Detecting and Contextualizing Harmful Language in Cultural Heritage Collections

Orfeas Menis Mastromichalakis
National Technical University of Athens

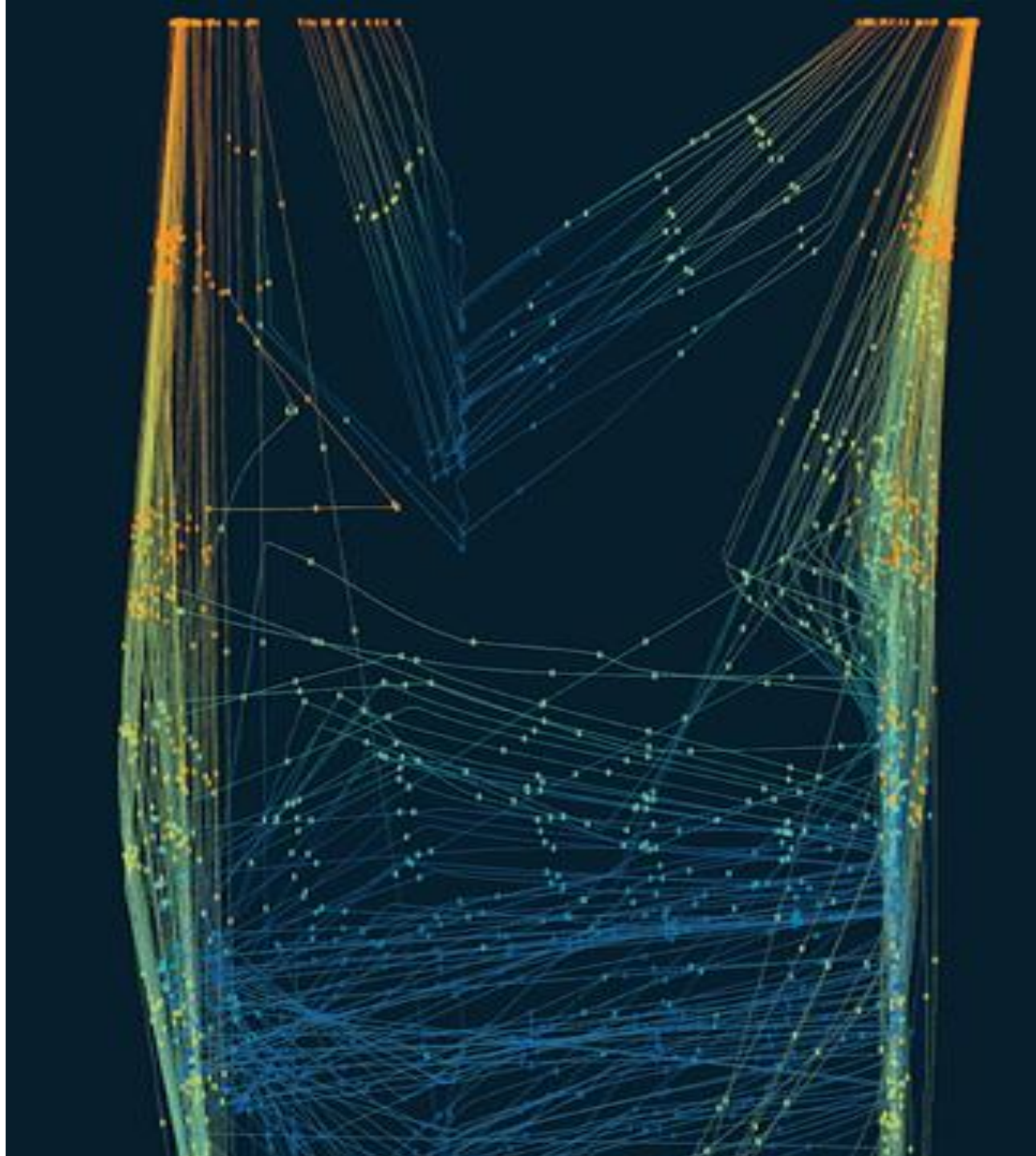
10 APRIL 2025

interoperable europe

1

Introduction

Introduce the DE-BIAS project, its main objectives and outcomes. Outline of our work, focusing on the 2 key points: the vocabulary, and the tool.



2

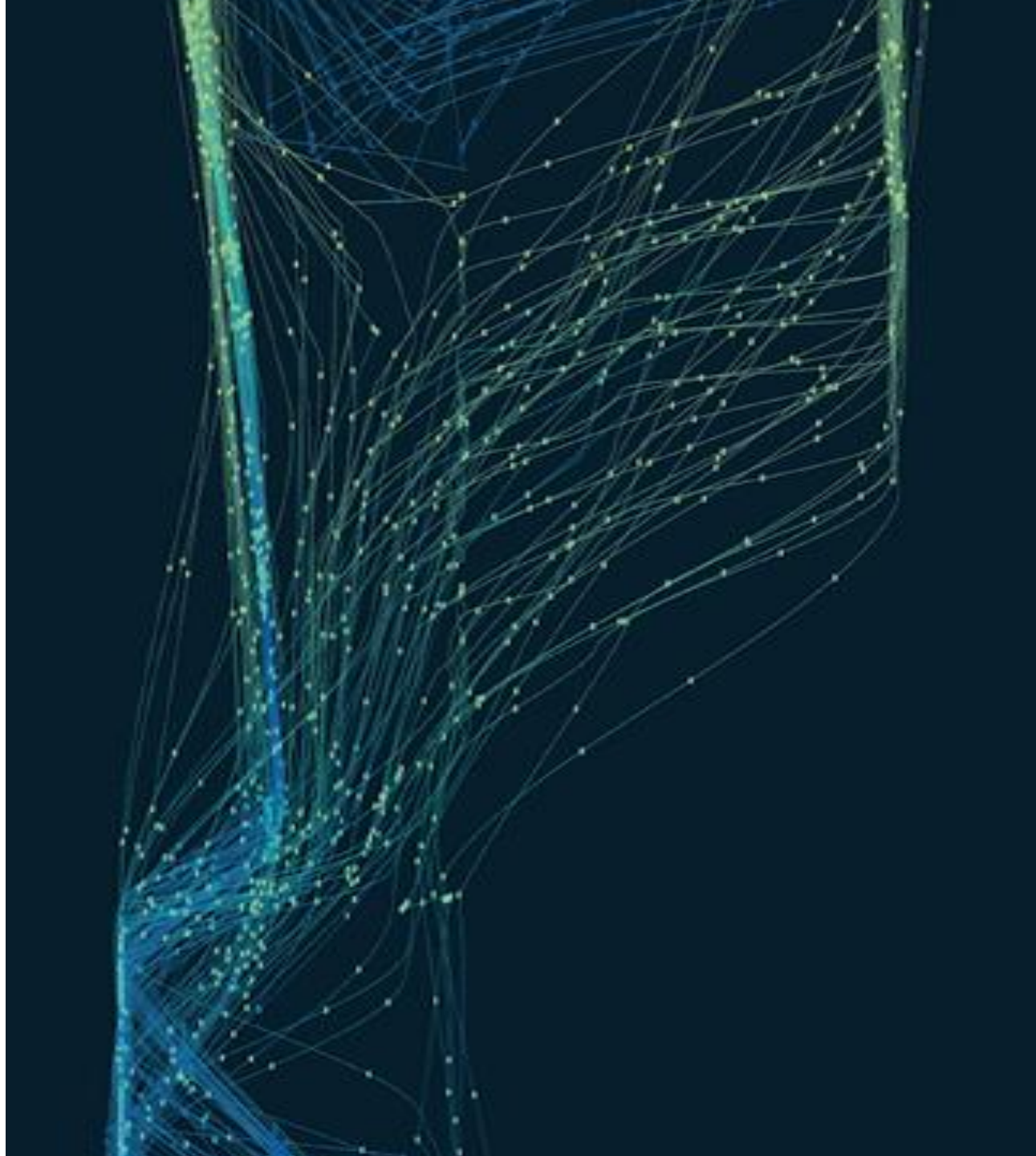
The Vocabulary

Briefly describe the vocabulary creation procedure, and the vocabulary structure.

3

The Tool

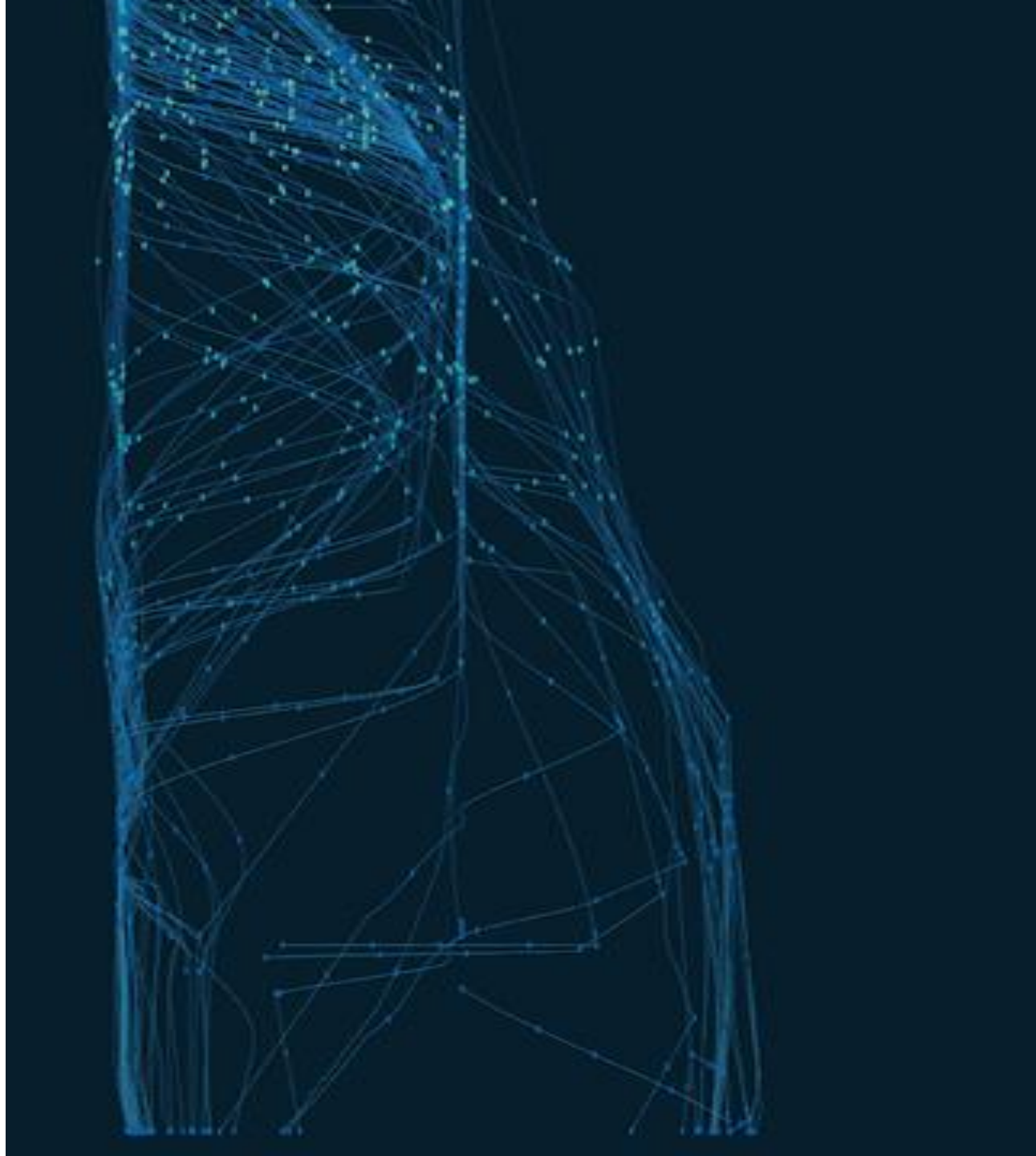
Describe the overall architecture and the technical details of the tool.



4

Conclusions

Sum-up our work, and discuss challenges, impact and future directions.



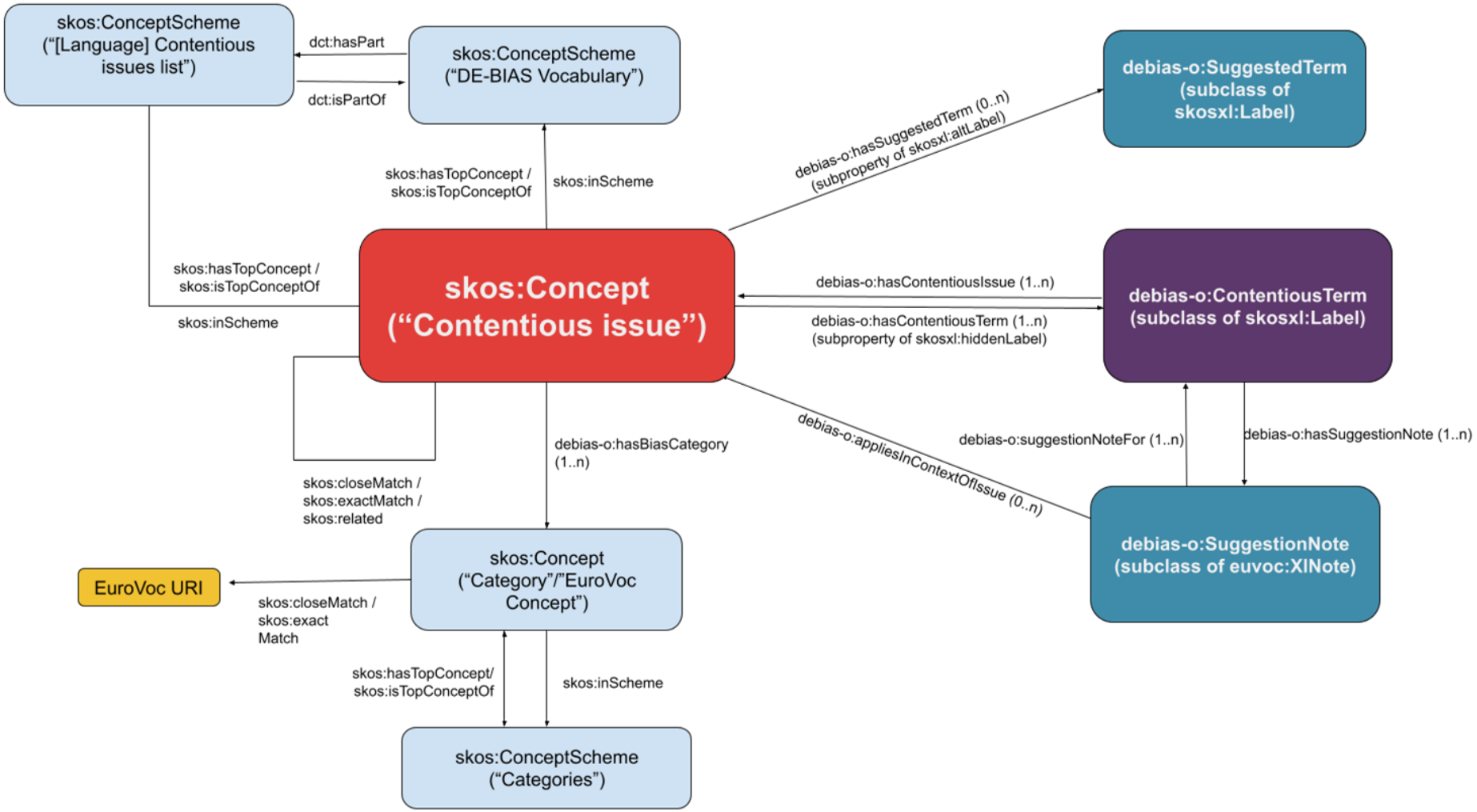
DE-BIAS: Detecting and cur(at)ing harmful language in cultural heritage collections

Aiming to promote a more inclusive and respectful approach to the description of digital collections and the telling of stories and histories of minoritised communities. DE-BIAS took a bottom-up approach, working with marginalised communities to improve the representation and participation of people who feel that museums and archives don’t speak of them or, worse, are not for them.

The DE-BIAS Vocabulary

- Enables automated bias detection and improves representation of marginalised narratives
- Covers 5 languages: English, German, French, Italian, and Dutch
- Almost 700 terms
- Major topics: migration, colonial history, gender, sexual identity, and ethnicity
- Co-created with marginalized communities and supported by academic research
- Incorporating revised and updated pre-existing vocabularies while clearly indicating sources to enhance the visibility of other initiatives' work
- Published in EU vocabularies

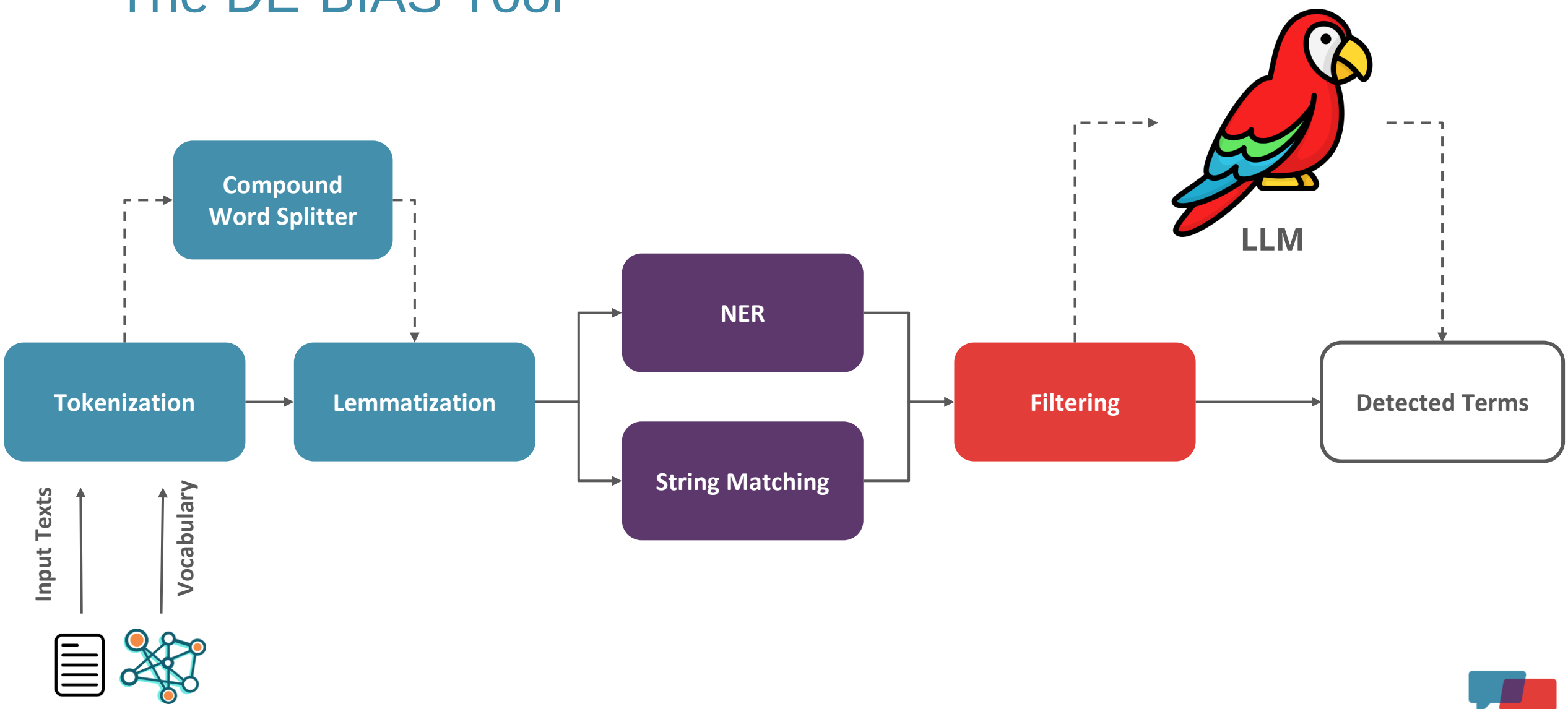
The DE-BIAS Vocabulary



The DE-BIAS Tool

- Hybrid system utilizing the DE-BIAS vocabulary to detect harmful language, combining traditional NLP techniques (lemmatization, string matching, named entity recognition) with state-of-the-art LLMs for disambiguation.
- Integrated into the Europeana Core Service Platform, processing 7.9 million records
- Open API allows independent adoption by external organizations
- Stand-alone web interface for accessibility by less tech-savvy users

The DE-BIAS Tool



LLM	NER	Precision↑	Throughput↑
✗	✗	0.70	15112
✗	✓	0.71	15092
✓	✗	0.86	787
✓	✓	0.87	813

Table 1: The effect of NER and LLM on performance.

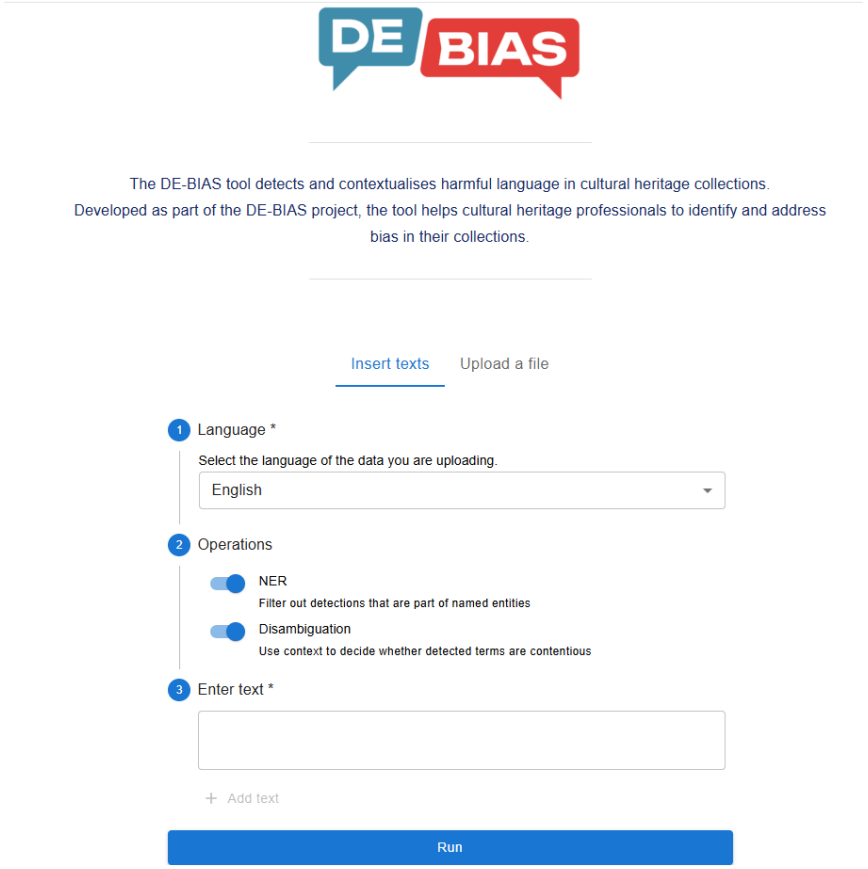
Model	Precision ↑					
	nl	en	fr	de	it	Aggregated
EuroLLM-9B	0.42	0.74	0.97	0.83	0.97	0.73
Llama-3.1-8B-Q8	0.76	0.83	0.97	0.88	0.97	0.88
Ministral-8B	0.73	0.89	0.97	0.87	0.97	0.88
Mistral-Small-24B-Q8	0.66	0.89	0.97	0.87	0.98	0.86
Mixtral-8x7B-Q3	0.72	0.83	0.97	0.89	0.98	0.87
StableLM-2-1.6B	0.60	0.72	0.96	0.82	0.95	0.80
StableLM-2-12B	0.72	0.82	0.96	0.87	0.97	0.86

Table 2: Precision of the tool, utilizing different LLMs

The DE-BIAS (standalone) tool webpage

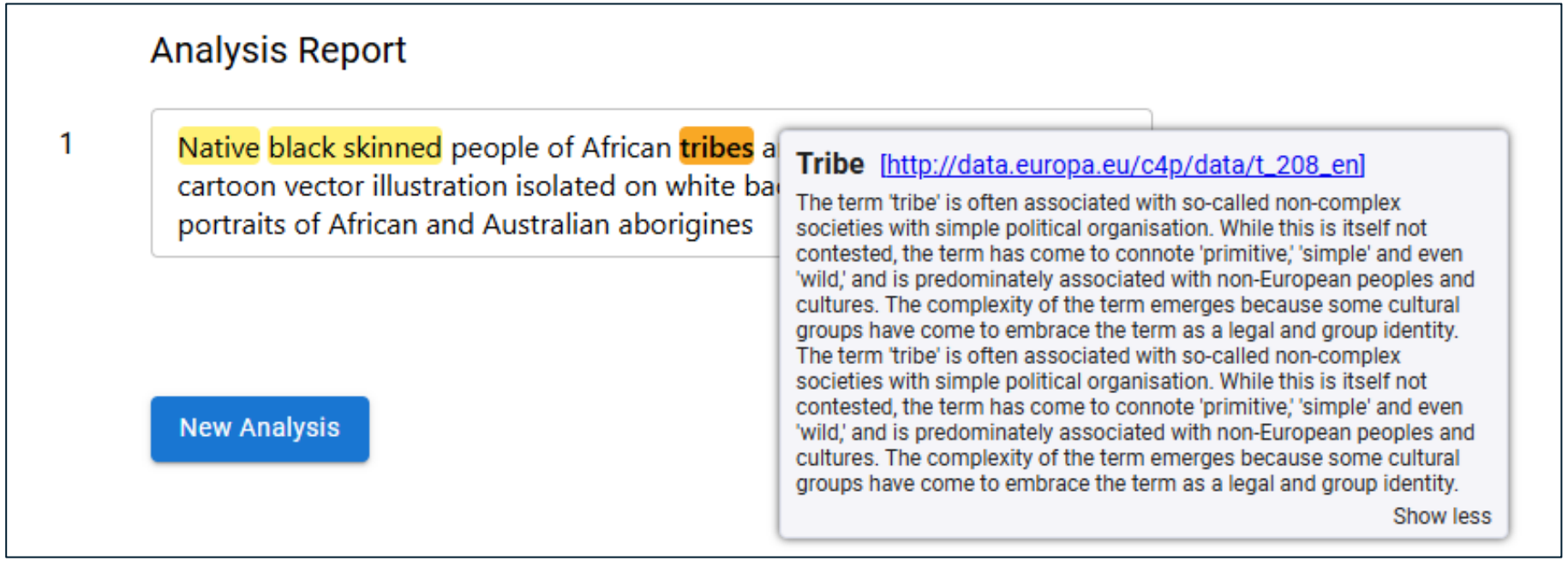
Intuitive and user-friendly interface for bias detection and analysis offering 2 functionalities:

- **Custom Input Testing**: Users can input plain text directly into the web app to test for bias and instantly view the detection results.
- **Batch Processing via File Upload**: Upload a ZIP file containing multiple text files for bias detection, and receive an email with a JSON containing all the detections, along with a report with a brief statistical analysis summarising the results.



https://debias-tool.ails.ece.ntua.gr

The DE-BIAS (standalone)



europeana

HOME COLLECTIONS STORIES SHARE YOUR COLLECTIONS LOG IN/JOIN

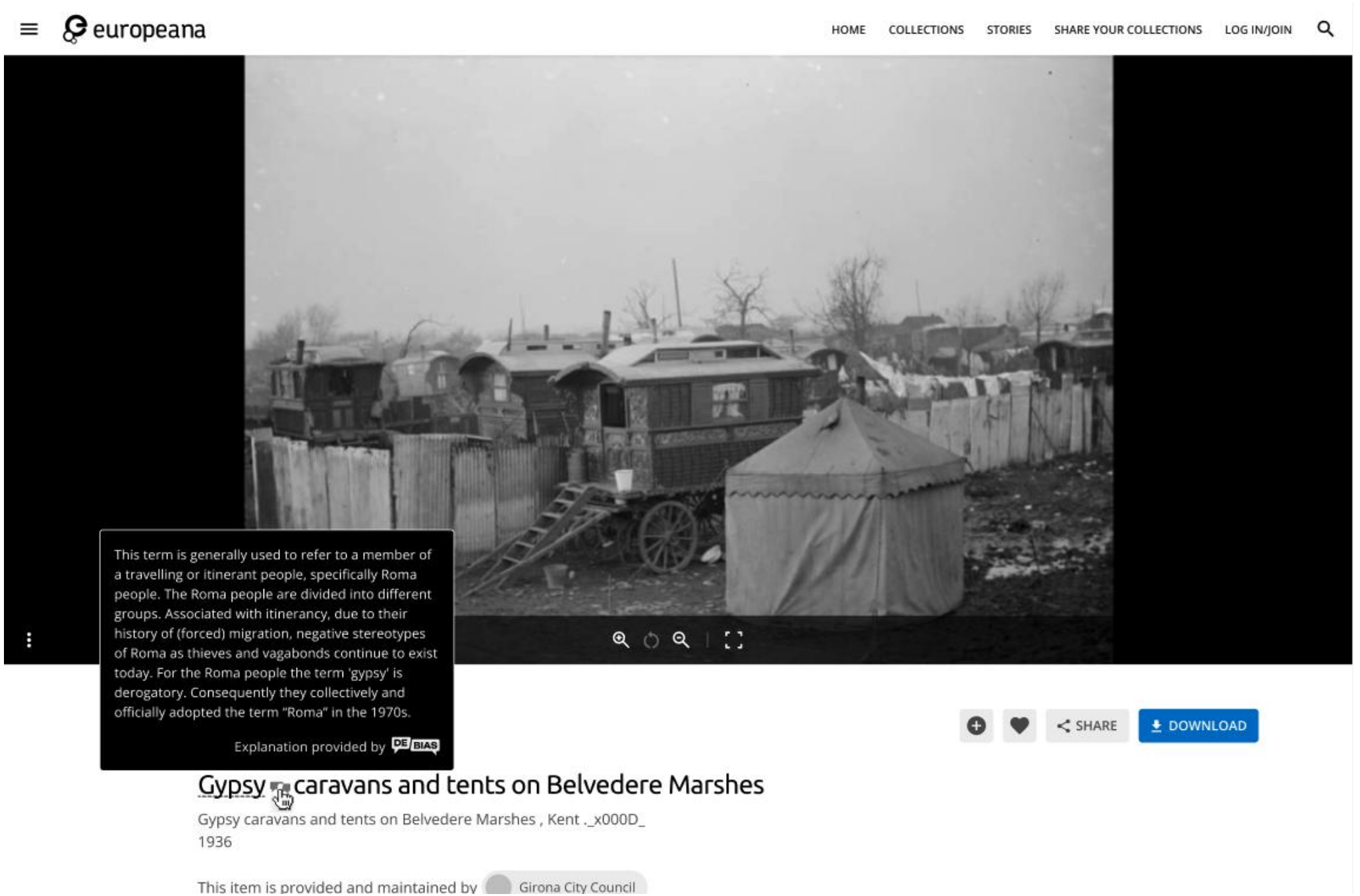
Public Domain

SHARE DOWNLOAD

Gypsy caravans and tents on Belvedere Marshes

Gypsy caravans and tents on Belvedere Marshes , Kent ._x000D_
1936

This item is provided and maintained by Girona City Council


europeana
HOME
COLLECTIONS
STORIES
SHARE YOUR COLLECTIONS
LOG IN/JOIN
This term is generally used to refer to a member of a travelling or itinerant people, specifically Roma people. The Roma people are divided into different groups. Associated with itinerancy, due to their history of (forced) migration, negative stereotypes of Roma as thieves and vagabonds continue to exist today. For the Roma people the term 'gypsy' is derogatory. Consequently they collectively and officially adopted the term "Roma" in the 1970s.
Explanation provided by DE BIAS
SHARE
DOWNLOAD
Gypsy caravans and tents on Belvedere Marshes
Gypsy caravans and tents on Belvedere Marshes , Kent ._x000D_
1936
This item is provided and maintained by
Girona City Council

Challenges and lessons learned

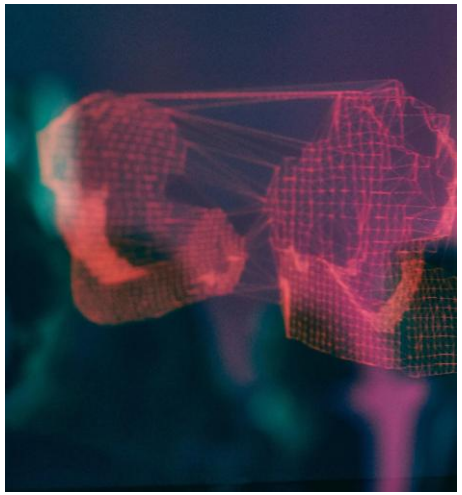
Multilinguality

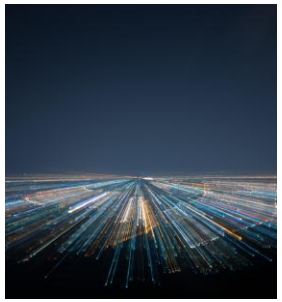
- There is no consistent behavior of the libraries and LLMs among different languages.
- Certain problems may require language-specific solutions, like compound word splitting for German and Dutch.
- Difficulty to assess the results without knowledge of the language. Contribution of native speakers is crucial.

Working with LLMs

- Controlling the output format is not always easy, especially with smaller LLMs, making parsing the results challenging. There are tools and techniques that can help.
- Incorrect reasoning / hallucinations can sometimes lead to a correct decision, especially if the task is a Yes/No question.

Broader Applicability and Future Steps



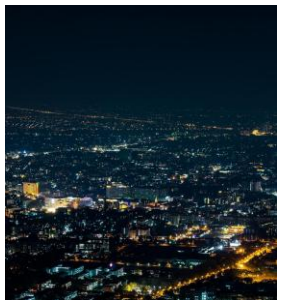


Accessibility

By annotating/enriching metadata with non-contentious alternatives we can improve searchability of the CH data as it is rather rare to use contentious terms in our text queries.

Beyond Cultural Heritage

Expand to other domains, beyond cultural heritage. Currently exploring Political Toxic Speech.

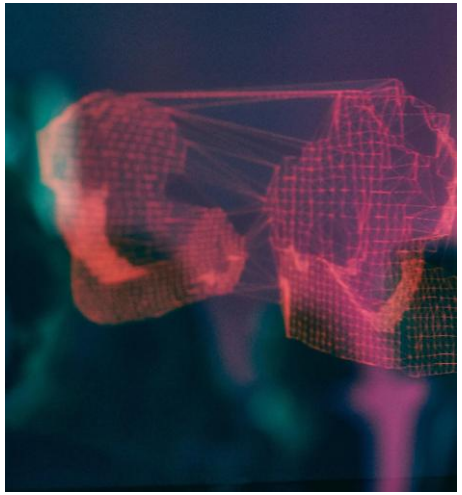


Beyond Bias Detection

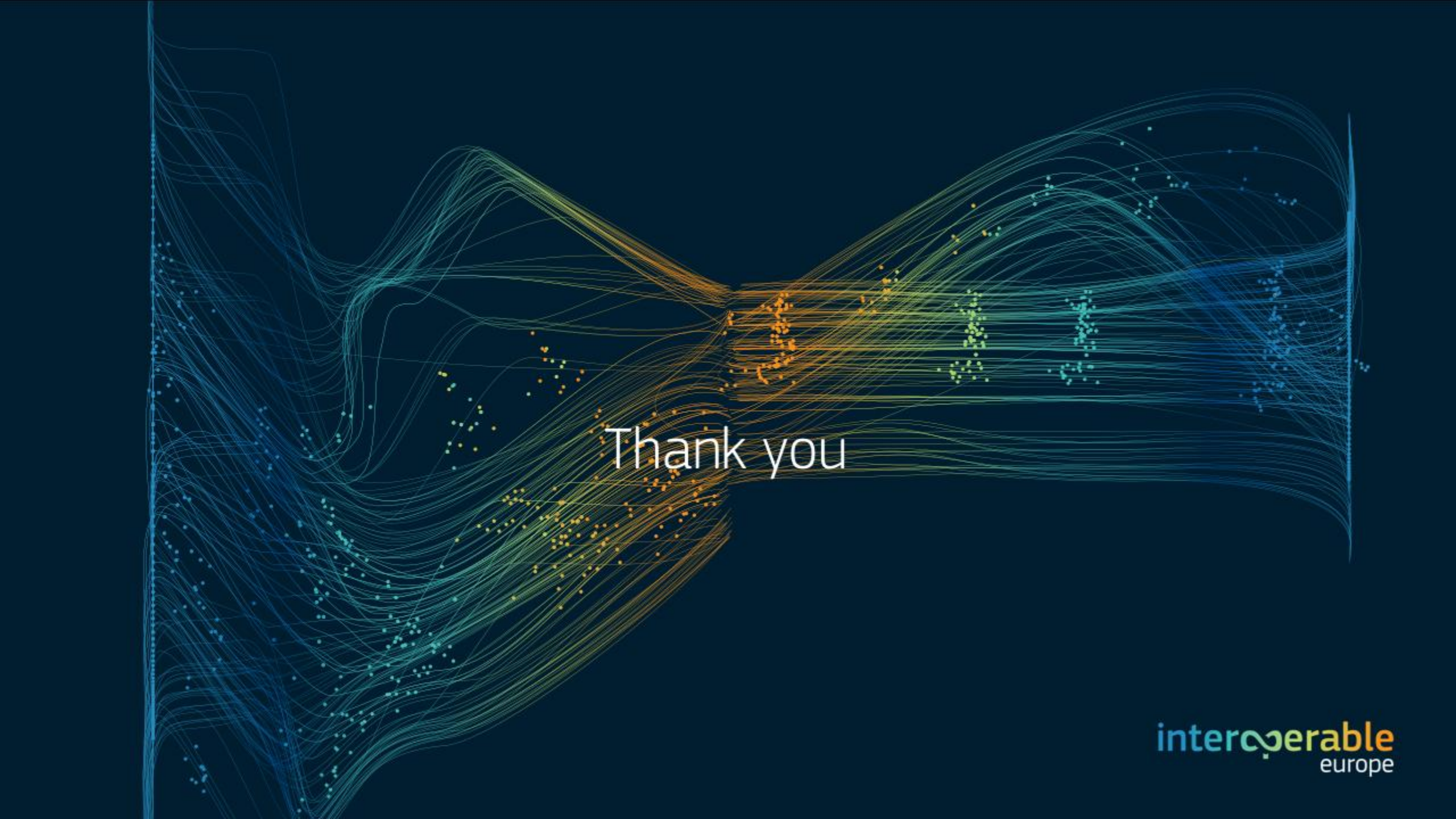
Make the tool vocabulary-agnostic, allowing for a broader application.

Explore greater LLM utilization
Explore zero-shot and few-shot setups, as well as RAG-based approaches to data auto-tagging.

A Modular and Accurate Data Enrichment and Auto-tagging System



A hybrid system that combines the control and transparency of semantics, Knowledge Graphs, and traditional NLP techniques, with the power of LLMs to provide a modular, scalable solution to data enrichment and auto-tagging.



interoperable europe