Mediterranean Digital Media Observatory

D 2.4

Multimedia Forensics Toolbox and Deepfakes Knowledge base set up

Project Title	MedDMO
Project Number	101083756
Thematic Priority	DIGITAL-2021-TRUST-01-EDMO
Start of Project	1 December 2022
Duration	30 months

Deliverable title	Multimedia Forensics Toolbox and Deepfakes Knowledge base set up
Deliverable number	D2.4
Deliverable version	V1.0
Contractual Date of delivery	31.05.2023
Actual Date of delivery	30.05.2023
Nature of deliverable	Report
Dissemination level	Public (PU)
Partner Responsible	CERTH
Author(s)	Efthimios Hamilos (CERTH), Nikos Sarris (CERTH), Zoi Palla (CERTH), Symeon Papadopoulos (CERTH)
Reviewer(s)	Danae Tsabouraki (ATC), Andreas Veglis (AUTH)
EC Project Officer	Kyriaki Tragouda

Abstract	This deliverable explains the operation and usage of the Multimedia Forensics Toolbox and the Deepfakes Knowledge base.
Keywords	Media Asset Annotation and Management (MAAM), Multimedia forensics Toolbox, Deepfake Knowledge base.

Copyright

© Copyright 2023 MedDMO Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the MedDMO Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

MedDMO is a Digital Europe SME Support Action project co-financed by the EC under Grant Agreement with project ID: 101083756. The content of this document is © the author(s).
For further information, visit www.meddmo.eu

Revision History

VERSION	DATE	MODIFIED BY	COMMENTS
V0.1	19/04/2023	Hamilos Efthimios (CERTH)	First Draft Table of Contents
V0.2	21/04/2023	Hamilos Efthimios (CERTH), Nikos Sarris (CERTH), Zoi Palla (CERTH)	Final Table of Contents
V0.3	12/05/2023	Hamilos Efthimios (CERTH)	First Draft
V0.4	15/05/2023	Nikos Sarris (CERTH), Symeon Papadopoulos (CERTH)	Second Draft
V0.5	16/05/2023	Hamilos Efthimios (CERTH), Nikos Sarris (CERTH)	Final Draft
V0.6	18/05/2023	Zoi Palla (CERTH), Symeon Papadopoulos (CERTH)	Final Draft Review
V0.7	23/05/2023	Andreas Veglis (AUTH), Danae Tsabouraki (ATC)	Review
V1.0	24/05/2023	Hamilos Efthimios (CERTH), Nikos Sarris (CERTH)	Final Document

Glossary

ABBREVIATION	MEANING
EC	European Commission
EDMO	European Digital Media Observatory
FAQ	Frequently Asked Questions
MAAM	Media Asset Annotation and Management
MedDMO	Mediterranean Digital Media Observatory
MeVer	Media Analysis Verification and Retrieval
NDD	Near-Duplicate Detection
NSFW	Not Safe For Work
SFW	Safe for Work
UI	User Interface

Table of Contents

Revision History 3
Glossary 4
Index of Figures 6
Executive Summary 8
1 Introduction 9
2 Media Asset Annotation and Management (MAAM) platform 10
2.1 Create an account and access your Workspaces 10
2.2 Uploading and organising your assets 12
2.3 Searching for assets 15
2.4 Manually annotating assets 19
2.5 The fact-checked multimedia dataset 23
2.6 Getting started guide and FAQ 24
3 Multimedia forensics tools 25
3.1 Multimedia analysis and automatic annotation 25
3.2 Deepfake detection 29
3.3 Image verification 33
3.4 Location estimation 34
3.5 Advanced search with visual similarity 35
4 Conclusions 38

Index of Figures

Figure 1: Account creation in MAAM platform 11
Figure 2: Accessing or leaving a MAAM workspace – inviting other potential users 11
Figure 3: Modifying your personal info – closing your account 11
Figure 4: Uploading multimedia content 12
Figure 5: Multimedia content in the Assets tab 12
Figure 6: Creating a new Project 13
Figure 7: Adding an asset to a project 13
Figure 8: Selecting assets 14
Figure 9: Adding selected assets to a project 14
Figure 10: Viewing the grouped assets of a project 15
Figure 11: Searching with keywords 'flag, people' 15
Figure 12: Near-Duplicate detection using the search bar 16
Figure 13: Find similar tool – example 16
Figure 14: Search and retrieve using the 'Action' filter 17
Figure 15: Search and retrieve using the 'Object' filter 17
Figure 16: Filter options location 18
Figure 17: Using the flags 18
Figure 18: Clicking the three dots to view the asset list menu 18
Figure 19: Sorting assets (by upload date) 19
Figure 20: Adding a comment 19
Figure 21: Creating an image annotation 20
Figure 22: Specifying the image annotation type 20
Figure 23: Specifying the annotation region 21
Figure 24: Clicking the tag and viewing the annotation info table 21
Figure 25: Creating a video annotation 22
Figure 26: Specifying the annotation type and region 22
Figure 27: Specifying the time segments 23
Figure 28: Clicking the tag and viewing the annotation info table 23
Figure 29: Accessing the getting started Guide and FAQ 24
Figure 30: Two asset cards with automatically generated captions 25
Figure 31: Search and retrieve with the 'object filter' - examples 26
Figure 32: Search and retrieve with the 'action filter' – examples 26
Figure 33: Asset with disturbing content - blur or reveal the actual content 27
Figure 34: Meme example 27
Figure 35: Asset details tab - Properties 28
Figure 36: Editing the asset description 28
Figure 37: Metadata analysis table 29
Figure 38: Deepfake detection prediction closer to 0% 30
Figure 39: Deepfake detection prediction closer to 100% 30
Figure 40: Video example with Putin having an interview 31
Figure 41: Deepfake detection in video and shot segmentation – one person 31
Figure 42: Video example with more than one person depicted in the video 32
Figure 43: Deepfake detection in video and shot segmentation – more than one person 32

Figure 44: Image Verification Assistance – Heatmaps 33
Figure 45: Fusion Heatmap 34
Figure 46: Location estimation results 34
Figure 47: Searching for similar images with the 'Find Similar' tool 35
Figure 48: Find similar' tool results 35
Figure 49: Region-based search operation with 'Find Similar' tool 36
Figure 50: 'Find Near Duplicate' – image example 37
Figure 51: Characteristic frames of the first video 37
Figure 52: Characteristic frames of the second video 37
Figure 53: 'Find Near Duplicate' – video example 37

Executive Summary

In this deliverable, we report the work that has been conducted in the context of T2.3 Collaboration environment and verification tools and more precisely the part of the Task that has to deal with setting up and making available the Multimedia Forensics Toolbox and the Deepfake Knowledge Base. Both form instrumental tools and services for researchers, fact-checkers, and media experts that aim to assist in the understanding of forgeries in multimedia content and detecting disinformation.

The multimedia forensics tools have been developed by the Media Analysis, Verification, and Retrieval (MeVer) group of CERTH in the course of previous EC-funded projects (REVEAL, InVID, WeVerify) and are continuously improved and evolved to meet current challenges of the disinformation landscape under the current projects vera.ai and AI4Media. In the context of MedDMO these tools have been packaged under a Multimedia Forensics Toolbox to be made available initially to members of MedDMO, (including the MedDMO community) and potentially to the whole EDMO community. A Deepfake Knowledge Base and several functionalities that assist deepfake detection have also been set up, multiplying the capabilities of fact-checkers and media experts in that field. Furthermore, it was decided to extend the knowledge base to not only include deepfakes but any multimedia (i.e. image and/or video) content that has been involved in fact-checks. Therefore, in the rest of this report, it will only be referred to as 'knowledge base'.

The toolbox and knowledge base have been packaged under a Media Asset Annotation and Management (MAAM) platform which has also been developed by MeVer in the context of the MediaVerse project. MAAM is a web-based application that facilitates the management and annotation of multimedia collections and is capable of dealing with content-based image and video retrieval. The Multimedia Forensics Toolbox includes the following tools that can greatly assist fact-checkers and researchers in their work against disinformation:

- **Multimedia analysis and automatic annotation**
- **Deepfake detection**
- **Image verification assistant**
- **Location estimation**
- **Advanced search with visual similarity**

As a whole, MAAM integrated with all the above tools create a user-friendly interface that provides MedDMO's community with a novel instrument to: store and retrieve fact-checked multimedia content, annotate image and video in manual and automatic ways, detect forgeries in image and video, and thus create a large and richly annotated database of multimedia content that can greatly assist in the work against disinformation.

1 Introduction

In this deliverable, we describe the Multimedia Forensics Toolbox and the Knowledge Base that have been integrated into the Media Asset Annotation and Management (MAAM) platform. The complete platform now includes the following functionalities:

- **Multimedia analysis and automatic annotation:** Any image that is uploaded to the platform is automatically analysed and a short textual description of its content is generated, the basic objects and actions are recognised and added as tags and it is determined whether the content includes memes, disturbing or inappropriate scenes (detailed in Section 3.1).
- **Deepfake detection:** Images and videos are analysed to automatically determine (with a confidence level) whether faces have been manipulated and should be considered as deepfakes (detailed in Section 3.2).
- **Image verification assistant:** Images are analysed with 14 image tampering detection algorithms to determine possible forgeries (detailed in Section 3.3).
- **Location estimation:** Images are analysed to determine the location of the content based on visual similarity to landmarks or visual cues that are tell-tale of a specific region (detailed in Section 3.4).
- **Advanced search with visual similarity:** Images or videos that bear visual or semantic similarities are automatically identified (detailed in Section 3.5).
- **The fact-checked multimedia dataset:** A curated repository of media content that has already been identified as deepfake or has been used in fact-checks by the fact-checkers of MedDMO, has been assembled and provided as a separate workspace. Advanced manual annotation functionalities are available to users so that they can indicate the exact spatial and temporal locations of manipulations that have been detected (detailed in Section 2.5).

This report explains the operation and usage of the Multimedia Forensics Toolbox and Knowledge Base and can act as a detailed user manual for the community members of MedDMO. In Section 2 we describe the basic functionalities of MAAM, whereas in Section 3 we explain the more advanced multimedia forensics tools and provide examples of their operational usage.

2 Media Asset Annotation and Management (MAAM) platform

All the tools of the Multimedia Forensics Toolbox as well as the Knowledge Base are provided through the **Media Asset Annotation and Management (MAAM) platform** which was developed in its initial form for the MediaVerse project[1], and further developed in the course of MedDMO to satisfy the needs of its fact-checkers and researchers. MAAM provides media analysis and annotation services that help users organise media assets, search and retrieve content based on advanced filters and detect potential manipulations on image and video.

In this section, we describe the basic operation of MAAM and in Section 3 we explain the multimedia forensics tools that have been integrated in order to provide the advanced functionalities that can service a fact-checking workflow for users from MedDMO and its community. In brief, these functionalities include the following:

- **Automatic image captioning**
- **Object detection**
- **Action recognition**
- **Content moderation**
- **Image and video manipulation detection**

As a whole MAAM is a web-based platform with advanced technological tools that enable users to easily store and retrieve multimedia content, automatically and/or manually annotate such content, as well as detect forgeries, thus servicing the work of a fact-checker or researcher.

2.1 Create an account and access your Workspaces

Users can request the creation of a new account through the MAAM entry page at https://maam.mever.gr/, completing the requested details shown in Figure 1. The MedDMO system administrator will receive the request and approve or deny it, also granting access to specific workspaces if needed.

Workspaces are areas containing specific content which is only visible to members of that workspace. They are meant to service specific groups of users. So, for example, there is a workspace called MedDMO which has been created for all MedDMO members, but more workspaces can be created to service subgroups of MedDMO, e.g., MedDMO-FactCheckers. Workspaces ensure that MAAM can be used by different groups of users potentially for entirely different purposes, maintaining the absolute privacy of content and annotations.

By default, a new user has access to a personal empty workspace and can create new workspaces where s/he can invite other users to join. When users are invited to other workspaces, they receive a notification allowing them to accept or deny the invitation. Users can also remove other users from workspaces they own, or leave workspaces they have previously joined. These options are illustrated in Figure 2.

Furthermore, by clicking on their avatars on the top right users can modify their personal information or close their accounts as shown in Figure 3.

[1] www.mediaverse.eu

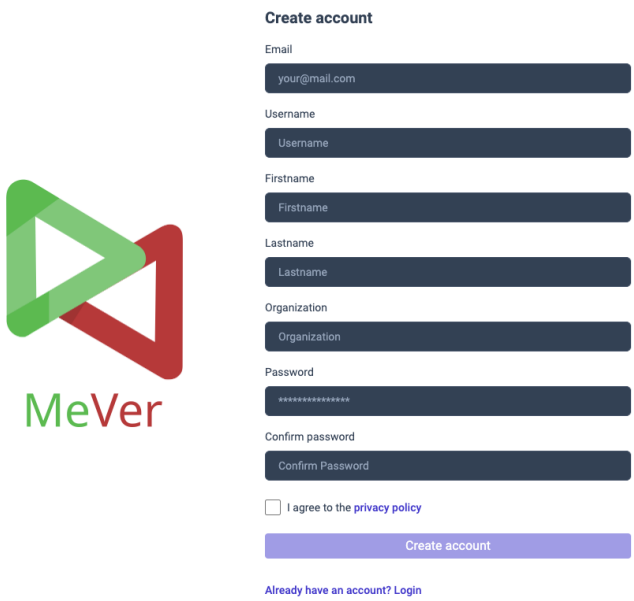


Figure 1: Account creation in MAAM platform

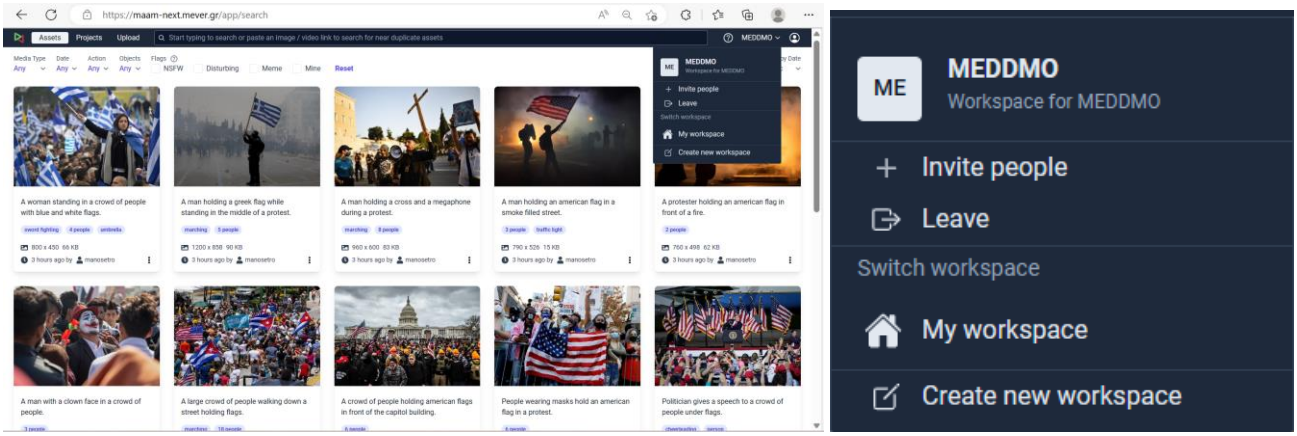


Figure 2: Accessing or leaving a MAAM workspace – inviting other potential users

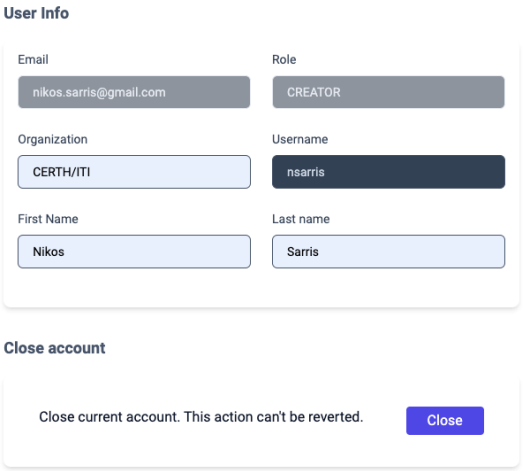


Figure 3: Modifying your personal info – closing your account

2.2 Uploading and organising your assets

Users can upload multimedia content through the 'Upload' tab either by dropping it into the box or by clicking to add files or add folders and provide the corresponding path as illustrated in Figure 4. Content can also be imported via CSV files. The supported file formats are jpeg, png, webp, mp4, quicktime, and CSV. Assets are uploaded into the user's current workspace.

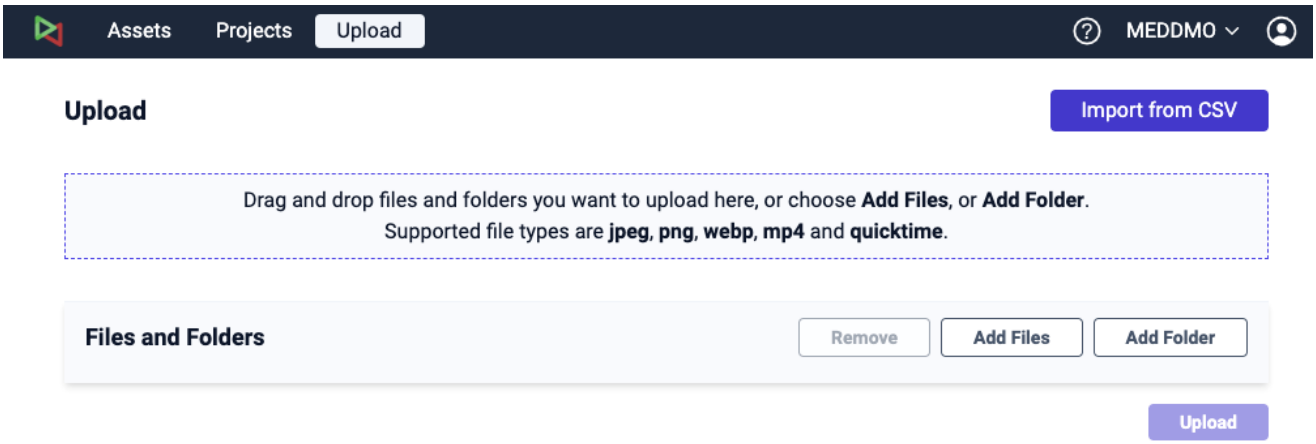


Figure 4: Uploading multimedia content

The uploaded multimedia content is available in the 'Assets' tab as illustrated in Figure 5.

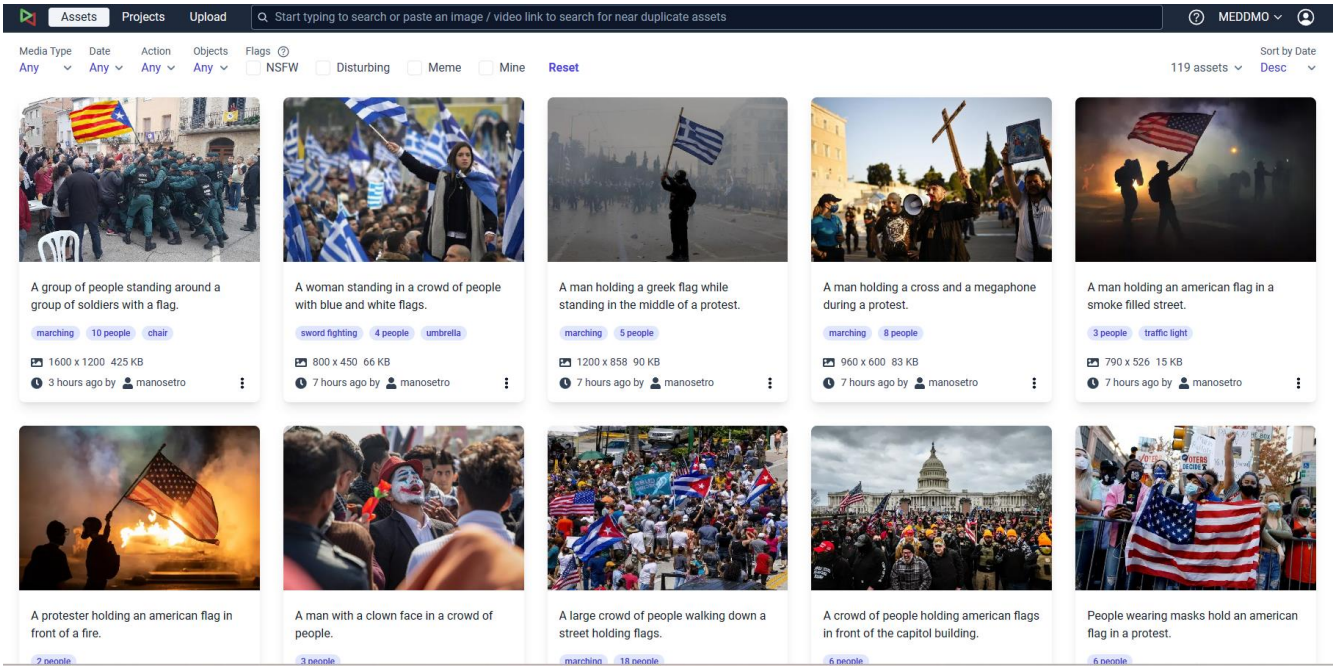


Figure 5: Multimedia content in the Assets tab

To better organise their content, users can create 'Projects' where they can include assets they wish to group together, as shown in the example in Figure 6. Projects are personal, i.e. users only see their own projects. Projects are also workspace specific, i.e. when a user switches to a different workspace the projects seen will be

different and any asset that is uploaded in a workspace can be only added to a project that belongs to that workspace.

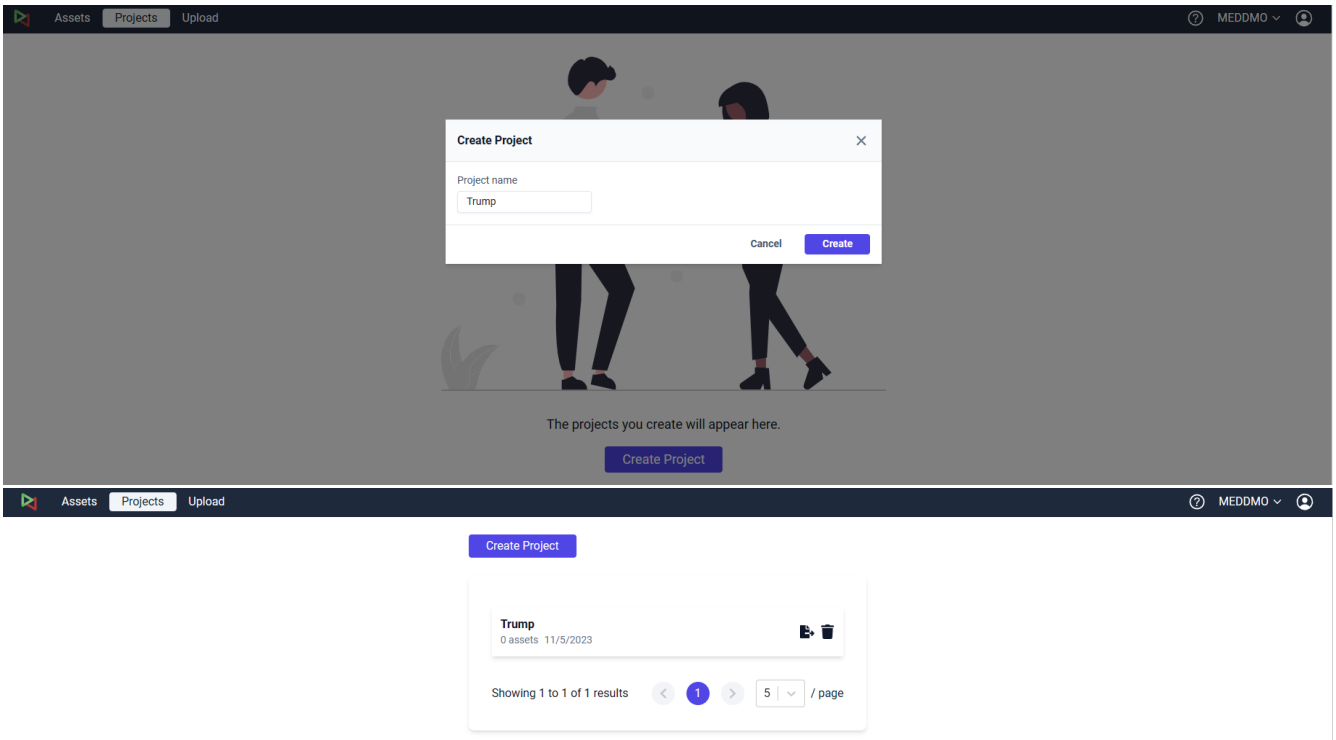


Figure 6: Creating a new Project

There are various ways to add assets to a project. For example:

1. Click on an asset, navigate to the Project section, and click Add, as shown in Figure 7.

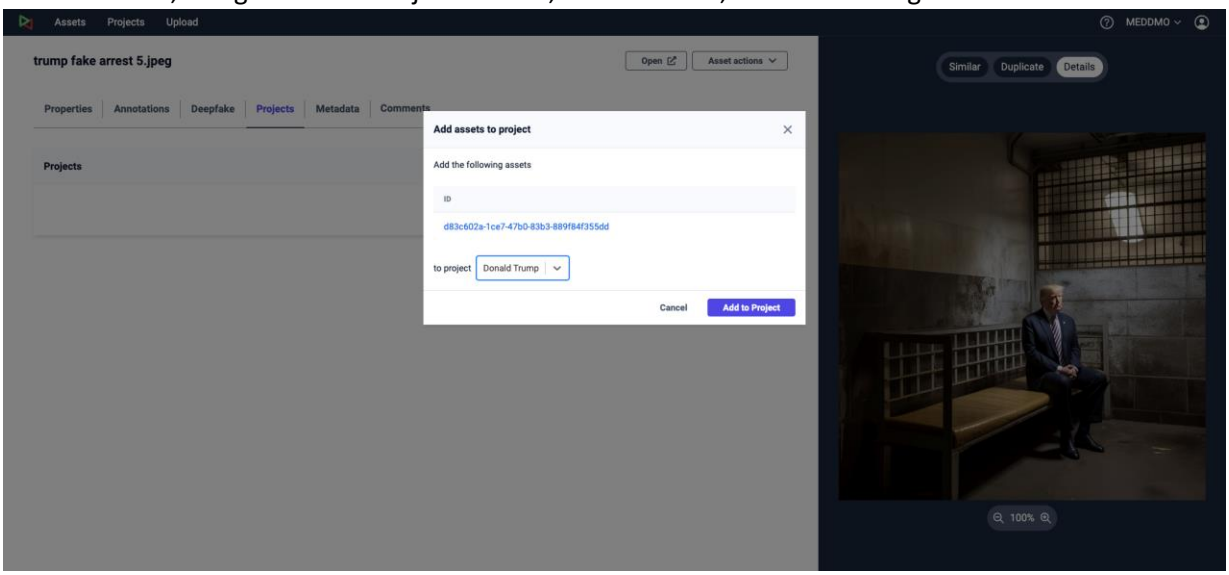


Figure 7: Adding an asset to a project

2. Use the select option which appears when clicking on the three dots icon in the lower right corner of each asset as illustrated in Figure 8. Choose one or more assets as illustrated in Figure 9 and then click on the

arrow at the top right of the window next to the word ‘selected’ and select ‘+ Add to Project’ picking the required project name.

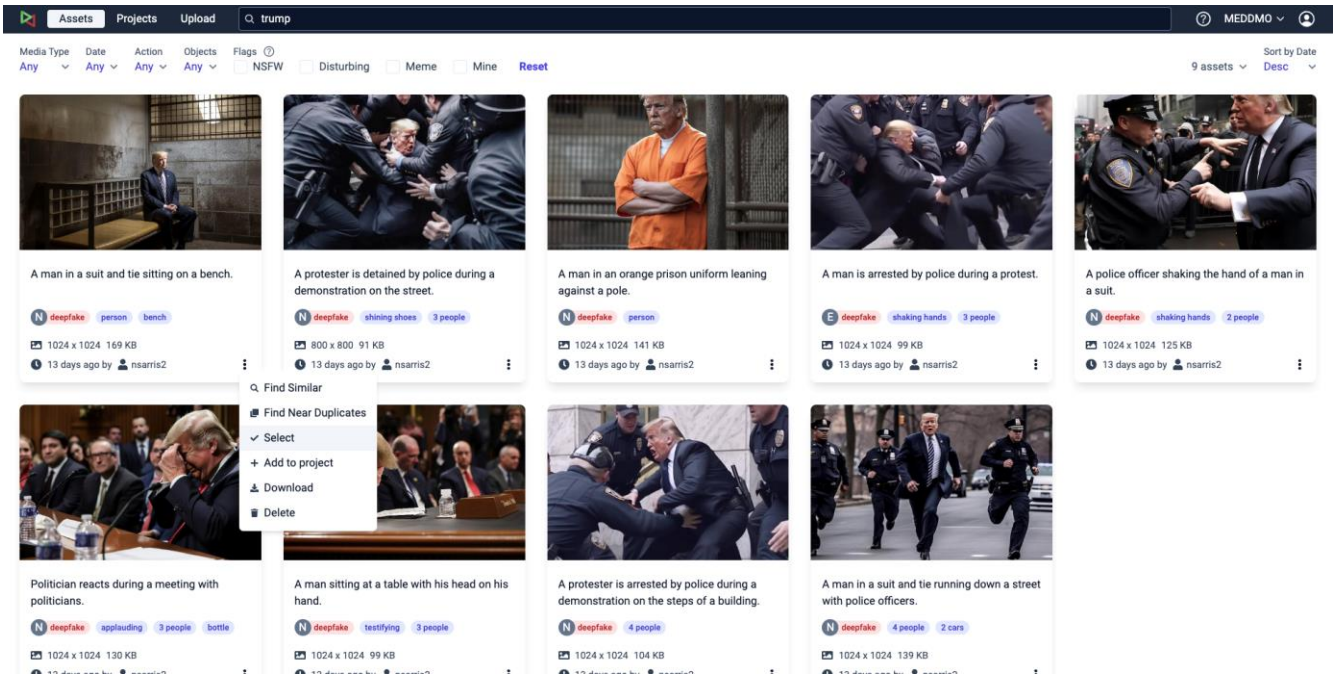


Figure 8: Selecting assets

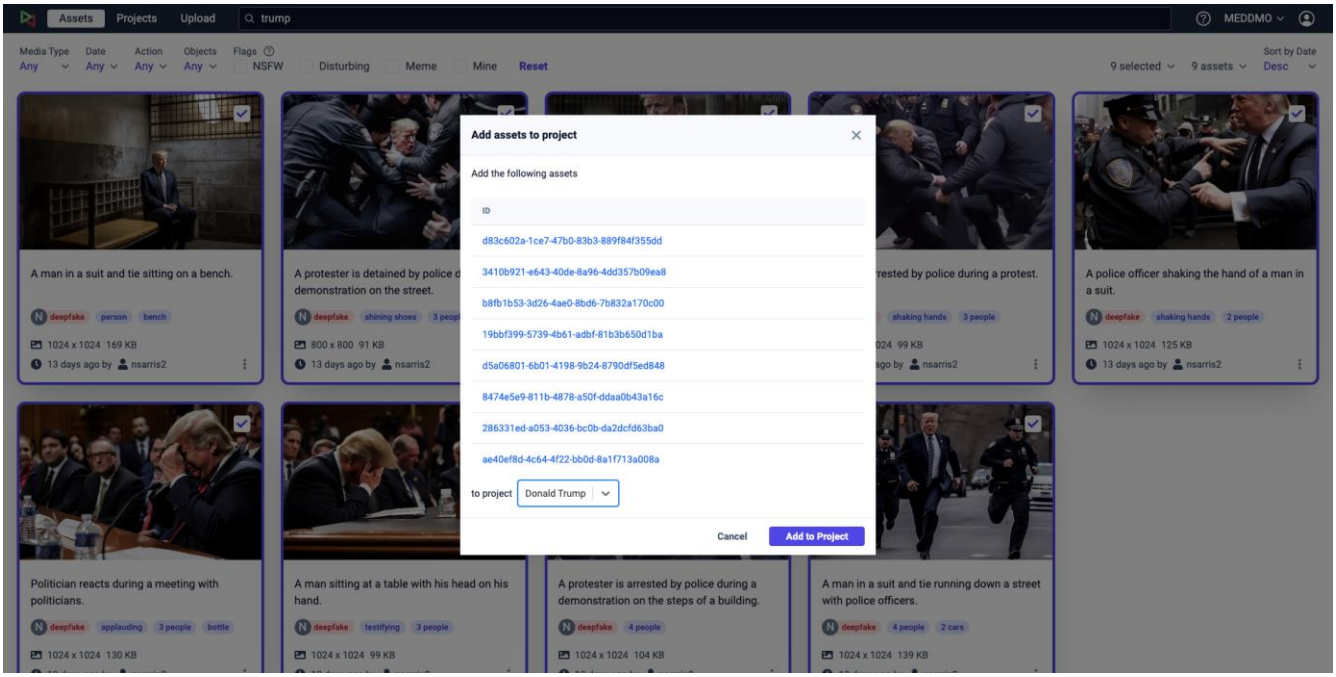


Figure 9: Adding selected assets to a project

From the ‘Projects’ tab clicking on the name of a project brings the assets that have been grouped under this project as illustrated in Figure 10.

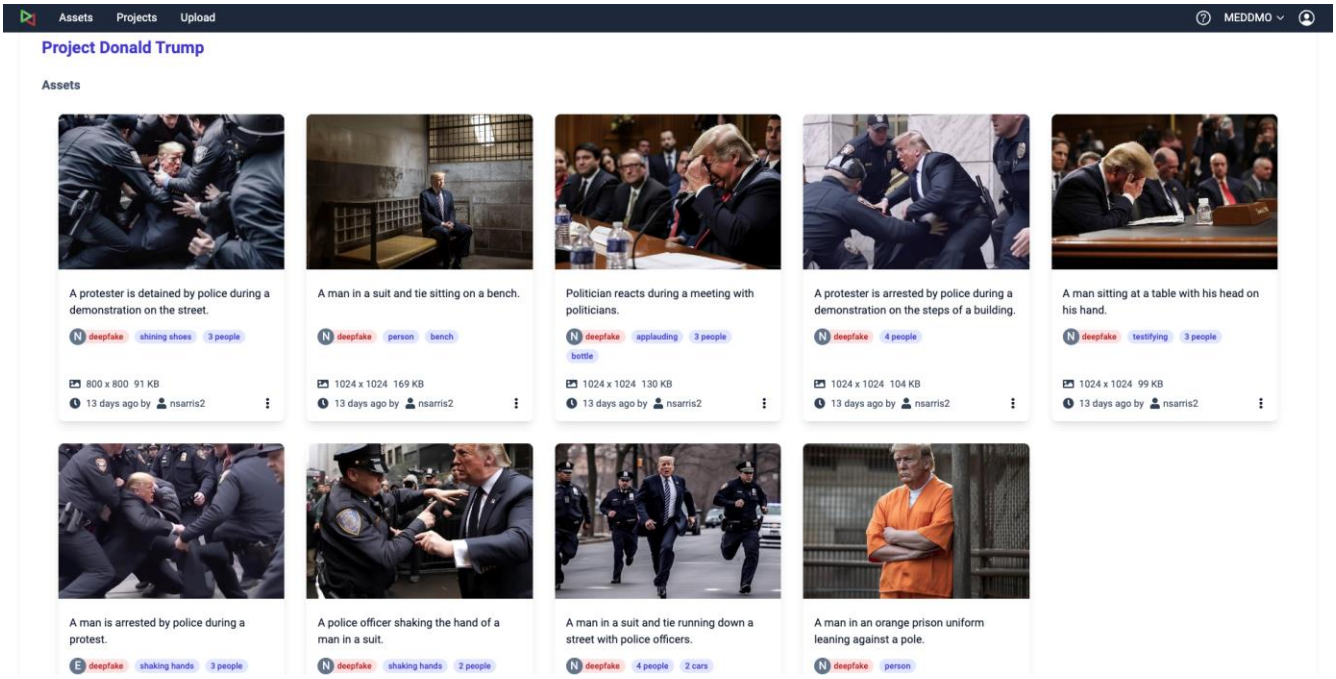


Figure 10: Viewing the grouped assets of a project

2.3 Searching for assets

Users can search for assets with the help of various tools:

1. Typing keywords on the search bar brings back content that contains the keywords in the (automatically generated) captions or tags as depicted in Figure 11.

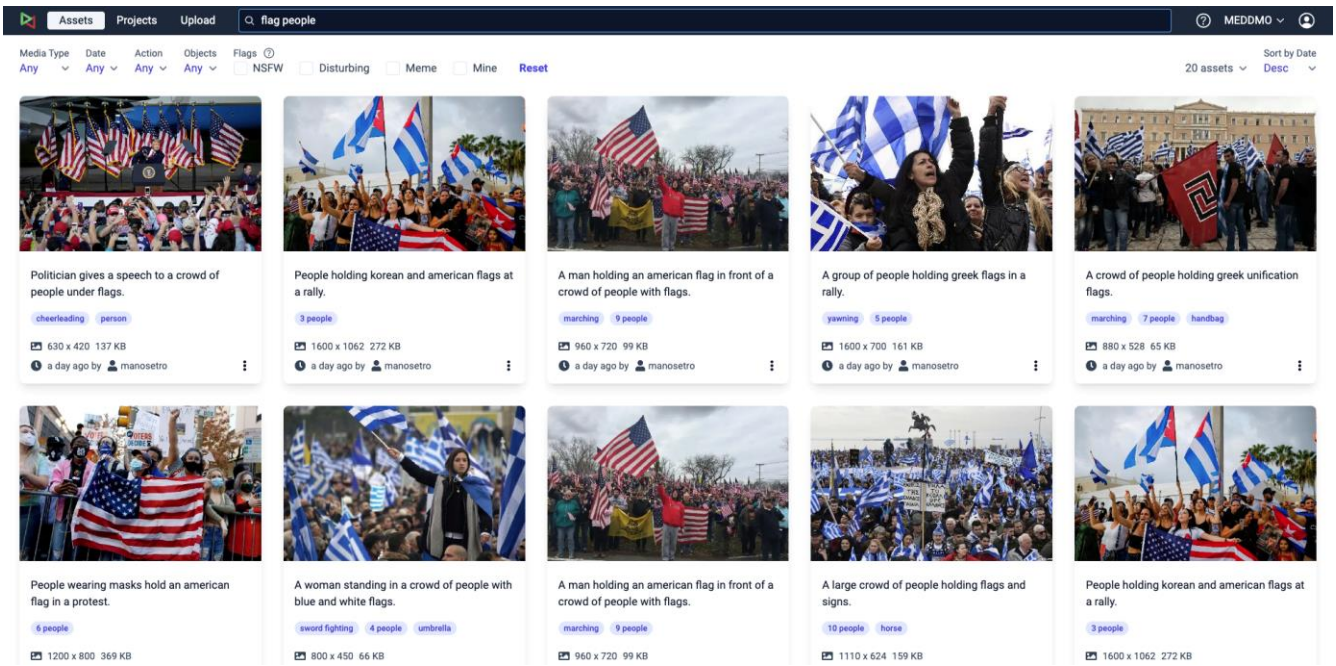


Figure 11: Searching with keywords 'flag, people'

2. Inserting an image or video link in the search bar brings back images or videos that visually resemble the one found at that link, as shown in Figure 12. This functionality is described in more detail in Section 3.5.

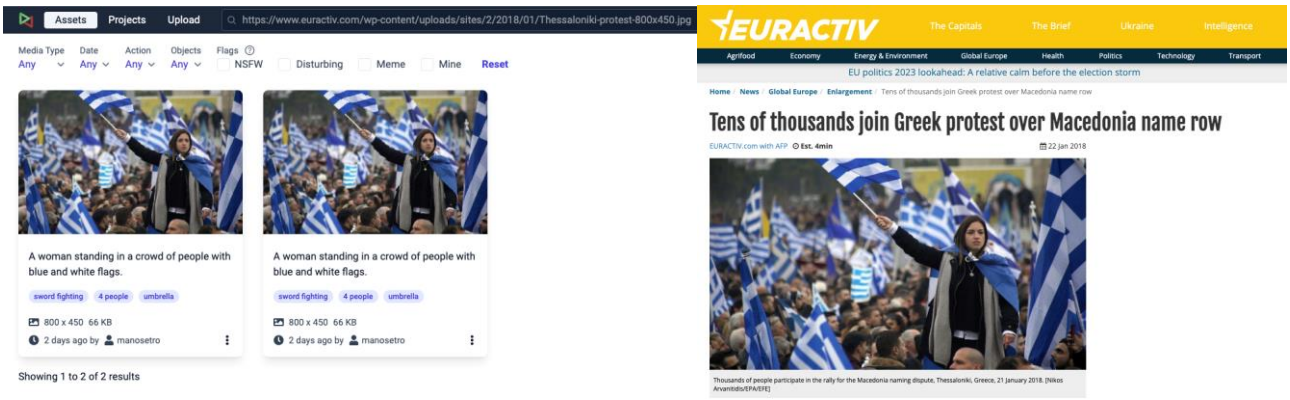


Figure 12: Near-Duplicate detection using the search bar

3. Clicking on the three dots in the lower right corner of an asset and picking 'Find Similar' the user is provided with assets that have some visual semantic similarity with the one in question, as shown in Figure 13 where an image containing Donald Trump is queried and the search window is constrained to include his face, other images containing Donald Trump are returned. This functionality is described in more detail in Section 3.5.

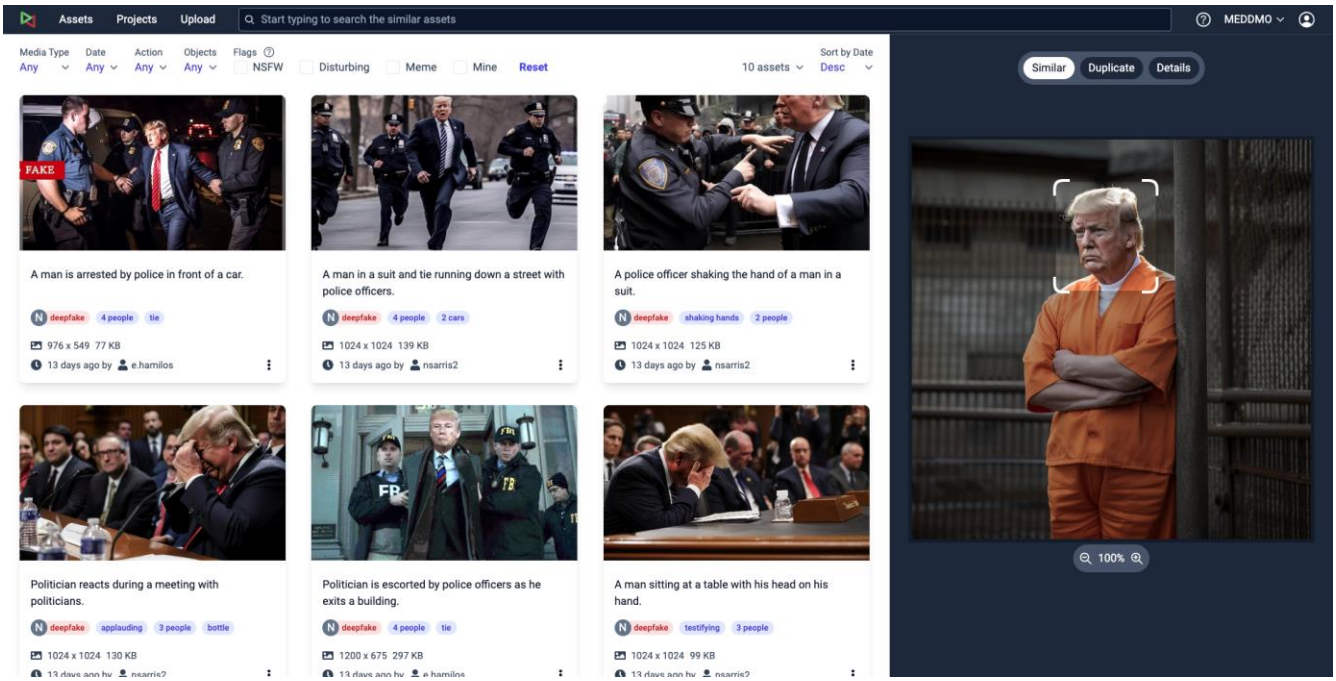


Figure 13: Find similar tool – example

4. Using the 'Media Type' filter restricts content to being image, video, or both.
5. Using the 'Date' filter restricts content according to the asset uploaded date.
6. Using the 'Action' filter restricts content according to a predefined set of about 400 activities detected in images and videos through automatic visual analysis. In Figure 14 we show an example with the action 'driving car'.

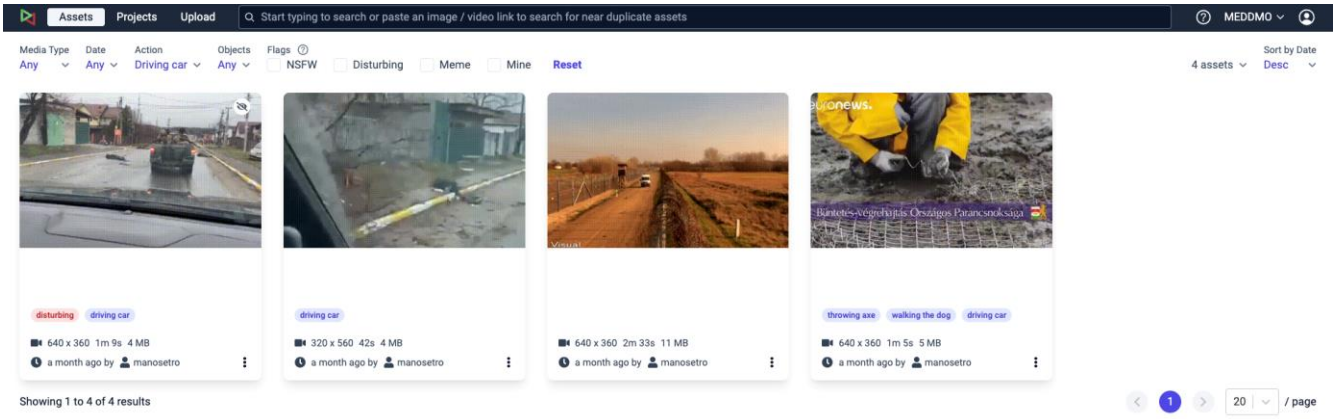


Figure 14: Search and retrieve using the 'Action' filter

7. Using the 'Object' filter restricts content according to a predefined set of about 100 common objects detected in images and videos through automatic visual analysis. In Figure 15 we show an example with the object 'laptop'.

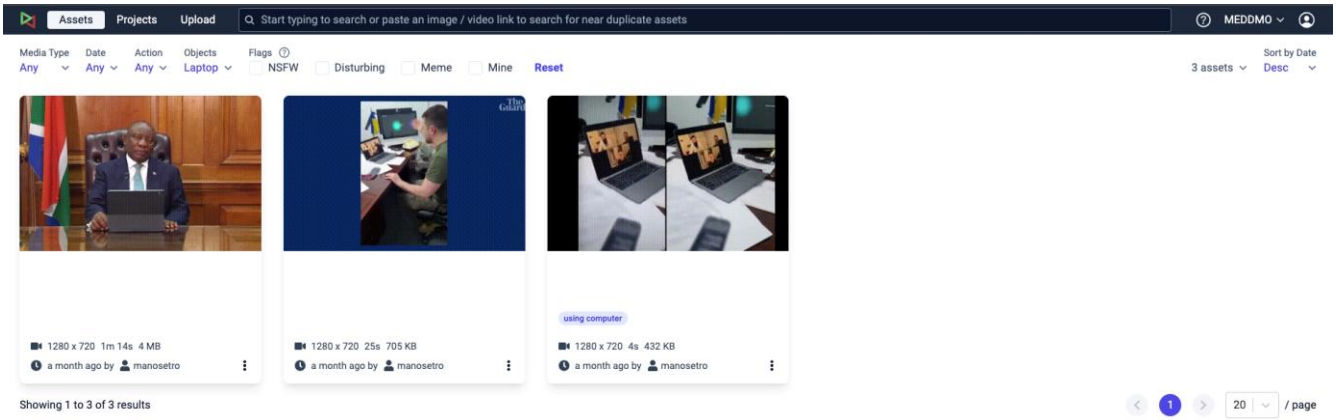


Figure 15: Search and retrieve using the 'Object' filter

8. Using the 'NSFW' (Not Safe For Work) filter restricts content to assets having been automatically classified to contain inappropriate or offensive scenes, which are automatically blurred.
9. Using the 'Disturbing' filter restricts content to assets having been automatically classified to contain inappropriate scenes such as graphic violence, gore, horror, explicit depictions of harm or suffering, disturbing imagery, or content that deals with sensitive topics such as abuse, self-harm, or psychological trauma. These assets are also automatically blurred.
10. Using the 'Meme' filter restricts content to assets having been automatically classified to contain Memes.
11. Using the 'Mine' filter restricts content to assets having been uploaded by the user.

For NSFW, Disturbing, meme, and mine filters, the user can choose between three flags to specify the content to be retrieved as depicted in Figure 17.

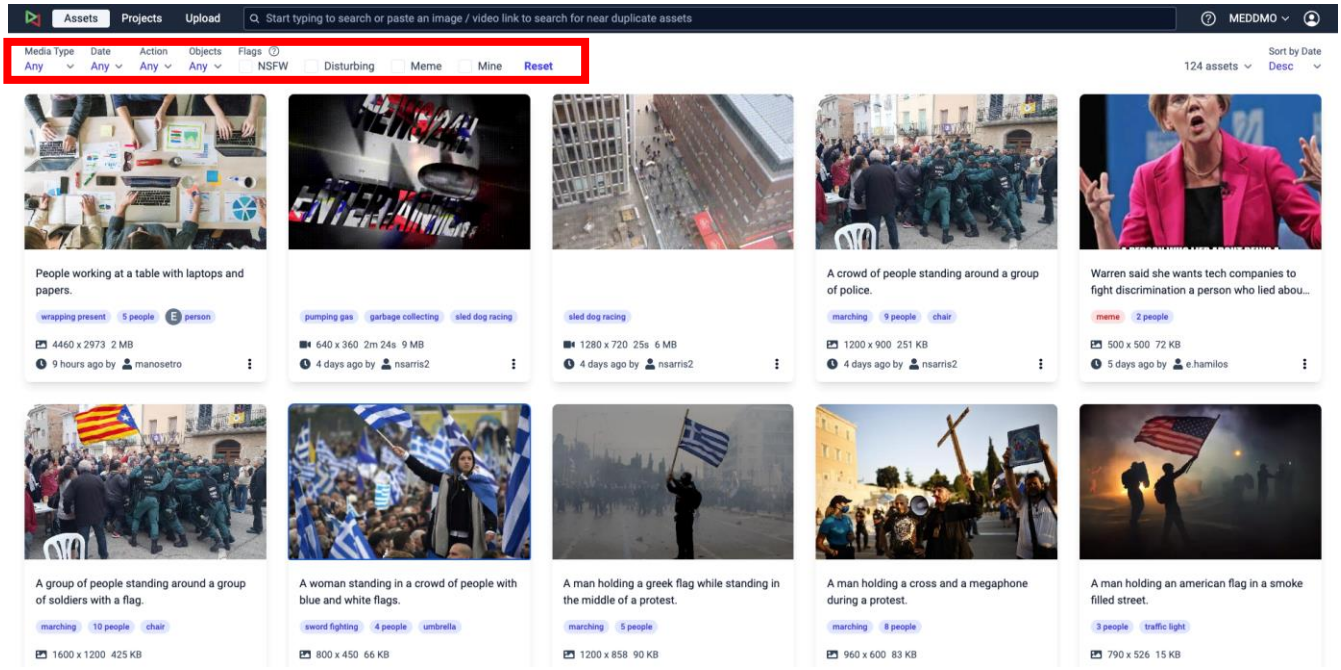


Figure 16: Filter options location

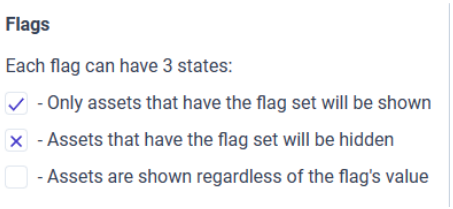


Figure 17: Using the flags

Clicking on the three dots icon at the lower right corner of each asset a menu list appears as illustrated in Figure 18, which allows users to find similar (in content) assets, or visually identical ones (near duplicates). For assets that have been uploaded by the user, download and delete functionalities are also offered under the same menu.

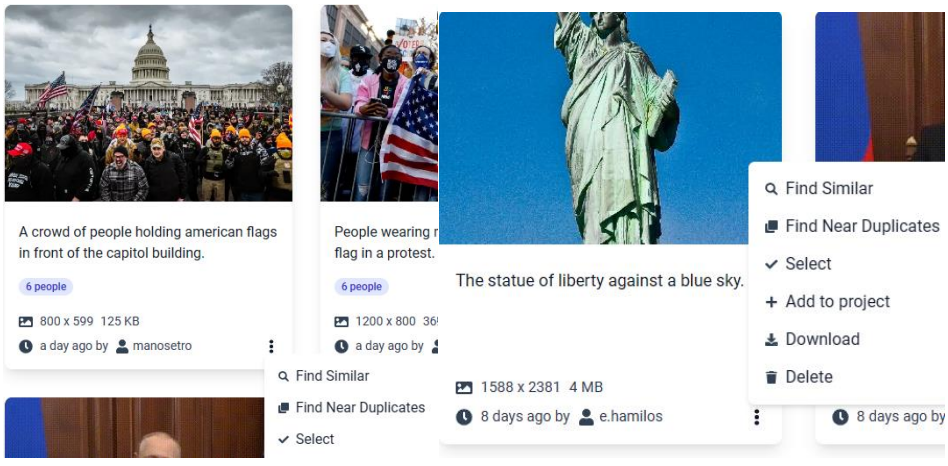


Figure 18: Clicking the three dots to view the asset list menu

Users may sort assets by (upload) date in descending or ascending order, as illustrated in Figure 19.

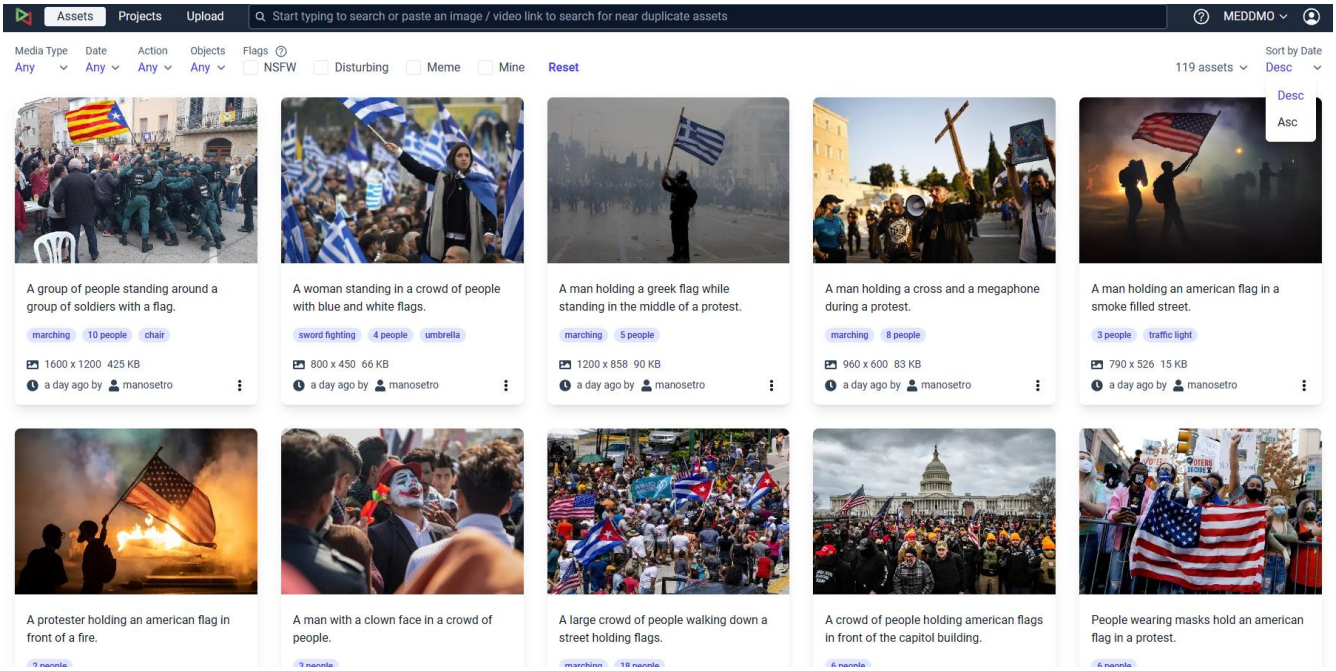


Figure 19: Sorting assets (by upload date)

The total number of assets is depicted on the left of the ''sort by date'' icon. Clicking on this arrow users can export all assets to a CSV file or add them to a project of their choice.

Users may also add free text comments to assets, as shown in Figure 20.

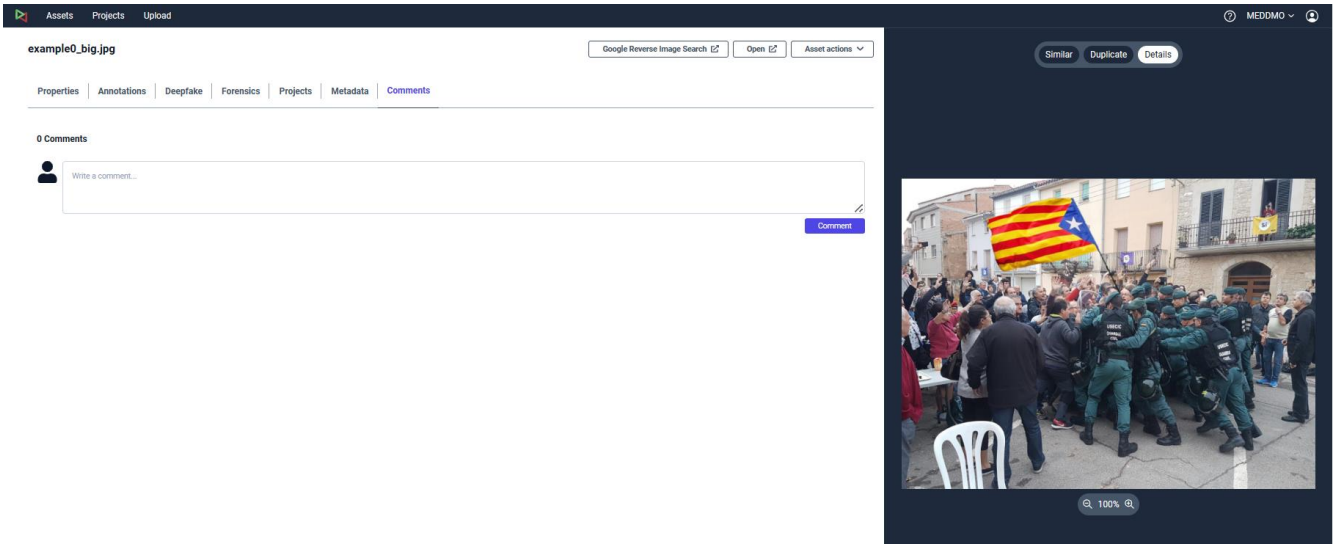


Figure 20: Adding a comment

2.4 Manually annotating assets

A manual annotator tool allows authenticated users to annotate areas (in images and videos) and/or time segments (in videos). The purpose of the manual annotator is to provide the MedDMO user with a tool that enables efficient and effective creation of personal models for annotations (specific to user's needs) in images and videos through a variety of options that are provided and are the same for image and video annotations.

Also, the tool enables the user to create an annotation for the whole image or video, but also allows to annotate specific regions by defining a bounding box. Furthermore, for video annotations, the user can specify, if it is necessary, the time segment for which the annotation is observed in the video.

To create an image annotation the user has to select the asset and then add an annotation as depicted in Figure 21. Then he has to select the type of annotation among the predefined options that are provided and specify the region if necessary, as illustrated in Figure 22.

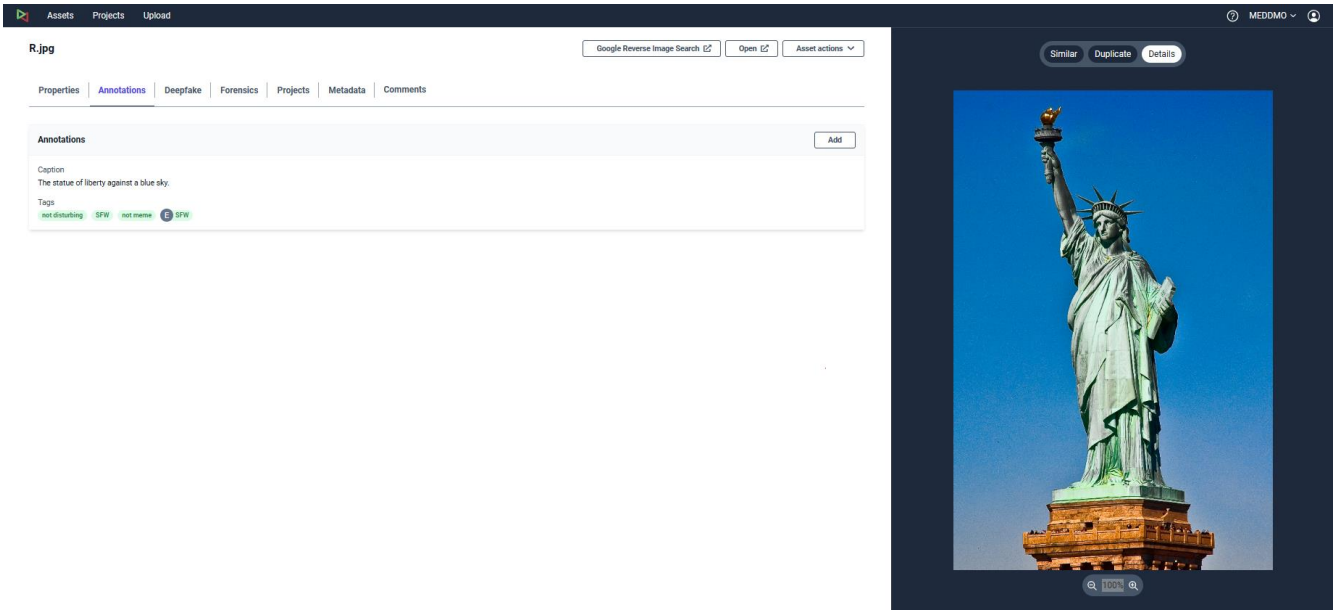


Figure 21: Creating an image annotation



Figure 22: Specifying the image annotation type

The manual image annotator provides a predefined list of options for annotations that include:

- **Object annotation:** A predefined list of objects among which the user can select is provided but the user can also define a new object.
- **Action annotation:** A predefined list of actions among which the user can select is provided but the user can also define a new action that describes better what is depicted in an image.
- **NSFW annotation:** The user can annotate if the depicted image contains inappropriate content
- **Disturbing annotation:** The user can annotate if the depicted image contains disturbing content
- **Meme annotation:** The user can annotate the depicted image as a meme
- **Deepfake annotation:** The user can annotate the depicted image as a deepfake

Furthermore, users can specify regions of the image which correspond to the annotation they want to attach, defining the bounding box as illustrated in Figure 23.

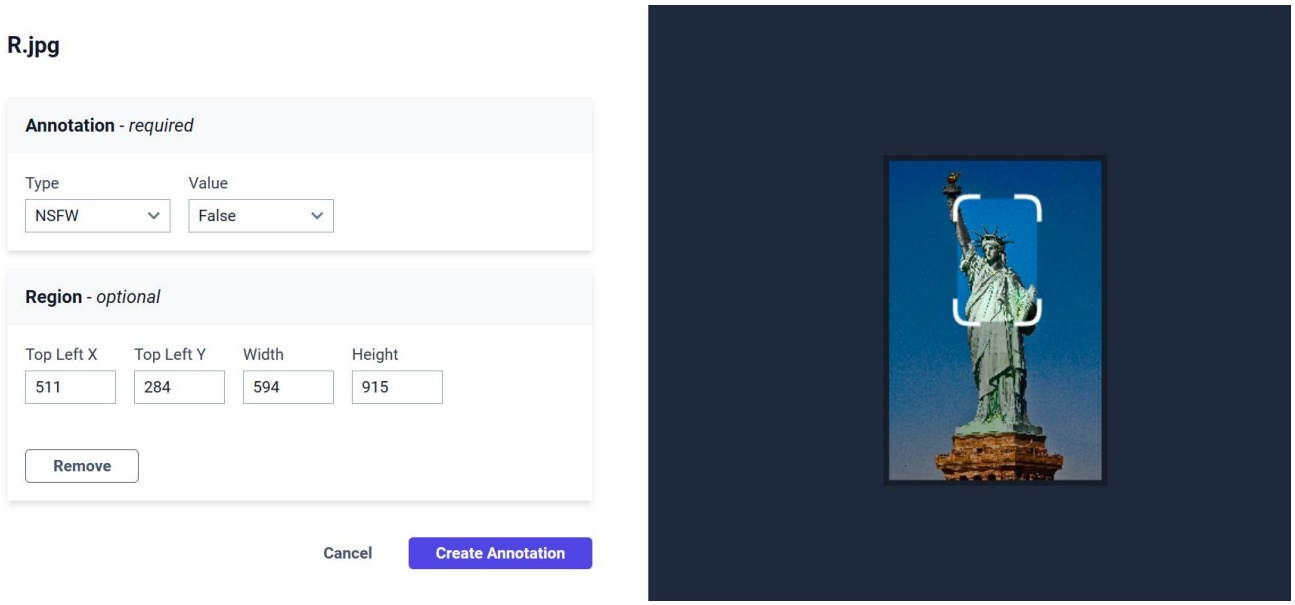


Figure 23: Specifying the annotation region

Manually created image annotations are added as tags next to the automatic annotations with the difference that the avatar of the user (creator) is also included on the left side of the tag. Clicking on the manually created annotation tag a table is provided with the annotation type, the date and time and creator as depicted in Figure 24. An overlay on the image also highlights the spatial region of the annotation, if any. Only the creator can modify or delete a manual annotation. The manual annotation tags are also visible on the assets home page.

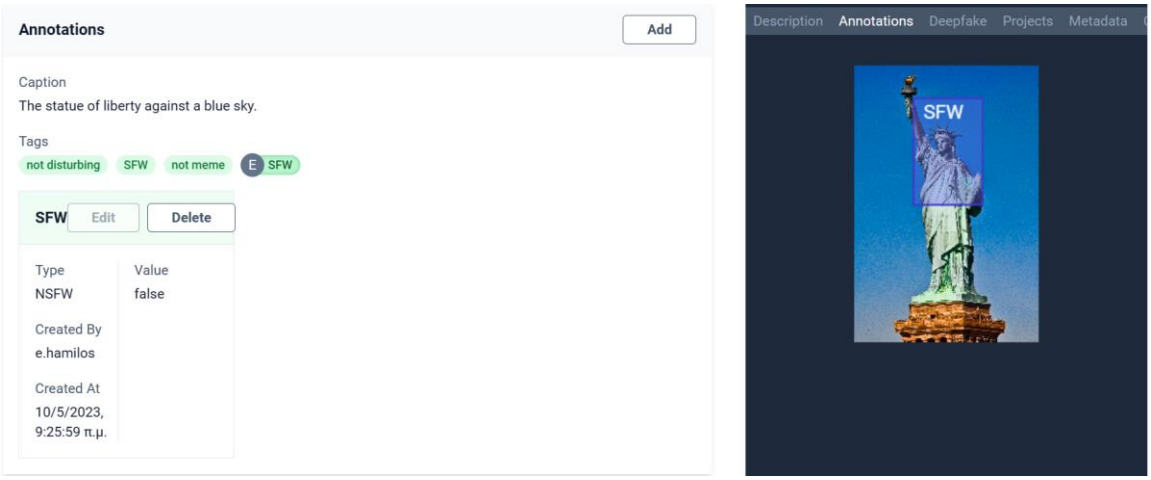


Figure 24: Clicking the tag and viewing the annotation info table

To create a manual annotation on a video the user has to select the asset, navigate to the annotations section and click the add button, as illustrated in Figure 25.

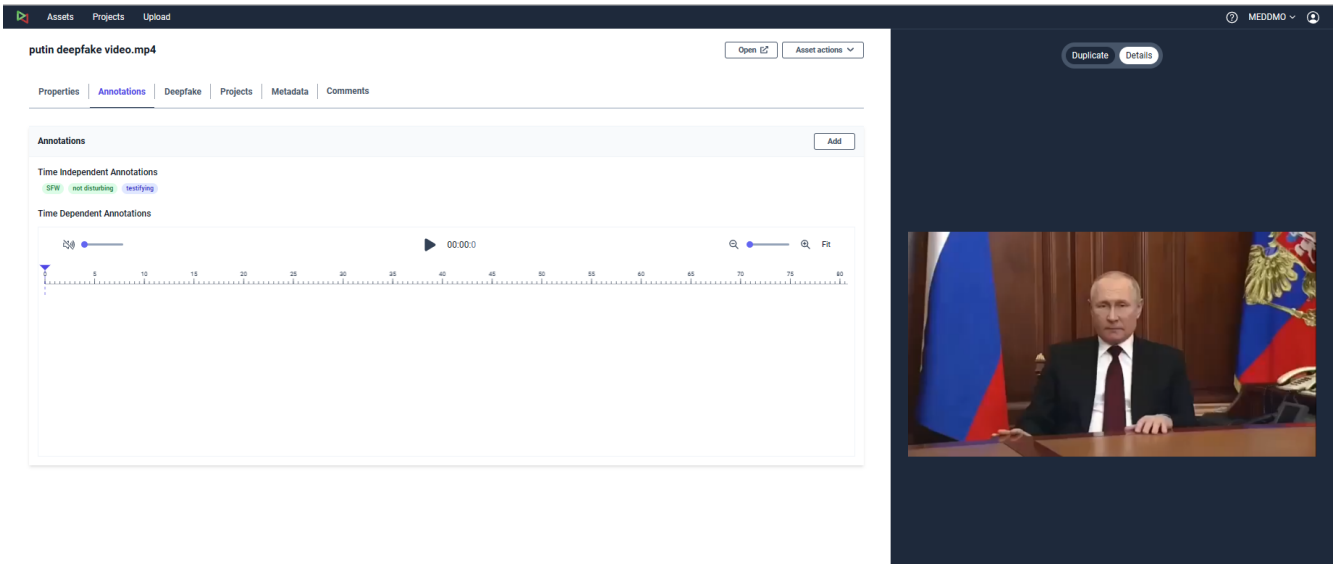


Figure 25: Creating a video annotation

Video annotations can be attached to a (temporal) segment of the video or to its whole duration. In the second case, the process is similar to creating an image annotation: the user has to select the type of annotation among the predefined options and specify the (spatial) region, if necessary, as illustrated in Figure 26. The predefined list of options for video annotations is exactly the same with images. In this case, the user doesn't have to 'add duration' and specify the time segments.

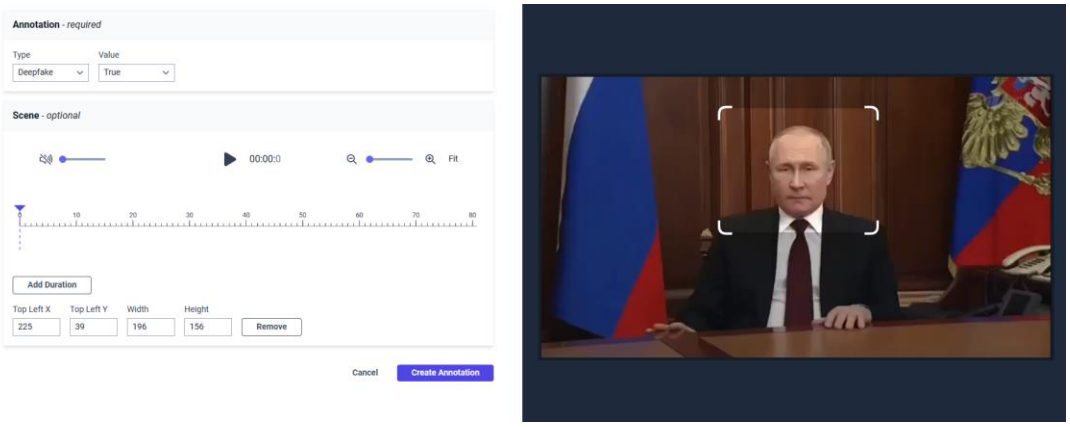


Figure 26: Specifying the annotation type and region

To create a time dependent annotation the user has to click on the 'add duration' button and then specify the time segments he wants to annotate. The user is able to specify the scene start and the scene end and, in that way, specify the duration for the annotation as illustrated in Figure 27.

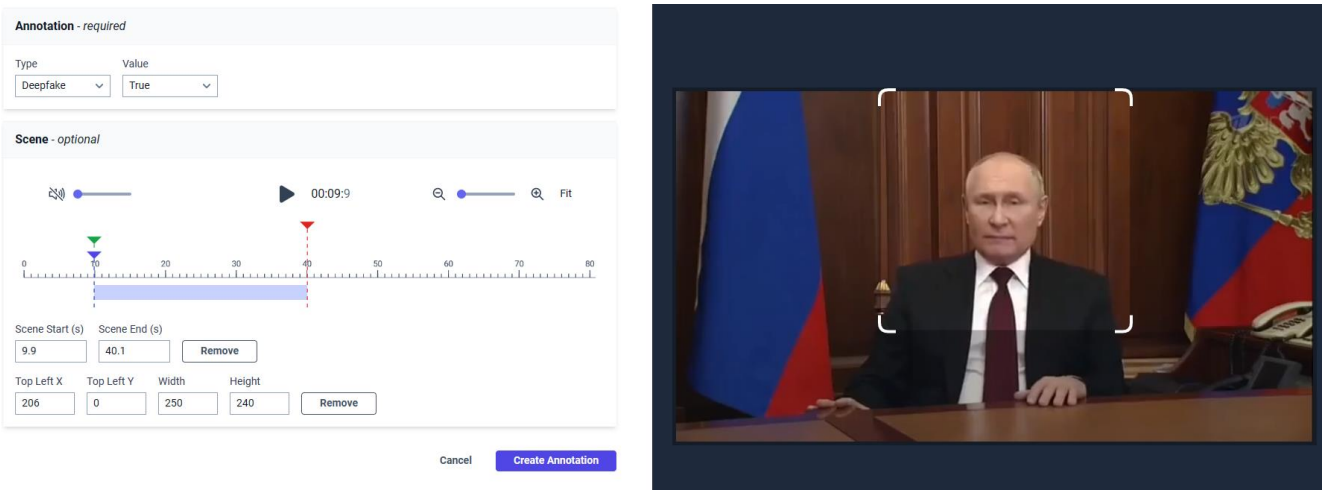


Figure 27: Specifying the time segments

Manually created video annotations are again added as tags next to the automatic annotations with the difference that the avatar of the user (creator) is also included on the left side of the tag. Time dependent or independent annotations are separated into two different sections. Clicking on the manually created time independent annotation tag a table is provided with the annotation type, the date and time, and the creator as depicted in Figure 28. An overlay on the image also highlights the spatial region of the annotation, if any. For time dependent annotations, annotated segments are shown on a timeline, and clicking on them a table is provided with the annotation type, the date and time, the creator and the 'start - end scene' as depicted in Figure 28. Also in this case an overlay on the video shows the annotated regions, if any.

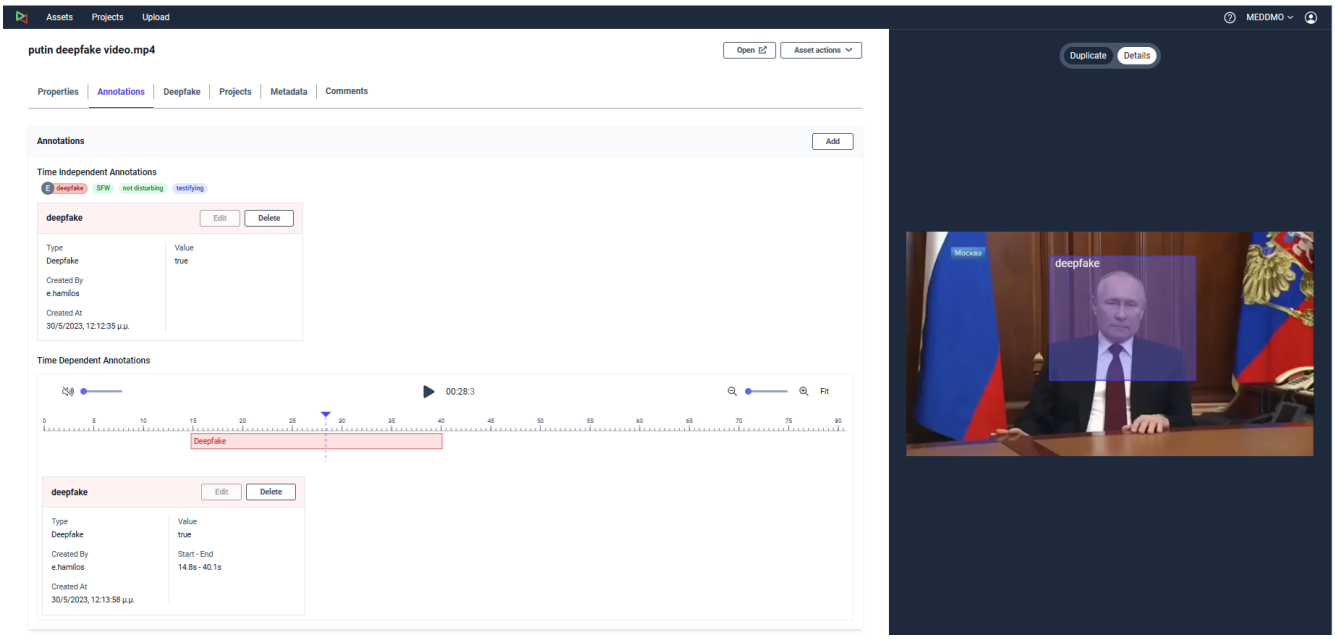


Figure 28: Clicking the tag and viewing the annotation info table

2.5 The fact-checked multimedia dataset

For the purposes of MedDMO a dedicated workspace has been created containing multimedia content from a large number of fact-checks provided by MedDMO fact-checkers. Initially, 500 assets have been provided by

Ellinika Hoaxes from their archives of fact-checks. These will be augmented with multimedia material from the other MedDMO fact-checkers (AFP and Times of Malta), while discussions are to be initiated with organisations and experts that join the MedDMO community to also include content from their archives. After streamlining the process and setting up an operation that will continuously feed this database with new content, discussions are planned with the other EDMO hubs, to include their content and eventually make the platform available to them as well.

The aim is to create a large database of fact-checked images and videos that will enable users to easily find multimedia content that has already been investigated by other fact-checkers in the past and avoid repetition of work, while also allowing collaboration across fact-checkers and EDMO hubs. The advanced tools for searching and retrieving content based on visual analysis are expected to ease the work of fact-checkers and researchers to a great degree.

2.6 Getting started guide and FAQ

By clicking on the question mark icon on the top right users can access a comprehensive getting started Guide and quick answers to Frequently Asked Questions to get familiarised with MAAM features and tools, as depicted in Figure 29.

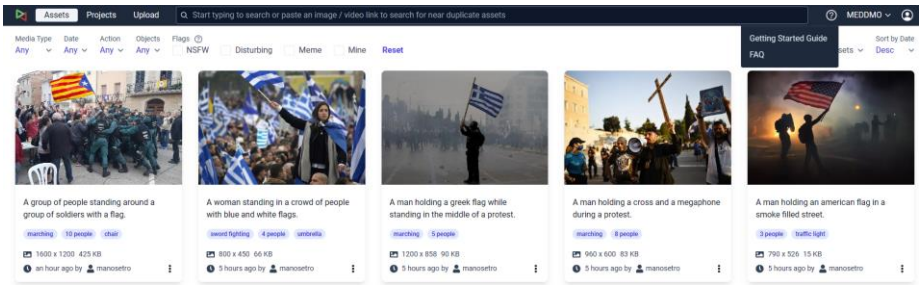


Figure 29: Accessing the getting started Guide and FAQ

3 Multimedia forensics tools

3.1 Multimedia analysis and automatic annotation

One of the core functionalities of MAAM is the automatic visual analysis of every asset at the time it is uploaded and its annotation with descriptive text and tags so that it can be easily retrieved and filtered. By applying annotations as filters in asset search and retrieval, users can easily find content in large-scale collections that fits their needs and organise it in semantically coherent groups (the Projects). For the asset annotation process, several models have been integrated to provide information ranging from objects and actions recognized in assets, to descriptive text captions that provide a user-friendly description of the assets. The automatic multimedia analysis and annotation process that MAAM employs in each asset includes the following:

Automatic image captioning: Captioning can greatly enhance image retrieval by generating a free-text description for each image, providing additional context for the visual content. Figure 30 shows two asset cards with the corresponding generated captions. By indexing these captions, we are able to retrieve assets even if the user has not provided any relevant metadata or textual description. For example, a user can retrieve the first asset at the left of Figure 30, which depicts the Statue of Liberty, even though the uploader provided no such information. The same applies to the second asset at the right of Figure 30 which depicts the ruins of the Colosseum.

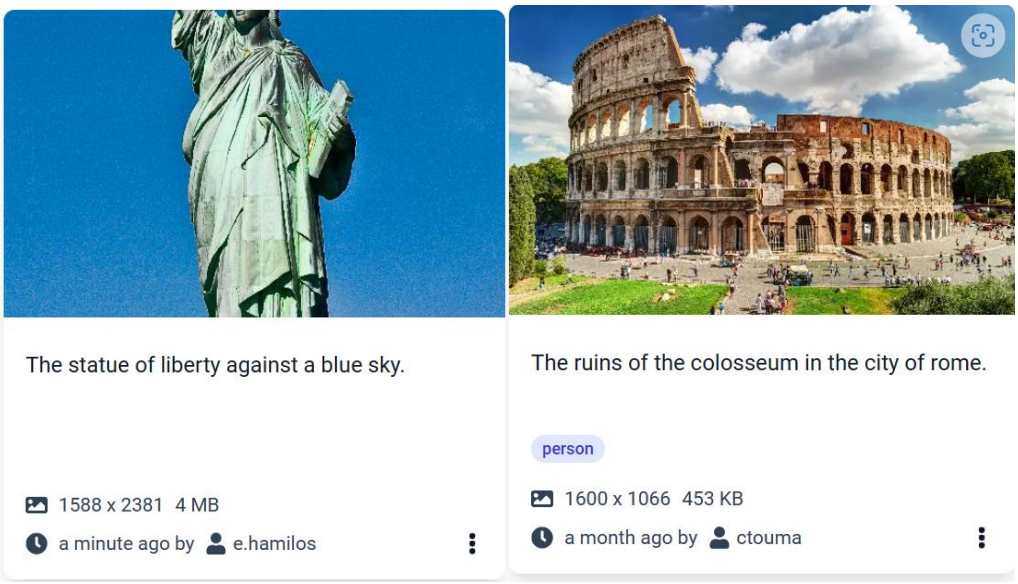


Figure 30: Two asset cards with automatically generated captions

Object detection: The object detection model is used to identify objects within images and video frames. In the case of images, we detect and store the bounding box containing the corresponding object, while in videos we also provide time information. If a user wants to find images or videos containing a specific object, they can filter assets by using the objects filter and the platform will return all images and videos that have been annotated with that label. Figure 31 shows three asset cards two images and one video with the corresponding identified objects.

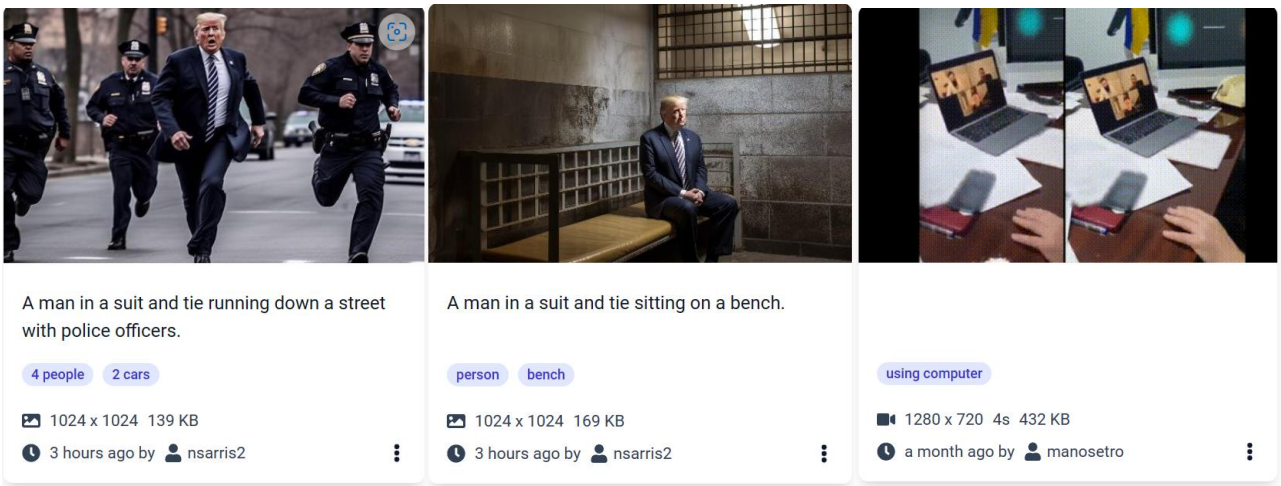


Figure 31: Search and retrieve with the 'object filter' - examples

Action recognition: The action recognition model is used to identify actions within images and video frames. Action recognition in images and videos indicates another category that a user could use to filter multimedia content. If a user wants to find images and/or videos containing a specific action, they can filter assets by using the action filter and the platform will return all images and/or videos that have been annotated with that label. In Figure 32 we provide three examples, two videos, and one image with the corresponding recognized actions.

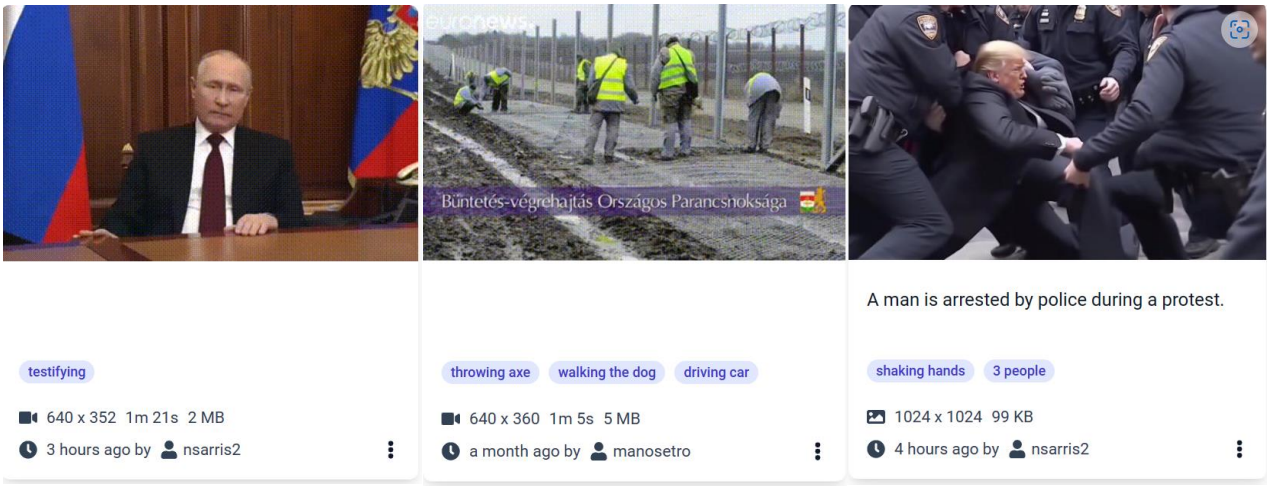


Figure 32: Search and retrieve with the 'action filter' – examples

Content Moderation: The MAAM platform employs two moderation models to filter content, ensuring the platform remains safe and appropriate for all users. We trained two moderation models that are able to detect disturbing and Not Safe for Work (NSFW) content. For images, the moderation results, including a confidence score, are stored and indexed. For videos, the models are applied to key-frames, and a video is considered NSFW or disturbing if at least one scene receives such a tag. At retrieval time, users have the option to include or exclude that type of content, while the UI uses these tags to blur the corresponding assets to prevent user exposure to potentially inappropriate content. For example, at the left of Figure 33, the user has enabled the Disturbing filter to get only those that have been tagged as such. Users can choose to reveal the actual content at their own discretion by clicking the reveal button at the upper right corner of each asset, as it is depicted at the right of Figure 33.

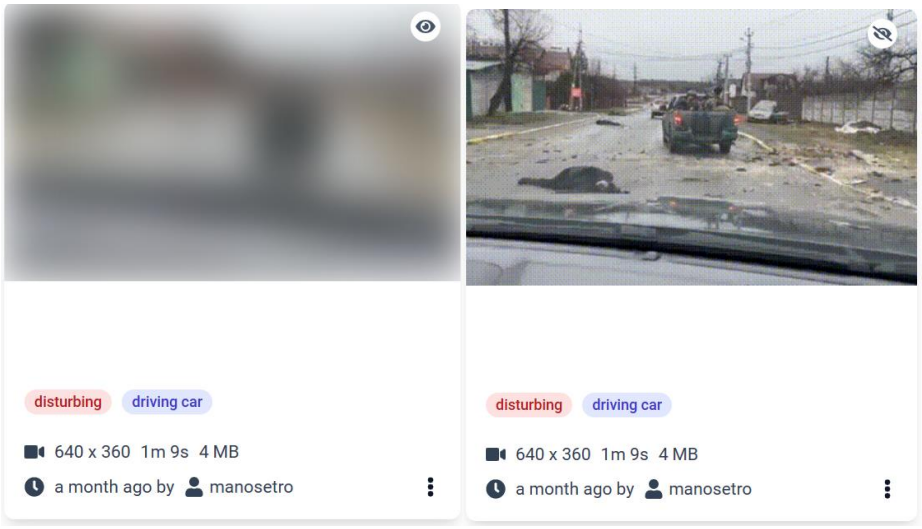


Figure 33: Asset with disturbing content - blur or reveal the actual content

Meme Detection: For images, we apply a meme detection model to each image to determine whether it is a meme or not. The result is stored and indexed. Figure 34 presents one image that has been categorised as a meme in the MedDMO workspace.

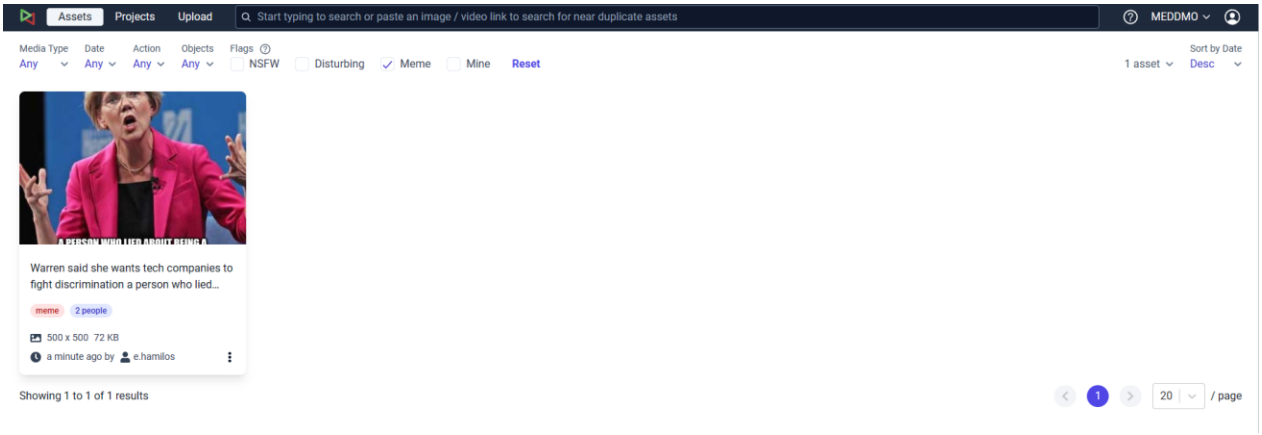


Figure 34: Meme example

Furthermore, the automatic multimedia analysis creates for each asset a table of properties, an asset description, and a metadata analysis table. The user can see the general properties in the asset details tab that appears when clicking an asset as illustrated in Figure 35,while, when clicking on the Metadata tab the user can browse through the complete image metadata as shown in Figure 37. The user can edit the description of an asset by clicking the edit button and providing a new asset description as depicted in Figure 36.

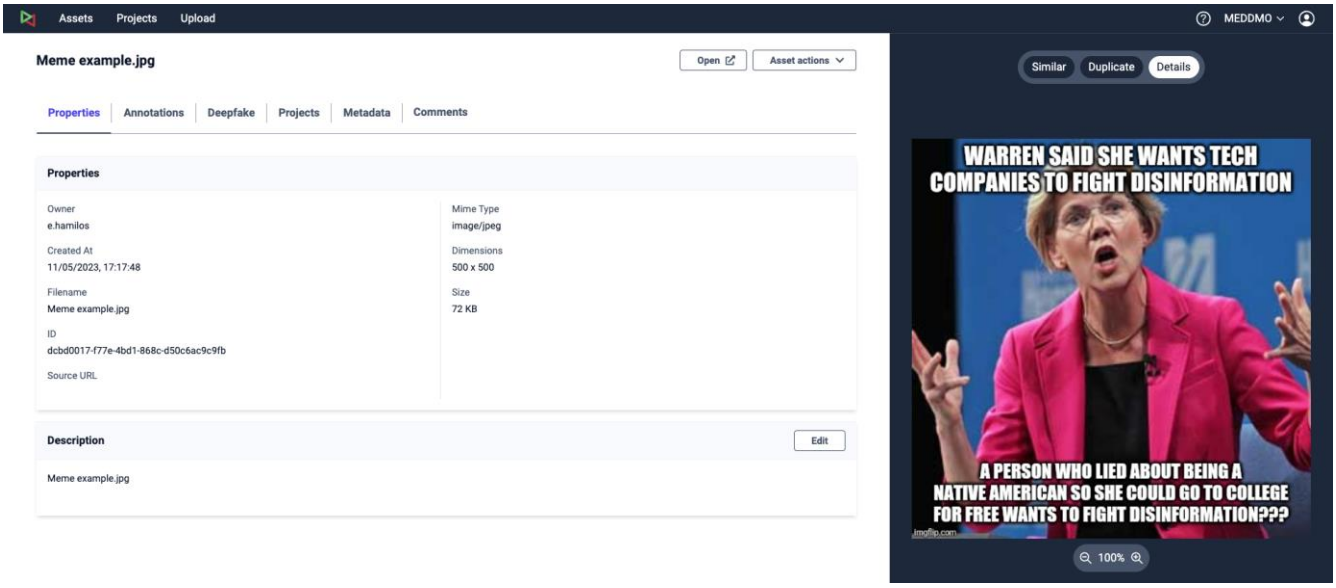


Figure 35: Asset details tab - Properties

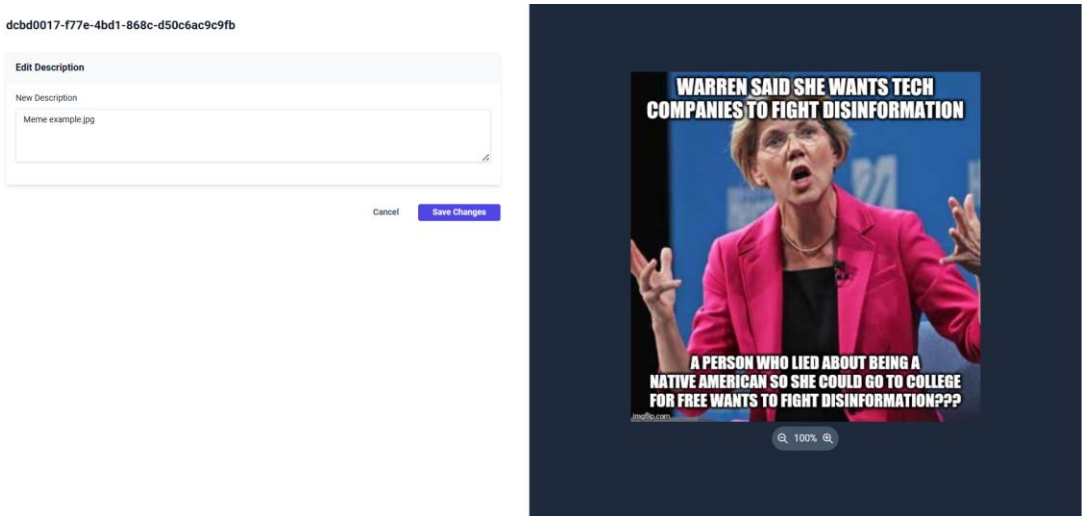


Figure 36: Editing the asset description

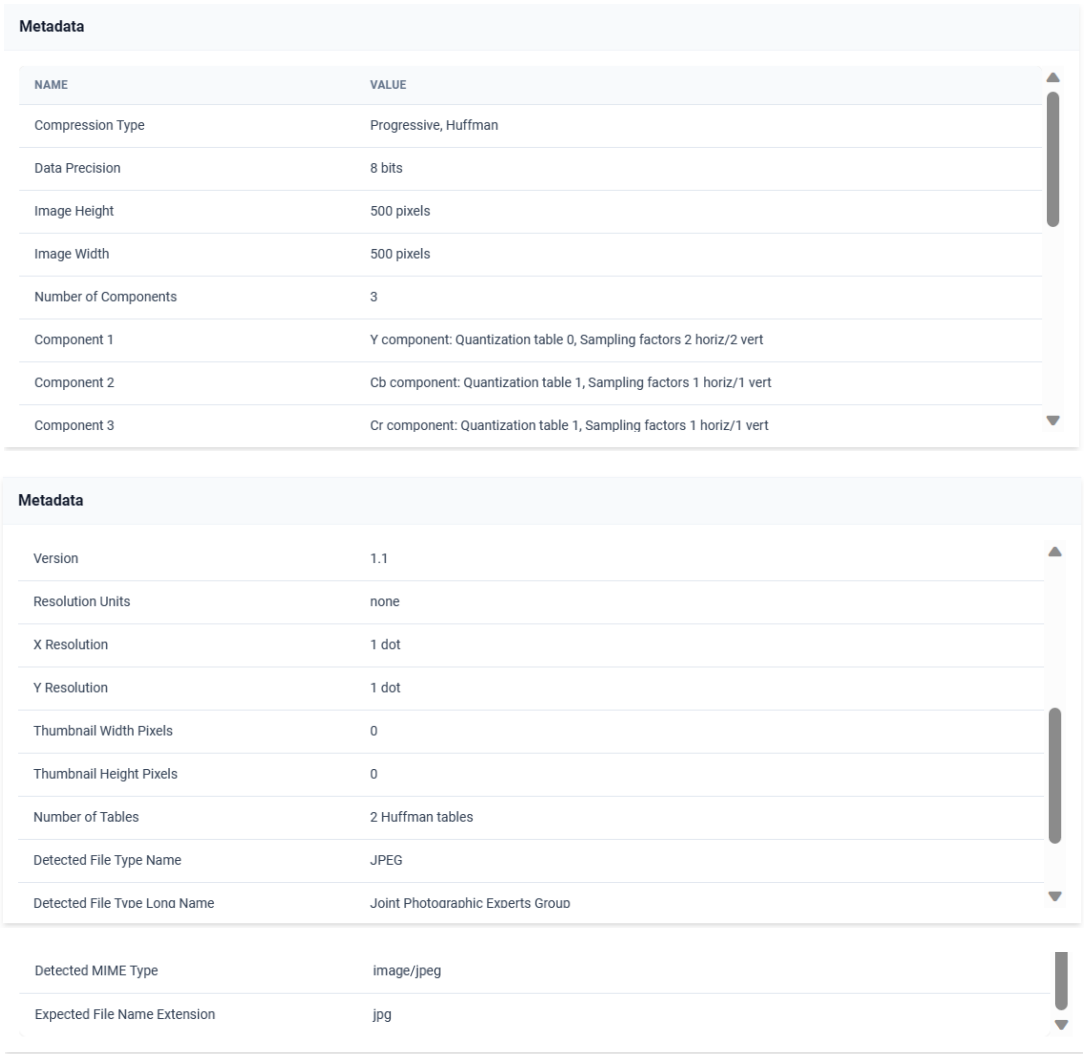


Figure 37: Metadata analysis table

3.2 Deepfake detection

The deepfake detection module is a tool that helps the user detect facial manipulations in images and videos. The tool can detect manipulations that are generated by many state-of-the-art deepfake generation approaches.

For an input image, the tool applies a face detection process and calculates a prediction for each detected face. Predictions closer to 0% mean that the corresponding faces are considered real by the model, on the contrary predictions closer to 100% indicate a manipulated face. The uncertainty of the model is increased for predictions closer to 50%.

In its current version, the tool supports JPEG, PNG, and TIF image formats and MP4, MOV, MKV, WEBM video formats. For an input image, the tool detects all faces on the frame and clusters them to face components. These components correspond to different people in the frame. For each face component, the deepfake detection model calculates a prediction. The final image prediction is an average among face component predictions. Below, there are two examples that show the results of deepfake analysis for two images. In Figure 38, we demonstrate an example with predictions closer to 0% which means that the corresponding faces are considered

real by the model, on the contrary in Figure 39, we illustrate another image with predictions closer to 100% which indicates a manipulated face.

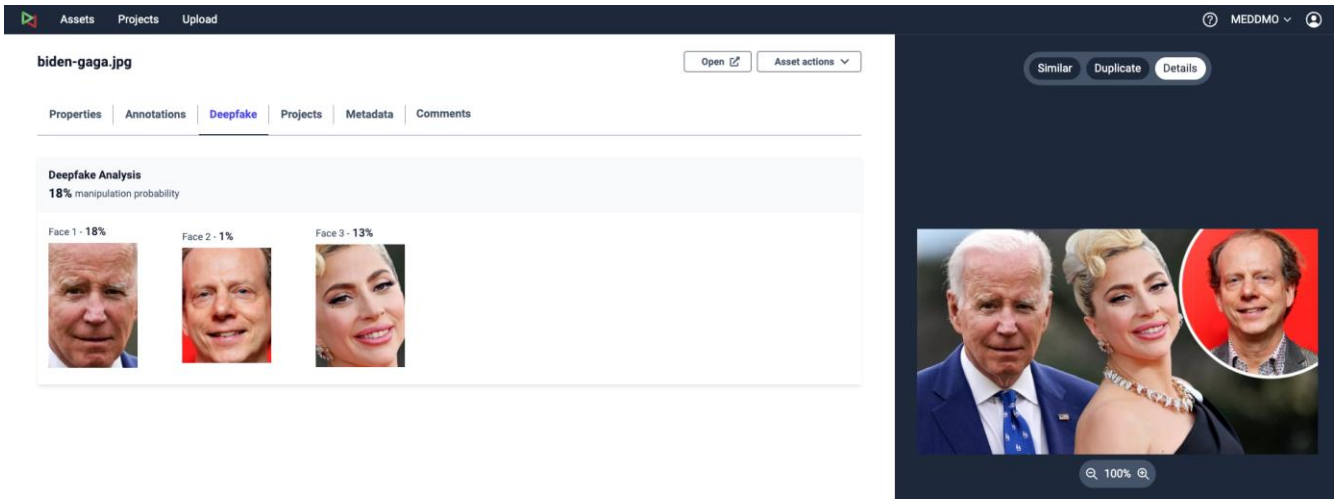


Figure 38: Deepfake detection prediction closer to 0%

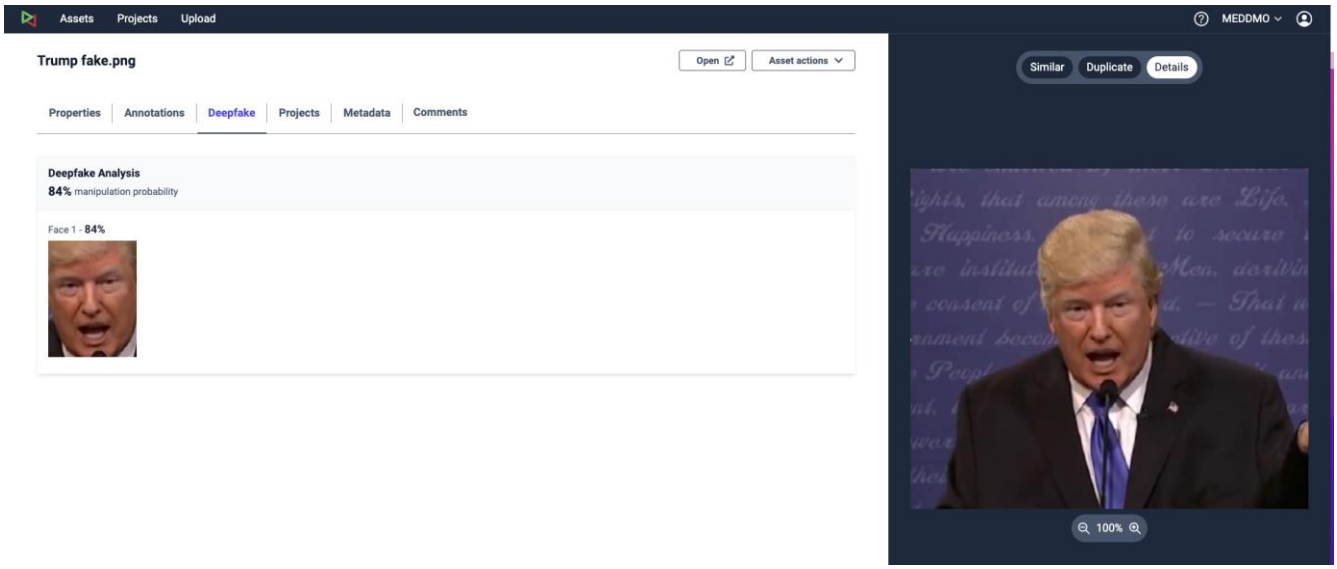


Figure 39: Deepfake detection prediction closer to 100%

For an input video, the tool applies a shot segmentation process that segments the video into parts that correspond to a continuous view filmed without interruption. For each frame in a shot, the tool detects all faces on the shot frames and clusters them to face components as depicted in Figure 41. These components correspond to different people in the shot. For each face component, the deepfake detection model calculates a prediction. The shot prediction is the maximum among face component predictions. The final video prediction is an average of the shot predictions.

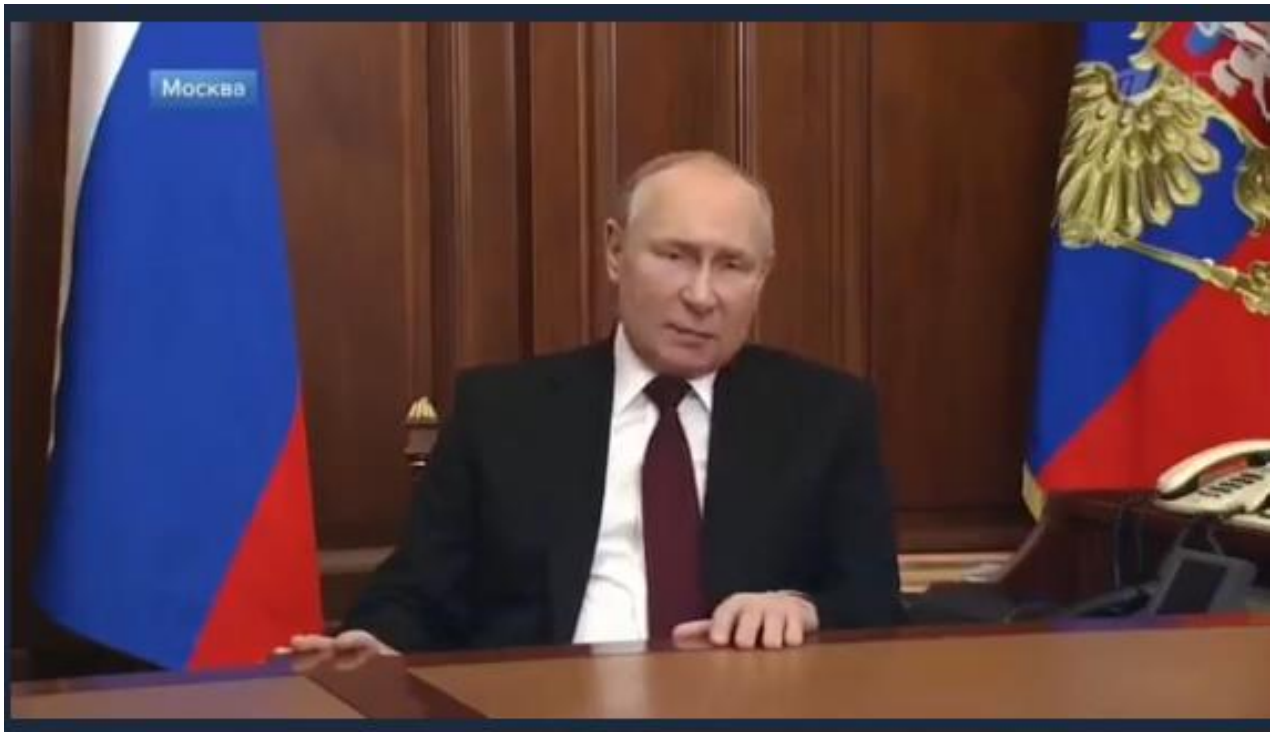


Figure 40: Video example with Putin having an interview



Figure 41: Deepfake detection in video and shot segmentation – one person



Figure 42: Video example with more than one person depicted in the video

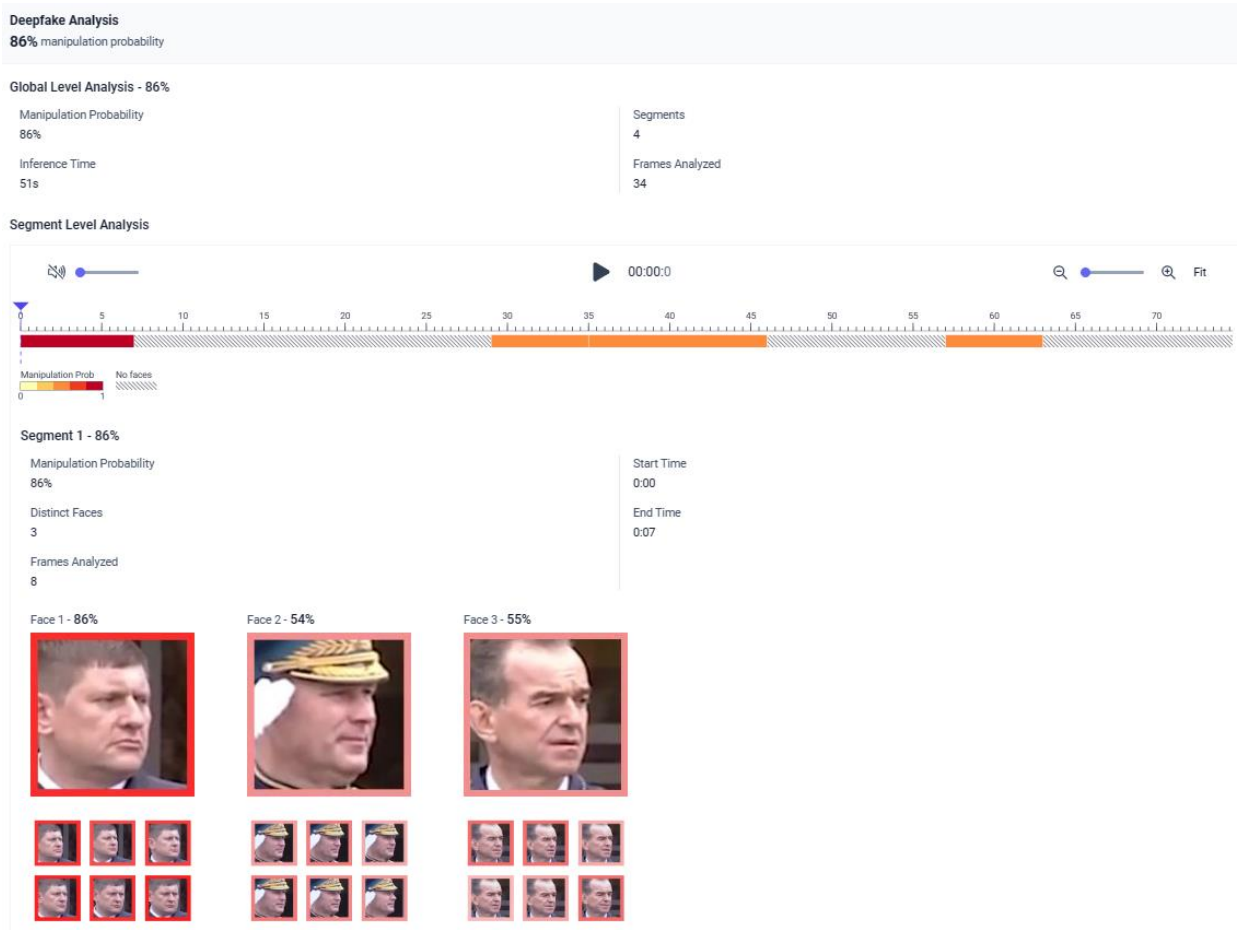


Figure 43: Deepfake detection in video and shot segmentation – more than one person

3.3 Image verification

The Image Verification Assistant is a tool that helps users detect various manipulations of the visual content in images. This is achieved through image analysis with various types of forensics algorithms to uncover possible forgeries along with explainable heatmaps which are overlayed on the image to highlight areas where tampering may have occurred. The tool supports JPEG, PNG, and TIF image formats and employs 15 different forensics algorithms along with a fusion algorithm that combines all outputs in a single heatmap.

The results of the image verification assistant are displayed under the asset details tab as illustrated in Figure 44.

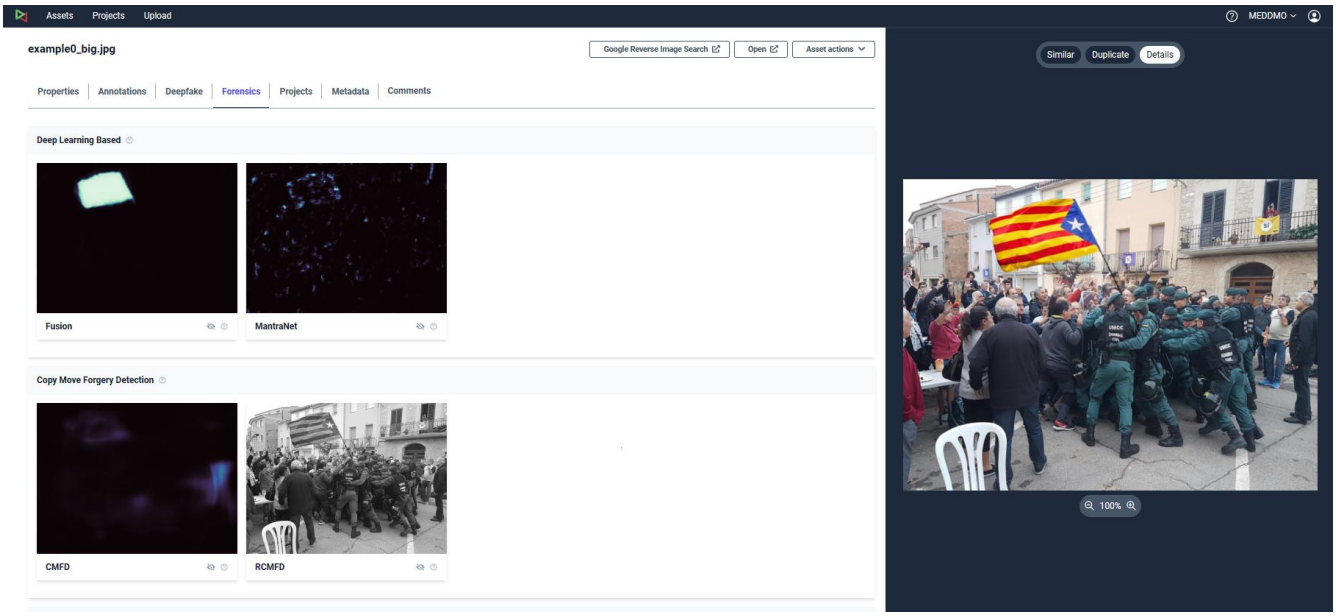


Figure 44: Image Verification Assistance – Heatmaps

The user can browse through the results of all forensic analysis on the left side of the screen while examining a single algorithm result in the right part by clicking on any heatmap.

The fusion algorithm combines all outputs from the forensics algorithms and produces a more easy-to-interpret localisation map as illustrated in Figure 45. In this example, the user can see that the fusion model has localised a forgery at the exact position of the flag (tampered areas appear as high values on a low-valued background). Furthermore, by clicking on the ‘’What does it mean?’’ button under any heatmap the user can read detailed information about the operation of each algorithm and have a better understanding of the visual output.

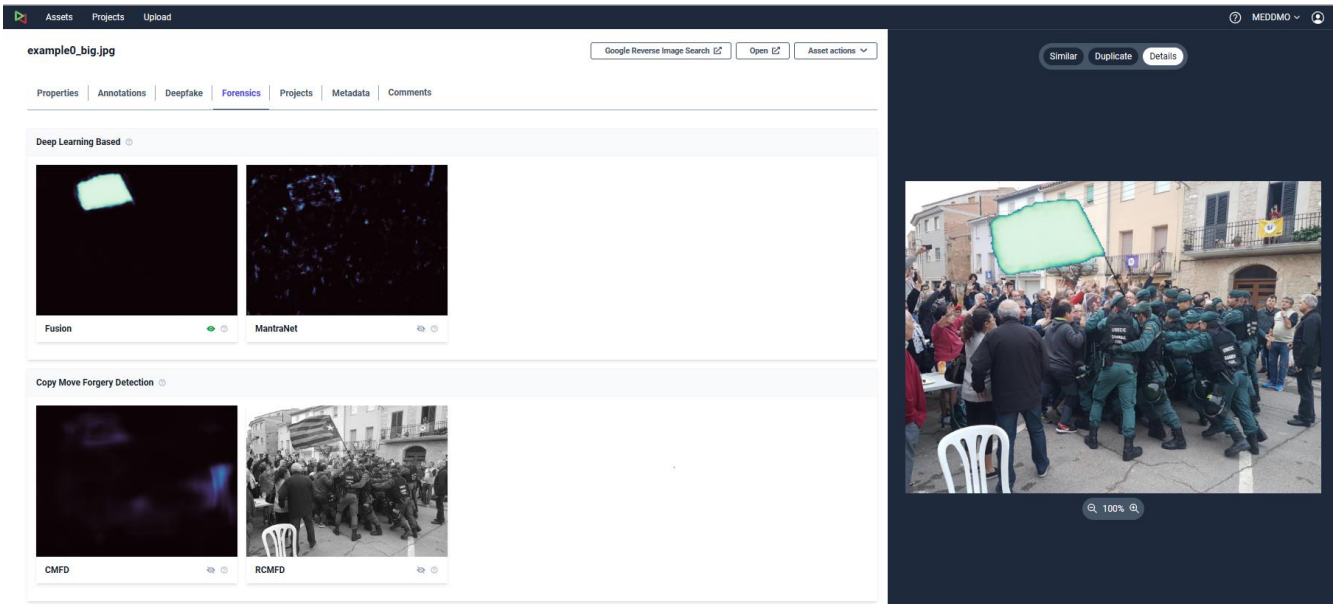


Figure 45: Fusion Heatmap

3.4 Location estimation

As there have been many cases where photographs have been presented out of context in order to mislead public opinion, someone must have some clues on the true location displayed in an image. The location estimation tool is a novel service, which, given an image, provides an estimate of its location solely based on the visual analysis of its content (e.g., based on landmarks or visual cues that are tell-tale of a specific region).

The estimated location information is available in the asset details. The tool returns the most likely location for the image visual content, shows the confidence level of the estimation, identifies the location name, shows a pin with the location in a Google map, and also displays some photos from the identified location as illustrated in Figure 46.

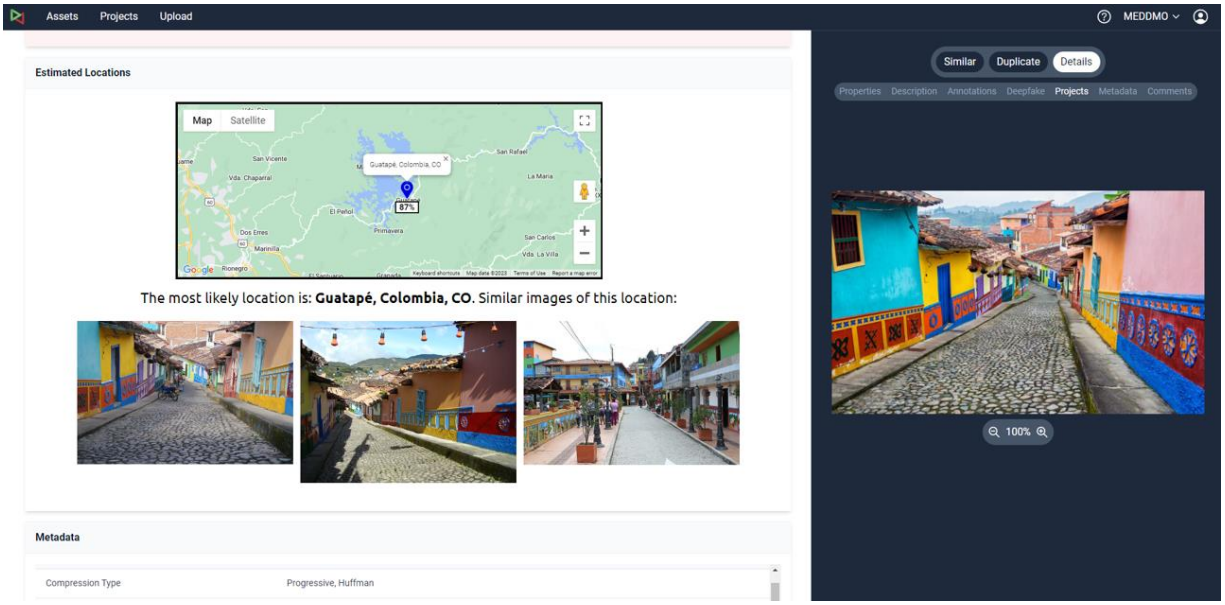


Figure 46: Location estimation results

3.5 Advanced search with visual similarity

Within the search operations described in Section 2.3 there are two operations that are based on visual similarity: the 'Find Similar' and the 'Find Near Duplicates', which are described in more detail in this section.

The 'Find Similar' tool allows users to find assets that are visually or semantically similar to the one under question. This tool is available either by clicking on the three dots at the lower right of any asset, or by clicking on the 'Similar' button, at the top right within the asset details tab (see Figure 47). The results of this example search are shown in Figure 48, which illustrate that this tool returns results that are similar both visually and semantically.

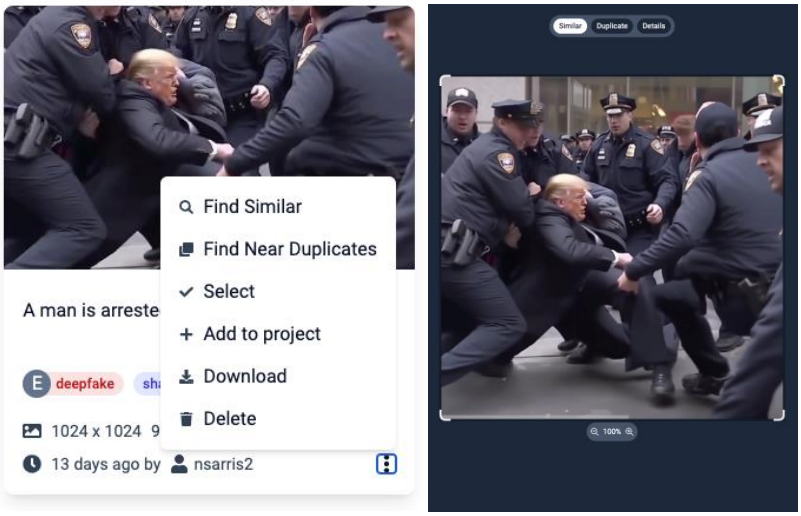


Figure 47: Searching for similar images with the 'Find Similar' tool

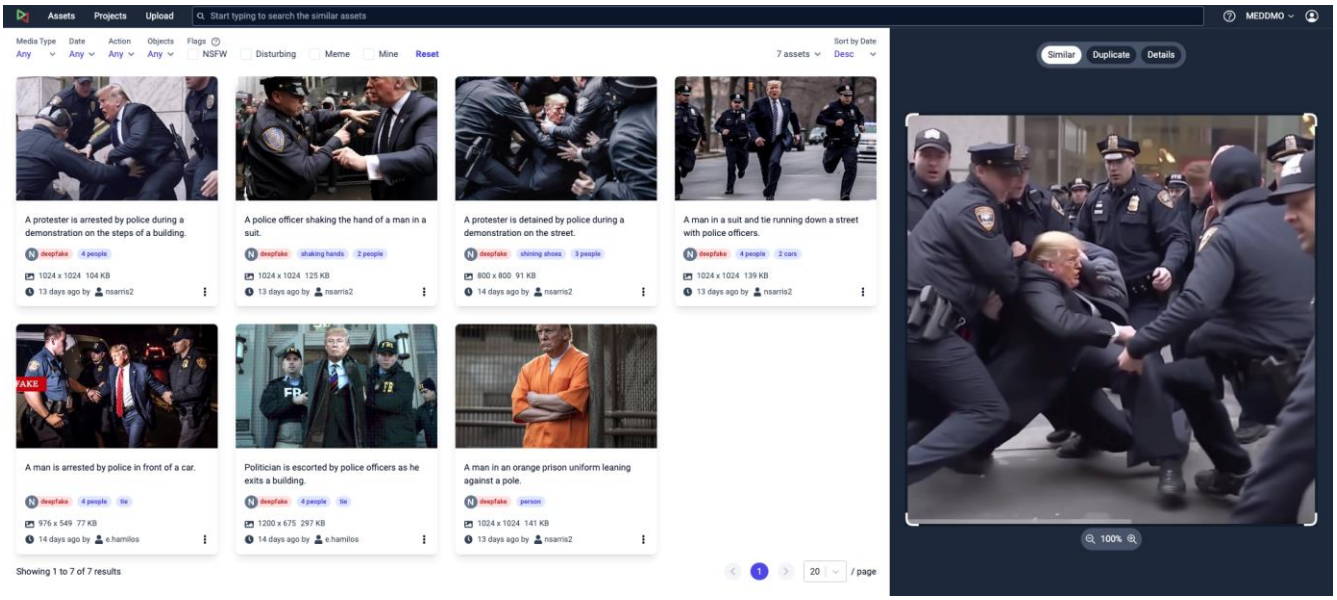


Figure 48: Find similar' tool results

The 'Find Similar' tool also supports region-based search operations. This allows users to select a specific region in an image by defining a bounding box, to retrieve content that is visually similar to that region in a more focused

manner. Figure 49, illustrates an example where the user selects a rectangular area containing a rainbow flag and retrieves results containing similar flags.

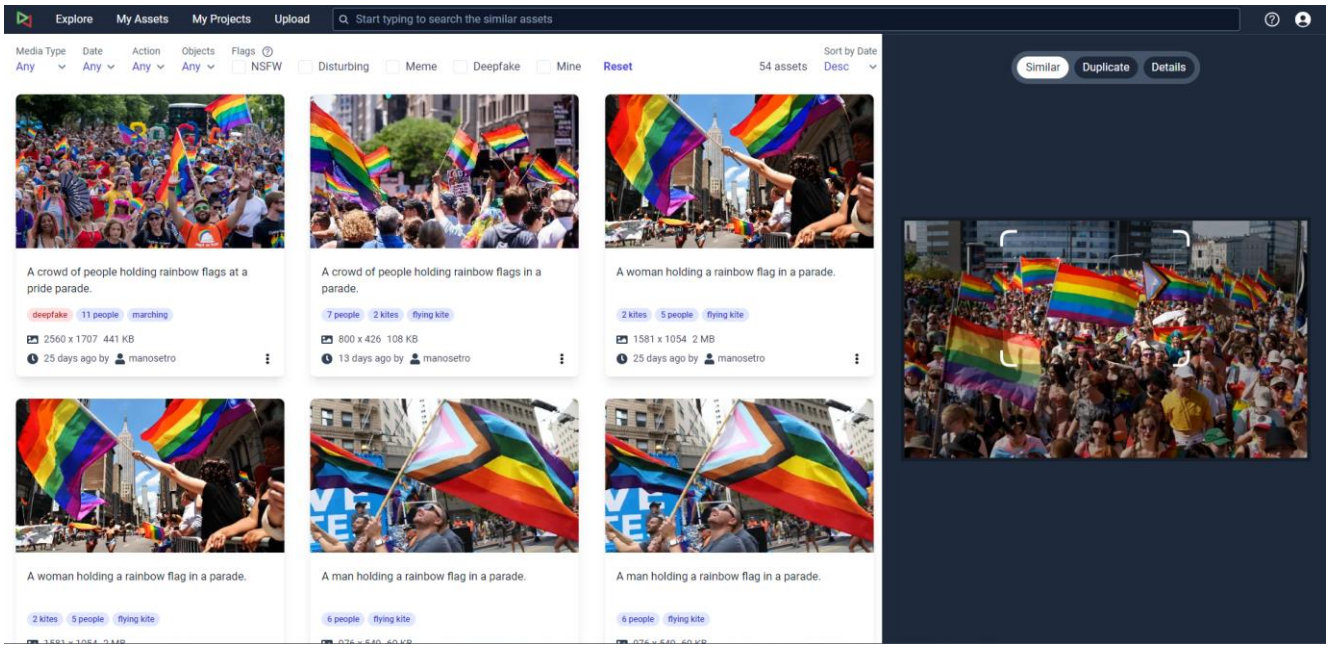


Figure 49: Region-based search operation with 'Find Similar' tool

The 'Find Near Duplicates' tool offers reverse search capabilities that enable users to find image or video assets that bear significant visual similarities. In the case of images, this tool will retrieve images that are the same but may have been cropped, resized, or overlayed with textual or other graphical content. In the case of video, users can find other videos that have some segments bearing significant visual similarities, including picture-picture cases.

As with the 'Find Similar' the tool is available either by clicking on the three dots at the lower right of any asset, or by clicking on the 'Duplicate' button, at the top right within the asset details tab. The results of an image duplicate example search are shown in Figure 50, which illustrates that this tool returns results that are visually similar but not necessarily identical, which allows in some cases to easily detect manipulations. In this case, although the Catalan flag has been pated on the original image the 'Find Near Duplicate' identifies the similarity.

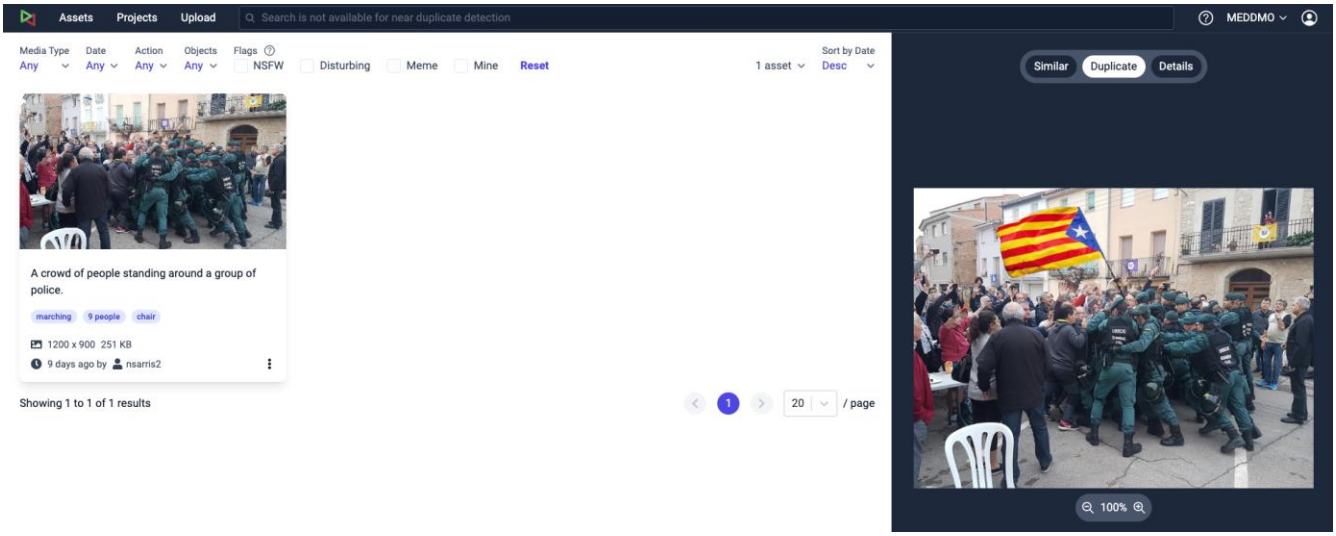


Figure 50: 'Find Near Duplicate' – image example

An example of a video case, where two videos that contain a common segment are found as near duplicates is shown in Figure 53.



Figure 51: Characteristic frames of the first video



Figure 52: Characteristic frames of the second video

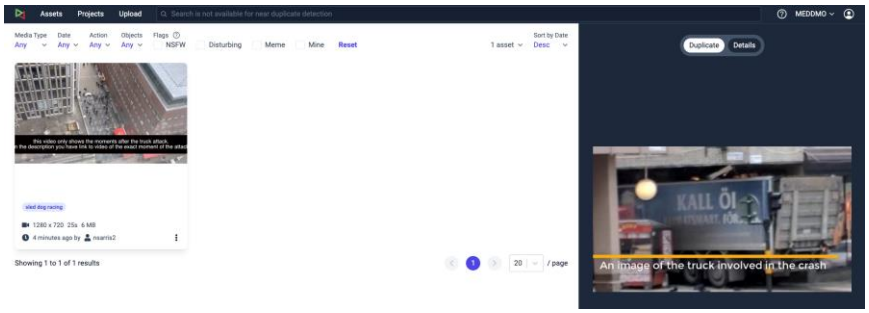


Figure 53: 'Find Near Duplicate' – video example

4 Conclusions

This deliverable describes the Multimedia Forensics Toolbox and the Knowledge Base that have been integrated into the Media Asset Annotation and Management (MAAM) platform providing a user-friendly interface to: store and retrieve fact-checked multimedia content, annotate image and video in manual and automatic ways, detect forgeries in image and video, and thus create a large and richly annotated database of multimedia content that can greatly assist in the work against disinformation.

This report has been assembled so that it may act as a detailed user manual for the community members of MedDMO, providing practical examples for every functionality.

This Multimedia Forensics infrastructure is expected to enable a significant upscale of research capabilities using technological tools, even though the main responsibility of disinformation research will remain in the hands of media researchers and fact-checkers, both from the consortium and the community.



Mediterranean Digital Media Observatory



MedDMO is a Digital Europe SME Support Action project co-financed by the EC under Grant Agreement with project ID: 101083756. The content of this document is © the author(s).
For further information, visit www.meddmo.eu