



# 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>

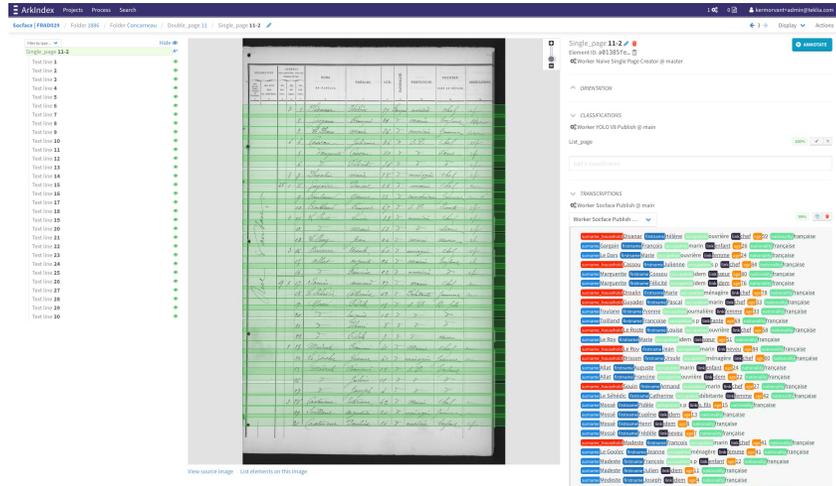
# Digitized humanities

*« 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.)*

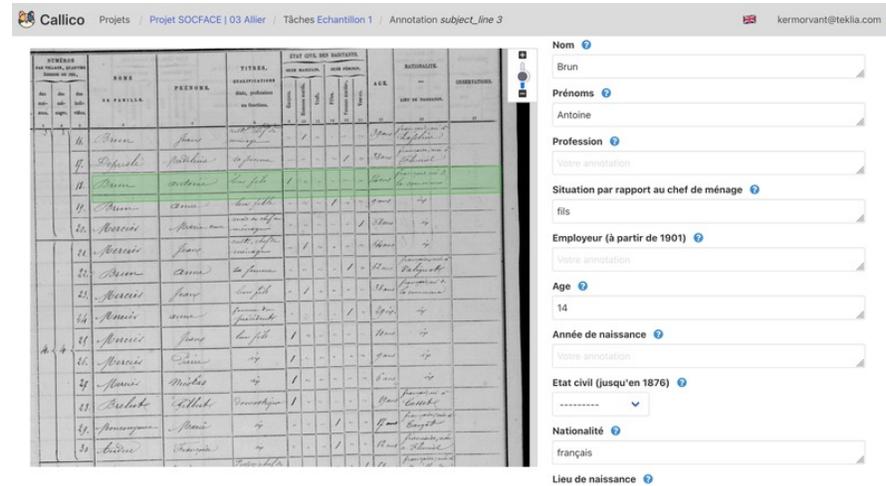
➤ *Création de la plateforme open-source Arkindex*

# 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- angehörig- keit	Nr. der Ur- teilsmarke	Name	Vorname	Geburts- tag	Geburts- ort	Vor- name des Vaters	Familien- name des Vaters	Name und Wohnort der zu benachrichtigenden Person	Dienstgrad	Truppenteil (in Nr. ufm.)	Matrikel- Nr.	Ort und Tag der Gefangennahme	Übergangenen, Todes (Beerdigungsort)	Bemerkungen (z. B. Angaben von anderen Quellen, bei Unklarheit oder Unvollständigkeit)
Frank- reich	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 gefangen	gefangen 8. 6. 1940 man bei Flécles
"	10261	Hubert	Roger	26. 8. 1907	Dijon	Emile	Theremin	M <sup>me</sup> Hubert Camille chez Bossinier St Simeon Saint-Martin	S. 2 Kl.	58. Pionnier Ecole 1911	1583	Bellenfleur 20. 11. 40	"	"
"	10263	Verinot	Robert	4. 7. 1913	Relichien	Jules	Bossard	M <sup>me</sup> Jules Verinot Bossard Quersweg 7/Deule Nord	S. 2 Kl.	4. Motor- Dragoner 10 Escad.	5688	Castres Lille 29. 5. 40	"	"
"	10265	Batalin	Alexandre	23. 6. 1904	Petro- grad Russie	Alexandre	Chernyshev	M <sup>me</sup> Batalin 33 Rue Henri-Martin Vannes Seine	Offizier	54 Pionnier 502-	4238	Arras Seine 20. 5. 40	"	"
"	10264	Lemckayer	Francois	5. 7. 1917	Pierre	Pierre	Niquel	M <sup>me</sup> Lemckayer Pierre Thembloy Mont Coba du Nord	Geführer	131. R. stabil.	1866	Versailles 21. 5. 40	"	"
"	10265	Steur	Georges	4. 8. 1894	Palunet	Victor	Leccmie	M <sup>me</sup> Steur Gaston Rue du Puits Boulay 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 Derise Rue de Conite Hoozen Nouvelle	S. 2 Kl.	151. R. 10 Kl.	1211 11122	C. Lestry 25. 5. 40	"	"
"	10267	Tramaille	Raymond	20. 4. 1915	Sturima la Nieme	Clasidie	Philippe	M <sup>me</sup> C. Tramaille Jardin la Nieme Saine et Loire	S. 2 Kl.	22 Pariz- 10 Escad.	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

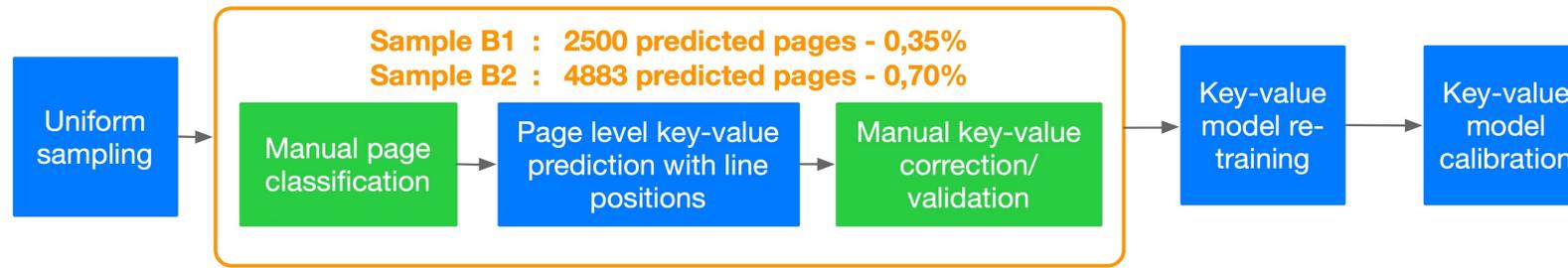
Transcription clé-valeur dans Callico



CICR

# CICR : Déroulement du projet

## Phase 2 : itérations



Page [RF\\_325288\\_325314\\_015.jpg](#) ANNOTATE

Element ID: a68ae8e2\_1  
Worker File import @ 68ed53

ORIENTATION

CLASSIFICATIONS

TRANSCRIPTIONS

Worker DAN Generic Arkindex models @ bump-submodule, main with model DAN CICR (HTR + N...)

Worker DAN Generic Ark... 80%

Barbedette	Robert	Date de naissance	19.7.07	Numero RI	325205
Francfort	Jean	Date de naissance	19.12.11	Numero RI	325205
Mainot	Alfred	Date de naissance	10.10.14	Numero RI	325205
Grenet	François	Date de naissance	23.4.02	Numero RI	325205
Diebold	Henri	Date de naissance	10.2.05	Numero RI	325205
Dénézergues	Jean	Date de naissance	22.14	Numero RI	325205
Mondon Marin	Joseph	Date de naissance	19.1.06	Numero RI	325205
Human	Alexis	Date de naissance	13.2.97	Numero RI	325205

MANAGE

Callico Projects CICR (List Annotation Campaign Line transcription Tools Annotation table\_new 3)

Numero de référence / Reference number: 10113

Nom / Last name / Name: Marcel

Prénom / First name(s) / Vorname: Marcel

Date de naissance / Date of birth / Geburtstag: 19/07/1907

SEP lock SUBMIT

Prédiction clé-valeur + position dans Arkindex

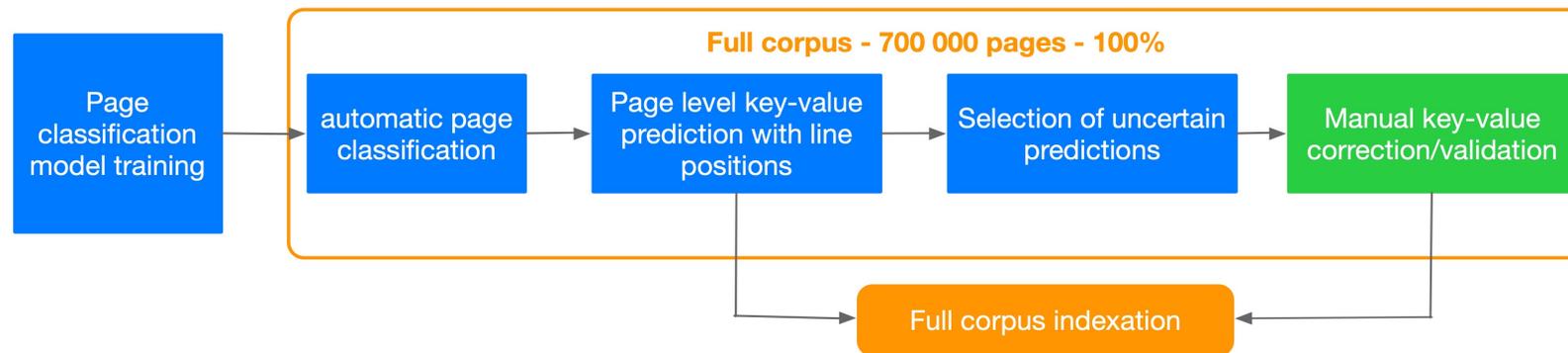
Correction clé-valeur dans Callico



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



### Workers activity

SELECT ALL FAILED ELEMENTS

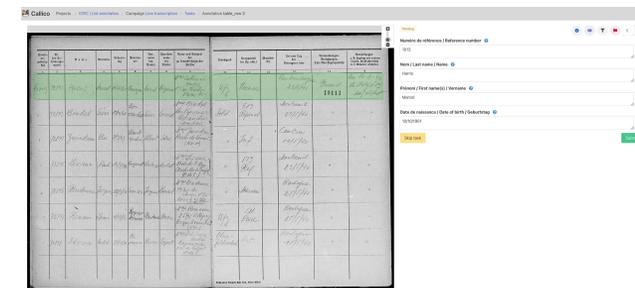
▼ DAN Generic Arkindex models with model DAN CICR (HTR + NER)

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

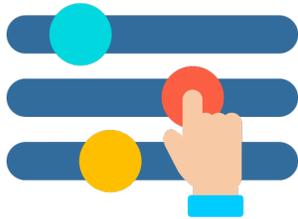
Temps d'annotation

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

Temps d'ingénieur : 200 heures

# Arindex : 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

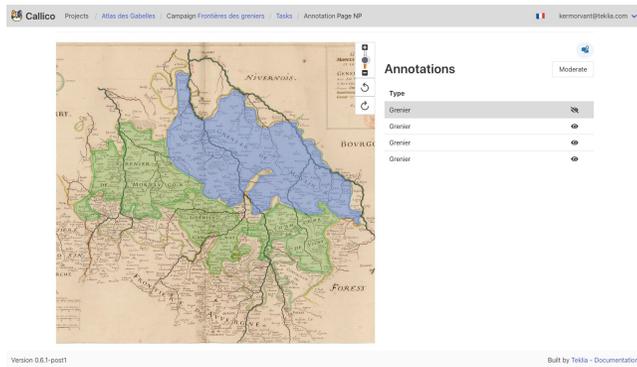
The screenshot shows the ArkIndex web interface. At the top, there's a navigation bar with 'ArkIndex', 'Projects', 'Process', and 'Search'. Below that, the breadcrumb path is 'HORAE | Complete / Volume France, Paris, Bibliothèque nationale de France, Manuscrits, nouv. acq. lat. 326...'. The main content area shows a grid of 12 thumbnail images of manuscript pages, labeled '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 'Volume France, Paris, Bibliothèque nationale de France, Manuscrits, nouv. acq. lat. 3260'. The metadata includes:

- CLASSIFICATIONS:** Add a classification
- METADATA:**
  - Date: 1480-1490
  - Digitised by: Bibliothèque nationale de France
  - Digitization Type: single page
  - 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.fr/gallica/ark:/12148/btv1b55013837f](http://oai.bnf.fr/oai2/OAIHandler?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:bnf.fr/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 Images:** <https://gallica.bnf.fr/ark:/12148/btv1b55013837f>

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

# Spécification par les annotations : Callico

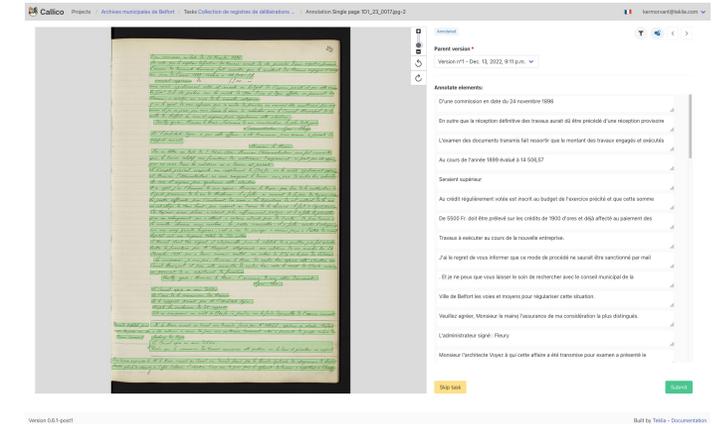
## Zonage



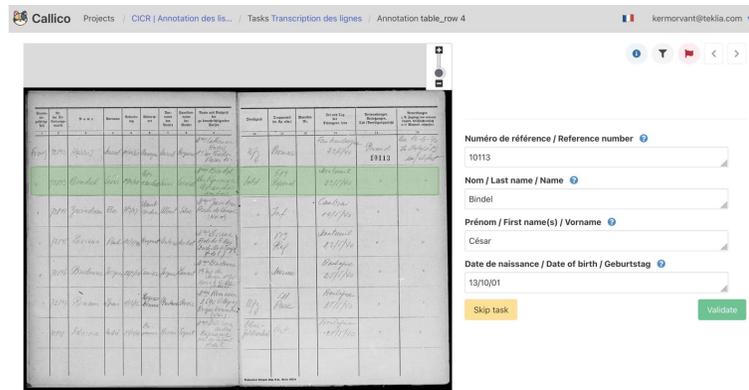
## Classification



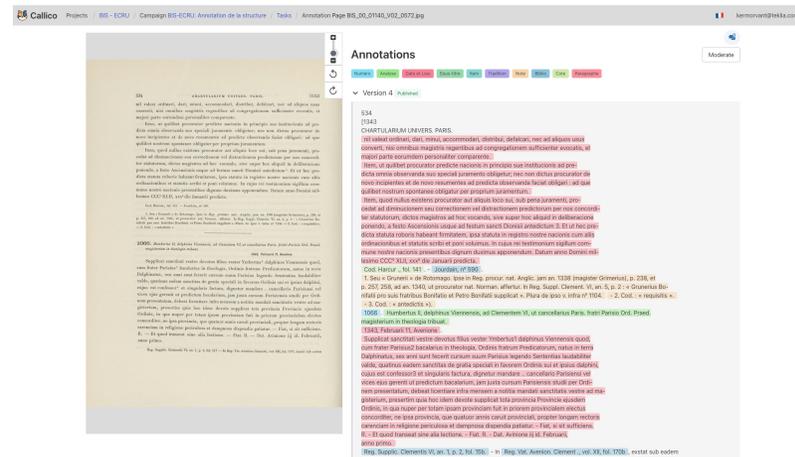
## Transcription



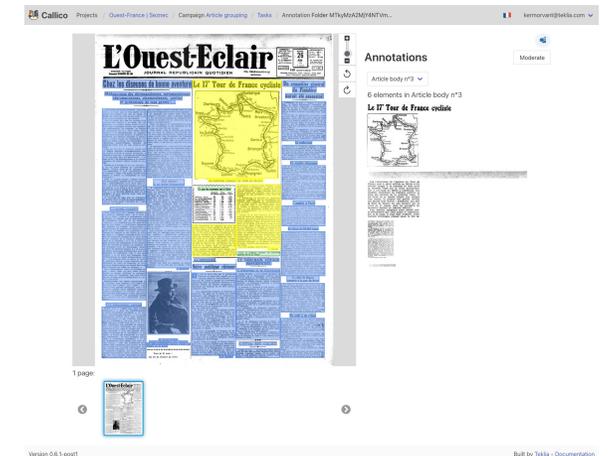
## Clé-valeur



## Entités nommées

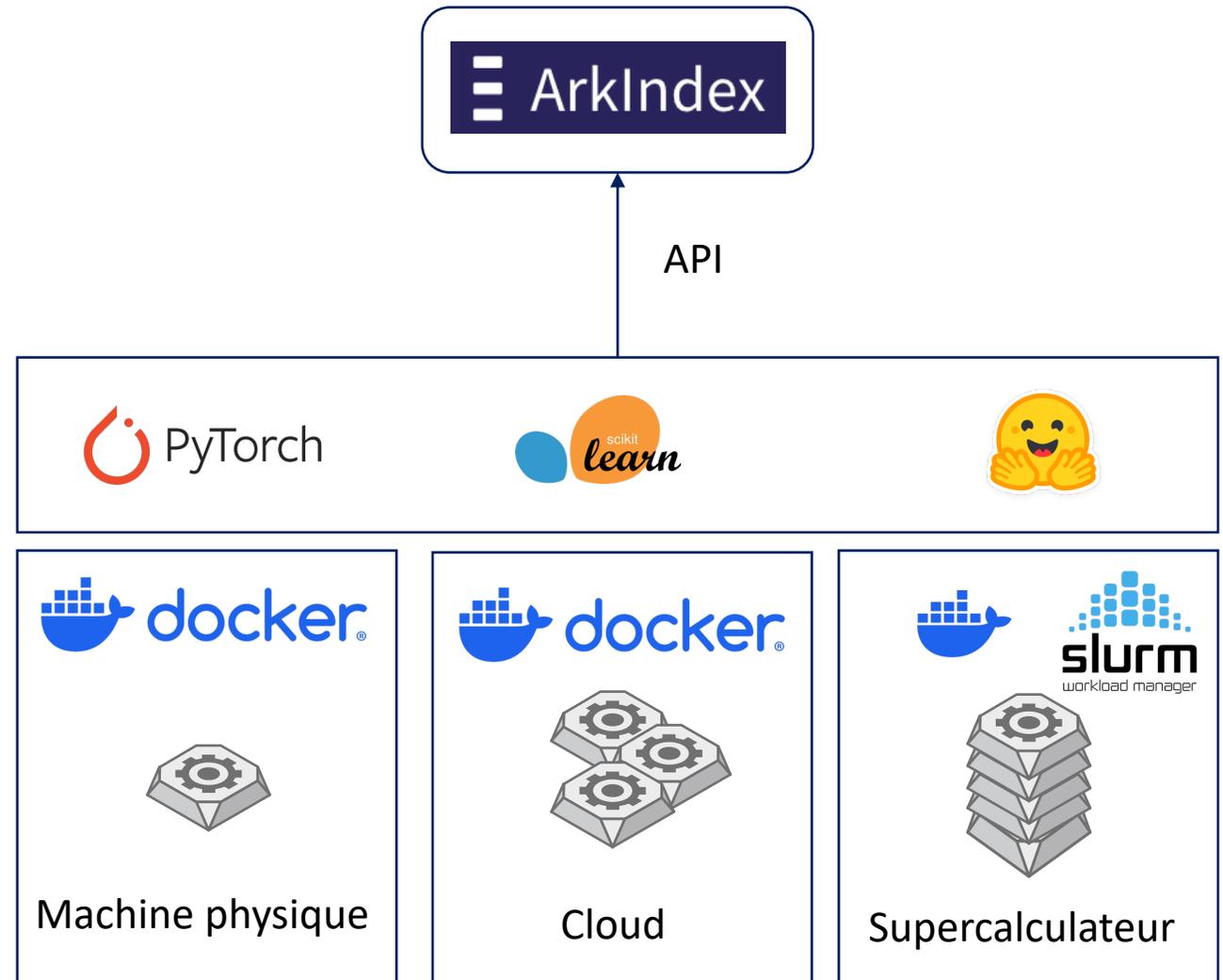


## Groupement



# 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)