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

Journée IA et Humanités Numériques - 2024

Christopher Kermorvant

TEKLIA, Paris, France

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

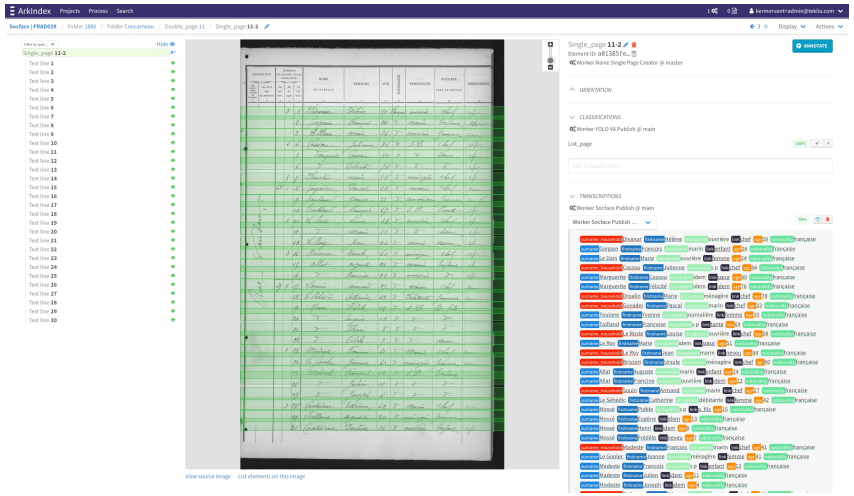
« développement d'outils informatiques pour numériser, stocker , traiter, collecter, connecter, organiser, diffuser, explorer et visualiser des corpus de textes, images, etc. »

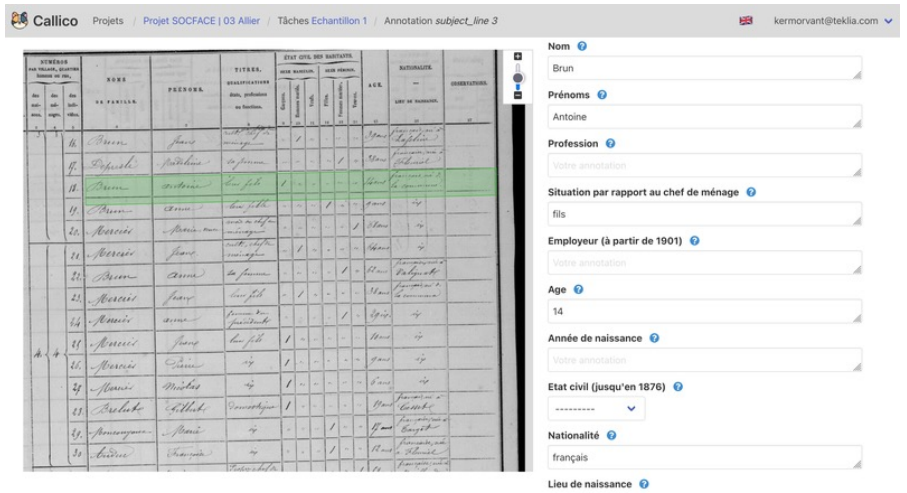
- *Ces outils sont communs à de nombreuses problématiques de recherche*
- *Ces outils sont soumis aux mêmes contraintes que les autres logiciels pour assurer leur pérennité (bonnes pratiques, maintenance, interopérabilité, documentation, open-source, financement, communauté, etc.)*

➢ *Création de la plateforme open-source Arkindex*

Roth, C. (2019). Digital, digitized, and numerical humanities. Digital Scholarship in the Humanities, 34(3), 616–632. https://doi.org/10.1093/llc/fqy057

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/

AGPLv3 Free Software
Free as in Freedom

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

docker python django Vue.js
iiif scikit learn slurm workload manager Amazon S3

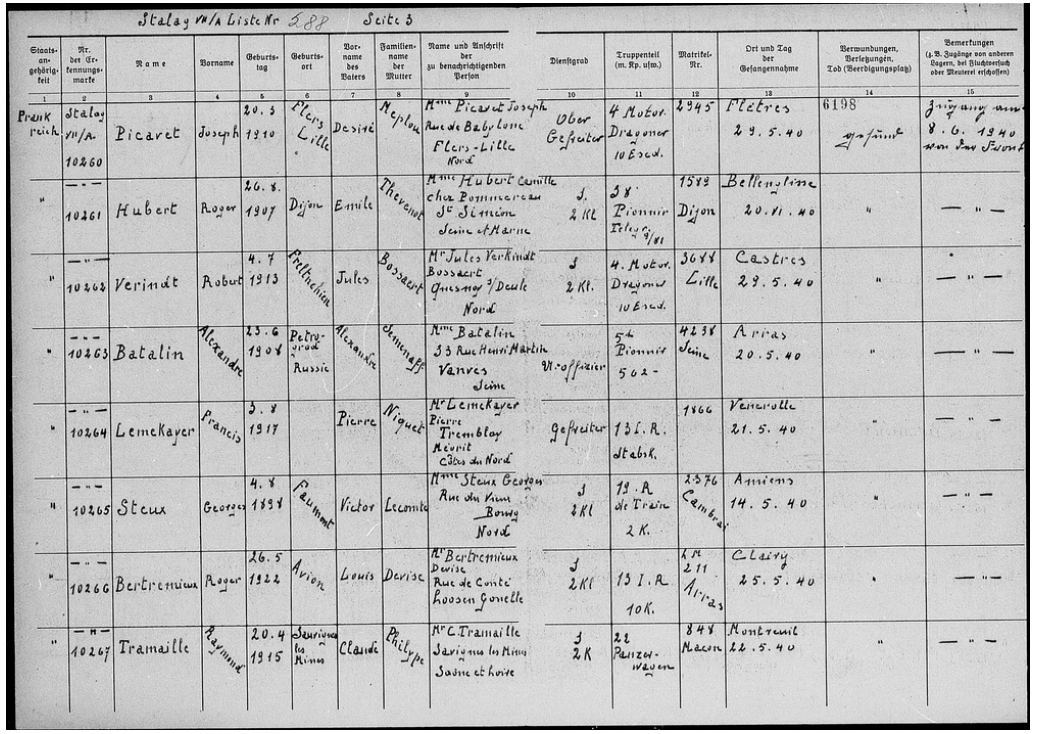
CICR : Listes de prisonniers français

CICR

Archives du Comité International de la Croix Rouge

36 M de noms dans les listes de prisonniers de la 2nd Guerre Mondiale

700 000 pages, principalement manuscrites



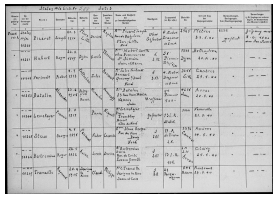
Stalag VII/A Liste Nr 288 Seite 3

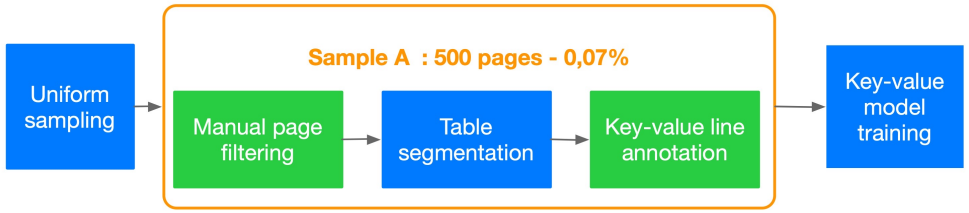
Staatsangehörigkeit	Nr. der Erkennungsmarke	Name	Vorname	Geburtstag	Geburtsort	Vorname des Vaters	Familienname der Mutter	Name und Anschrift der zu benachrichtigenden Person	Dienstgrad	Truppenteil (m. Kp. usw.)	Matrikel-Nr.	Ort und Tag der Gefangennahme	Verwundungen, Verletzungen, Tod (Beerdigungsplatz)	Bemerkungen
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frankreich	Stalag VII/A 10260	Picaret	Joseph	20.3 1910	Flers Lille	Desiré	Neplow	Mme Picaret Joseph Rue de Babylone Flers-Lille Nord	Ober Gefreiter	4 Motor. Dragoner 10 Esch.	2945	Flers 29.5.40	6198	
"	10261	Hubert	Roger	26.8. 1907	Dijon	Emile	Thevenot	Mme Hubert Camille chez Pommereau St Simeon Seine et Marne	J. 2 Kl.	38. Pionnir Telegr. 3/81	1583	Dijon Bellenglise 20.VI.40	"	— " —
"	10263	Verindt	Robert	4.7 1913	Frelinghien	Jules	Bossaert	Mr Jules Verkindt Bossaert Quesnoy s/Deule Nord	J 2 Kl.	4. Motor. Dragoner 10 Esch.	3648 Lille	Castres 29.5.40	"	— " —
"	10263	Batalin	Alexandre	23.6 1904	Petrograd Russie	Alexandre	Semenoff	Mme Batalin 33 Rue Henri Martin Vanves Seine	Unteroffizier	5te Pionnir 502 -	4238 Seine	Arras 20.5.40	"	— " —
"	10264	Lemeskayer	Francis	3.8 1917		Pierre	Niguet	Mr Lemeskayer Pierre Tremblay Mesnil Côtes du Nord	Gefreiter	131. R. Stabsk.	1866	Vermelle 21.5.40	"	— " —
"	10265	Steux	Georges	4.8 1894	Feluy Mont	Victor	Lecomte	Mme Steux Georges Rue du Vieux Bourg Nord	J 2 Kl	19. A. de Train 2 K.	2576 Cambrai	Amiens 14.5.40	"	— " —
"	10266	Bertremieux	Roger	26.5 1922	Avion	Louis	Denise	Mr Bertremieux Denise Rue de Comté Roosen Gonelle	J 2 Kl	13 I. R 10 K.	1re 211 Arras	Clairy 25.5.40	"	— " —
"	10267	Tramaille	Raymond	20.4 1915	Savigny les Mines	Claude	Philippe	Mr C. Tramaille Savigny les Mines Saône et Loire	J 2 K	22 Panzer-Division	848 Mâcon	Montreuil 22.5.40	"	— " —

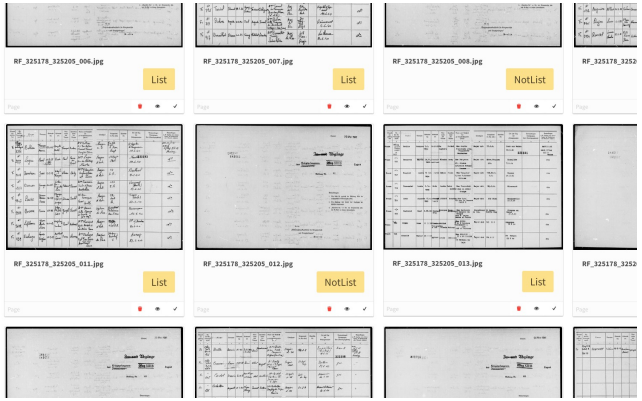
CICR : Déroulement du projet

CICR

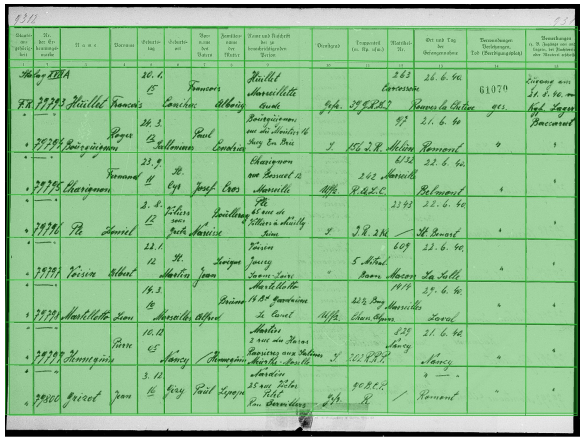
Phase 1 : initialisation



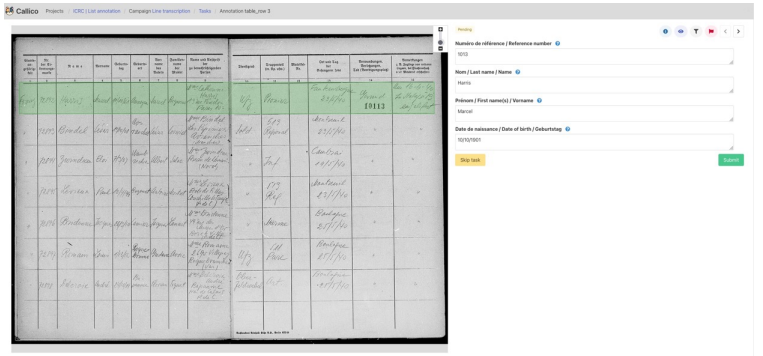




Annotation des classes en batch dans Arkindex



Détection automatique des lignes de tableau dans Arkindex

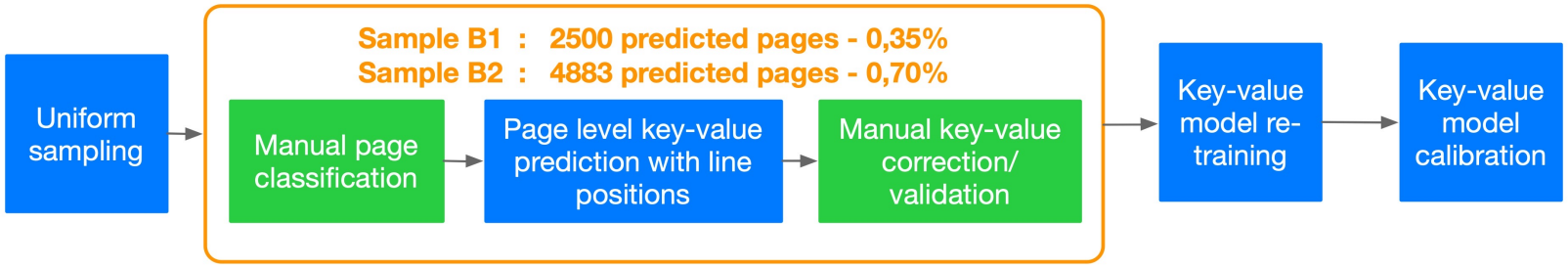


Transcription clé-valeur dans Callico

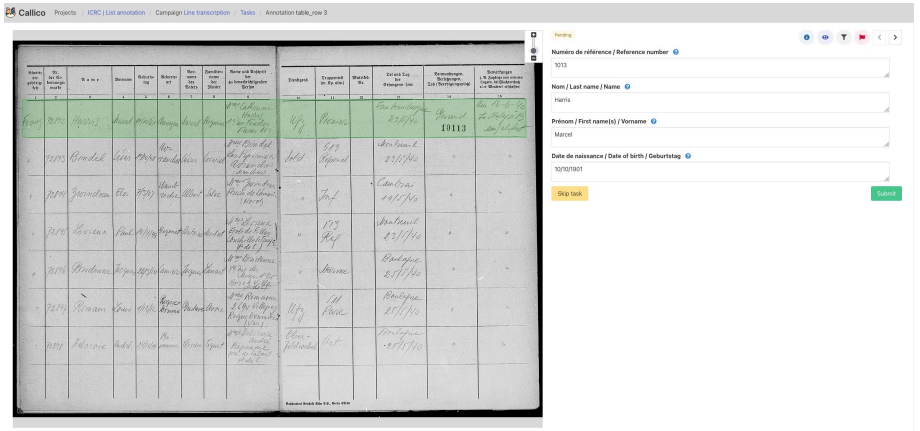
CICR : Déroulement du projet

CICR

Phase 2 : itérations



Prédiction clé-valeur + position dans Arkindex

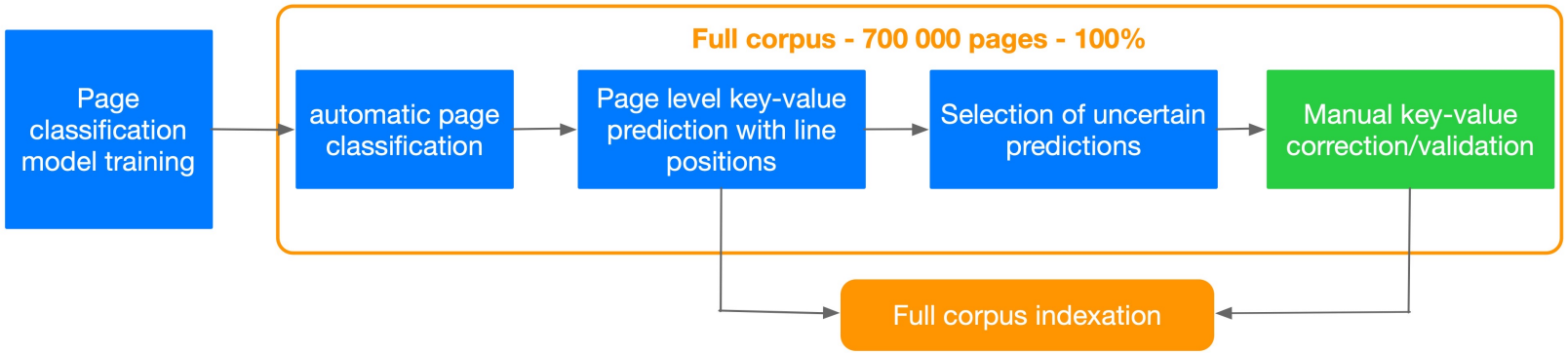


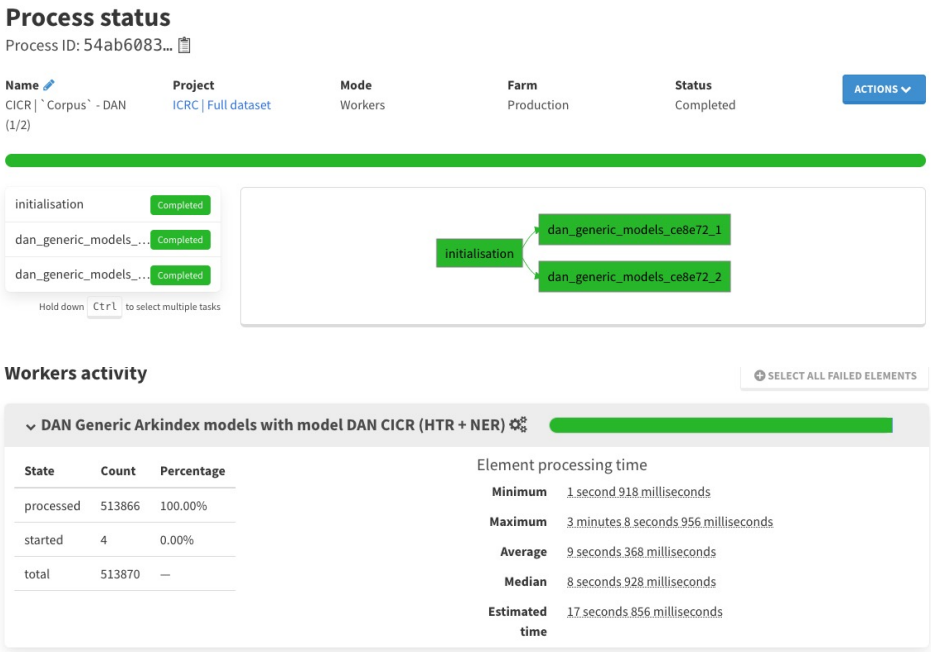
Correction clé-valeur dans Callico

CICR : Déroulement du projet

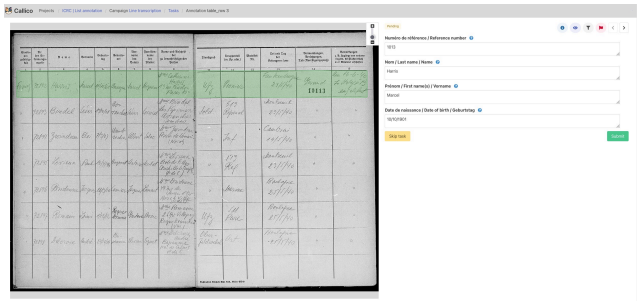
CICR

Phase 3 : Production





Traitement distribué dans Arkindex
Classification
+
Extraction Clé-Valeur



Validation/correction des incertains dans Callico

CICR : Métriques

CICR

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

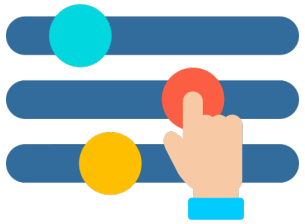
Temps d'annotation

- 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

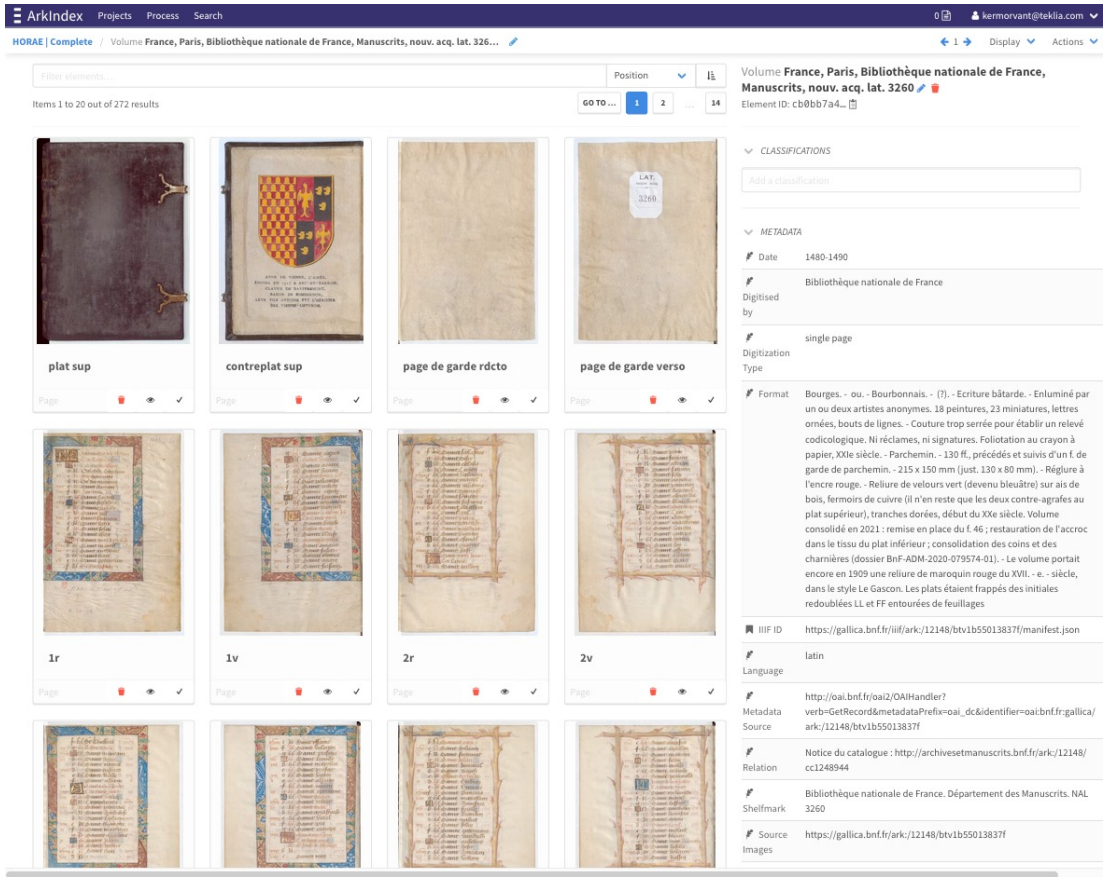
Traiter 1000 ou 10 millions de pages

Open-source

Diffusion et participation de la communauté

Stockage et gestion des documents

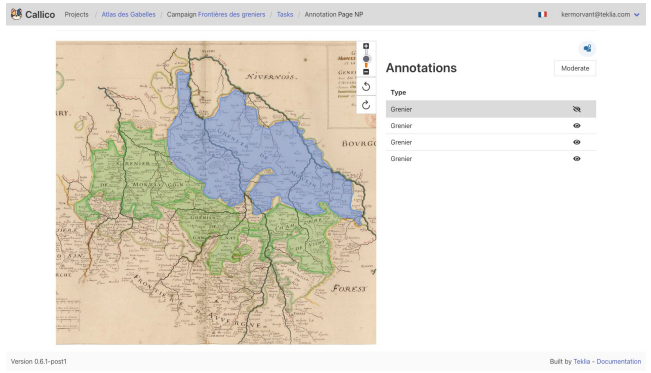
- Import web (petit corpus), S3 (gros corpus) ou par manifest IIIF
- Support des formats images et PDF
- Structuration hiérarchique des corpus complètement adaptable
- Gestion des méta-données à tous les niveaux



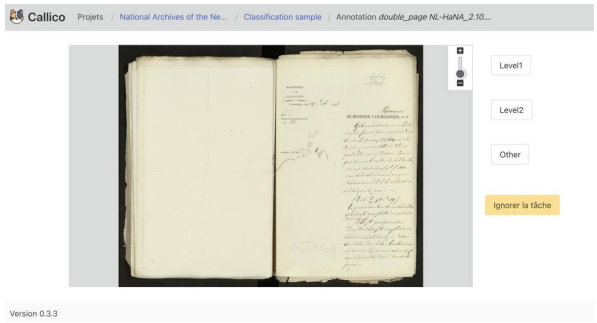
https://gallica.bnf.fr/iiif/ark:/12148/btv1b55013837f/manifest.json

Spécification par les annotations : Callico

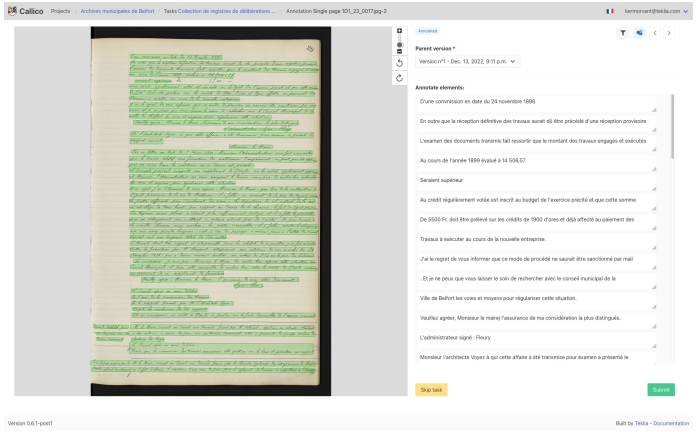
Zonage



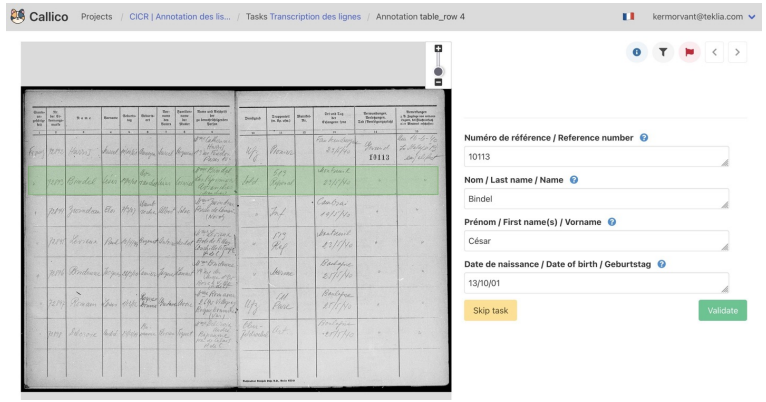
Classification



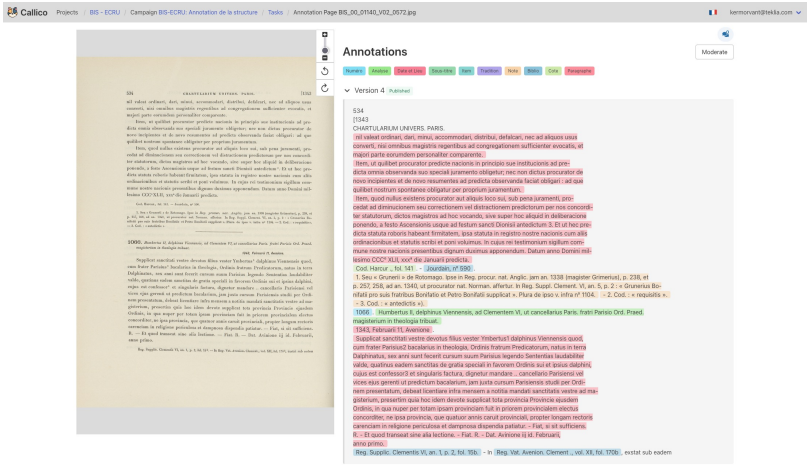
Transcription



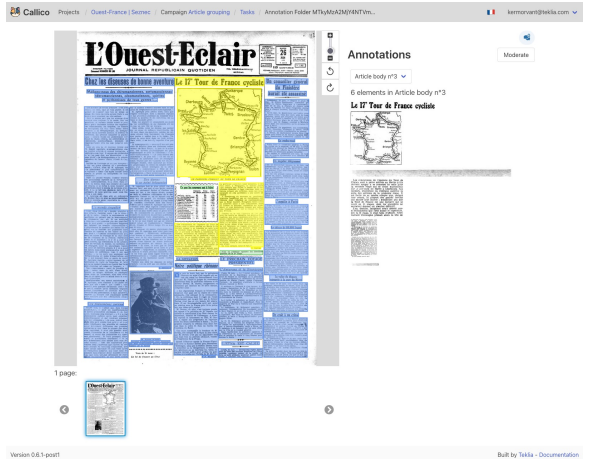
Clé-valeur



Entités nommées

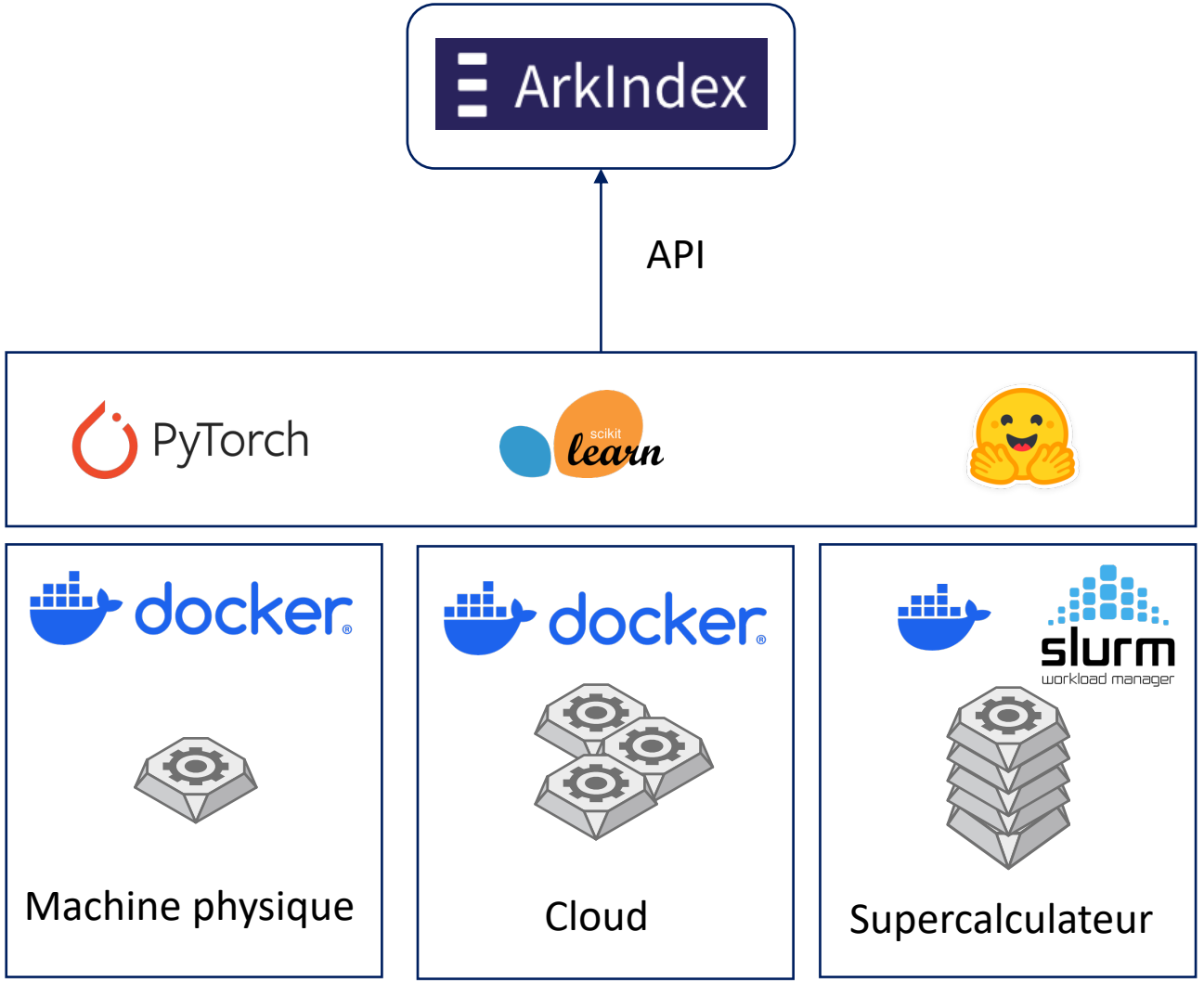


Groupement



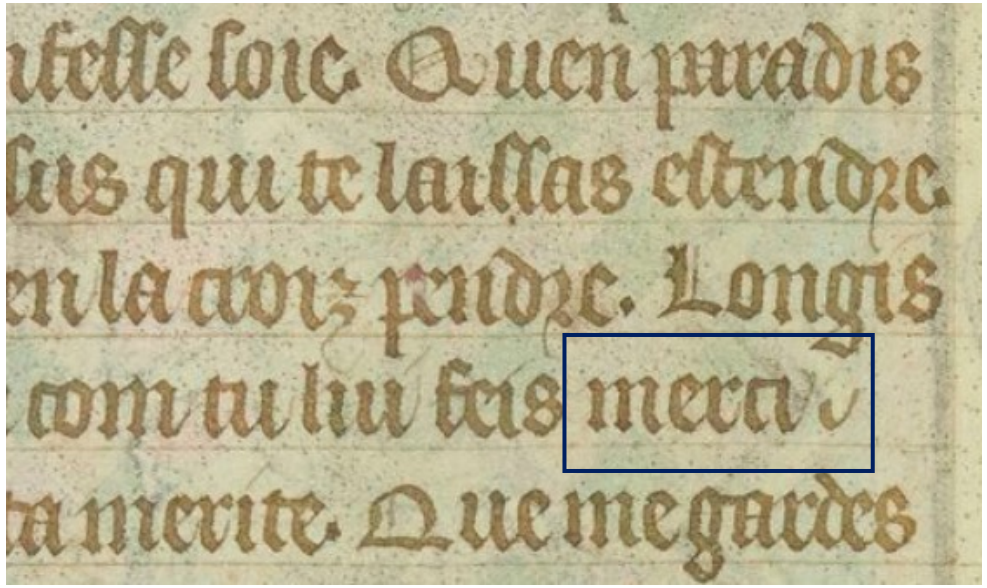
Intégration de modèles/algorithmes

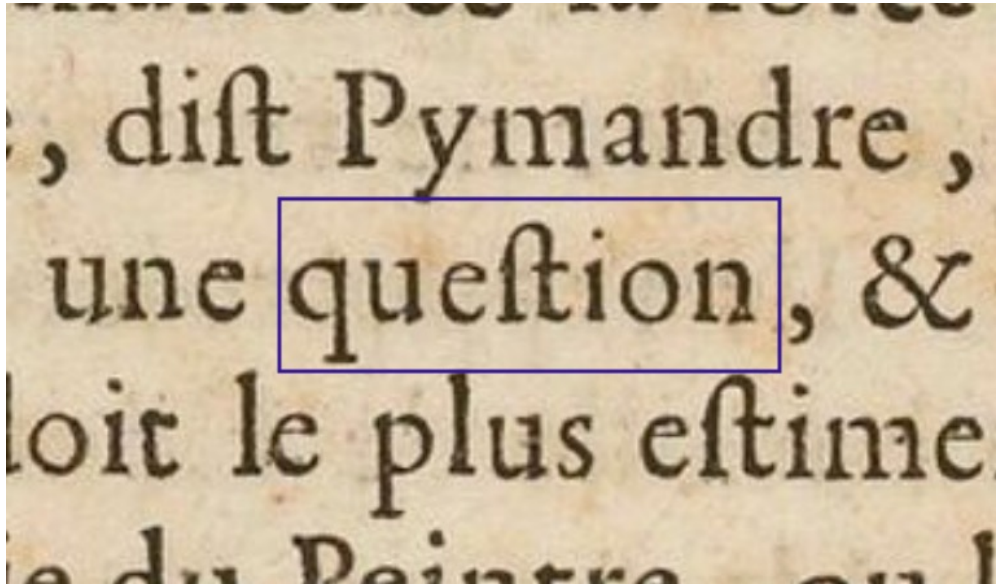
- Intégration de n'importe quel langage/code/modèle
- Code de base python fourni
- Intégration par API
- Déploiement par Docker
- Entrainement et inférence



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





TEKLIA

kermorvant@teklia.com