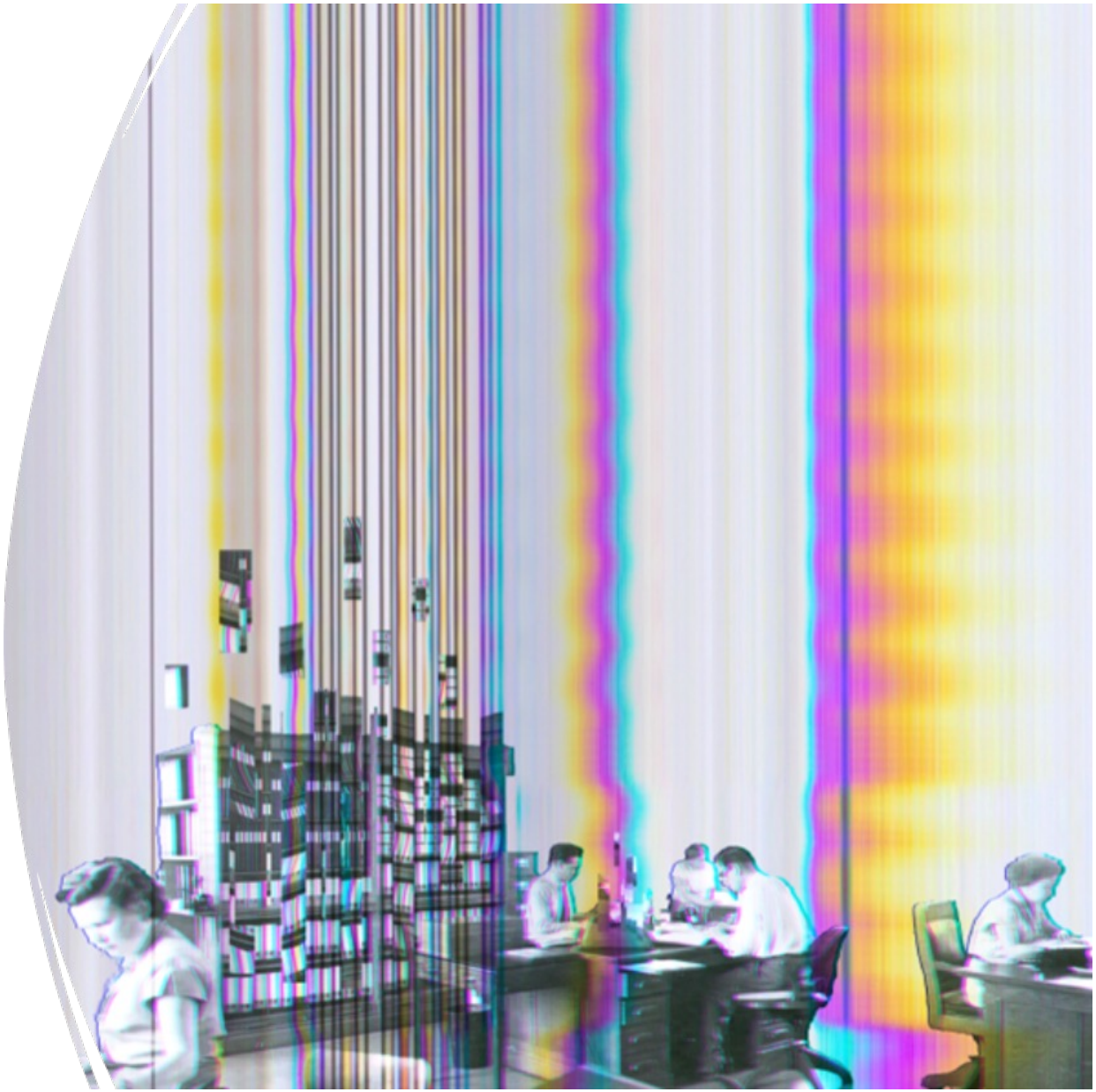

THE NATIONAL
ARCHIVES OF FINLAND

Demonstration of AI-driven Developments Linked to Digitization

Introduction

- DALAI-project
- Automatic validation of image quality
- Development of content recognition
- Document structure recognition and automatic metadata generation
- Project outcomes and deployment



DALAI-project

- 1.9.2021–31.8.2023
- EU Regional Development Fund
- 4 co-partners
  - National Archives of Finland (leader)
  - ELKA – Central Archives for Finnish Business Records
  - XAMK – South-Eastern Finland University of Applied Sciences
  - Disec Oy – a company related to processing of digital material
- Project objectives:
  - to advance the deployment of artificial intelligence and machine learning methods in the memory institutions
  - to publish these components as open source applications → project results are open to all users

Vipuvoimaa EU:lta 2014–2020

Euroopan unioni
Euroopan aluekehitysrahasto

THE NATIONAL ARCHIVES OF FINLAND

XAMK
Kaakkois-Suomen ammattikorkeakoulu

elka
SUOMEN ELINKEINOELÄMÄN KESKUSARKISTO

DISEC

DALAI-project

WP 1 Automatic validation of image quality

(National Archives of Finland)

WP 2 Development of content recognition (National Archives of Finland)

WP 3 Document structure recognition and automatic metadata generation

(National Archives of Finland)

WP 4 Business cooperation and network-like development

(Disec Oy)

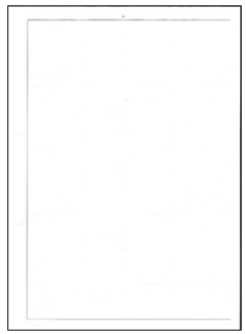
WP 5 User interface, where WP 1-3 components were integrated

(XAMK)

Automatic validation of image quality (WP 1)

- **Blank page recognition**
- **Post-it note recognition**
- **Bent corner recognition**

- The need to develop these components came from **mass digitization (National Archives of Finland)**
  - Efficient digitization of documents that are in the posession of state authorities and are preserved permanently
  - After digitizing, the original paper documents will be disposed
  - In mass digitization almost 38 % of the digitized material so far consits of blank pages
  - Various flaws on digital images have been recognised, e.g. post it notes or bent corners that might hide information underneath.



THE NATIONAL ARCHIVES OF FINLAND

Automatic validation of image quality (WP 1)

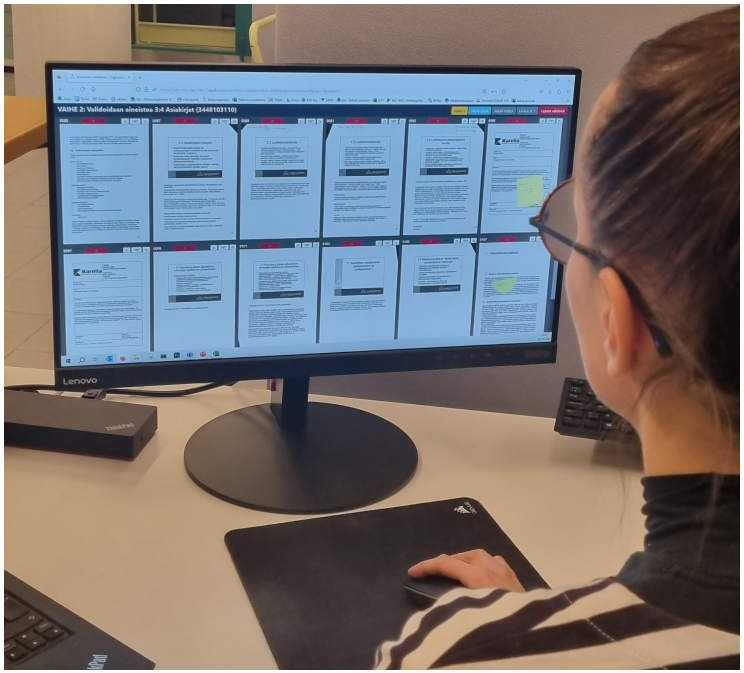
- **Blank page recognition**
  - Training data: 116 000 images
- **Post-it note recognition**
  - Training data: 55 657 images (4318 with post it notes)
  - Rescan recognition was also developed.
- **Bent corner recognition**
  - Training data: 54850 images (6551 with bent corner)

All the training data was gathered from different kind of digitized archival material and it was also produced by data annotators

THE NATIONAL ARCHIVES OF FINLAND

Automatic validation of image quality (WP 1)

- **Benefits:**
  - Blank page recognition → unnecessary data files are not created → save processing power and storage
  - Improve the validation process
  - Catch flaws better than in manual validation
  - Provide a better user experience
- **Limitations**
  - Small errors in recognition e.g. text visible on the other side of the page or colourful square text elements or images.

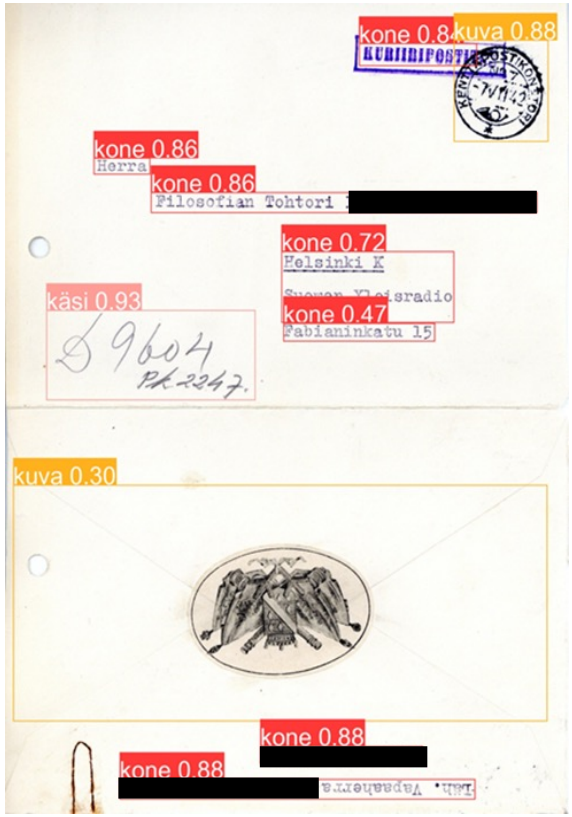


THE NATIONAL ARCHIVES OF FINLAND

Development of content recognition (WP 2)

- **Text type recognition**
  - The component classifies pages into 3 different categories:
    - page contains only typewritten text
    - page contains only handwritten text
    - page contains both handwritten and typewritten text (combined class)
  - Training data: 65 275 images

- **Segmentation tool**
  - The component searches, delimits and classifies handwritten areas, textwritten areas, images and tables from the document.



THE NATIONAL ARCHIVES OF FINLAND

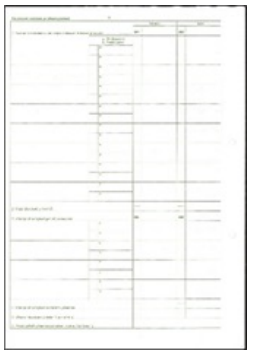
Development of content recognition (WP 2)

- **Benefits**
  - the text areas from the scanned document can be directed through right process (OCR or HTR) and the text can be better utilized for other purposes (NER, automated subject indexing etc.)
  - Segmentation tool can be used as a filter e.g. the image elements can be directed to a separate image recognition algorithm for content recognition
- **Limitations**
  - Small text written in the margin or dim text can make identification difficult
  - Processing power increases significantly (segmentation tool)

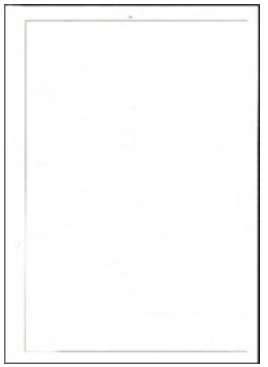
Document structure recognition and automatic metadata generation (WP 3)

- **<u>Case file recognition</u>**
  - Inheritance tax documents (Tax administrator's records)
  - Component recognises five categories in a case file
    - ➢ four-page inheritance tax form (classes 1-4)
    - ➢ estate inventory deed and it's appendices (class 5).
  - Training data: 21 888 images

Class 1



Class 2



Class 3

Class 4

THE NATIONAL ARCHIVES OF FINLAND

Document structure recognition and automatic metadata generation (WP 3)

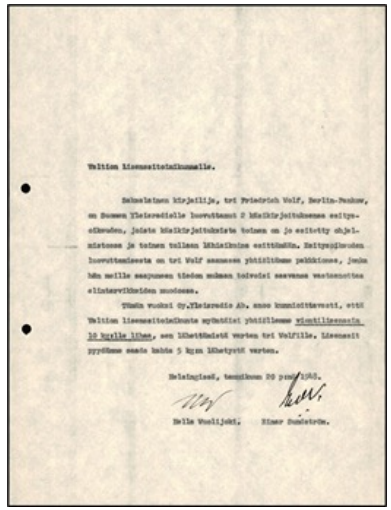
Case file recognition

- **Benefits:**
  - Easy to find and use certain documents
  - Can be further developed to recognise other kind of material/case files

- **Limitations:**
  - The model cannot be directly applied to different document types
  - We can find out where the case file starts and where it ends, but we do not know who it concerns
  - Some case files might include correction forms → there are multiple tax inheritance forms in one case file. The component does not recognise the difference.
  - Expensive and difficult to deploy

Document structure recognition and automatic metadata generation (WP 3)

- **Named entity recognition**
  - developed in collaboration with the FIN-CLARIAH research consortium
  - Survey for necessary entities was carried out
    - journal numbers, dates, business identity numbers, persons, events, organisations, locations, products
  - Training data: 160 602 entities
- **Tool for subject indexing**
  - Annif software (open source code/developed mainly by the Finnish National Library)



PERSON 1 DATE 2 ORG 3 GPE 4

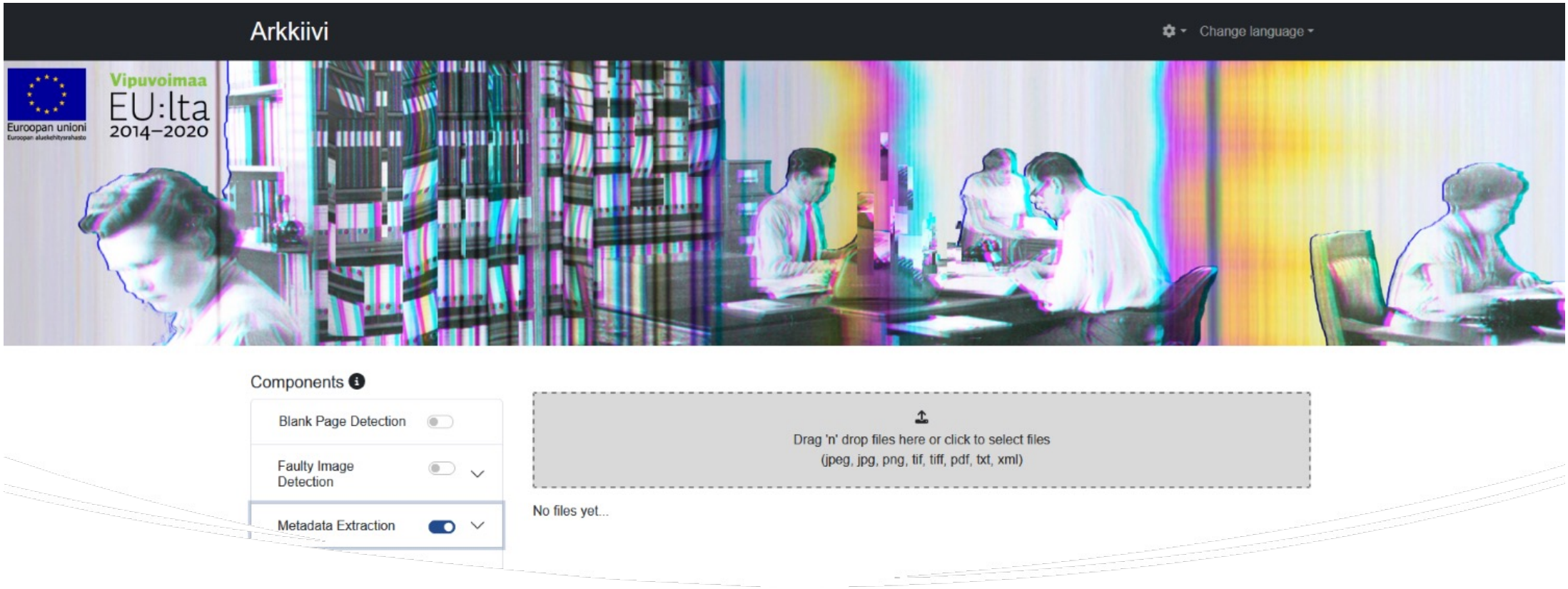
Valtion Lisenssitoimikunnalle. Saksalainen kirjailija, tri Friedrich Wolf, Berlin-Pankow, on Suomen Yleisradiolle luovuttanut 2 käsikirjoituksensa esitysoikeuden, joista käsikirjoituksista toinen on jo esitetty ohjel- mistossa ja toinen tullaan lähiaikoina esittämään. Esitysoikeuden luovuttamisesta on tri Wolf saamassa yhtiöltämme pakkkionsa, jonka hän meille saapuneen tiedon mukaan toivoisi saavansa vastaanottaa elintarvikkeiden muodossa. Tämän vuoksi Oy.Yleisradio Ab. anoo kunnioittavasti, että Valtion lisenssitoimikunta myöntäisi yhtiöllemme vientilisenssin 10 kg:lle lihaa, sen lähettämistä varten tri Wolfille. Lisenssit pyydämme saada kahta 5 kg:n lähetystä varten. Helsingissä, tammikuun 20 p:nä 1948. Hella Wuolijoki. Einar Sundström.

THE NATIONAL ARCHIVES OF FINLAND

Document structure recognition and automatic metadata generation (WP 3)

- **Benefits**:
  - better findability
  - new types of subject-based content units
  - recommendations of similar content
  - subject-based personalization of services

- **Limitations:**
  - Because the named entity recognition and subject indexing is based on OCR output, OCR-quality has a key role



Project outcomes

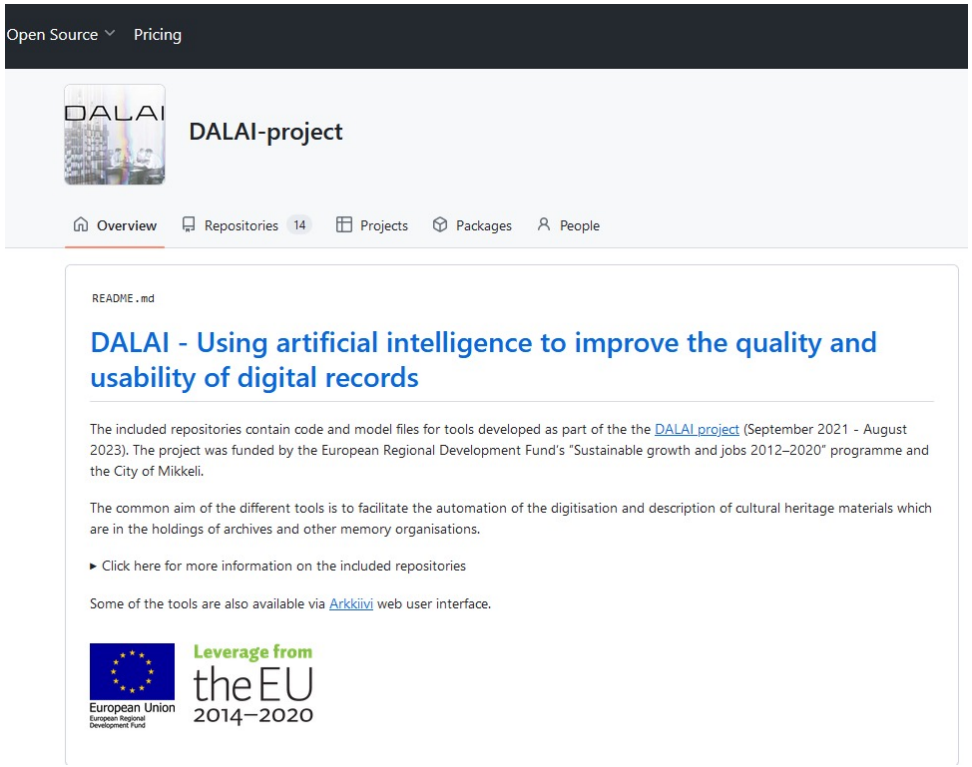
- User interface Arkkiivi (Mira Kolari/XAMK) was launched in the end of August 2023 and it is free for everyone to use (also in English): https://arkkiivi.fi/

THE NATIONAL ARCHIVES OF FINLAND

Project outcomes

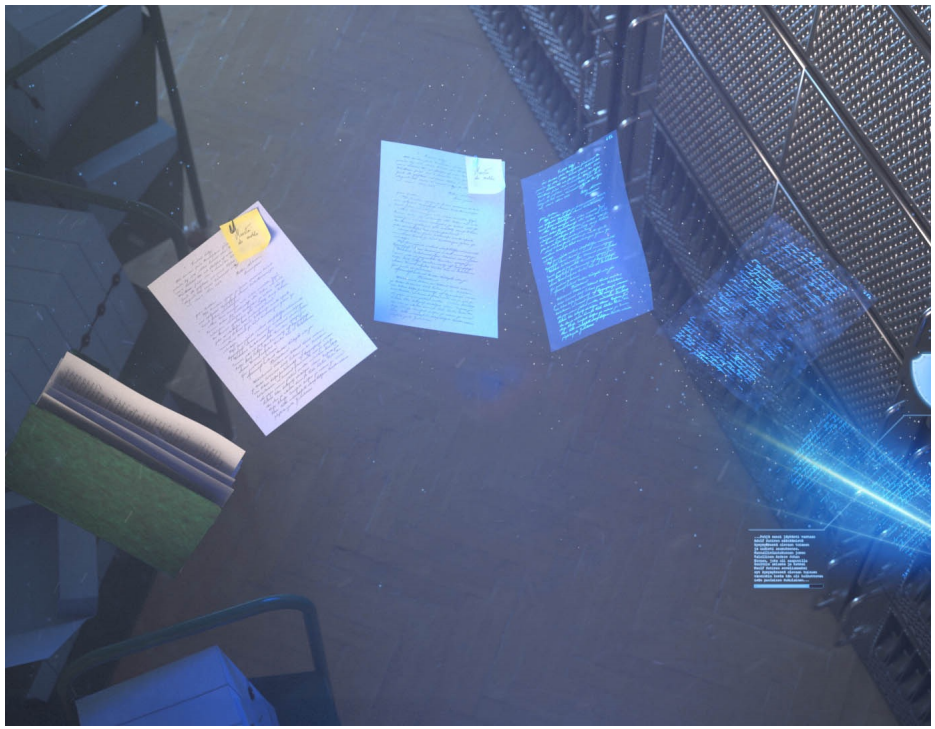
- Github: open code access https://github.com/DALAI-project
  - APIs can be used to create softwares suitable for different kind of needs

- Follow up project AIDA (*Potential of Artificial Intelligence for Digital Archives Users*, 1.9.2023-31.8.2024)
  - Value in the end users point of view



Deployment – challenges and opportunities

- **Challenges:**
  - Multiple systems, multiple dependencies → changes related to individual components
  - Components are not finished products → they require maintenance and further development
- **Focus on building an analytics infrastructure (National Archives of Finland)**
  - Components can be used separately if needed.
  - Components can be further developed independently
  - Processing pipeline
  - Selective use
  - Possibility to use retroactively



THE NATIONAL
ARCHIVES OF FINLAND

THE NATIONAL ARCHIVES OF FINLAND

www.kansallisarkisto.fi

@kansallisarkisto

@kansallisarkist