

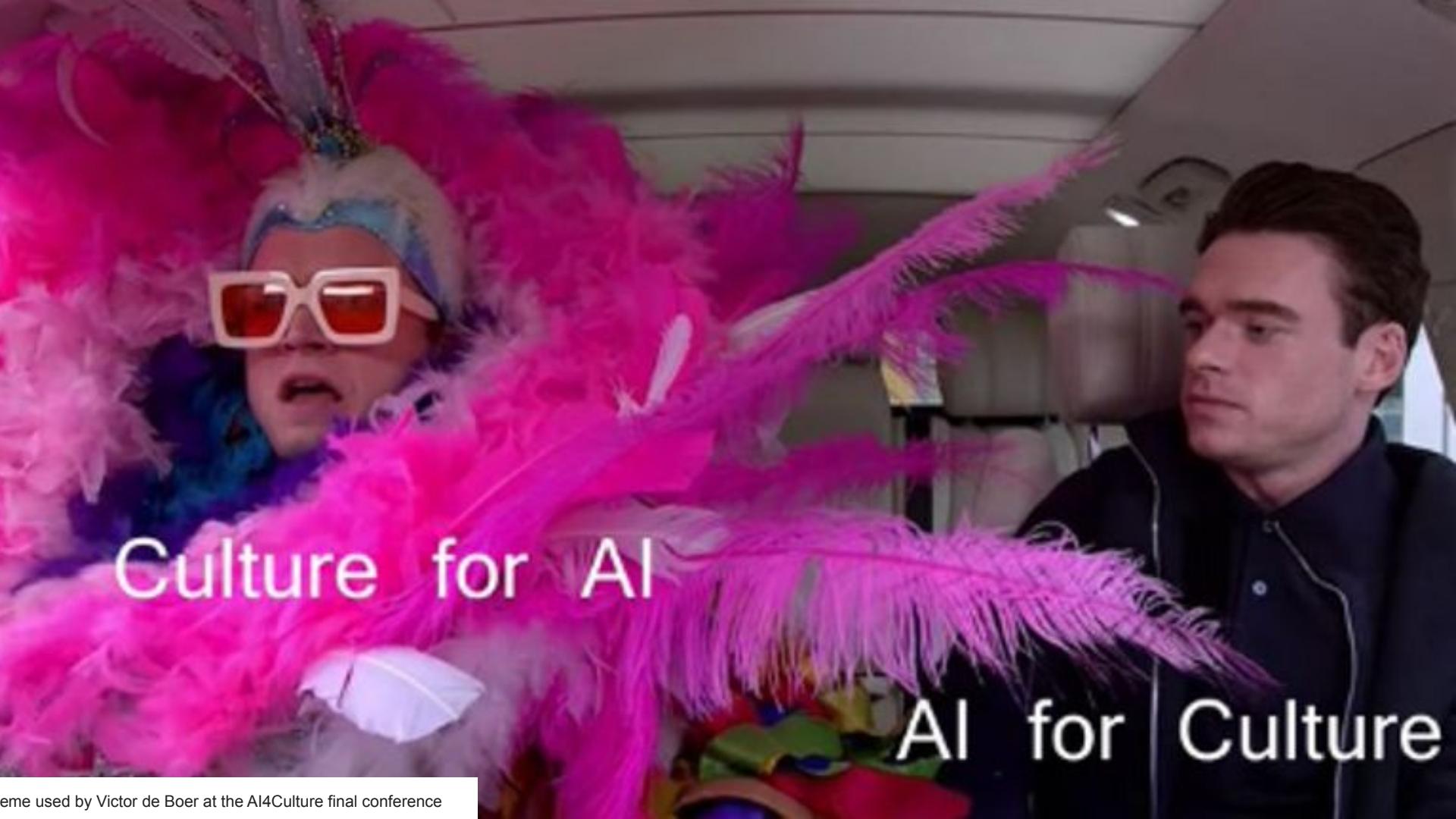


# An AI platform for the cultural heritage data space

---

**Eirini Kaldeli**

Senior Researcher at the National Technical University of Athens  
Co-founder of Datoptron

A woman with blue hair and white sunglasses is wearing a large, colorful feathered costume, including a headpiece and a long train of pink and purple feathers. She is looking towards the right. A man in a dark suit is standing to her right, looking in the same direction. The background is an indoor setting with a white ceiling.

Culture for AI

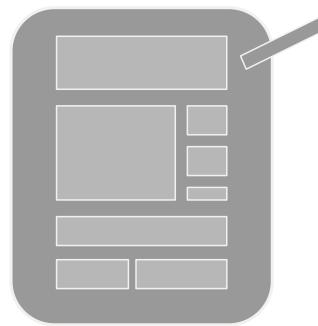
AI for Culture

# Agenda

- The importance of good-quality metadata: how can AI support metadata enrichment?
  - Automatic linking of textual metadata
  - Bias detection
  - Validation mechanisms
- Promoting broader AI applicability across diverse scenarios in the Cultural Heritage sector
  - Making AI technologies more accessible
  - Supporting the practical uptake of AI by CH organisations
  - Capacity building

**ai4culture**

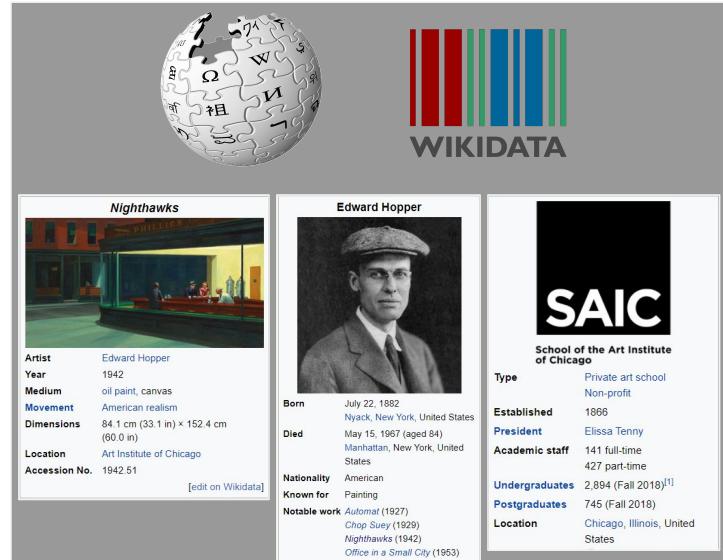
# Semantic linking



Description: Nighthawks is a 1942 oil on canvas painting by Edward Hopper that portrays people in a downtown diner late at night as viewed through the diner's large glass window. The artwork is displayed at the School of the Art Institute of Chicago.

## Semantic linking process

Description: **Nighthawks** is a 1942 oil on canvas painting by **Edward Hopper** that portrays people in a downtown diner late at night as viewed through the diner's large glass window. The artwork is displayed at the **School of the Art Institute of Chicago**.



**Nighthawks**  
Wikidata ID: Q83872

**Edward Hopper**  
Wikidata ID: Q203401

**School of the Art Institute of Chicago**  
Wikidata ID: Q7432601

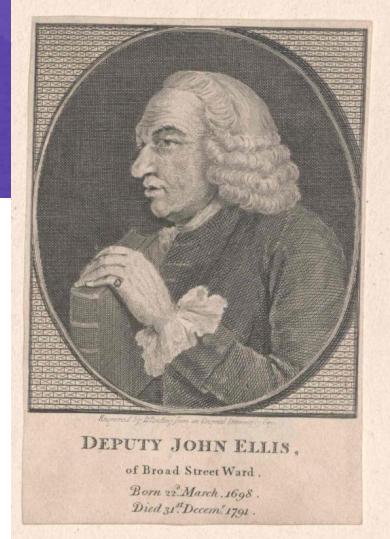
# Semantic linking

- Homogenisation

```
<dc:creator>John Ellis</dc:creator>  
<dc:creator>Ellis, John</dc:creator>  
<dc:creator>Jon Ellis</dc:creator>
```

- Unique meaning (disambiguation)

```
<dc:description> [...] In this video, British guitarist  
and songwriter John Ellis is performing live with his  
band, the Vibrators. [...].</dc:description>  
<dc:description> [...] This portrait depicts English  
clergyman and theologian John Ellis.  
[...]</dc:description>
```



# Semantic linking

- Contextualisation and multilinguality

URI: <http://semantics.gr/authorities/craft-item-types/1214765664>

RDF/XML  JSON-LD 

URI <http://semantics.gr/authorities/craft-item-types/1214765664>  
@rdf:about

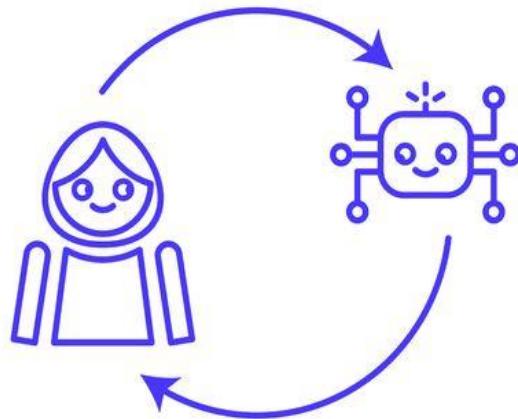
Προτεινόμενη ονομασία Πολυθρόνα   
skos:prefLabel  
Armchair 

Ευρύτερη έννοια (από ίδιο λεξιλόγιο) Εργαλεία και εξοπλισμός ► Οικιακός εξοπλισμός ► Επίπλωση ► Κάθισμα   
skos:broader  
Tools and equipment ► Household accessories ► Furniture ► Seat   
(<http://semantics.gr/authorities/craft-item-types/226871072>)

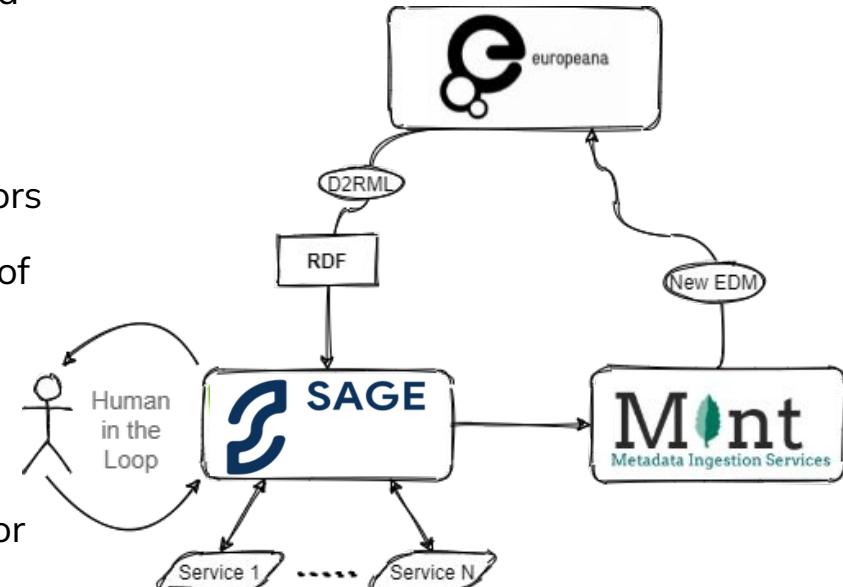
Ίδιο με armchair   
skos:exactMatch  
(<http://www.wikidata.org/entity/Q11285759>) wikidata

# Metadata enrichment methodology

Combine the power of Artificial Intelligence tools with human intelligence for improving the quality of cultural heritage metadata



- Import heterogeneous data and transform them to a knowledge graph (RDF)
- Publish, organise, and provide access to data
- Automatically enrich published data using annotators
- Create validation campaigns for human inspection of automatic enrichments
- Aligned with the Europeana Annotation Model
- Interlinked with the MINT metadata aggregation platform, component of the European Data space for Cultural heritage



# SAGE validation environment

- Human validation by experts in the field
- Grouping of annotations and string values together to speed up validation
- Humans can add their own annotations

Validations: Gennadius [20240917] (Place+of+Production)

Remaining time: 29:46      Validations: 8990   Annotations: 8990   Additions: 753   Progress: 100.00%

All   **Annotated**   Non Annotated

Rank	Place	Count	URI	Annotations
1	London	3003	<a href="http://semantics.gr/authorities/geonames-places-earth/2643743">http://semantics.gr/authorities/geonames-places-earth/2643743</a>	London export as: Prefix, Property
2	Αθήνα	1438	<a href="http://semantics.gr/authorities/geonames-places-earth/2643711">http://semantics.gr/authorities/geonames-places-earth/2643711</a>	Athens export as: Prefix, Property
3	Paris	482	<a href="http://semantics.gr/authorities/geonames-places-earth/2988307">http://semantics.gr/authorities/geonames-places-earth/2988307</a>	Paris export as: Prefix, Property
4	New York	184	<a href="http://semantics.gr/authorities/geonames-places-earth/5128638">http://semantics.gr/authorities/geonames-places-earth/5128638</a>	New York State <a href="http://semantics.gr/authorities/geonames-places-earth/5128581">http://semantics.gr/authorities/geonames-places-earth/5128581</a> New York export as: Prefix, Property
5	Τεργεστη	184	<a href="http://semantics.gr/authorities/geonames-places-earth/3165185">http://semantics.gr/authorities/geonames-places-earth/3165185</a>	Trieste export as: Prefix, Property
6	Κωνσταντινούπολη	148	<a href="http://semantics.gr/authorities/geonames-places-earth/745044">http://semantics.gr/authorities/geonames-places-earth/745044</a>	Istanbul export as: Prefix, Property
7	Ερμούπολη	110	<a href="http://semantics.gr/authorities/geonames-places-earth/262603">http://semantics.gr/authorities/geonames-places-earth/262603</a>	Ermoupoli export as: Prefix, Property

Page 1 of 8      Save      Cancel

# Lessons learnt: enrichment

- Precision versus recall
- No one fit-for-all solution: Different approach depending on available data, format, objective and required expertise.
- The quality of the original textual description affects the quality of the automatic enrichment
- Generic models (for NERD and LLMs) by themselves perform poorly (lack of sufficient context; training on unrelated corpora - domain-specificity; wrongly disambiguated)
- Fine-tuning: patterns for metadata values (e.g. “surname, name”, “city/region/country” etc); restrict to subset of thesaurus hierarchy; check other properties’ values etc.

## Lessons learnt: validation and filtering

- Involve validators who are fit for the task (domain experts, language proficiency, public campaigns)
- Be pragmatic when setting your validation targets
- The sample to be validated does not need to be large but it must be balanced and representative
- Establish thresholds for final inclusion in the metadata: majority vote; based on precision; automatic confidence scores

## SAGE in numbers



Annotations  
3.900.000+



Enriched  
Records  
1.500.000+



Users  
60+



Europeana  
Aggregators  
6



## DE-BIAS: Detecting and cur(at)ing harmful language in cultural heritage collections

Promoting a more inclusive and respectful approach to describing cultural heritage.

# Harmful Language in CH

- Title: "Home for *Invalids* Annual Report" Term for people with **disability**. It suggests a person is 'less than' their contemporaries and that they are weak, inferior or useless.
- Description: "List of *invalids* residing at the home, with notes on their physical and mental deficiencies." "Oriental" homogenizes diverse Asian identities and reflects Western **exoticism**.
- Title: "*Oriental* Lady in Traditional Costume" This term has long been used to dehumanize and harm **Native Americans** and **Indigenous people**, labeling them as **uncivilized** or **primitive**.
- Description: "Photograph of an *Oriental* woman wearing exotic attire, likely taken in a bazaar setting." This term refers to the **Khoikhoi people**, who live in the western part of South Africa. It is a Dutch **colonial term**, first used in the 17th century, and was based on an imitation of the sound of the Khoikhoi language.
- Title: "Portrait of a *Hottentot* Woman" This term refers to the **Khoikhoi people**, who live in the western part of South Africa. It is a Dutch **colonial term**, first used in the 17th century, and was based on an imitation of the sound of the Khoikhoi language.
- Description: "An ethnographic study of a *Hottentot* woman for educational purposes."

# The DE-BIAS Approach

- **Curated a multilingual vocabulary of harmful terms**

We built a structured vocabulary by working with CH professionals, researchers, and affected communities. Each entry includes a curated explanation of why the term may be considered harmful today, ensuring cultural and historical specificity and supporting accurate, respectful detection. **5 languages, 700 terms**

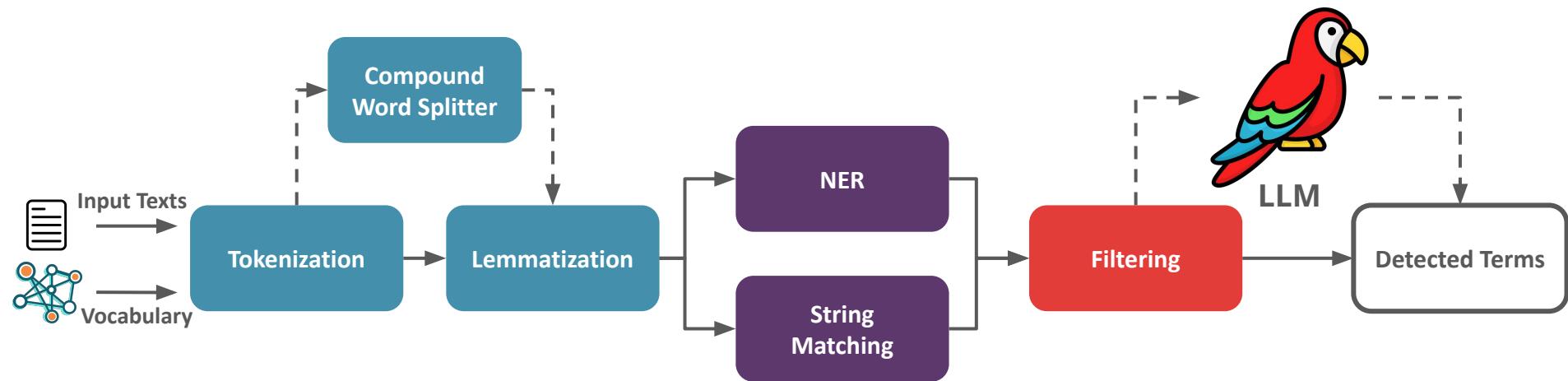
- **Developed a harmful language detection tool**

- Using the vocabulary, we created a tool that flags harmful terms in metadata using both traditional NLP techniques and large language models (LLMs).

- **Applied the tool at scale and integrated it into real-world platforms**

- We deployed our tool both as a standalone web app and via integration with Europeana.

# The Tool



TL;DR vocabulary-based fuzzy string matching and LLM for disambiguation

# Human validation



- **3,700 annotated texts** covering all 687 terms of the vocabulary
- **145 participants** contributed in CrowdSourcing campaigns

## BIAS IN THE METADATA

Start typing your keyword and select the appropriate term from the list. Then select a property to tag and a text fragment from the value.

ⓘ Hover on the labels of the list to display the context of each annotation

Transvestite 6 U.S. historical term used to describe individuals who wear clothes associated with a gender other than their gender assigned at birth; generally considered to be a derogatory term;

Comment

---- Select Error Type ----

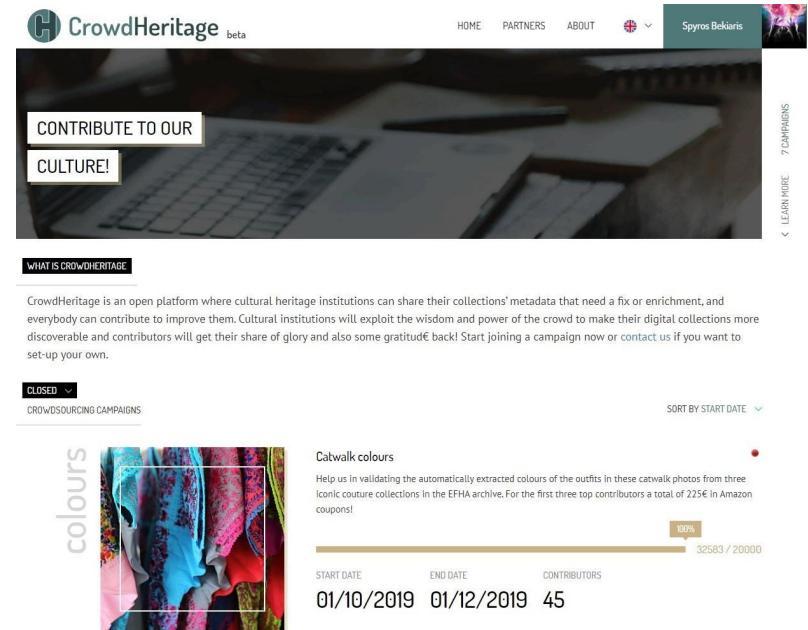
## Crowd- and niche-sourcing

- Distribute an online task to a group of people
- Broadly used in the CH domain: e.g. tagging, transcribathons, subtitl-a-thons
- Benefits:
  - For CHIs: improve the quality of their collections and engage with target audiences
  - For participants: interact with cultural heritage items



# Using crowd- and niche-sourcing

- A campaign leads to the enrichment of CH metadata
- Users asked to inspect an item (image, video, text, metadata) and fulfill a task
  - (a) Add annotations from scratch
  - (b) Validate/evaluate annotations produced by automatic algorithms
- Target audiences
  - Citizen Science (wide audiences)
  - Education (pupils, students)
  - Expert campaigns (closed groups)



The screenshot shows the CrowdHeritage website interface. At the top, there is a navigation bar with links for HOME, PARTNERS, ABOUT, and a user profile for Spyros Bekiaris. A sidebar on the right shows a profile picture and the text '7 CAMPAIGNS' and '< LEARN MORE'. The main content area features a large image of a computer keyboard and a hand pointing at it. Overlaid on this image are two white buttons: 'CONTRIBUTE TO OUR' and 'CULTURE!'. Below this is a section titled 'WHAT IS CROWDHERITAGE' with a description of the platform's purpose. A 'CLOSED' button is visible above a section titled 'CROWDSOURCING CAMPAIGNS'. One campaign listed is 'Catwalk colours', which is described as helping to validate automatically extracted colors of outfit photos from the EFHA archive. It includes a progress bar showing 100% completion, 32583 out of 20000 contributors, and a timeline from 01/10/2019 to 01/12/2019 with 45 contributors.

# Using crowd- and niche-sourcing

- Semantic enrichment: add terms representing various entities and concepts (from instruments and emotion tagging to color- and geo-tagging)
- Evaluate and post-edit automatic translations
- Validate the results of color and object detection algorithms
- Validate the results of bias detection algorithms and mark biased terms
- Rank automatically produced super-resolution images



# What AI can't do

- Photos of excavation places and findings in Ancient Corinth
- Connect archaeological items with people from the local community connected to their finding
- Participants: Inhabitants of the Ancient Corinth region with good knowledge of local people

ΜΕΤΑΔΕΛΟΜΕΝΑ ΑΝΤΙΚΕΙΜΕΝΟΥ

**CrowdHeritage**

APXIKH APPLICATIONS ΣΧΕΤΙΚΑ ΜΕ ΕΜΑΣ Maria Ralli

S. Stoa Shrine, Y. Metaxas



ΟΔΗΓΙΕΣ ΚΑΜΠΑΝΙΑΣ

Επιθεωρήστε την εικόνα και εντοπίστε άτομα που γνωρίζετε. Εάν το όνομα του ατόμου(ων) που προσδιορίσατε δεν βρίσκεται στη λίστα κάτω από το πεδίο "Επισημέωση", πληκτρολογήστε το στο ελεύθερο κείμενο παρακάτω.

Επισημέωση

Αρχίστε να πληκτρολογείτε μία λέξη κλειδί και επιλέξτε από τη λίστα.

Αρχίστε να πληκτρολογείτε έναν όρο και στη συνέχεια επιλέξτε

Επισημέωση Ελεύθερου Κειμένου

Περιγράψτε το αντικείμενο με δικά σας λόγια.

Γράψτε την περιγραφή σας εδώ...

ΠΡΟΣΘΗΚΗ

TITΛΟΣ  
S. Stoa Shrine, Y. Metaxas

ΑΝΤΙΚΕΙΜΕΝΟ  
Forum southwest

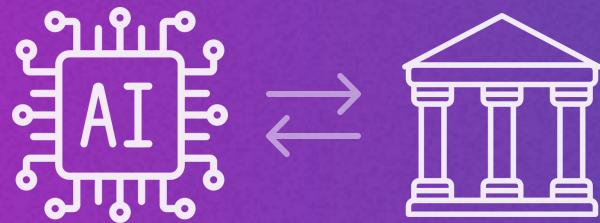
XΩΡΑ  
Greece

ΠΟΛΗ  
Ancient Corinth

Δείτε το στο: ASKSA

## Moving from Task-Specific to Cross-Domain: AI4Culture

Online capacity building hub for making AI technologies more accessible and usable by the CH sector



# Barriers to the uptake of AI in the Cultural Heritage sector

---



Lack and fragmentation of knowledge about how AI can be applied in the CH sector



Limitations of existing tools



Complexities of interconnecting AI tools with other digital components and with the Europeana platform



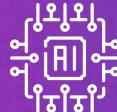
Lack of labelled datasets for domain adaptation



Difficulty to convincingly demonstrate ethical applications of AI with clear benefits

# Ai4culture

## AI Tools



Open source or free SaaS

- Data space-compliant
- Curation of CH-relevant AI tools
- Contributed by end users

## Open Datasets



Open labelled datasets for training and testing AI models



## Upskilling Material

Technical material (e.g. API documentation) and resources for professionals with lower IT skills (e.g. interviews, user tutorials)

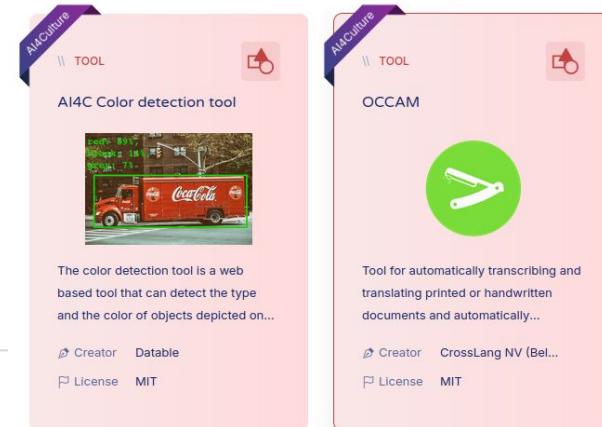


## Recipes

Step-by-step instructions about how to perform a complex task by combining different AI tools

# AI tools and datasets

- **Open source tools** (or offered as free services) relevant to CH domain
  - Documented with respect to various parameters (e.g., AI category, CH task type, required skills, intended users, etc) to facilitate discoverability and understandability
- **80 tools** in total
  - Tools developed by AI4Culture partners for scenarios with high interest for CHIs: offered as Docker containers and as online services; interoperable with Europeana
  - Existing tools curated by the consortium
  - Tools contributed by end users
- **15 datasets** that can be used for AI training and validation



# Recipes

- Step-by-step instructions about how to use and combine different tools to solve complex tasks



Resource Details Recipe Steps

1 Install the AI4C FBK Subtitler tool X

Description

You need the AI4C FBK Subtitler tool which can be used to generate the subtitles (in SRT format) of a given audio.

Clone the AI4C FBK Subtitler repo and follow the steps described in the README.md file to install and run the subtitler tool.

Once installed, the AI4C FBK Subtitler works as described in this [10 minute video](#). Note that the FBK Subtitler is a web service without a GUI: to observe the functionalities shown in the video, it needs the platform described in the step 2.

Resources Used

 // TOOL  
AI4C FBK-subtitler

2 Install the AI4C TLT Subtitle Manager tool X

Description

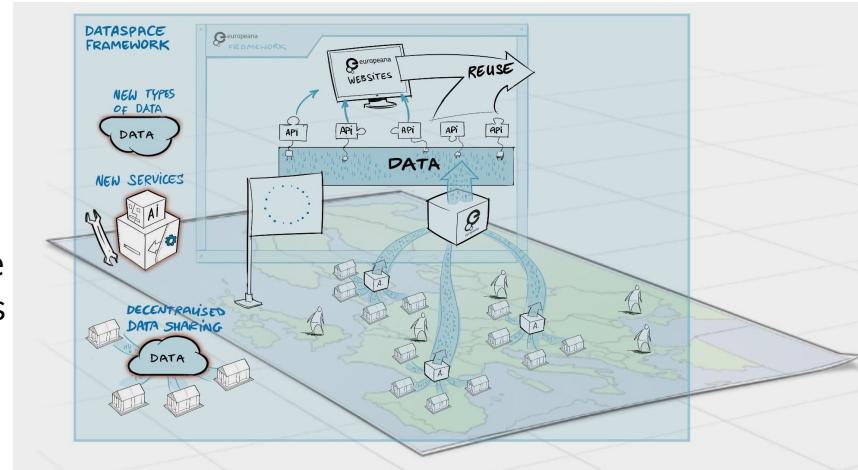
Clone the AI4C TLT Subtitle Manager repo

The AI4C TLT Subtitle Manager tool is based on three components (Frontend, Sub-Player and Backend), each implemented as a docker image.

The README.md file contains the instructions to download the three docker images (ai4europeana-frontend.tar ai4europeana-subplayer.tar and ai4europeana-backend.tar) and run them.

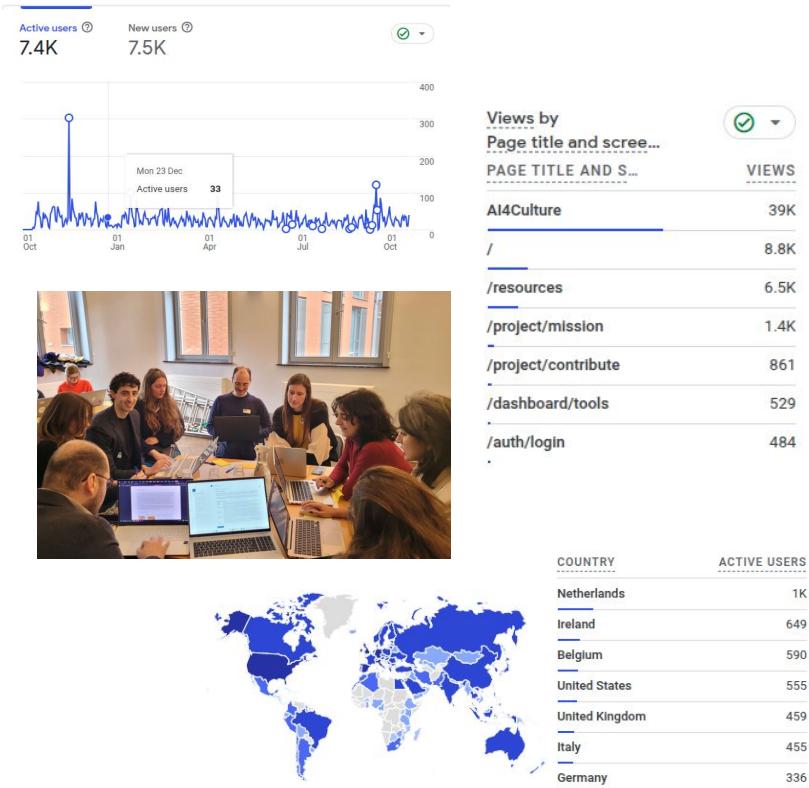
# Potential to support synergies and interoperability at the catalogue level

- The AI4Culture platform can act as an online reference point where stakeholders can register their resources (tools, datasets, recipes, success and failure stories etc)
  - Existing synergies with De-bias, Flemish AI Academy (VAIA), AI4LAM, and EuropeanaTech
- Interoperability
  - **Catalogue level:** use of standards for tools' and datasets' descriptions based on DCAT-AP with extensions to serve the sector needs
  - Data level: Mapping between different data models is use case specific and a big challenge
  - Technical: Loose interoperability based on APIs but also consider tighter integration, depending on use case



# Ongoing momentum

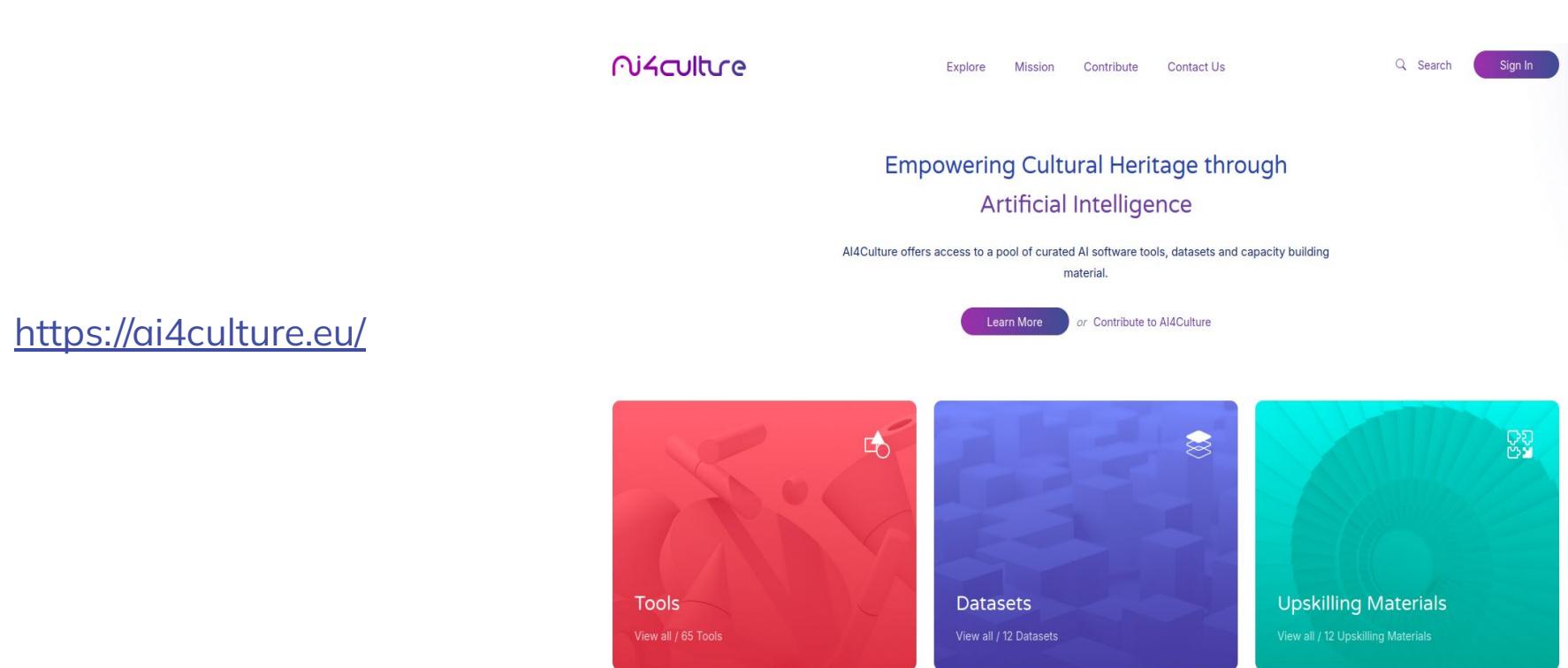
- >7.5K unique platform visitors
  - 3.9K after the project's completion (Apr 2025)
- > 120 registered resources
- > 1.2k followers on social media
- Upskilling activities attracted high interest (>470 attendees)
- 12 blog posts on Europeana Pro with >12K views
- 



# Capacity building

- Lessons learnt
  - High interest in gaining hands-on experience
  - CHIs often lack the necessary expertise and resources for adopting AI tools
  - AI tools cannot be applied as ready-to-use black boxes but need to be *adapted* to the needs of the sector and *integrated* with existing digital tools used by CHIs
  - Scepticism and legitimate concerns (e.g. about CH datasets being exploited by big tech with no real gain for the sectors) often hinders willingness to adopt AI technologies

# Let's explore the platform!



The image shows the homepage of the AI4Culture platform. At the top, there is a navigation bar with the AI4Culture logo, a search bar, and a 'Sign In' button. Below the navigation bar, the platform's mission is stated: "Empowering Cultural Heritage through Artificial Intelligence". A brief description follows: "AI4Culture offers access to a pool of curated AI software tools, datasets and capacity building material." There are three main sections: "Tools" (red background), "Datasets" (blue background), and "Upskilling Materials" (teal background). Each section has a 'Learn More' button and a 'View all' link. The "Tools" section shows a 3D model of a classical statue.

<https://ai4culture.eu/>

AI4culture

Explore Mission Contribute Contact Us

Search Sign In

Empowering Cultural Heritage through Artificial Intelligence

AI4Culture offers access to a pool of curated AI software tools, datasets and capacity building material.

Learn More or Contribute to AI4Culture

Tools

View all / 65 Tools

Datasets

View all / 12 Datasets

Upskilling Materials

View all / 12 Upskilling Materials

- Do you have an IT tool that is useful for the CH sector?
- Do you have an AI success (or failure) story to recount?
- Do you have a dataset that can be useful for in-domain training?
- Do you wish to experiment with AI to solve a problem?
- Do you wish to share your ideas and concerns about the application of AI in the CH sector?

Register on [ai4culture.eu](https://ai4culture.eu) to co-shape the AI4Culture platform!

---

