DE BIAS

Detecting and Cur(at)ing Harmful Language in Cultural Heritage Collections

Worldwide wall - Associazione Culturale GoTellGo, Italy - CC BY

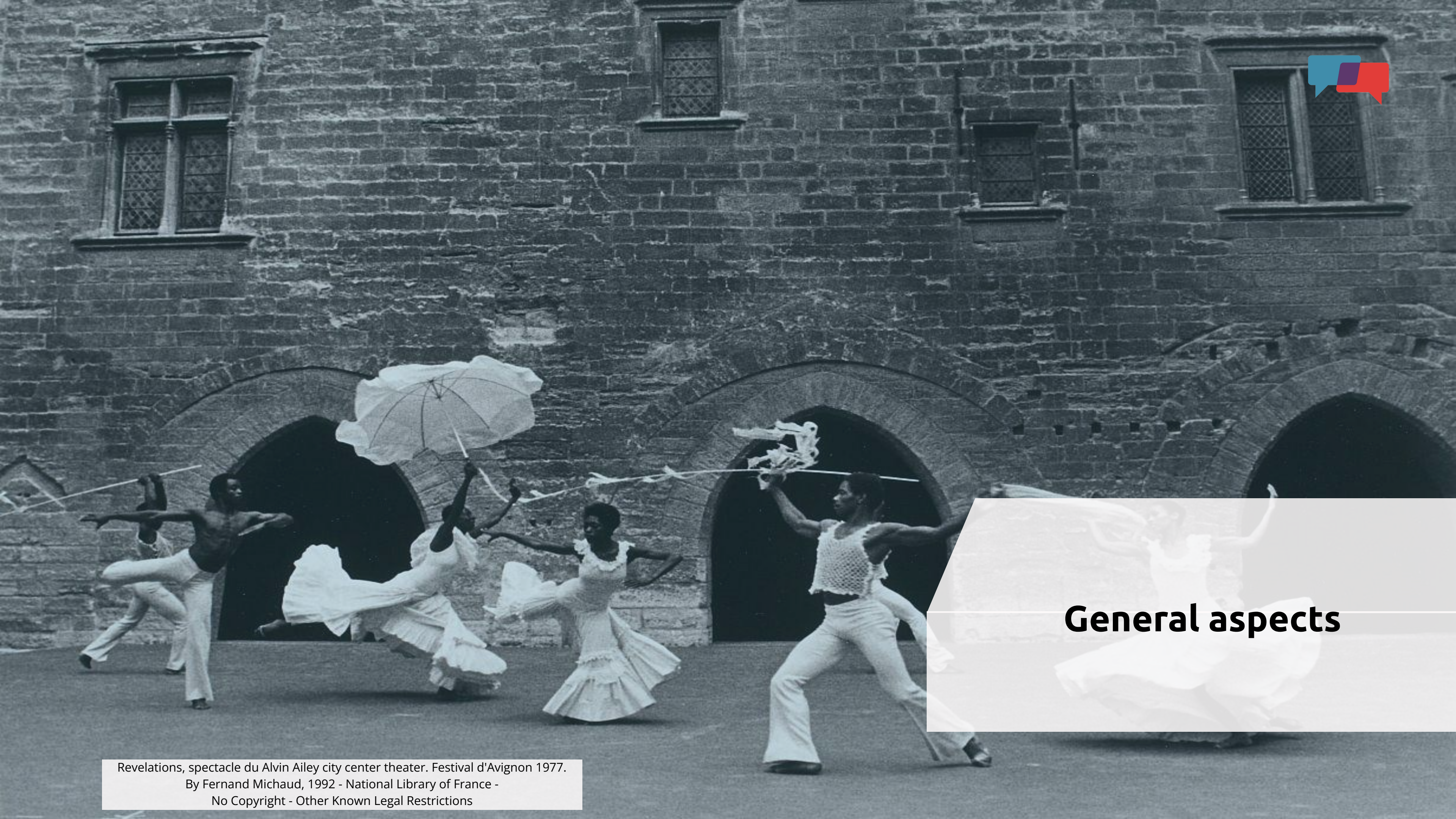
Co-funded by the European Union

The DE-BIAS Tool

Worldwide wall - Associazione Culturale GoTellGo, Italy - CC BY

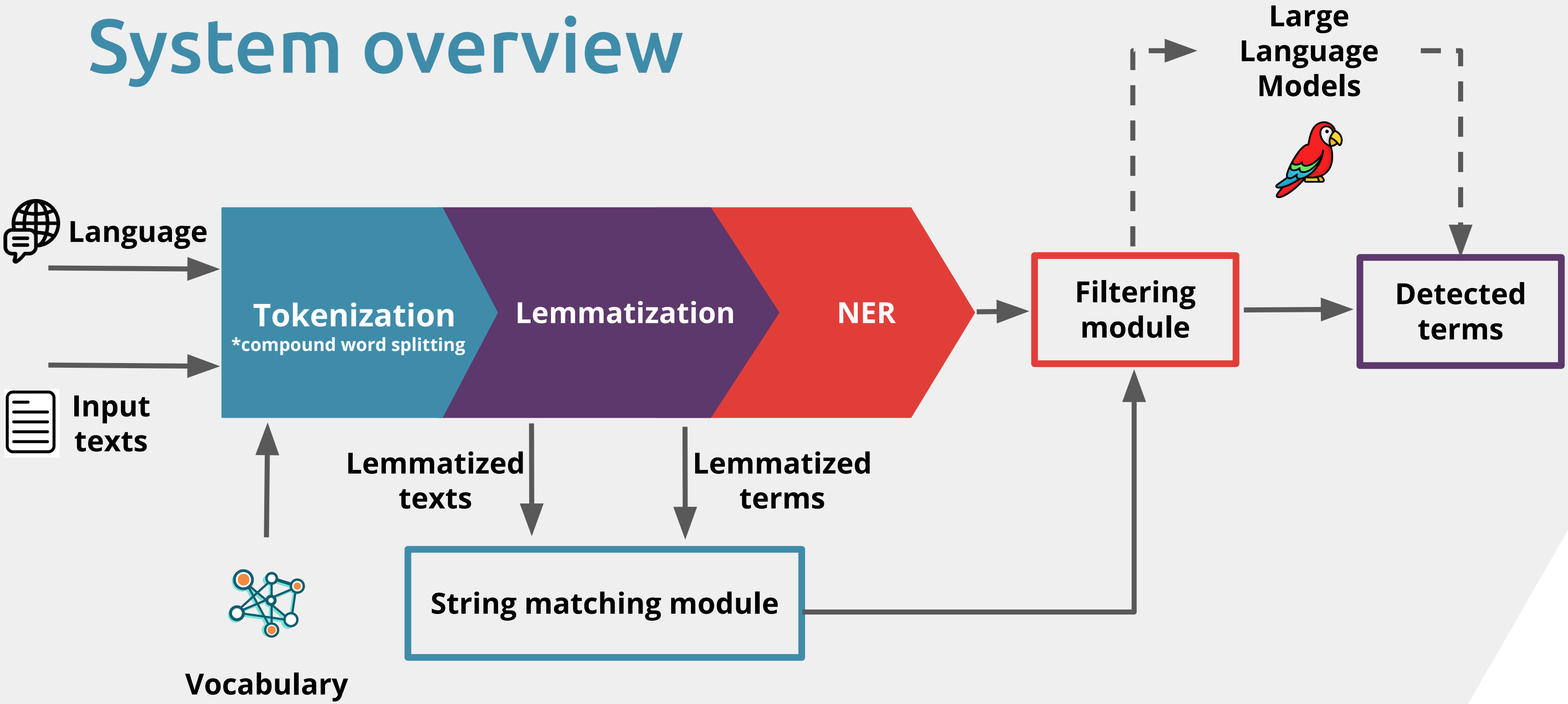
Co-funded by
the European Union

General aspects



Revelations, spectacle du Alvin Ailey city center theater. Festival d'Avignon 1977.
By Fernand Michaud, 1992 - National Library of France -
No Copyright - Other Known Legal Restrictions

System overview



Language

Input texts

Tokenization

*compound word splitting

Lemmatization

NER

Filtering module

Large Language Models

Detected terms

Lemmatized texts

Lemmatized terms

String matching module

Vocabulary

Processing steps

Tokenization → The plain text is split into words, phrases or other meaningful elements for analysis.

Lemmatization → The lemma of a word is identified. This enables matching a word from the vocabulary to a word in the analysed text, including different inflections.

NER = Named Entity Recognition → A word’s context and positioning within the text is used to determine whether a term qualifies as a proper noun (e.g. a person's name or a place name), which may then be excluded from the tagging process.

LLM = Large Language Model → Used to determine whether a term that can be contentious in one context, but appropriate in another is to be flagged.

The DE-BIAS Tool

- **Applies the DE-BIAS vocabulary** with contextual information and alternatives as a foundation.
- **Works language-specific** → Potential errors may occur if metadata is not labeled or mislabeled.
- Applies **lemmatization** → Words are reduced to their base or root form so that different forms of the same word are recognised.
- Can detect whether a problematic term is part of a proper name via **Named Entity Recognition**.
- Partially recognises ambiguous terms (e.g., "exotic", “dwarf”) by applying the contentious issue descriptions of the vocabulary and exploiting **Large Language Models**.

The tool's output

Detection and contextualisation

- The tool identifies potentially problematic language used in the metadata of heritage collections
- It enables the highlighting of such language by annotating the analysed metadata
  - Marks the contentious term, the metadata field it was found in, and its specific position within the text (helpful especially with longer descriptive data)
  - Makes the original context of words understandable to users and suggests (if applicable) possible alternatives

Example 1 - Terms detected in text

Indianerkapelle, Markgraf Christian von Brandenburg als **Indianer**, begleitet von Herolden. 4. Part, Blatt 16 aus: Abriß und Verzeichnis aller Inventionen und Aufzüge, welche an Fastnachten Anno 1609 im kurfürstl. Schlosse zu Dresden aufgeführt wurden. Deckfarben; ca. 289 x 527 mm. Dresden: SLUB Mscr.Dresd.J.18

(from https://www.europeana.eu/en/item/188/item_4NA4VY56ATII67YP2RLNT5WQN2SZIVGC)

The terms "Indianerkapelle" and "Indianer" were flagged by the tool as contentious.

Example 2 - Detection rejected by NER

The Queer Liberation March is an annual LGBTQAI+ protest march in Manhattan, organised by the Reclaim Pride Coalition as an anti-corporate alternative to the NYC Pride March.

The term "Queer" was detected in the text, but because the term is part of the "Queer Liberation March" named entity, it was not flagged as contentious.

Example 3 - Disambiguation with the context-aware module (LLM)

Prompt: The term queer can have a contentious meaning when used in some contexts. It is contentious when used in a derogatory manner to refer to a person's sexuality but not when used as an umbrella term for sexual interests and identities that challenge social norms for sexual behaviour.

Is the term used in a contentious manner in the following text, yes or no?

Queer studies are not the same as queer theory, which is an analytical viewpoint within queer studies (centered on literary studies and philosophy) that challenges the existence of "socially constructed" categories of sexual identity.

Example 3 - Disambiguation with the context-aware module (LLM)

Response:

No

Because the term queer has both a contentious and a non-contentious context, we use an LLM guided by the information provided by the vocabulary (in the prompt) to detect whether the use of the term in the provided text (red) is derogatory or not. In this example, the term is used when talking about different academic research areas.

Some more technical background

- Combination of the NLP tools provided by the Stanford Stanza Python library and a locally deployed LLM
  - Pipeline of various Natural Language Processing (NLP) tools, models, and algorithms, including tokenization, multi-word token (MWT) expansion, lemmatization, part-of-speech (POS) and morphological features tagging, dependency parsing, and NER
  - Covers all five of the DE-BIAS project languages with high performance levels of 95% accuracy and above (based on a variety of research data sets used for training)
- Adaptations specific for DE-BIAS, e.g. inclusion of compound words

Some more technical background

- For the LLM, DE-BIAS decided for the application of an Autoregressive LLM (AR LLM)
  - Allowing for some form of “creativity” in analysing texts
  - Better in “understanding” language, producing text that respects common facts about the world and contextual meaning
  - Have shown increased knowledge acquisition capabilities and have been used successfully for tasks other than text generation without having to adapt them
- Tested various models from the Mistral AI and Stable LM families
  - Decided for the Mixtral 8x7b Instruct Q3_K_L based on our quality requirements and available computing capacities

Availability of the tool

API Endpoint

DE-BIAS Platform

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

Use cases for the tool

- **Web interface**: fast checking of values
- **Data upload**: bulk checking of values
- **Custom API integration**: repeated checking of large amounts of data

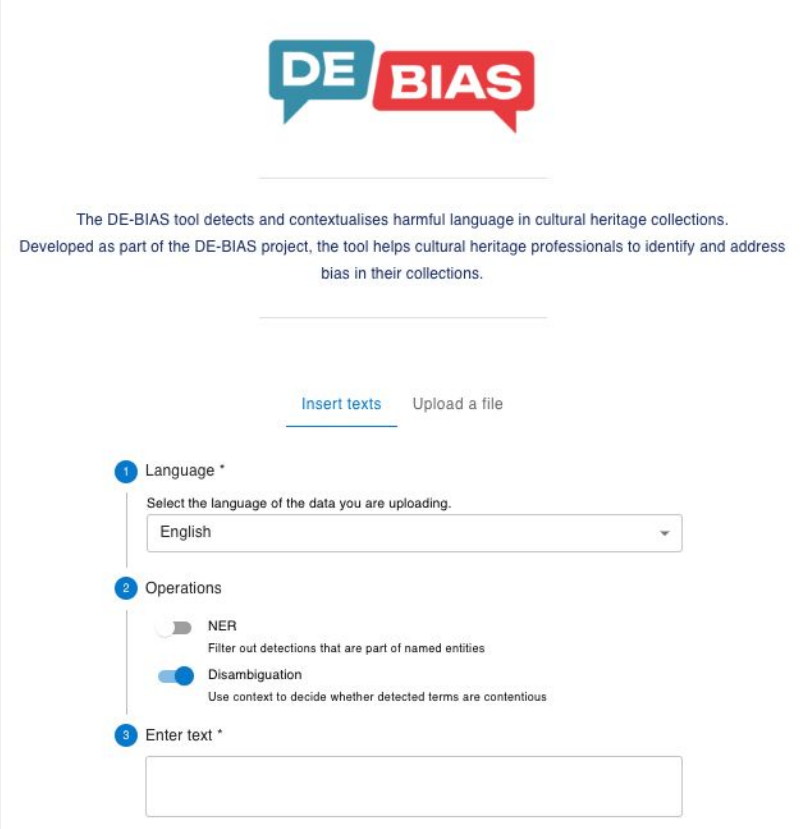
Documentation

https://pro.europeana.eu/files/Europeana_Professional/Projects/debias/DE-BIAS_D3.2_ReportOnTheFunctionalityOfTheTool.pdf

The standalone version

Revelations, spectacle du Alvin Ailey city center theater. Festival d'Avignon 1977.
By Fernand Michaud, 1992 - National Library of France -
No Copyright - Other Known Legal Restrictions

Option 1: Insert texts



The DE-BIAS tool detects and contextualises harmful language in cultural heritage collections. Developed as part of the DE-BIAS project, the tool helps cultural heritage professionals to identify and address bias in their collections.

1. Language *

Select the language of the data you are uploading.

English

2. Operations

NER

Filter out detections that are part of named entities

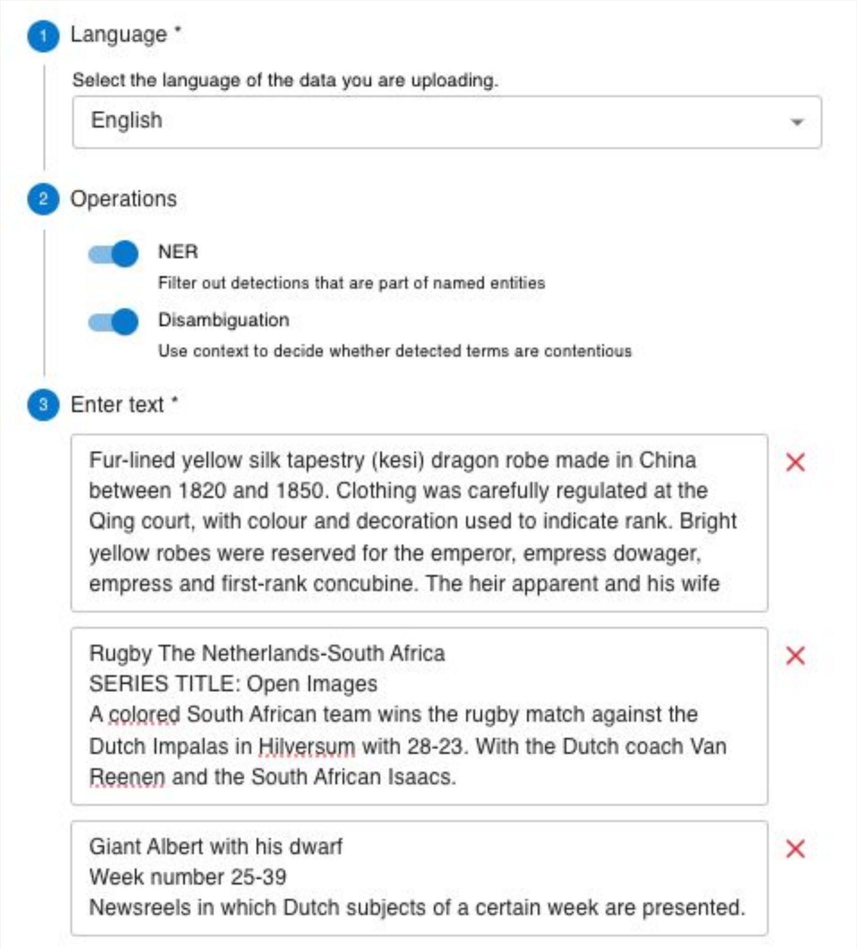
Disambiguation

Use context to decide whether detected terms are contentious

3. Enter text *

- Select the language of your text (Dutch, English, French, German, Italian)
- Decide whether to apply Named Entity Recognition and/or Disambiguation
- Enter your text (type or copy & paste)
- Add one or several texts and have them analysed together

Example analysis of several texts in English



1 Language *

Select the language of the data you are uploading.

English

2 Operations

● NER

Filter out detections that are part of named entities

● Disambiguation

Use context to decide whether detected terms are contentious

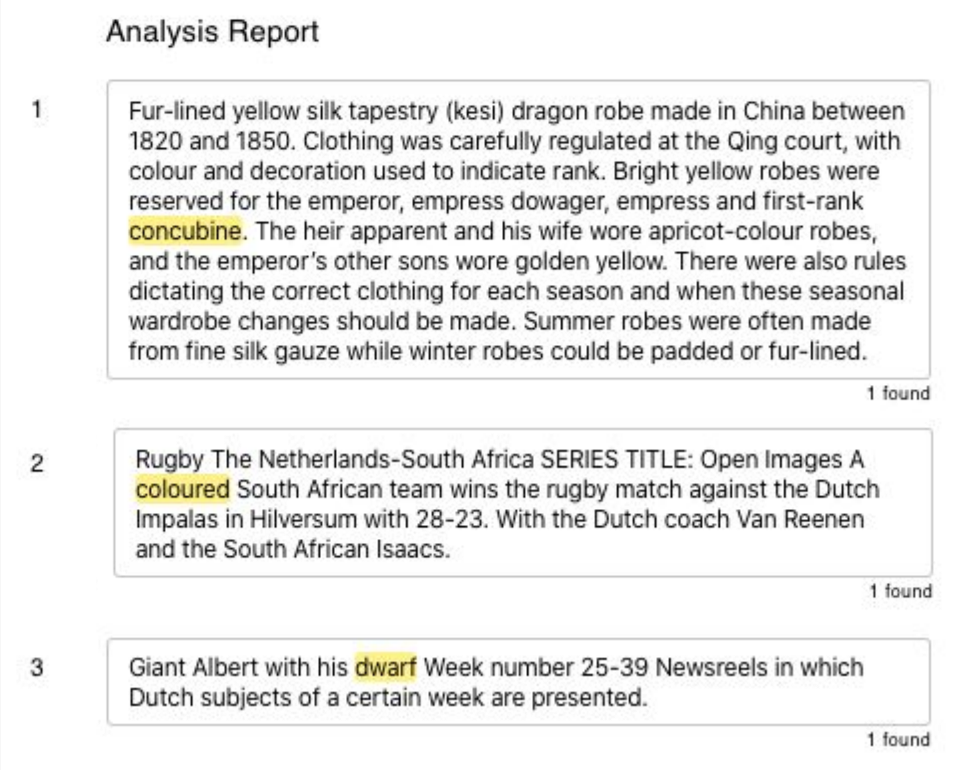
3 Enter text *

Fur-lined yellow silk tapestry (kesi) dragon robe made in China between 1820 and 1850. Clothing was carefully regulated at the Qing court, with colour and decoration used to indicate rank. Bright yellow robes were reserved for the emperor, empress dowager, empress and first-rank concubine. The heir apparent and his wife

Rugby The Netherlands-South Africa
SERIES TITLE: Open Images
A colored South African team wins the rugby match against the Dutch Impalas in Hilversum with 28-23. With the Dutch coach Van Reenen and the South African Isaacs.

Giant Albert with his dwarf
Week number 25-39
Newsreels in which Dutch subjects of a certain week are presented.

What does the result look like?



Analysis Report

1 Fur-lined yellow silk tapestry (kesi) dragon robe made in China between 1820 and 1850. Clothing was carefully regulated at the Qing court, with colour and decoration used to indicate rank. Bright yellow robes were reserved for the emperor, empress dowager, empress and first-rank concubine. The heir apparent and his wife wore apricot-colour robes, and the emperor's other sons wore golden yellow. There were also rules dictating the correct clothing for each season and when these seasonal wardrobe changes should be made. Summer robes were often made from fine silk gauze while winter robes could be padded or fur-lined.

1 found

2 Rugby The Netherlands-South Africa SERIES TITLE: Open Images A coloured South African team wins the rugby match against the Dutch Impalas in Hilversum with 28-23. With the Dutch coach Van Reenen and the South African Isaacs.

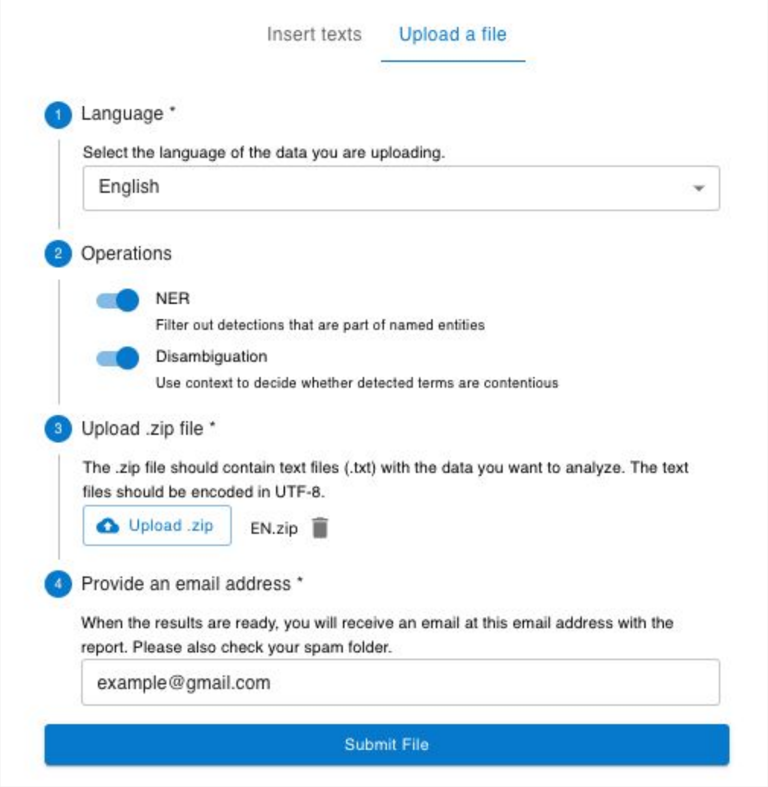
1 found

3 Giant Albert with his dwarf Week number 25-39 Newsreels in which Dutch subjects of a certain week are presented.

1 found

- Detected terms are highlighted
- Highlights link to the term in the DE-BIAS Vocabulary from where all further information can be accessed:
  - Contentious issue description
  - Alternative terms
  - Suggestions on language to use instead
  - Etc.

Option 2: Upload a file



1 Language *

Select the language of the data you are uploading.

English

2 Operations

NER: on

Filter out detections that are part of named entities

Disambiguation: on

Use context to decide whether detected terms are contentious

3 Upload .zip file *

The .zip file should contain text files (.txt) with the data you want to analyze. The text files should be encoded in UTF-8.

Upload .zip EN.zip

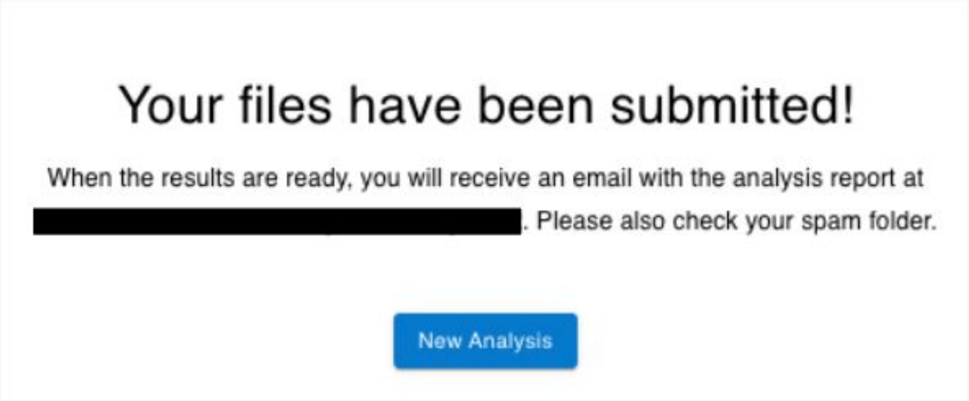
4 Provide an email address *

When the results are ready, you will receive an email at this email address with the report. Please also check your spam folder.

example@gmail.com

- Select the language of your text (Dutch, English, French, German, Italian)
- Decide whether to apply Named Entity Recognition and/or Disambiguation
- Attach ZIP file with the files to be analysed (stick with .txt files; make sure, their encoding is UTF-8)
- Add your email address

What does the result look like?



Your files have been submitted!

When the results are ready, you will receive an email with the analysis report at [redacted]. Please also check your spam folder.

- Email with summary report in PDF and annotations of detected bias terms in JSON
- The JSON file includes the URI of the detected terms in the DE-BIAS Vocabulary, via which more information can be retrieved
- It will also indicates where in the text a term has been detected (start character, end character, length)

How to prepare data for upload

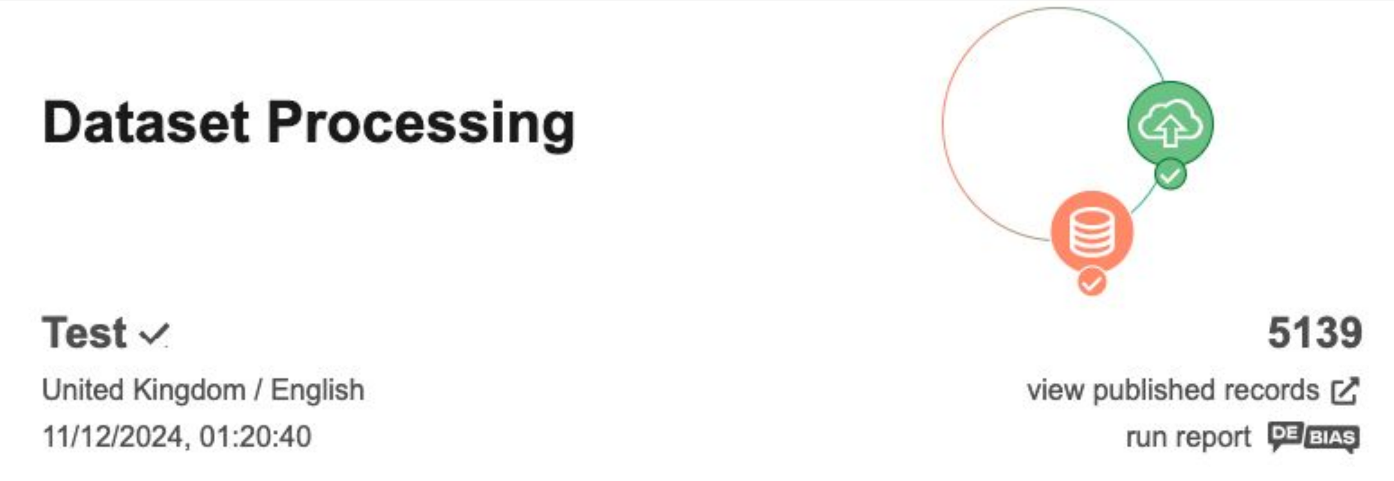
- Each record should be stored in a separate .txt file
  - The tool will treat each .txt file as one value to be analysed
  - To each value, the tool will the associate the annotations for detected biassed terms and reference the file name in the output
- All .txt files to be analysed - even if it is only one - need to be provided to the tool in .zip format
- Since the **language of the metadata is selected once per analysis for all records**, a .zip file should only contain records with metadata in one language
  - Prepare separate .zip files if you wish to analyse metadata in different languages

The DE-BIAS Tool in the Metis Sandbox

Revelations, spectacle du Alvin Ailey city center theater. Festival d'Avignon 1977.
By Fernand Michaud, 1992 - National Library of France -
No Copyright - Other Known Legal Restrictions

Preparing data for Europeana.eu

- Upload a dataset to the Metis Sandbox for data quality and other checks
- Follow the usual processing steps
- Once the dataset has been processed, the option to "Run report DE-BIAS" becomes available



Dataset Processing

Test ✓

United Kingdom / English

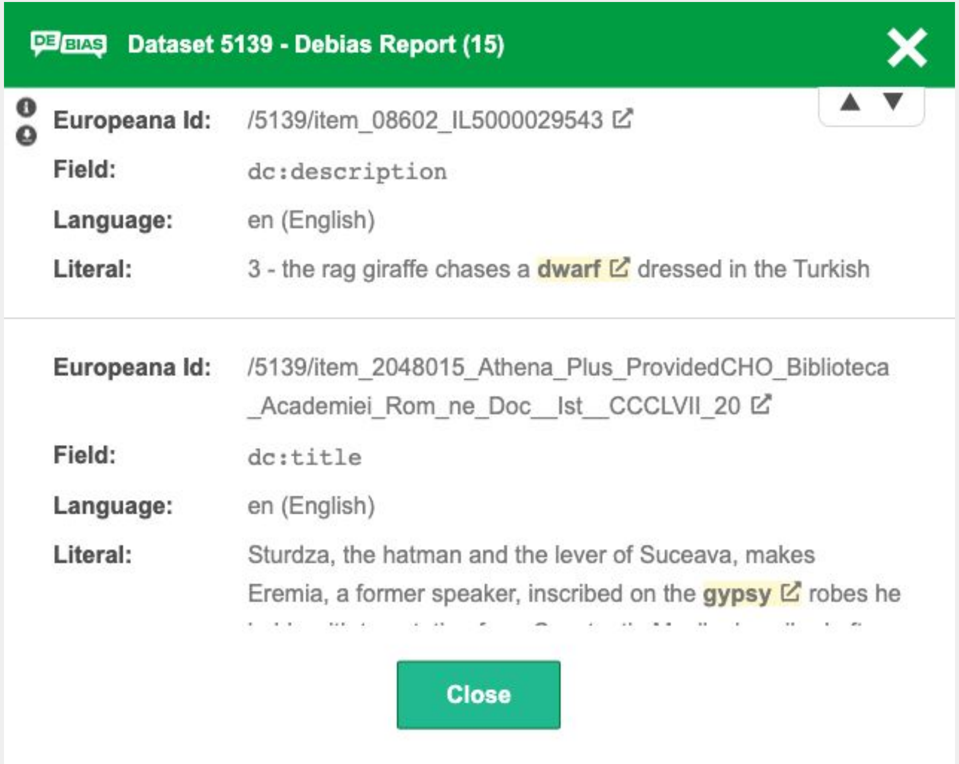
11/12/2024, 01:20:40

5139

view published records

run report DE BIAS

What does the result look like?



Dataset 5139 - Debias Report (15)

Europeana Id: /5139/item_08602_IL5000029543

Field: `dc:description`

Language: en (English)

Literal: 3 - the rag giraffe chases a **dwarf** dressed in the Turkish

Europeana Id: /5139/item_2048015_Athena_Plus_ProvidedCHO_Biblioteca_Academiei_Rom_ne_Doc__Ist__CCCLVII_20

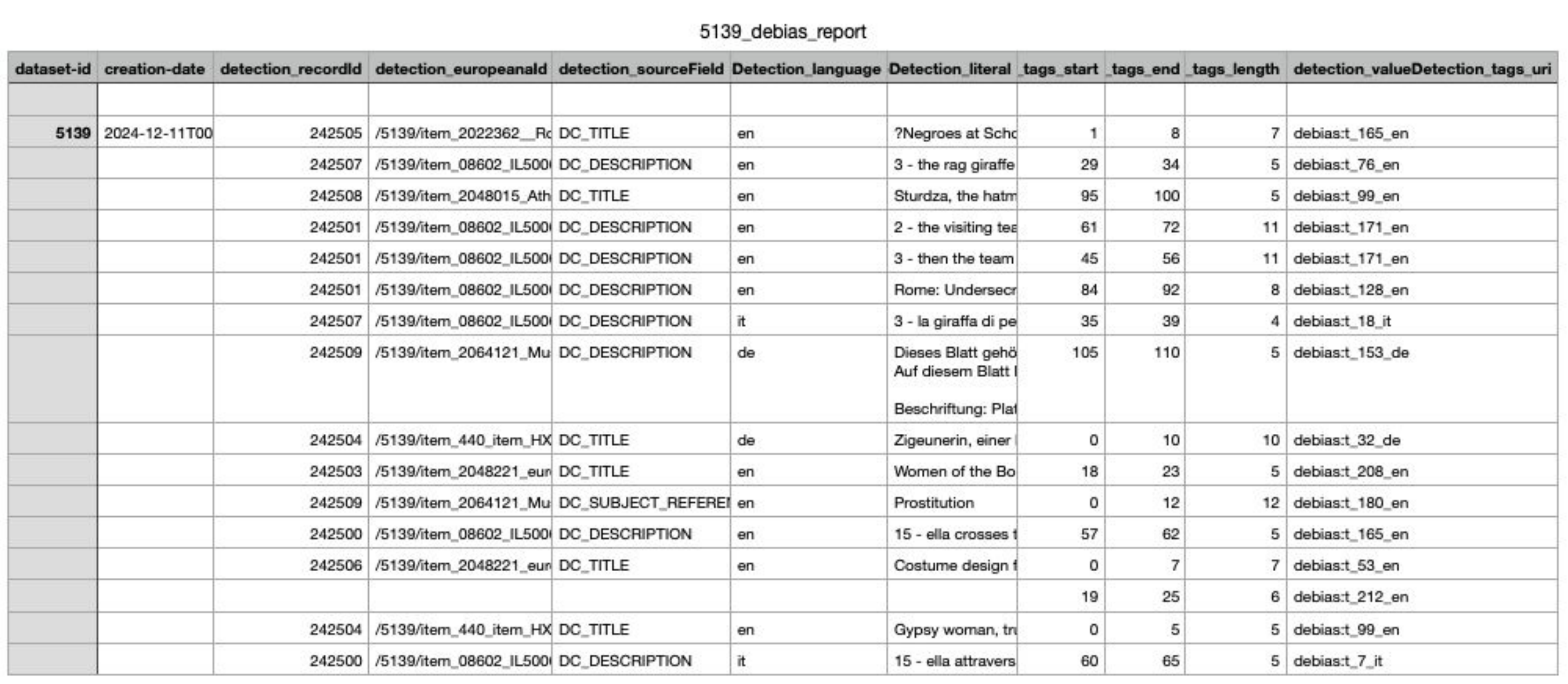
Field: `dc:title`

Language: en (English)

Literal: Sturdza, the hatman and the lever of Suceava, makes Eremia, a former speaker, inscribed on the **gypsy** robes he

- “Run report” changes to “View report” once done → click to open the report in a pop-up
- Detected term is highlighter in context (“literal”) and includes the record ID, the metadata field where the term was found and the language used for analysing
- Highlights link to the term in the DE-BIAS Vocabulary from where all further information can be accessed
- Download option in CSV

CSV download of the analysis report



5139_debias_report

dataset-id	creation-date	detection_recordId	detection_europeanaId	detection_sourceField	Detection_language	Detection_literal	_tags_start	_tags_end	_tags_length	detection_valueDetection_tags_uri
5139	2024-12-11T00	242505	/5139/item_2022362__R	DC_TITLE	en	?Negroes at Sch	1	8	7	debias:t_165_en
		242507	/5139/item_08602_IL500	DC_DESCRIPTION	en	3 - the rag giraffe	29	34	5	debias:t_76_en
		242508	/5139/item_2048015_Ath	DC_TITLE	en	Sturdza, the hatm	95	100	5	debias:t_99_en
		242501	/5139/item_08602_IL500	DC_DESCRIPTION	en	2 - the visiting tea	61	72	11	debias:t_171_en
		242501	/5139/item_08602_IL500	DC_DESCRIPTION	en	3 - then the team	45	56	11	debias:t_171_en
		242501	/5139/item_08602_IL500	DC_DESCRIPTION	en	Rome: Undersecr	84	92	8	debias:t_128_en
		242507	/5139/item_08602_IL500	DC_DESCRIPTION	it	3 - la giraffa di pe	35	39	4	debias:t_18_it
		242509	/5139/item_2064121_Mu	DC_DESCRIPTION	de	Dieses Blatt gehö Auf diesem Blatt Beschriftung: Pla	105	110	5	debias:t_153_de
		242504	/5139/item_440_item_HX	DC_TITLE	de	Zigeunerin, einer	0	10	10	debias:t_32_de
		242503	/5139/item_2048221_eur	DC_TITLE	en	Women of the Bo	18	23	5	debias:t_208_en
		242509	/5139/item_2064121_Mu	DC_SUBJECT_REFERE	en	Prostitution	0	12	12	debias:t_180_en
		242500	/5139/item_08602_IL500	DC_DESCRIPTION	en	15 - ella crosses	57	62	5	debias:t_165_en
		242506	/5139/item_2048221_eur	DC_TITLE	en	Costume design	0	7	7	debias:t_53_en
							19	25	6	debias:t_212_en
		242504	/5139/item_440_item_HX	DC_TITLE	en	Gypsy woman, tr	0	5	5	debias:t_99_en
		242500	/5139/item_08602_IL500	DC_DESCRIPTION	it	15 - ella attravers	60	65	5	debias:t_7_it

How to prepare data for upload

- Use the established workflow and guidelines for ingesting records into the Metis Sandbox
    - See https://europeana.atlassian.net/wiki/spaces/EF/pages/2104295432/Metis+Sandbox+User+Guide#3-How-to-prepare-your-dataset
- Make sure that any literal values in the metadata elements *dc:title, dcterms:alternative, dc:description, dc:subject, dc:type* and *skos:prefLabel* are accompanied by an *xml:lang* attribute in these elements
- Please note that only elements with the following language attributions can be analysed by the tool at the moment: en, nl, it, fr, de

The Chinese Market, François Boucher, 1742 - Rijksmuseum, The Netherlands - Public Domain

Website
pro.europeana.eu/project/de-bias

Contact
project.debias@gmail.com

Follow us
on social media using the hashtag
#DeBias

Co-funded by
the European Union