ArchXAI – AI-enhanced Cross-Border Archives

Using AI to unlock the past

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

Project funding & basic info

- Co-funded by the European Union - Interreg Central Baltic Programme
- PO7, improving public services and solutions for citizens
- Total budget: 2,922,830.80€
- Co-funding rate: 80%
- Timeline: 36-months, from 1 June 2025 to 31 May 2028
- Partners: NAE, NAF, NAL, Xamk

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

Why ArchXAI project?

- **Challenges in accessing archives**
  - Historical archives
    - handwritten, fragile, not digitised → difficult and time-consuming access
- **Pressure on archivists**
  - Rising workloads
  - Balancing between preservation and search tasks → Causes delays and inefficiencies.
- **AI-Driven archival solutions**
  - AI for automated text recognition, indexing, and smart AI assisted toolset to modernize archival services.
- **Benefits of ArchXAI**
  - Faster and more accurate user access.

Interreg Co-funded by the European Union
Central Baltic Programme
ArchXAI

Fundemendals

- **Cross-Border Historical Access**
  - The project digitizes Baltic heritage, enabling cross-border access to shared historical records.
- **Inclusive Research and Learning**
  - empowers historians, students, genealogists, and citizens to explore history with ease.
- **Preserving Cultural Heritage**
  - The project preserves valuable cultural knowledge while bridging past and digital futures responsibly.

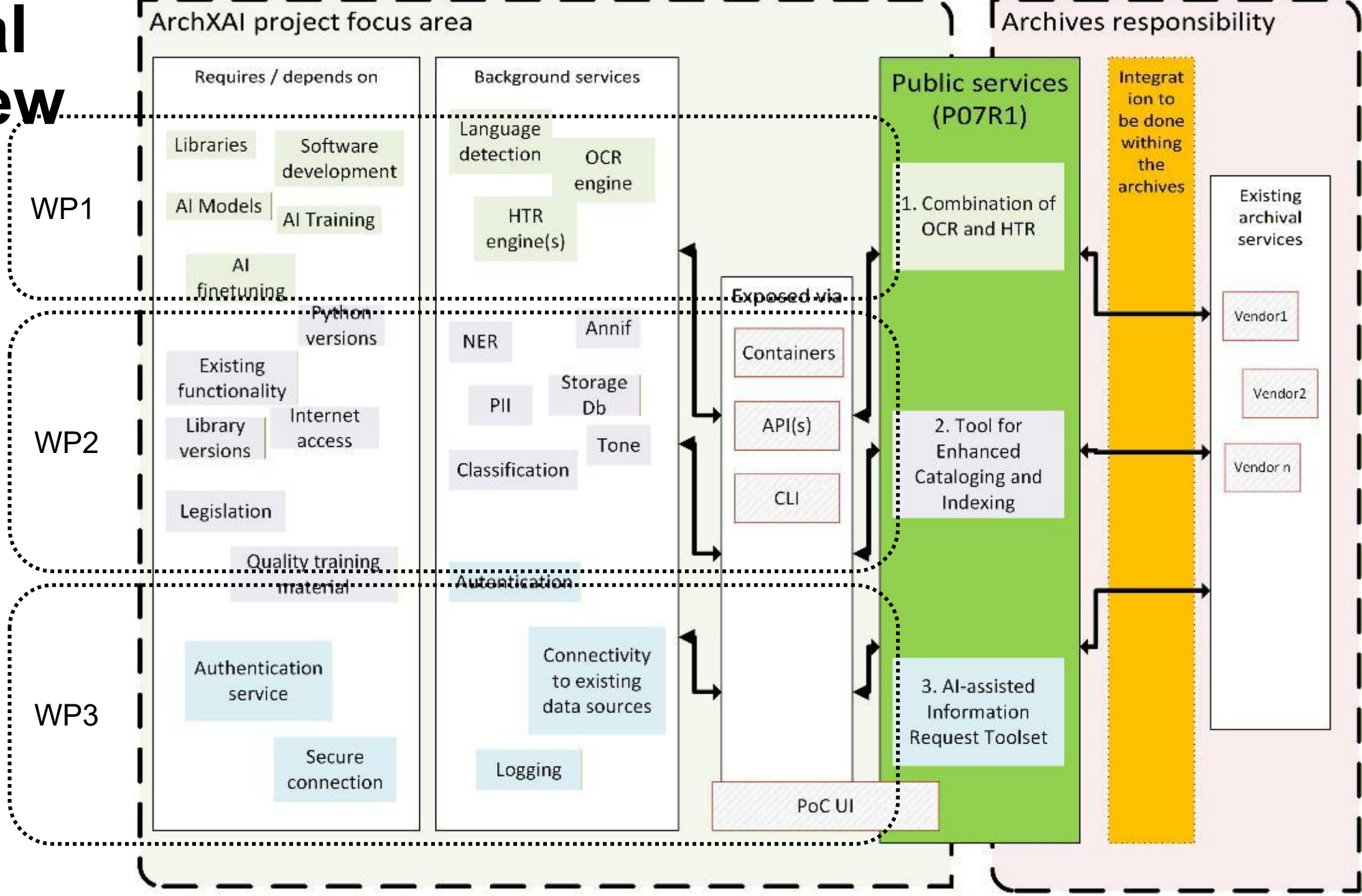
Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

Three WPs

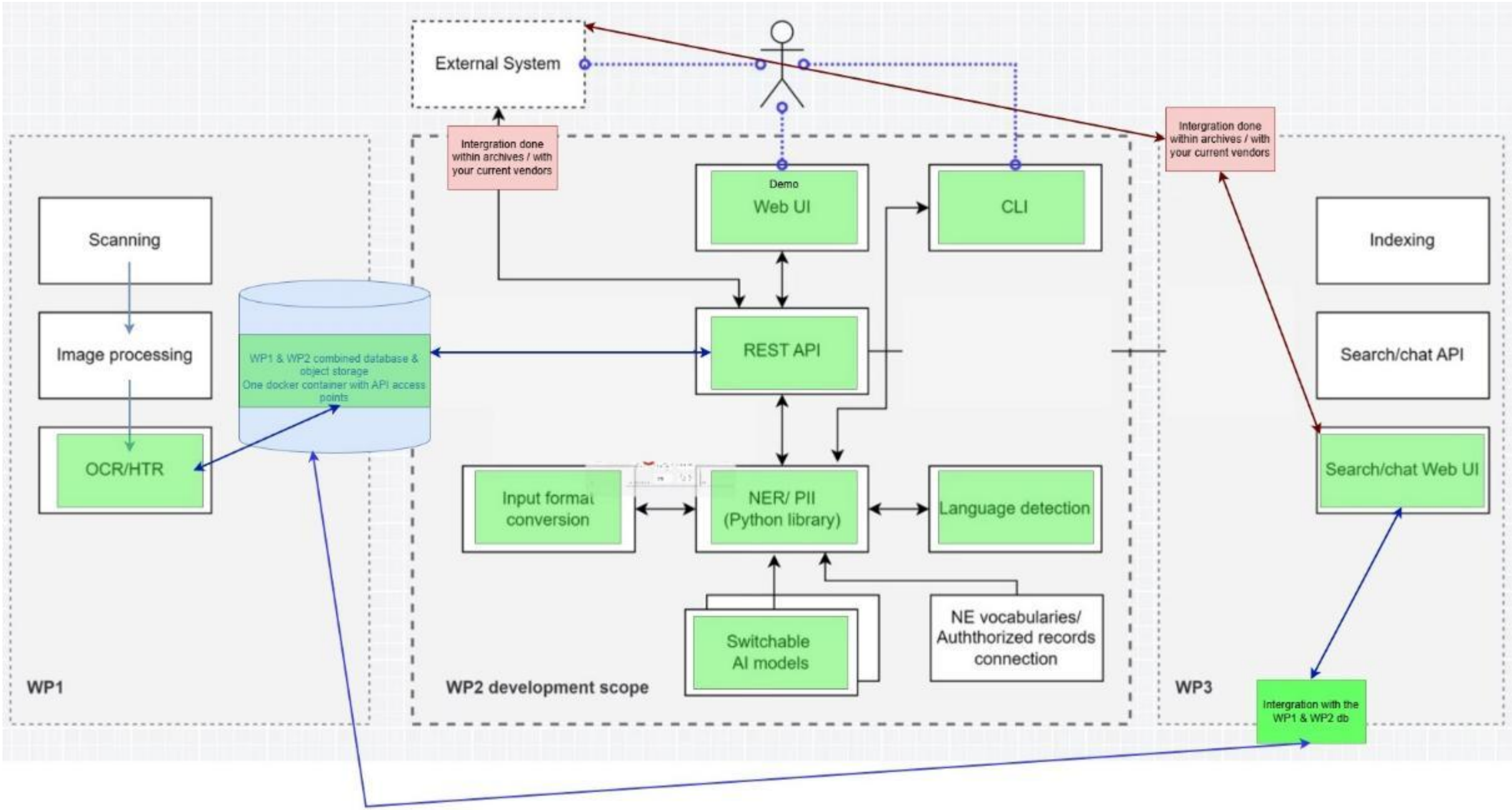
- **WP1: Analog to Digital Conversion**
  - Converts archival materials into high-quality digital images
  - Uses annotated material while training HTR (+OCR if needed) AI models
  - Trained models utilized in recognition process
- **WP2: AI-based Cataloguing and indexing**
  - Extracts general and sensite metadata, names, dates, places, tone, etc. to create rich metadata storage attached to document → Can be harvested/transferred/imported into existing archival systems.
- **WP3 AI-assisted Information Requests**
  - Design and implements intuitive ”AI helper for archives” which enable faster and more accurate information retrieval for archivists and general public.

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

General overview



A bit technical overview



Digitization & HTR & OCR (WP1)

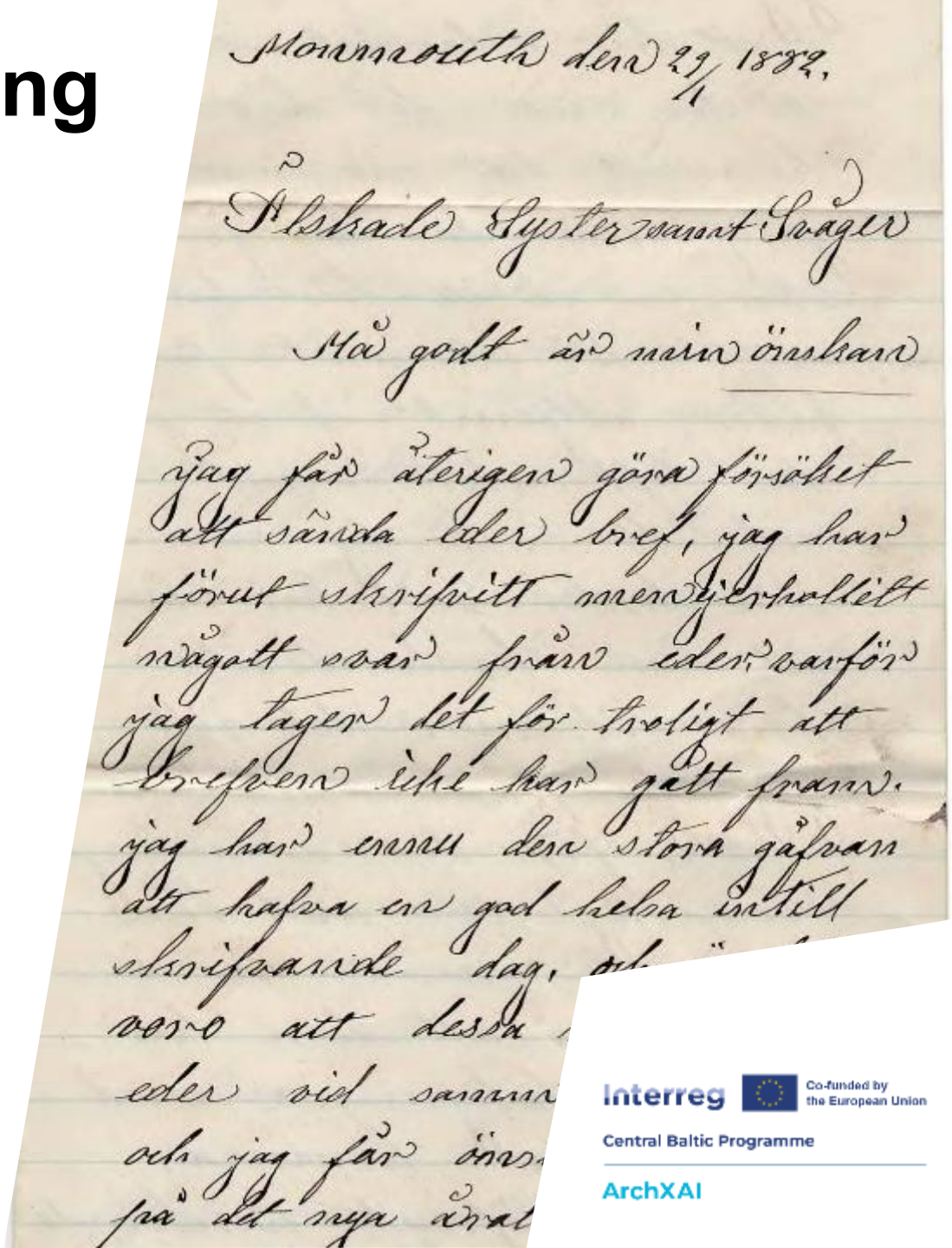
Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

Digitising the collections

- **Coordinated Digitization Plan**
  - A shared plan ensures consistent quality and prioritizes historically relevant cross-border materials for digitization.
- **Hand picked materials**
  - Common aspect, all should contain old hand writing with project languages + german, russian and swedish
  - Digitization & annotation
- **Dynamic training material generation**
- → Fuel for the AI training & accessible collections

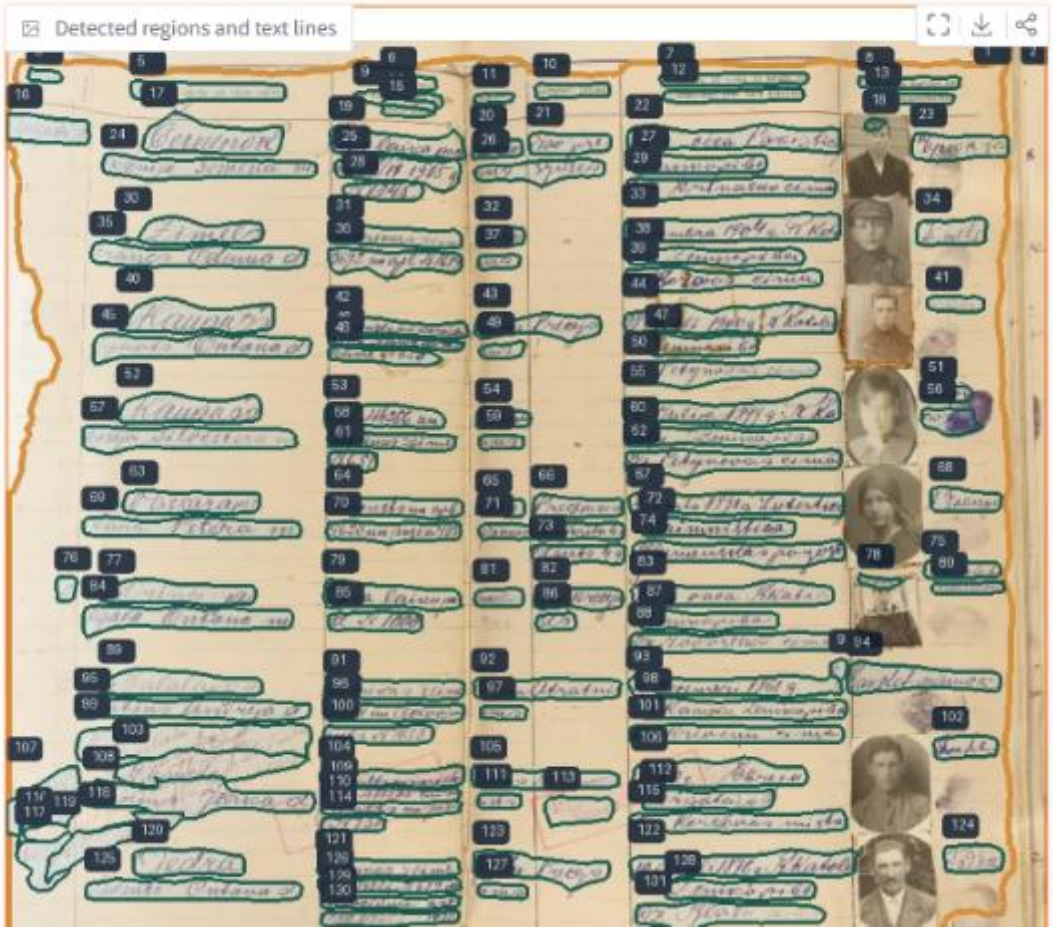
Teaching AI to Read Handwriting

- **Accurate Handwriting Recognition**
  - AI models achieve low character error rates across Estonian, Latvian, and Cyrillic scripts, reducing manual corrections.
- **Use of Synthetic Data**
  - Synthetic handwriting samples supplement real data, enabling faster learning and better generalization across styles.
- **Open Science Demonstrations**
  - Early models are publicly demonstrated to encourage feedback, transparency, and collaboration among researchers and professionals.



Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

Teaser



Transcript

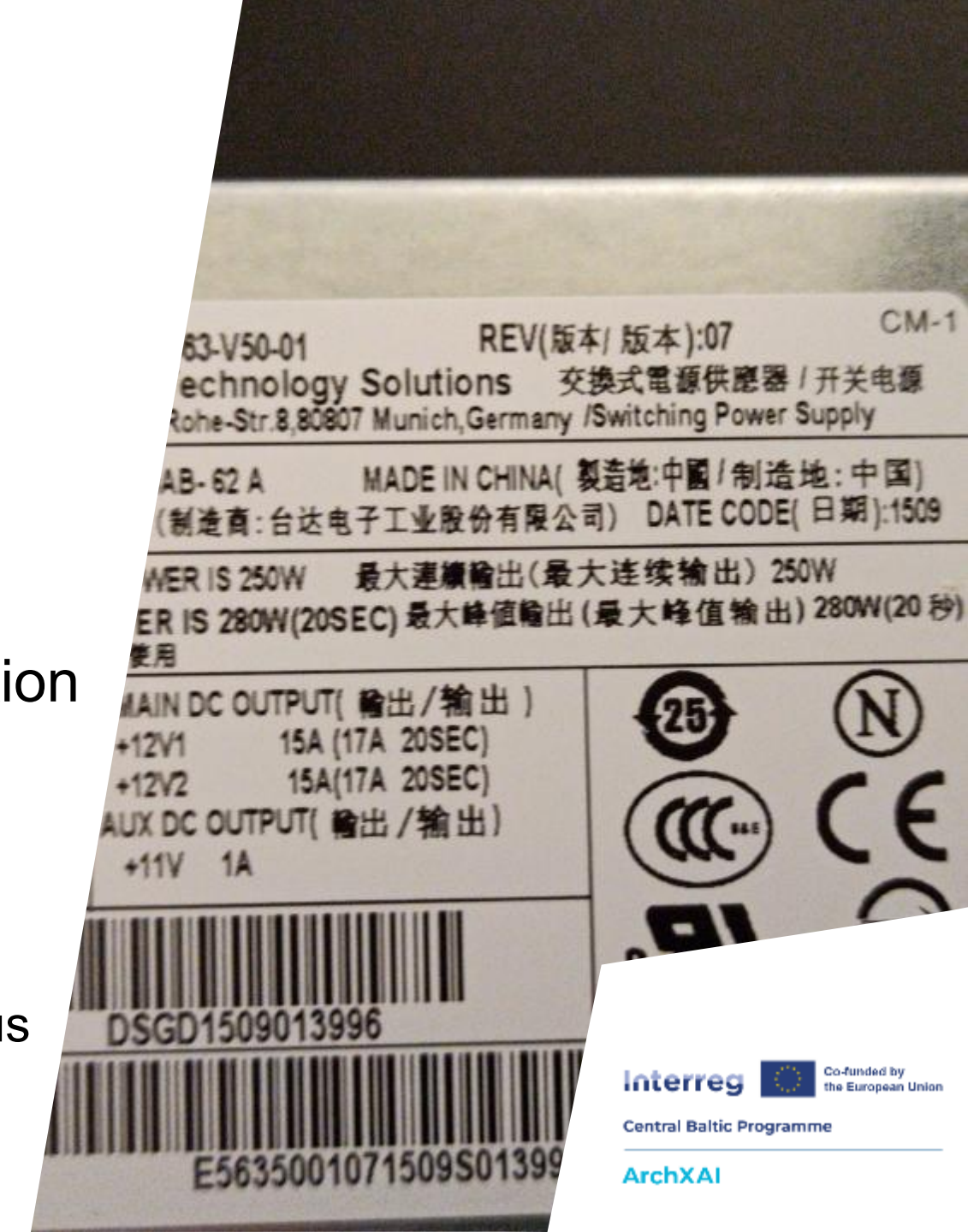
10. Gimenes sislars
11. Dzimis
12. 4) ariekšbas pret kara klausibu
13. pirksā nuspeedums
14. Jauns pils
15. Jaunis pārī
16. Augusta 10
17. Ģernenok
18. izsniegt manam brālim Tichvinskis Jānim Nikolaja d.
19. Caralaika par
20. Schasku
21. Nav pre-
22. 24 g. vēca Parciztici
23. Uzmanism
24. Iksenija Šemena m.
25. nu 18/17 1915 g
26. ams spuses
27. Zemkopība
28. Neras.
29. Oz Driņašku cimā
30. Zimels
31. Metrikas ceiņu
32. Irwas
33. (Eeembra 1904 g. R. Kato

Models

- https://huggingface.co/Kansallisarkisto/rfdetr_textline_textregion_detection_model **(Segmentation)**
- https://huggingface.co/Kansallisarkisto/latvian-large-handwritten **(Latvia)**
- https://huggingface.co/Kansallisarkisto/estonian-large-handwritten **(Estonia)**
- https://huggingface.co/Kansallisarkisto/multicentury-htr-model **(Finnish & Swedish)**
- https://huggingface.co/Kansallisarkisto/cyrillic-large-handwritten **(Cyrillic)**
  - Not included in demo but model is available already

OCR with basic Paddle OCR

- Enhancements later if required
- Converts images of printed text into machine-readable form
- Intention: To get accurate text representation for latter analysis
- Stored into Db
  - Can be used to retrieve needed data in various formats



Interreg Co-funded by the European Union
Central Baltic Programme
ArchXAI

Teaser

ArchXAI TEST OCR + NER Demo

OCR → NER pipeline

OCR | NER

Upload image

Run OCR

OCR extracted text

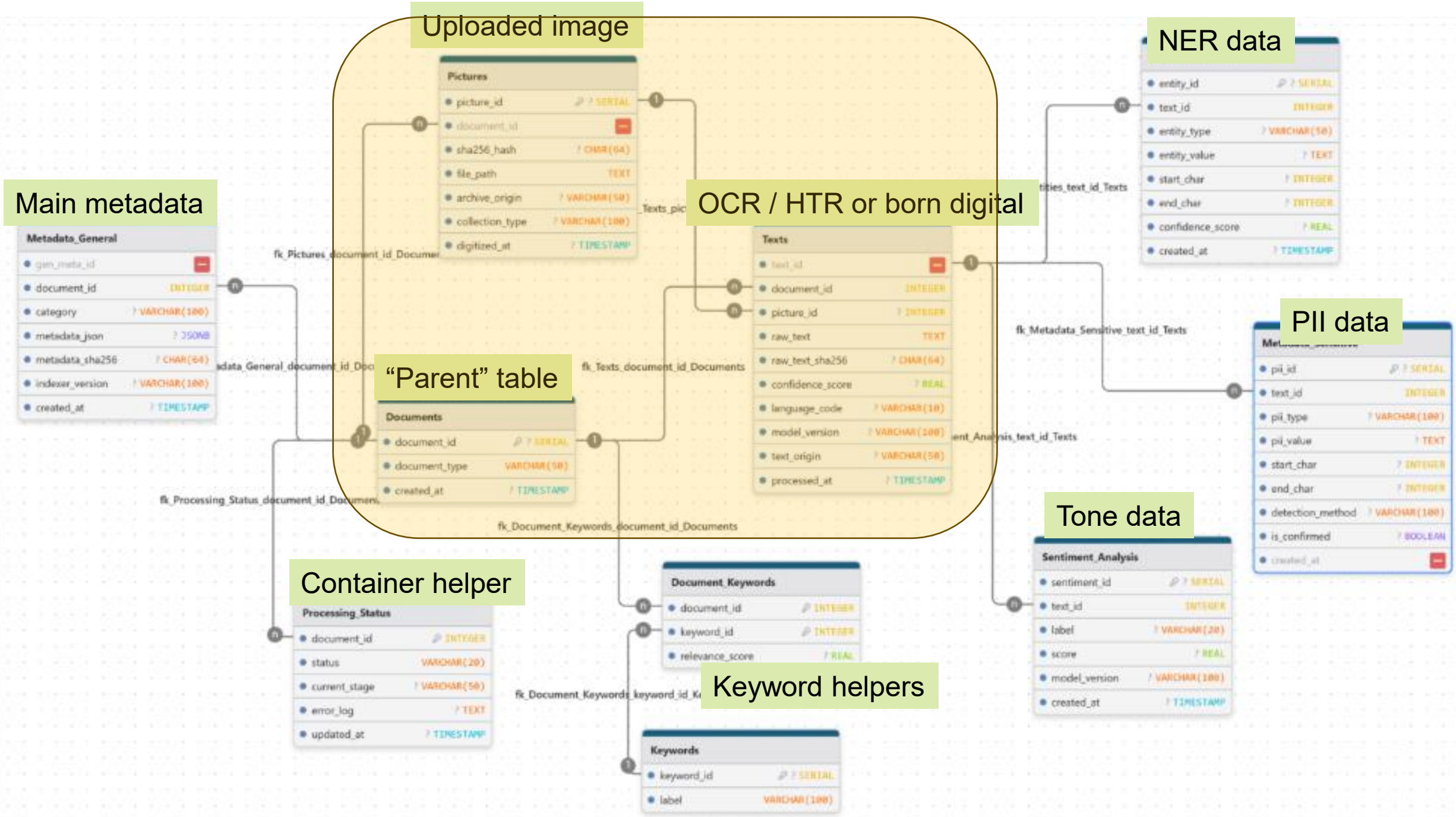
```
CM-1
83-V50-01
REV():07
echnology
Solutions
Rohe-Str.8,80807 Munich, Germany /Switching Power Supply
AB-62 A
MADE IN CHINA(
(:7I)
DATE CODE( ):1509
WER IS 250W
)250
ERIS280W(20SC))280W(20
EB
25
N
MAINDCOUTPUT/
+12V1
15A (17A 20SEC)
+12V2
```

OCR Performance

OCR time: 3.482 s

Indexer (WP2)

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI


Uploaded image
Pictures
picture_id
document_id
sha256_hash
file_path
archive_origin
collection_type
digitized_at
NER data
entity_id
text_id
entity_type
entity_value
start_char
end_char
confidence_score
created_at
Main metadata
Metadata_General
document_id
category
metadata_json
metadata_sha256
indexer_version
created_at
OCR / HTR or born digital
Texts
document_id
picture_id
raw_text
raw_text_sha256
confidence_score
language_code
model_version
text_origin
processed_at
PII data
pii_id
text_id
pii_type
pii_value
start_char
end_char
detection_method
is_confirmed
"Parent" table
Documents
document_id
document_type
created_at
Tone data
Sentiment_Analysis
sentiment_id
text_id
label
score
model_version
created_at
Container helper
Processing_Status
document_id
status
current_stage
error_log
updated_at
Keyword helpers
Document_Keywords
document_id
keyword_id
relevance_score
Keywords
keyword_id
label
fk_Pictures_document_id_Documents
fk_Texts_document_id_Documents
fk_Metadata_Sensitive_text_id_Texts
fk_Processing_Status_document_id_Documents
fk_Document_Keywords_document_id_Documents

Planned functionalities

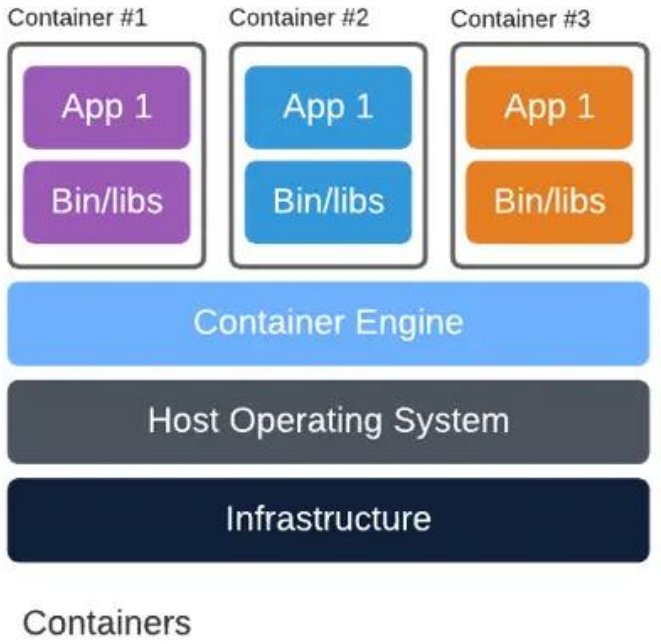
Model	mean precision	mean recall	mean f1	mean LOC precision	mean LOC recall	mean LOC f1	mean ORG precision	mean ORG recall	mean ORG
pierre-tassel/rapido-ner-entity	0.547	0.549	0.548	0.478	0.217	0.298	0.540	0.612	0.57
51la5/roberta-large-NER	0.397	0.669	0.455	0.615	0.632	0.613	0.511	0.580	0.54
Davlan/distilbert-base-multilingual-cased-ner-hrl	0.480	0.546	0.503	0.577	0.644	0.605	0.411	0.383	0.39
nicolauduran45/affilgood-ner-multilingual-v2	0.177	0.218	0.195	0.269	0.264	0.259	0.123	0.206	0.15

timestamp	model	dataset	sentence	precis	recall	f1	skipped_r	truncated_misalig	param_count	param_by
12/2/2025 11:19	pierre-tassel/rapido-ner-entity	Latvian-food-NE	744	0.1528	0.609865	0.244385	0	0	305773639	122309455
12/2/2025 11:19	51la5/roberta-large-NER	Latvian-food-NE	744	0.195	0.730942	0.307838	0	0	558849032	223539612
12/2/2025 11:19	Davlan/distilbert-base-multilingual-cased-ner-hrl	Latvian-food-NE	744	0.386	0.569507	0.460145	0	0	134741001	53896400
12/2/2025 11:20	nicolauduran45/affilgood-ner-multilingual-v2	Latvian-food-NE	744	0.1313	0.174889	0.15	0	0	277466129	110986451

timestamp	model	dataset	sentence	precis	recall	f1	skipped_r	truncated_misalig	param_count	param_by
12/10/2025 8:39	pierre-tassel/rapido-ner-entity	data/multileg_l	3120	0.5467	0.54911	0.547909	0	0	305773639	122309455
12/10/2025 8:40	51la5/roberta-large-NER	data/multileg_l	3120	0.5989	0.606629	0.602736	0	0	558849032	223539612
12/10/2025 8:41	Davlan/distilbert-base-multilingual-cased-ner-hrl	data/multileg_l	3120	0.5735	0.522312	0.546708	0	11	134741001	53896400
12/10/2025 8:41	nicolauduran45/affilgood-ner-multilingual-v2	data/multileg_l	3120	0.2231	0.260297	0.240243	0	0	277466129	110986451

Read Me | HF Est | HF Fin | HF Lat | HF Rus | LLM All | Combined All | Combined 744 | +

- Containerized solution
- NER& PII
  - https://huggingface.co/spaces/presidio/presidio_demo
- Tone / Sentiment analysis
  - ~10 different models evaluated
- Document classification
- Metadata extraction
  - With Annif for Finnish, Swedish and English
    - Possibility to train for other languages
  - https://annif.org/
- These all will write into database
  - There will be an API for retrieving the db content



NER

ArchXAI TEST OCR + NER Demo

OCR → NER pipeline

OCR | NER

Input Text

Ceļš uz mācībām latviešu valodā
2004. gads – Latvijas vēsturē viens no izteiksmīgākajiem gadiem, ja runā par valsts pārvaldes virzītu politiku, lai mazākumtautību skolās mācības notiktu ne tikai krievu, bet arī latviešu valodā. Tobrīd galvenais reformu varonis – izglītības ministrs Kārlis Šadurskis, pati reforma, ka vidusskolas posmā mācībām latviski jābūt ne mazāk kā 60 % apjomā, bet pretim teju sacelšanās: ministrs tiek saukts par Melno Kārli, galvenais protestētāju sauklis – "Rokas nost no krievu skolām".

Tobrīd mazākumtautību skolu Latvijā bija daudz, sevišķi lielajās pilsētās – krievu valodā mācījās aptuveni trešā daļa no teju 300 tūkstošiem Latvijas skolēnu, kas bija viens no Padomju Savienības mantojumiem, desmitiem gadu laikā Latvijā ievedot vairākos simtos tūkstošu mērāmu ne latviski runājošu darbaspēku.

Par spīti protestiem, reforma toreiz tika realizēta, savukārt par nākamo zīmīgo krustpunktu kļuva 2018. gads, kad tika noteikts, ka vidusskolas posmā mācībām jābūt tikai latviski. Tāpat ieviestas arī valodas proporcijas – proti, līdz 6. klasei 50 % mācību satura jābūt latviski, no 7. līdz 9. klasei latviešu valodas īpatsvaram jau jābūt 80 % apmērā. Protestu šoreiz mazāk, reformas pretinieki šo lēmumu apstrīdēja Satversmes tiesā, bet nesekmīgi.

Un visbeidzot 2022. gads, kad līdz ar Krievijas iebrukumu Ukrainā un Uzvaras pieminekļa krišanu Rīgā Latvijas politiki atvēzējas pēdējam solim – pilnīgai pārejai uz mācībām tikai latviski pakāpeniski pa noteiktām klašu grupām triju gadu laikā. Un visa rezultātā šis – 2025./2026. – mācību gads ir pirmais gads, kad nu visiem, visiem ir jāmācās tikai latviešu valodā. Par šo – protestu faktiski nekādu, masu mediju pārsvarā labas ziņas no Izglītības un zinātnes ministrijas, cik viss ir lieliski, cik visi šai reformai ir gatavi, un kā viss notiek.

Use OCR text | Analyze

Detected Language: lv Latvian

Performance (NER): NER time: 2.605 s

Entities Highlighted

Ceļš uz mācībām latviešu valodā
2004. gads – Latvijas LOCATION
vēsturē viens no izteiksmīgākajiem gadiem, ja runā par valsts pārvaldes virzītu politiku, lai mazākumtautību skolās mācības notiktu ne tikai krievu, bet arī latviešu valodā. Tobrīd galvenais reformu varonis – izglītības ministrs
Kārlis Šadurskis PERSON
, pati reforma, ka vidusskolas posmā mācībām latviski jābūt ne mazāk kā 60 % apjomā, bet pretim teju sacelšanās: ministrs tiek saukts par
Melno Kārli PERSON
, galvenais protestētāju sauklis – "Rokas nost no krievu skolām".
Tobrīd mazākumtautību skolu Latvijā LOCATION bija daudz, sevišķi lielajās pilsētās – krievu MISC
valodā mācījās aptuveni trešā daļa no teju 300 tūkstošiem Latvijas LOCATION skolēnu, kas bija viens no Padomju Savienības LOCATION
mantojumiem, desmitiem gadu laikā Latvijā LOCATION ievedot vairākos simtos tūkstošu mērāmu ne latviski runājošu darbaspēku.
Par spīti protestiem, reforma toreiz tika realizēta, savukārt par nākamo zīmīgo krustpunktu kļuva 2018. gads, kad tika noteikts, ka vidusskolas posmā mācībām jābūt tikai latviski. Tāpat ieviestas arī valodas proporcijas – proti, līdz 6. klasei 50 % mācību satura jābūt latviski, no 7. līdz 9. klasei latviešu valodas īpatsvaram jau jābūt 80 % apmērā. Protestu šoreiz mazāk, reformas pretinieki šo lēmumu apstrīdēja
Satversmes ties ORGANIZATION ā, bet nesekmīgi.
Un visbeidzot 2022. gads, kad līdz ar Krievijas LOCATION iebrukumu Ukrainā LOCATION un Uzvaras pieminekļa krišanu
Rīgā Latvijas LOCATION
politiki atvēzējas pēdējam solim – pilnīgai pārejai uz mācībām tikai latviski pakāpeniski pa noteiktām klašu grupām triju gadu laikā. Un visa rezultātā šis – 2025./2026. – mācību gads ir pirmais gads, kad nu visiem, visiem ir jāmācās tikai latviešu valodā. Par šo – protestu faktiski nekādu, masu mediju pārsvarā labas ziņas no
Izglītības un zinātnes ministrijas ORGANIZATION , cik viss ir lieliski, cik visi šai reformai ir gatavi, un kā viss notiek.

Text	Type	Confidence	Start	End
Latvijas	LOCATION	100.00%	45	53
Kārlis Šadurskis	PERSON	99.91%	283	299
Melno Kārli	PERSON	99.85%	437	448
Latvijā	LOCATION	100.00%	543	550
krievu	MISC	94.96%	591	597
Latvijas	LOCATION	100.00%	656	664
Padomju Savienības	LOCATION	99.96%	692	710
Latvijā	LOCATION	100.00%	745	752

Archivist toolset (WP3)

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

On going development

- **Bottle necks identification**
  - **Joint “covert ops” ☺ team visited NAF, NAE and NAL**
  - **Workshop 1 after lunch**
  - **How AI could assist in information requests handling**
- **Toolset requirements from end user point of view**
  - No technical things yet

68 313
INFORMATION REQUESTS IN 3 COUNTRIES (2025)

206
ARCHIVISTS WORKING IN 3 COUNTRIES

in 5 to 30 days
RESPONSE TO THE CUSTOMER

Information Request Bottlenecks

Estonia, Latvia, Finland

1. THE CLIENT SUBMITS AN INFORMATION REQUEST
2. REQUEST IS REGISTRATED
3. REQUEST IS ASSESSED AND ASSIGNED TO THE RIGHT UNIT
4. CLARIFICATION IS ASKED FROM THE CLIENT (IF NEEDED)
5. ARCHIVIST SEARCH FOR MATERIAL
6. MATERIAL IS DIGITISED AND INTERPRETED
7. RESPONSE IS PREPARED
8. PAYMENT IS PROCESSED
9. RESPONSE IS SENT TO THE CLIENT

FORTHCOMING DEVELOPMENT

- Archivist toolset implementation and rollout
- General public toolset implementation and piloting
- Multilingual cross-border access

Demo links

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI

OCR + NER

- https://demot.memorylab.fi/archxai/latvia-demo/

HTR

- https://demot.memorylab.fi/archxai/latvia-demo-htr/

Interreg
Co-funded by the European Union
Central Baltic Programme
ArchXAI