This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 787061

Advanced Tools for fighting Online illegal trafficking

D3.10 – Data Management Plan - Update

WP number and title	WP3 – Social, Ethical, Legal and Privacy issues of online sources analysis
Lead Beneficiary	CERTH
Contributor(s)	IIP
Deliverable type	ORDP: Open Research Data Pilot
Planned delivery date	31/10/2021
Last Update	05/11/2021
Dissemination level	PU

Disclaimer

The ANITA Consortium consists of the following partners:

Participant No	Participant organisation name	Short Name	Type	Country
1	Engineering Ingegneria Informatica	ENG	IND	IT
2	Centre for Research and Technology Hellas CERTH – ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	CERTH	RTO	GR
3	Centro Ricerche e Studi su Sicurezza e Criminalità	RISSC	RTO	IT
4	Expert System S.p.A.	EXPSYS	SME	IT
5	AIT Austrian Institute of Technology GMBH	AIT	RTO	AT
6	Fundacio Institut de BioEnginyeria de Catalunya	IBEC	RTO	ES
7	Istituto Italiano per la Privacy	IIP	NPO	IT
8	SYSTRAN SA	SYSTRAN	SME	FR
9	Stichting Katholieke Universiteit Brabant	TIU-JADS	RTO	NL
10	Dutch Institute for Technology, Safety & Security	DITSS	NPO	NL
11	VIAS Institute	VIAS	RTO	BE
Law Enforcement Agencies (LEAs)				
12	Provincial Police Headquarters in Gdansk	KWPG	USER	PL
13	Kriminalisticko-Policijska Univerzitet	UCIPS	USER	RS
14	Home Office	HO	USER	UK
15	National Police of the Netherlands	NPN	USER	NL
16	General Directorate Combating Organized Crime, Ministry of Interior	GDCOC	USER	BG
17	Local Police Voorkempen	LPV	USER	BE

To the knowledge of the authors, no classified information is included in this deliverable

Document History

VERSION	DATE	STATUS	AUTHORS, REVIEWER	DESCRIPTION
0.1	23/08/2021	Draft	IIP	First draft
0.2	25/10/2021	Final Draft	CERTH	Version ready for internal revision
0.3	27/10/2021	Final Version	CERTH, IBEC, EXPSYS	Final version, updated based on reviewers' feedback. Ready for SAB
0.4	31/10/2021	Final Version	CERTH	Update version based on SAB review
0.5	03/11/2021	Security check	Security Advisory Board	Documents considered as "not containing classified information" by Security Advisory Board
1.0	05/11/2021	Final	CERTH, ENG	Some minor changes. Document ready to be submitted

Definitions, Acronyms and Abbreviations

ACRONYMS / ABBREVIATIONS	DESCRIPTION
COVID-19	Coronavirus Disease 2019
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
JSON	JavaScript Object Notation
LEA	Law Enforcement Agency
ORD	Open Research Data
OSINT	Open-Source INTelligence
UC	Use Case
WP	Work Package

Table of Contents

List of Figures

List of Tables

Executive Summary

In EU-funded projects under Horizon 2020, a Data Management Plan (DMP) is necessary, especially for those projects that participate in H2020 Open Research Data Pilot (ORD Pilot). The present DMP is a document that outlines how research data will be stored after the project. ANITA is participating in ORD Pilot, therefore this DMP has been prepared to take this fact into account. Actions taken in order to participate in ORD Pilot have been reported in this deliverable.

This document constitutes the last update of the DMP of the ANITA project, i.e., it describes the set of data that will constitute the "legacy" of the project.

The objective of the ANITA DMP is to support the availability of the data used in order to carry out the research activity.

It will describe the datasets that it has to be made available and the ontology used to label the data. Datasets can be very diverse and of multiple typologies. On the one hand, we can have datasets of multimedia content, such as images or videos. On the other hand, we can have datasets of user feedback, ranging from signals such as the eye gaze to information such as the clicks they perform, as well as questionnaires regarding their experience. Initially, we joined efforts to form datasets from online and offline routes, more specifically many GBs of data have been crawled from the web and successfully examined by technical partners to drive our research. Apart from these, we have worked on datasets from user feedback. Capturing multiple types of user feedback can be extremely useful for multiple purposes. Understanding better how users interact with a system so that it can be improved. Improve artificial intelligence systems to better match human capabilities. Obtaining new scientific knowledge about human cognition and behaviour. We conducted multiple experiments collecting physiological signals, such as their eye gaze the pupil dilation, while they were exposed to different stimuli. Different datasets were produced from these studies, including eye activity (both gaze and pupil dilation) when visually scanning multiple images.

This deliverable will summarize these efforts, and report the concrete decisions regarding the possibility to provide open access to the datasets, their sharing and re-use by third parties.

1 Data protection perspective

Within the ANITA project, the datasets used did not, by design, contain personal data. In fact, in order to train the algorithms feeding ANITA, it was not deemed necessary to use personal data.

Within the processed datasets, only an extremely small amount of data that could potentially be considered as personal data is present. In this sense, some reflections of a legal nature were carried out in order to implement the principles of data protection by design and by default.

To respect general fundamental principles and to ensure valid legal grounds/lawfulness conditions of data processing could not be enough. A general rule, coming from Articles 13-14 of the GDPR, imposes to adequately inform data subjects about their personal data processing. This is the “privacy notice” or “privacy policy” that we are used to receive as data subjects and users, almost in all cases our data are going to be collected and processed (some exceptions occur, for instance, just in justice/law enforcement scenarios).

Collecting data from the (dark, deep, surface) web, such as collecting data from publicly available records and documents, implies that data are not obtained directly from the data subject. In this case (collection not from the data subject, but third parties), Article 14.5.b) of the GDPR applies and justify at certain conditions even the exclusion of the obligation for the data controller to inform data subjects, where and insofar as “the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases, the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available.”

In conclusion, the possible inclusion of personal data within the datasets may not be taken into account. Nevertheless, the publication of the datasets will take into account the presence of such data in order to exclude their dissemination. It should also be noted that none of the Uses Cases considered, nor the ontologies used to classify the data, have taken into account the use of personal data, so the possibility that personal data may be subject to further data processing activities is excluded.

2 ANITA Open-Access Datasets

Datasets can be very diverse and of multiple typologies. On the one hand, we can have datasets of multimedia content, such as images or videos. On the other hand, we can also have datasets of user feedback, ranging from signals such as their heart rate to information such as the clicks they perform, as well as questionnaires regarding their experience.

In order to train efficient analysis algorithms and to develop robust investigation tools the following datasets were formed, during the ANITA project:

- GAze on Target (GATA) dataset
- TEll me you SEe (TESE) dataset
- Polish places dataset
- COVID-19 Affective Ratings dataset

The datasets were designed and produced to fulfil the requirements for the whole ANITA project and more specifically technical tasks from WP6 and WP8 as they derived from the Use Cases analysis. The following subsections provide a brief technical description of the datasets to enable their utilization in future projects and experiments, as well as decisions made regarding the possibility to provide open access to the datasets, their sharing and re-use by third parties.

2.1 GAze on Target (GATA) dataset

Capturing multiple types of user feedback can be extremely useful for multiple purposes. Understanding better how users interact with a system so that it can be improved. Improve artificial intelligence systems to better match human capabilities. Obtaining new scientific knowledge about human cognition and behaviour.

DESCRIPTION

GAze on TArget (GATA) dataset is a large-scale annotated gaze dataset, tailored for training deep learning architectures. It was created following the “target search” paradigm where subjects were asked to visually search for a specific object class. Forty-eight different subjects participated in the recording procedure using myGaze capturing sensor.

Figure 1: Gaze annotation process

In order to create an annotated gaze dataset, a simple interface (visual stimulus) that included 6 images (centre-aligned in two rows), denoted as image-group session, was projected on a 23''monitor. The gaze sensors capturing frequency was 30Hz. Additionally, a wireless mouse sensor was handed to the user, in order to freely indicate the beginning and the end of the capturing image-group session. The utilized images constituted general-purpose ones and were randomly selected from the publicly available COCO dataset[1]. Before the beginning of each image-group session, the human subject was given a keyword (object type) as a query term and was subsequently asked to search for instances of this category of objects in the depicted images. The set of supported objects consisted of 80 types of everyday ones, such as persons, cars, cats, etc. Each subject was positioned approximately ~ 60-70cm away from the computer monitor, as depicted in Fig.1, and was informed about the capturing procedure prior to the execution of the experiments. For each human subject, the sensor was calibrated to the subject's eyes, before any gaze capturing took place. Moreover, each human subject underwent subsequent capturing image-group sessions (i.e., Different sets of 6 images included in the projected interface with different query terms defined each time), where the total duration of all sequential capturing image-group sessions did not exceed the limit of 15 −20 minutes (which corresponded to a targeted number of approximately ~ 85 image-group sessions per experimental session). Overall, forty-eight (48) individual subjects, 41 males and 7 females were involved in the gaze recordings, ageing from 22 to 45, while a total number of 238 experiments.

The assembled dataset comprises a large-scale benchmark of approximately 120.000 object instances with associated gaze signal captured, given a query object class. In general, the COCO dataset includes images with a varying number of objects with different sizes and occlusion levels. Therefore, the created dataset comprises a challenging one for interpreting the human gaze signal and also investigating its possible integration in image analysis methods. In terms of gaze characteristics, the minimum, maximum and average number of gaze points (captured at a frequency of 30Hz) per image are **1, 1.842 and 42.251**, respectively. In terms of fixations, the corresponding numbers are **1, 104** and **3.166**, respectively. Indicative examples of the projection of the captured gaze signal on the corresponding COCO image are presented in Figure 2. The proposed dataset was utilized for building a deep learning model capable of predicting objects in an image as relevant or non-relevant, based on gaze, according to the users' preferences.

Dataset Statistics	
Base dataset	MSCOCO 2014 dataset
Number of gaze annotations	120000
Number of object classes	80
Number of subjects who participated	48 (41M/7F)
Number of experimental sessions recorded	238
Average execution time	20min
Number of image-groups per session	85
Number of images per image-group	6
Number of total image-groups recorded	20000

Table 1: GATA dataset statistics

[1] https://cocodataset.org/

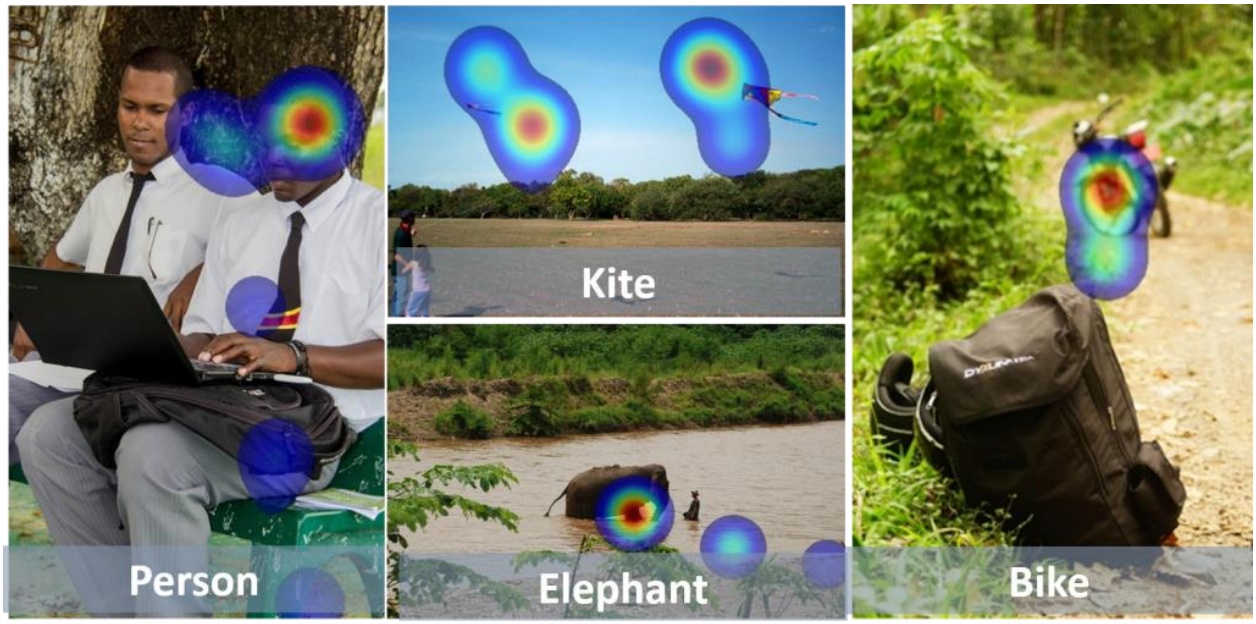

Figure 2: Visualization of the implicit human response (various classes heatmaps)

Open Access Format

The gaze annotations are provided with the JSON file format using the following naming convention.

objectID_imageID.json

The first part (**objectID**) denotes the target object class id and the second part (imageID) the COCO image id respectively. Inside the JSON file gaze points annotations, timestamp and x,y coordinates, are included as presented below.

[{"Time":12510830749,"X":343,"Y":285},
{"Time":12510864083,"X":343,"Y":285},
{"Time":12510897394,"X":343,"Y":285},
{"Time":12510930745,"X":343,"Y":286},
{"Time":12510964081,"X":343,"Y":287}]

Key Publication

- Stavridis, K., Psaltis, A., Dimou, A., Papadopoulos G. Th., & Daras, P. (2019). Deep Spatio-Temporal Modeling for Object-Level Gaze-Based Relevance Assessment. In 2019 27th European Signal Processing Conference (EUSIPCO). IEEE.

Open Access Link

https://www.kaggle.com/athanasiospsaltis/gaze-on-target-dataset-gata

2.2 TELL me what you See (TESE) dataset

DESCRIPTION

TEll me what you SEe (TESE) dataset is a gaze annotated dataset, tailored for training deep learning architectures and extracting useful viewing patterns. It was created following the "description" paradigm where subjects were asked to orally describe visual scenes. Fifty-six different subjects participated in the recording procedure using Tobbi Pro eye tracker sensor. The publicly available Visual Genome (VG) dataset[2] is a generic content image database with object bounding box, class and relation dense annotations. Therefore, it was selected as the image pool in order to enable comparisons with other state-of-the-art methods in the field. In the preliminary analysis phase, 10.000 images out of about 108.000 of VG were picked in a way that satisfies the requirements of 80\% representation of the relation classes and randomness. During the second phase, 5.000 out of 10.000 were manually selected based on the following criteria: a) objects and relations to be easily identified by the human annotator, and b) the number of objects present in an image to be equal or greater than 4. By this, it was ensured that each image stimulus included enough information to trigger a sufficient description from the subjects. VGis densely annotated using a crowd-sourcing method. In addition, the annotator is not restricted to choose from a pre-defined closed set of classes and can determine the classes of objects and relations with free natural language. In this way, a plethora of unique classes are present in the dataset. However, many of them refer to the same conceptual entity using different `lexicalizations' (synonym words). Regarding the selected set of **5.000** images, there are 2.912 unique object categories. Therefore, normalizing and reducing the number of object classes was required.

- At the first step of normalization, a reduction scheme based on frequencies of appearance was applied. The object categories with a frequency of appearance equal to or greater than 10 were kept, reducing the number of unique categories to 1.842.
- In the second phase, semantically similar classes were merged in one. In this way, the number of unique object classes was further reduced to 645. The mapping scheme, from old to new object classes, was applied to the whole VG dataset to enable training with more than 5.000 images. Objects with classes that were out of the scope of the mapping scheme were filtered out.

During the viewing process, every subject provided annotation on an average of 67 images, while each image was projected on the screen on an average of 30 seconds until the subject moved forward to the next image. The latter resulted in the generation of approximately 2266 gaze points for each image, and a total of 61 fixations. A cumulative fact sheet of our experiment can be found in Table 2. The formed dataset contains the following information sources: a) gaze point annotation in the form of gaze fixations and b) speech to text image captioning data. Indicative examples of the projection of the captured gaze signal on the corresponding Visual Genome image are presented in Figure 4. The proposed dataset can be used but not limited to scientific fields of neuro-science, knowledge extraction from humans, object detection, visual recognition and knowledge enhanced deep learning.

[2] https://visualgenome.org/

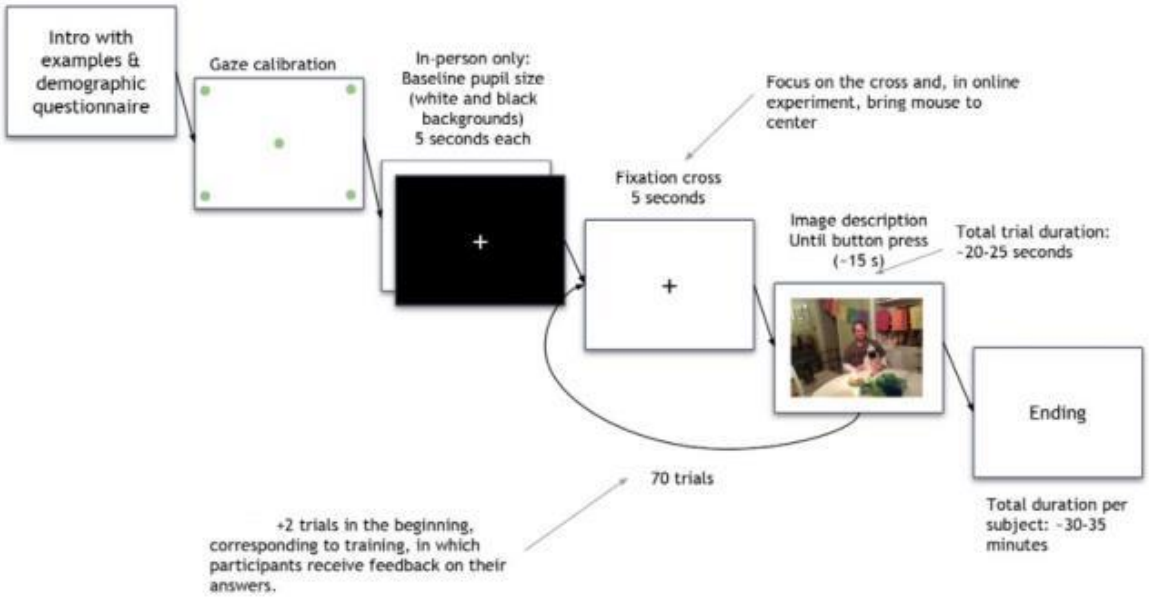

Figure 3: The TESE experimental protocol

Dataset Statistics	
Base dataset	Visual Genome dataset
Number of gaze annotated images	3765
Number of object classes schemes	80/645/1000/3000
Number of subjects participated	56 (45M/11F)
Average number of trials per subject	67.23
Average number of gaze points per image	2266,26
Average number of fixations per image	61

Table 2: TESE dataset statistics

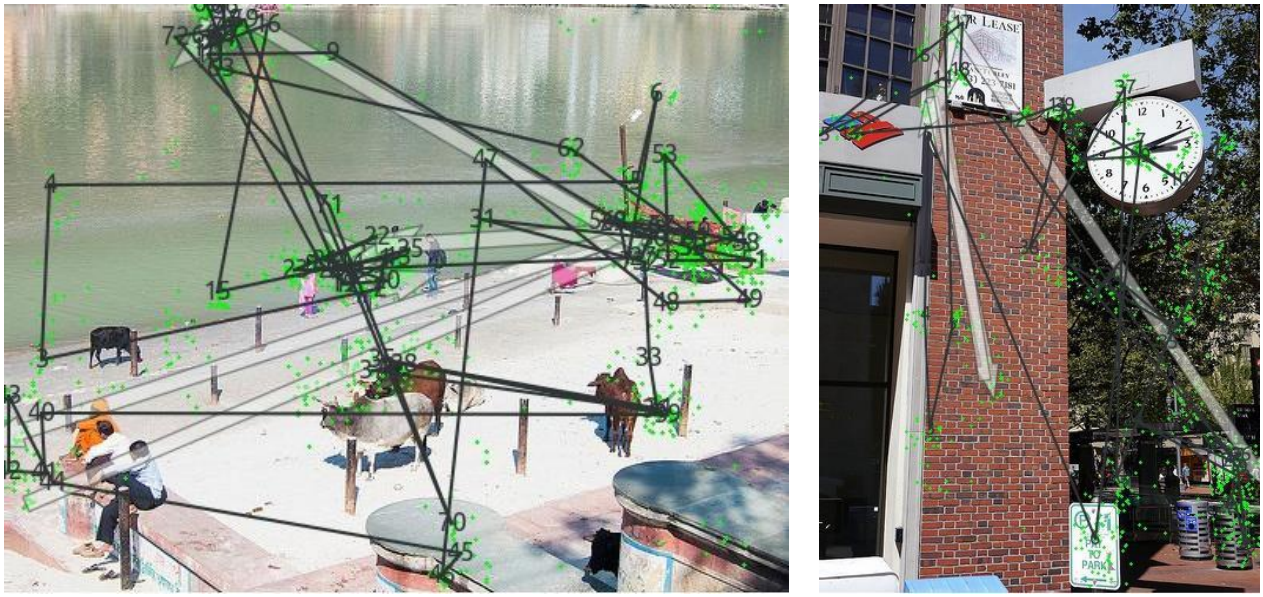

Figure 4: Fixation scan-paths

Open Access Format

The gaze annotations are provided with the JSON file format using the following naming convention.

imageID.json

The **imageID** denotes the Visual Genome image id which is actually the filename of the image. Inside the JSON file gaze points annotations, timestamp and x,y coordinates and transcribed text from voice description are included as presented below.

{“subjectID”:1,
“gaze”:[{“Time”: 6105.218916422819, “X”: 0.5127838850021362, “Y”: 0.4649979770183563},
{“Time”: 6105.2274152988775, “X”: 0.507378339767456, “Y”: 0.4624013900756836}]
}
In addition, the subjects.json file provides information about the subjects as described below.
[{“id”: “1”, “gender”: “Male”, “age”: “38”},
{“id”: “2”, “gender”: “Female”, “age”: “29”}]

Open Access Link

https://www.kaggle.com/athanasiospsaltis/tell-me-what-you-see-datasetZ

2.3 Polish places dataset

DESCRIPTION

The Polish places dataset consists of images captured from several places in Poland. 30000 images were collected from the internet using the google search engine. The collected dataset was analyzed manually. Several duplicate images were detected and removed among different categories. Moreover, downloaded images that had a name denoting a place different from its category were also removed. Images with irrelevant content (other than places) were removed as well. After the above annotation procedure, 6781 valid images remained for the fine-tuning process. The introduced dataset can be used for image classification/geo-localization and retrieval tasks.

Class ID	Classes (location)
1	Bartoszylas POLAND
2	Białka Tatrzańska POLAND
3	BORSK POLAND
4	Gdańsk street Okopowa POLAND
5	Gdynia POLAND
6	Konarzyny pomerania region POLAND
7	Kościerzyna pomerania region POLAND
8	Malbork POLAND
9	Owidz pomerania region POLAND
10	Poronin street Krośne Hamry POLAND
11	sHertogehbosh HOLLAND
12	Szczytno street Spacerowa POLAND
13	Szklarska Poręba mountain house POLAND

Table 3: List of Polish Places concept categories

Figure 5: Example images from the collected Polish places dataset

Dataset Statistics	
Image origin	Google images
Number of images annotated	6781
Number of concept classes (places)	13
Average number of images per class	521

Table 4: Polish places dataset statistics

Open Access Format

The dataset annotations are provided with the JSON file format using the following name.

polishPlacesAnns.json

The image **id**, the **class** and the **filename** are included as presented below.

[{"id":1, "filename": "image1.jpg", "class": 8}
{"id":2, "filename": "image18.jpg", "class": 10}
{"id":3, "filename": "image158.jpg", "class": 3}]

Open Access Link

https://www.kaggle.com/athanasiospsaltis/polish-places-dataset

2.4 COVID-19 Affective Ratings

Description

Dataset of affective ratings of images, obtained in the context of the COVID-19 pandemic, and in particular the lockdown situation that took place in Spring of 2020. The dataset consists of three types of information: demographic data, affective ratings, and answers to a questionnaire related to subjective experiences during the lockdown period. The affective ratings comprise a total of 46 images, obtained from the OASIS (Open Affective Standardized Image Set[3]) dataset, rated for arousal and valence using the Affective Slider[4] (also with timing information). The dataset comprises 434 JSON files corresponding to the data of over 110 participants. The data was obtained between the 9th and the 20th of April of 2020.

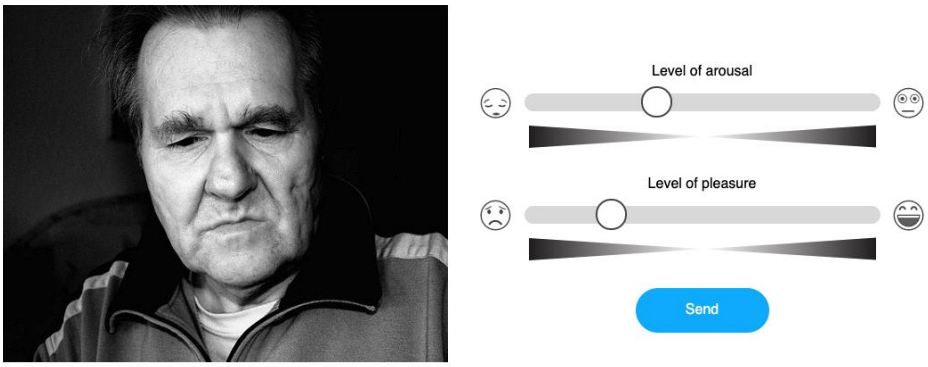

Figure 6: An example image of the COVID-19 Affective Ratings dataset

Dataset Statistics	
Image origin	OASIS
Number of images used	46
Number of participants	110
Number of feedback files collected	434

Table 5: COVID-19 Affective Ratings dataset statistics

Open Access Format

The demographics, questionnaire, and ratings are provided with the JSON file format using the following naming convention:

SessionID_ demographics.json

SessionID_ questionnaire.json

SessionID_ ratings.json

[3] Kurdi, B., Lozano, S. & Banaji, M.R. Introducing the Open Affective Standardized Image Set (OASIS). *Behav Res 49*, 457–470 (2017). https://doi.org/10.3758/s13428-016-0715-3

[4] Betella A, Verschure PFMJ (2016) The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE 11*(2): e0148037. https://doi.org/10.1371/journal.pone.0148037

An indicative example of user demographics and questionnaire responses can be seen in Figure 7 and Figure 8, while the affective image used and values of **Valance** and **Arousal** extracted in Figure 9 below.


```
"root" : { 1 item
  "0" : { 7 items
    "userAgent" :
    string "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/13.1
    Safari/605.1.15"
    "language" : string "en"
    "gender" : string "female"
    "age" : string "34"
    "education" : string "3"
    "countryFrom" : string "PL"
    "countryIn" : string "PL"
  }
}
```


Figure 7: Example of user demographics


```
"root" : { 1 item
  "0" : { 13 items
    "q1" : string "yes"
    "q2" : string "yes"
    "q3" : string "4"
    "q4" : string "yes"
    "q5" : string "children"
    "q6" : string "0.58"
    "q7" : string "0.04"
    "q8" : string "no"
    "q9" : string "0.71"
    "q10" : string "0.2"
    "q11" : string "1"
    "q12" : string "0"
    "q13" : string ""
  }
}
```


Figure 8: Example of a questionnaire JSON file

Figure 9: Example of user ratings

Open Access Link

https://www.kaggle.com/hectorlopezcarral/covid19-affective-ratings

3 ANITA Ontology

3.1 Illegal Trafficking Ontology

DESCRIPTION

In this section we will discuss the structure and the different components of the final ontology developed, modelling the illegal trafficking. Starting from the concept maps representing the main features of the topic (in fact, we have defined a different concept map on the several subdomains of interest and all of them has been considered as a base for the global ontology), entities and relationships are presented in tabular form, which fields are defined according to Protégé notation and they are reported below.

In order to develop the required ANITA tools and services, the collected knowledge has been modelled as an OWL ontology; the key-concepts are Class, Object Property and Data Property. Therefore, all the elements included in the domain knowledge can be represented in the adopted standard format: entities are modelled as classes, relationships as object properties, attributes as data properties.

Any ontology can be exhaustively described by its taxonomies, which are hierarchies of concepts. In particular, the most meaningful one is the Class Hierarchy: it is more relevant than the others because it is easily understandable.

Class	SubClass
Activity	• Dissemination • Donation • Finding • Production • Sale • Shipment
Actor	• Organization • Person
Ideology	
Money	

Product	• Accessory
	• Precursor
	• Substance
	  ○ Drug
	  ○ Medicine
	  ○ NPS
	  ○ Other
	• Weapon

Shop	• CryptoMarket • eShop • SingleVendor

Table 6: Classes and Subclasses table

Object Property	Domain	Range	Characteristics
hasActivity	Actor	Activity	Inverse Functional
hasActor	Activity	Actor	Functional
hasAdministration	Actor	CryptoMarket	Inverse Functional
hasAvaiability	Product	Shop	
hasConnection	Activity	Product	
hasExpression	Dissemination	Ideology	
hasProduct	Actor	Product Money	Inverse Functional
hasRelatedProduct	Accessory Precursor	Product	
hasSite	Activity	Shop	Inverse Functional
hasTransfer	Donation	Money Product	

Table 7: Object properties table

Data Property	Domain	Range (Data Type)
accessoryType	Accessory	rdfs:Literal
Channel	Activity	rdfs:Literal
Concealment	Shipment	rdfs:Literal
Currency	Money	rdfs:Literal
currencyType	Money	rdfs:Literal
deadDrop	Person	xsd:boolean
drugType	Drug	rdfs:Literal
Effect	Substance	rdfs:Literal
eShopType	eShop	rdfs:Literal
Format	Substance	rdfs:Literal
ideologyType	Ideology	rdfs:Literal
intakeMethod	Substance	rdfs:Literal
isCamouflaged	eShop	xsd:boolean
Keyword	Product	rdfs:Literal
medicineType	Medicine	rdfs:Literal

Mode	Finding	rdfs:Literal
Nature	Donation	rdfs:Literal
paymentMethod	Sale	rdfs:Literal
productiomMode	Production	rdfs:Literal
productionPlace	Production	rdfs:Literal
Role	Actor	rdfs:Literal
Service	CryptoMarket	rdfs:Literal
shipmentFrom	Shipment	rdfs:Literal
shipmentID	Shipment	rdfs:Literal
shipmentMethod	Shipment	rdfs:Literal
shipmentTo	Shipment	rdfs:Literal
Source	Shop	xsd:anyURI
terrorismFunding	Activity	xsd:boolean
Tipology	Dissemination	rdfs:Literal
weaponType	Weapon	rdfs:Literal

Table 8: Data properties table

As stated in the description step of modelling process, the tables just presented make it easy to construct the ontology. Its taxonomy of classes and subclasses is the following:

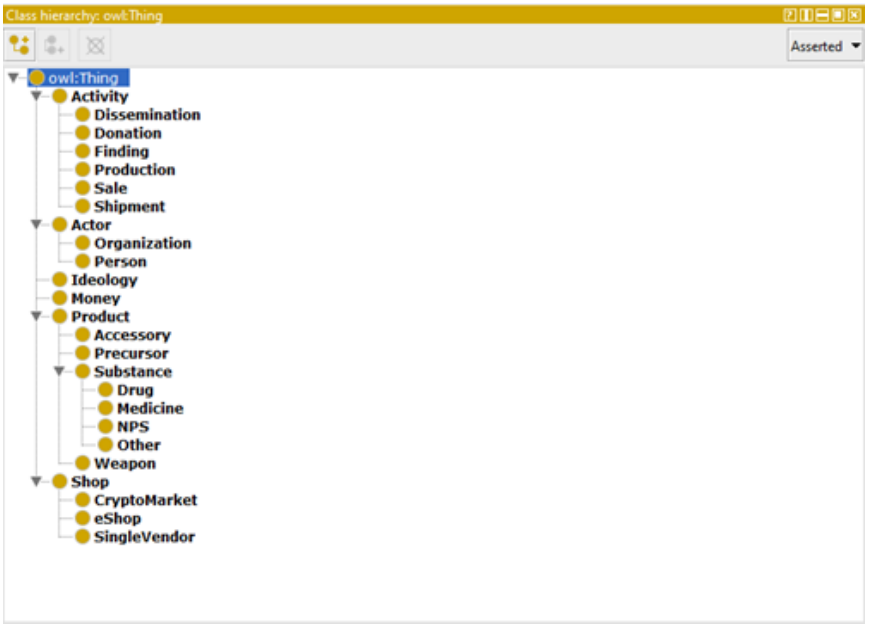

Figure 10: Classes taxonomy

To clarify the meaning of all classes and their corresponding subclasses, some specifications are proposed below.

- ***Activity*** *Class* models the concept of illegal event/activity; as its main properties, two attributes have been defined: the channel, on which the event actually takes place, that admits Surface Web, Deep Web, Dark Web as values; the purpose: setting the terrorism funding attribute (of boolean type) as TRUE we mean that the event has the purpose of financing terrorism, while it will be set the same attribute as FALSE otherwise.

- ***Dissemination*** *Class* (***Activity*** *SubClass*) models the concept of information dissemination, through advertising and/or propaganda.

- ***Donation*** *Class* (***Activity*** *SubClass*) models the concept of any product (like weapon or money) transfer without official and traceable payment.

- ***Finding*** *Class* (***Activity*** *SubClass*) models the concept of achievement of any requested product; it can occur in different ways: purchase, misappropriation, scam and so on.

- ***Production*** *Class* (***Activity*** *SubClass*) models the concept of goods production.

- ***Sale*** *Class* (***Activity*** *SubClass*) models the concept of goods sale.

- ***Shipment*** *Class* (***Activity*** *SubClass*) models the concept of goods shipment.

- ***Actor*** *Class* models the concept of person who participates in the illegal event; this person could involve in one or more activities. Since the actor of an activity can be a single subject or a group, ***Person*** *Class* (***Actor*** *SubClass*) and ***Organization*** *Class* (***Actor*** *SubClass*) have been defined. The attribute role on ***Actor*** *Class* expresses the role of the participant in the action.

- ***Ideology*** *Class* models the concept of political and/or religious ideology that can be disseminated through propaganda.

- ***Money*** *Class* models the concept of money that can be transferred.

- ***Product*** *Class* models the concept of product which is the object of the illegal action. In addition to drugs, NPS, medicines and weapons, products also include accessories and precursors. The attributes of these classes define their type and the product to which they refer. In particular, ***Other*** *Class* represents all the substances not yet recognised as drugs or NPS or any other mentioned substance, but from text/video evidence can be clear that their body effect resembles the one of the proven narcotic substances. ***Drug*** *Class* (***Product*** *SubClass*), ***Medicine*** *Class* (***Product*** *SubClass*), ***NPS*** *Class* (***Product*** *SubClass*) and ***Other*** *Class* (***Product*** *SubClass*) share the same attributes so they have been generalized in the ***Substance*** *Class* (***Product*** *SubClass*), of which we specify the format, the possible methods of assumption and the type of effects they provide. The keyword attribute on the ***Product*** *Class* admits as values all the slang expressions with which the various products are known on the web.

- ***Shop*** *Class* models the concept of online markets. It is the generalization of ***CryptoMarket*** *Class* (***Shop*** *SubClass*), ***eShop*** *Class* (***Shop*** *SubClass*) and ***SingleVendorShop*** *Class* (***Shop*** *SubClass*).

In the next figure, the Object Properties list is shown.

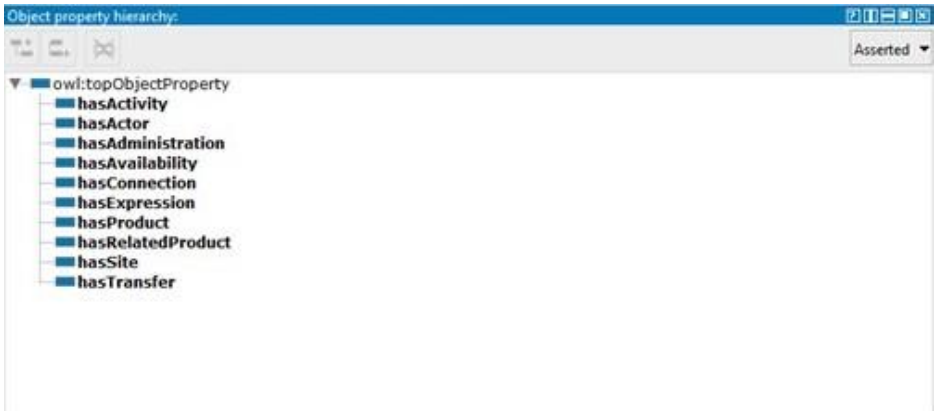

Figure 11: Object properties list

Some specifications are proposed below.

- ***hasActivity*** *ObjectProperty* expresses the relationship between ***Actor*** Class and **Activity** *Class*. It means that all the actors take part in at least one event.

- ***hasActor*** *ObjectProperty* expresses the relationship between **Activity** *Class* and ***Actor*** Class. It means that, for all the actions present in the class, exists at least one actor who is the subject of the action.

- ***hasAdministration*** *ObjectProperty* expresses the relationship between ***Actor*** *Class* and ***CryptoMarket*** *Class*. It means that every cryptomarket requires an administrator.

- ***hasAvailability*** *ObjectProperty* expresses the relationship between ***Product*** *Class* and ***Shop*** *Class*; in this way, we analyze the availability of a selected product in a selected shop.

- ***hasConnection*** *ObjectProperty* expresses the relationship between ***Activity*** *Class* and ***Product*** *Class*, since every event is related to some product.

- ***hasExpression*** *ObjectProperty* refers to the relationship between ***Dissemination*** *Class* and ***Ideology*** *Class.*

- ***hasProduct*** *ObjectProperty* expresses the relationship between ***Actor*** *Class* and ***Product*** *Class*, since every actor, being the subject of an action, is related to some product (taking in consideration *hasConnection*).

- ***hasRelatedProduct*** *ObjectProperty* refers to the relationship between ***Accessory*** *Class* (***Product*** *SubClass*), ***Precursor*** *Class* (***Product*** *SubClass*) and ***Product*** *Class*, meaning that any accessory and any precursor corresponds to at least one specific product.

- ***hasSite*** *ObjectProperty* refers to the relationship between ***Activity*** *Class* and ***Shop*** *Class*, meaning that in this context we can consider the shop as the *site* where the event takes place.

- ***hasTransfer*** *ObjectProperty* expresses the relationship between ***Donation*** *Class* and ***Product*** *Class*, ***Money*** *Class*.

Using OntoGraph plugin, it is possible to provide a graphical representation of the ontology, proposed below.

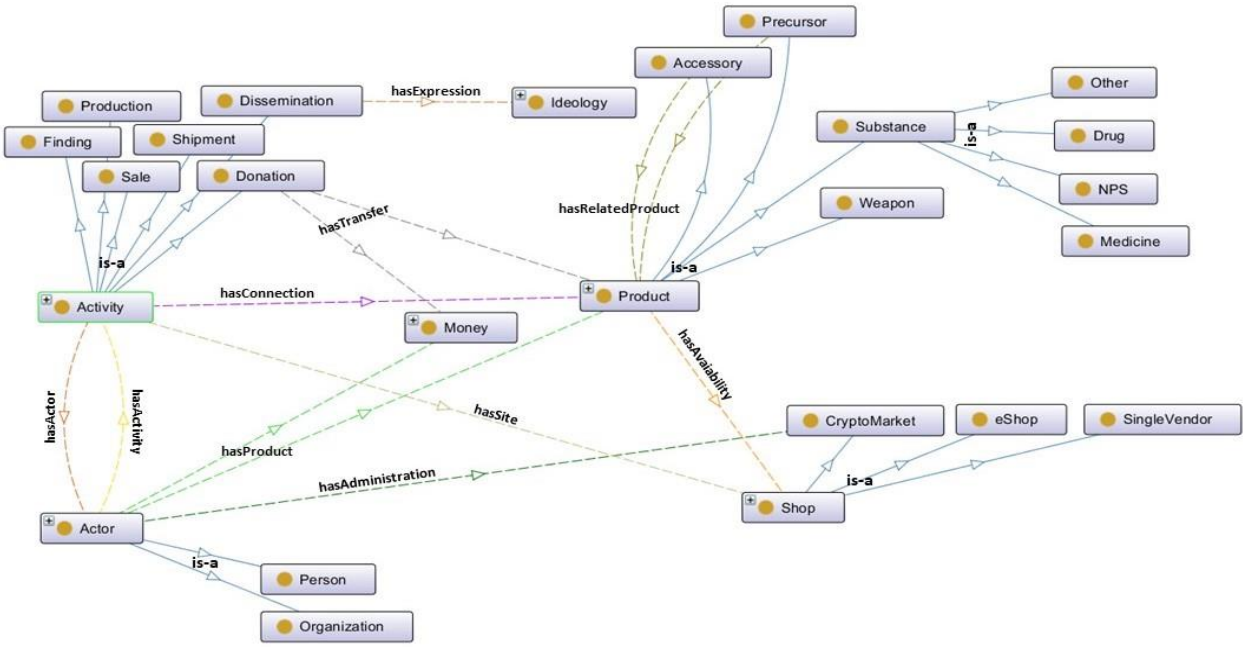

Figure 12: Ontology graph

REQUIREMENTS

One of the objectives of the WP7 is knowledge modelling for illegal trafficking. This goal requires a specific module that is able to model all crime aspects including activities, people, organizations, places, black-markets and illegal shops, products and their relationships. The knowledge stored in the system can be accessed through functional and supportive tools and in order to facilitate such an access, all entities and relationships have been represented properly. Moreover, the use of common taxonomies, ontologies and metadata enable analysis modules to represent their outcomes in a unified way, which facilitates integration and reasoning processes.

RECOMMENDED MODULE

Protégé 5, as ontology editor. Available at http://protege.stanford.edu

OPEN ACCESS LINK

https://www.anita-project.eu/assets/anitaOntology.owl

3.2 Illegal Trafficking Knowledge Base

DESCRIPTION

After modelling process, started the population of the KB; for this aim, more than 5000 documents have been analyzed and this TOP 3 Classification of the most populated entities just comes from these analyses.

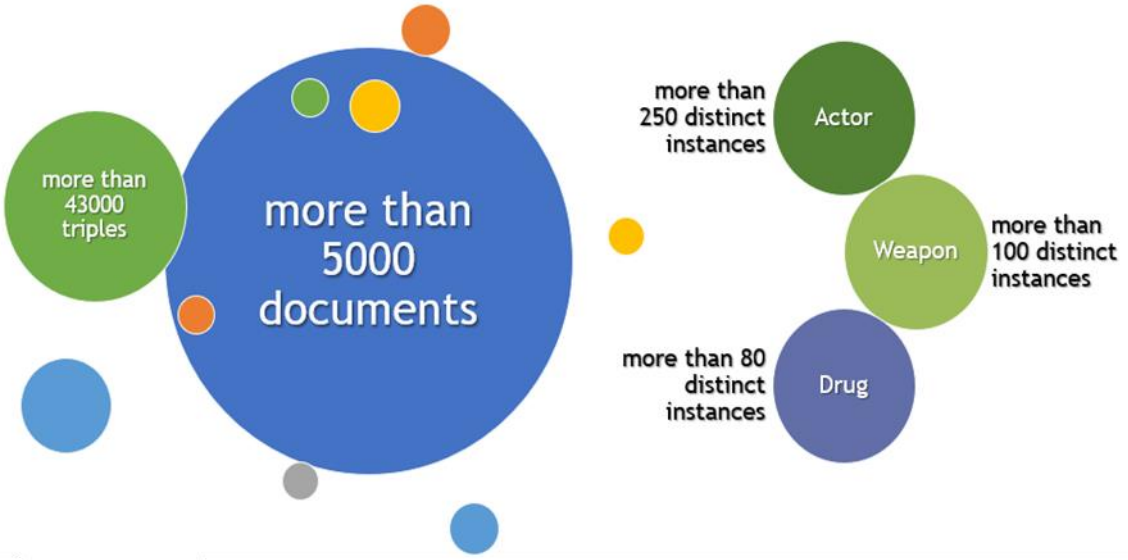

Figure 13: Knowledge Base Instances

To date there is this number of instances for each entity:

- Weapons: 176
- Drugs: 96
- Substances: 42
- Medicines: 32
- Locations: 31
- NPS: 27
- Accessories: 19
- Precursors: 13

OPEN ACCESS LINK

https://www.anita-project.eu/assets/ANITA_knowledge_base_instances.zip

4 Conclusions

This DMP has been prepared to take into consideration that the ANITA project is participating in the ORD Pilot. In this deliverable, we have highlighted how research data will be stored after the project, while all the necessary actions taken in order to participate in ORD Pilot have been reported. This document constitutes the last update of the DMP of the ANITA project, i.e., it describes the set of data that will constitute the "legacy" of the project. As part of the ANITA DMP objectives, this material will support the availability of the data used in order to further facilitate the research activity on the ANITA-related domain. The contributions within the ANITA Project are very diverse in terms of datasets, ranging from user feedback (2.1, 2.2 and 2.4) of many types to different multimedia content (2.3). This has led to diverse improvements to the ANITA system, as well as new scientific knowledge. The collection of these datasets has resulted in multiple scientific publications of several categories. The datasets formed during the project as well as the illegal trafficking ontology and knowledge base have been listed here, while detailed information about its availability and use have been reported.