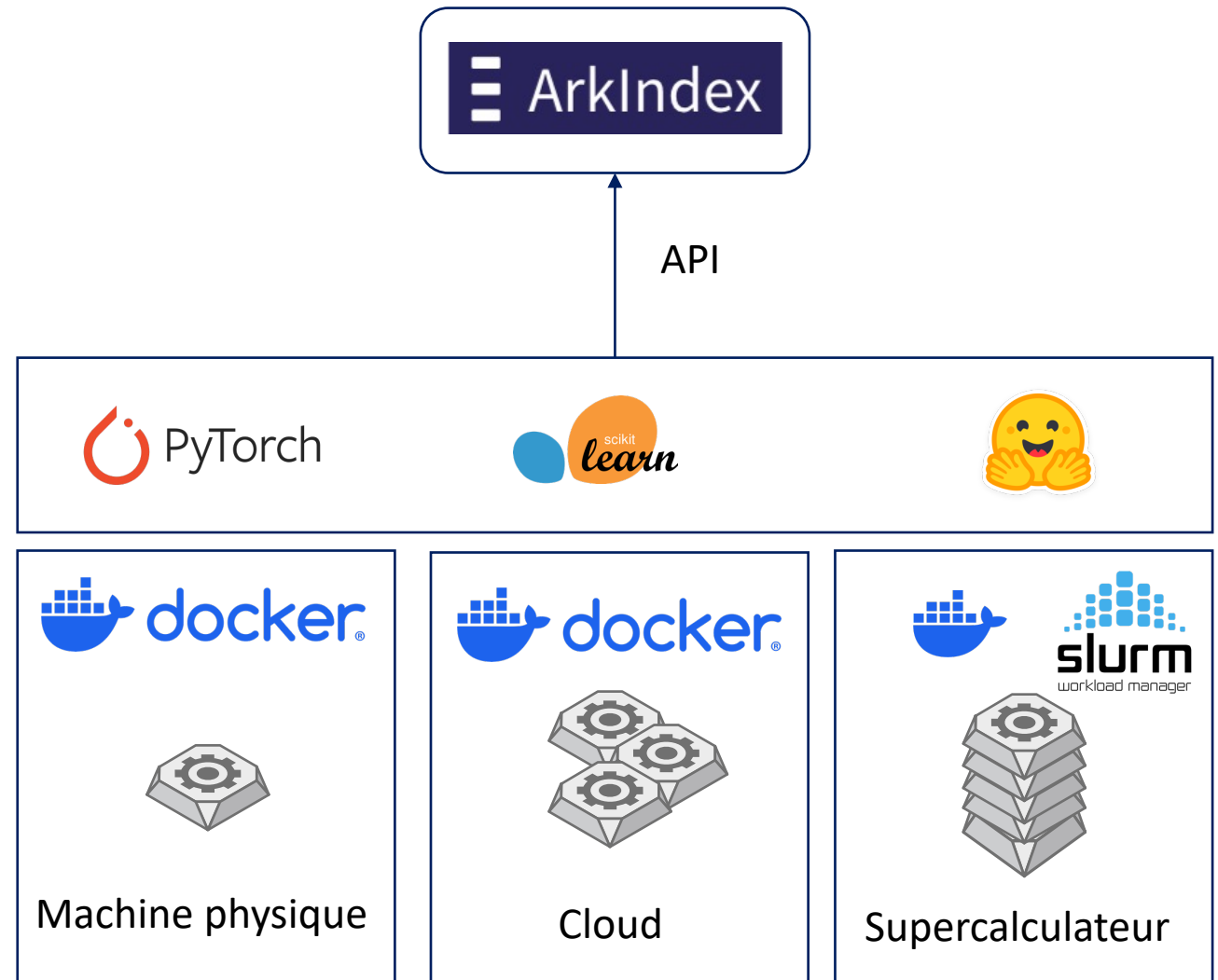


## Groupement



# Intégration de modèles/algorithmes

- Intégration de n'importe quel langage/code/modèle
- Code de base python fourni
- Intégration par API
- Déploiement par Docker
- Entraînement et inférence





# Projets open-source avec Arkindex-Callico

- Comité International de la Croix Rouge (Genève) :
  - Utilisation de Callico pour valider 4, 201,357 prisonniers
- Projet ANR CollabScore - CEDRIC-CNAM (Paris)
  - Développement d'un mode d'annotation de partitions musicales pour Callico
- Projet ANR TypoReF – LIFAT & CESR (Tours)
  - Intégration d'algorithmes d'analyse des matériels typographiques

utelle loie. O uen paradis  
suis qui te laissas estendre  
en la croiz pendre. Longis  
com tu lui feis **merci**  
ta merite. Que me gardes

, dist Pymandre,  
une **question**, &  
loit le plus estime  
e du Peintre

**TEKLIA**

[kermorvant@tekliia.com](mailto:kermorvant@tekliia.com)