This project that has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-FCT-2015, under grant agreement no 700381.

Analysis System for GAthered Raw Data

ASGARD

Instrument: Research and Innovation Action proposal

Thematic Priority: FCT-1-2015

D5.2. Report on datasets acquisition and/or creation

Deliverable number	D5.2	
Version:	1.0	
Delivery date:	31/08/2019	
Dissemination level:	Public	
Classification level:	Non-classified	
Status	FINAL	
Nature:	REPORT	
Main author(s):	Bernardo Pacheco	INOV
	Gabriele Ranco	IBM
Contributor(s):	CAST, GUCI, PJ	

DOCUMENT CONTROL

Version	Date	Author(s)	Change(s)
0.1	30/01/2019	Bernardo Pacheco	TOC
0.2	01/06/2019	Bernardo Pacheco	INOV contribution
0.3	28/06/2019	Bernardo Pacheco	Merge partners' contribution
0.4	10/07/2019	Bernardo Pacheco	Ready for peer review
0.5	21/08/2019	Bernardo Pacheco	Document updated addressing peer review from AIT (Refiz Duro) and CERTH (Apostolos Axenopoulos)
0.6	22/08/2019	Romaios Bratskas (ADI)	Quality review of the deliverable
1.0	31/08/2019	Juan Arraiza	All changes approved. Final version submitted

DISCLAIMER

Table of Contents

Annexes

Tables

Figures

1. Introduction

1.1. Overview

The DoA describes this deliverable as:

> *D5.2* - Report on datasets acquisition and/or creation. *[month 36]*
>
> *The main objective of this document is to describe the work done on dataset acquisition and/or creation of simulated (realistic) data for research purposes. It will include a data set gap matrix.*

1.2. Relation to other deliverables

This deliverable is related to the following other ASGARD deliverables:

Receives inputs from:

Deliv. #	Deliverable title	How the two deliverables are related
D1.1	Data Protection Policy Report	D1.1 describes the EU legal framework regarding Data Protection and Privacy Rights, and details the strategy, actions put in place in the ASGARD project to comply with EU ethical, data protection and privacy principles.
D3.1	Use cases definition and end-user requirements report	The content of these deliverables was the major support in the tools definition and needed datasets.
D3.2	System specifications	
D3.3	System Architecture	

Table 1 – Relation to other deliverables – receives inputs from

Provides outputs to:

Deliv. #	Deliverable title	How the two deliverables are related
D10.5	ASGARD technical validation	D10.5 requires the data generated from D5.2 to be functionally tested

Table 2 – Relation to other deliverables – provides outputs to

1.3. Structure of the deliverable

This document includes the following sections:

- Section 1, which provides a brief description of this deliverable and its relations with other deliverables of the project;
- Section 2, which describes the process of datasets acquisition;
- Section 3, describing the initial dataset collection;

- Section 4, describing the datasets created during the project, purposely to cover the gaps from collected datasets and the needs of the relevant stakeholders (i.e., developers and LEAs);
- Section 5, describing how the datasets were maintained and kept available;
- Section 6, presenting the list of datasets used in the hackathons, and
- Section 7, which concludes the main text of the report and makes recommendations for future work.

The full list of datasets, including its ID, SELP Unit approval status, availability for ASGARD, name, source, use case and description is presented in ANNEX III.

2. Datasets acquisition process

During the course of ASGARD, partners have used a variety of datasets to test and develop their tools. Generally, these datasets should be open source/public data that comply with EU regulations, but sometimes data is not available, and it was necessary to create some datasets. To fulfil their needs, developers and LEAs have used one or more of the following:

- Open source/public datasets;
- Datasets designed/created in previous projects and owned by one of the partners, and shared with the other ASGARD partners for project purposes;
- Dataset created within ASGARD project to fulfil identified needs not covered by available datasets.

Section 2.1 describes the acceptance criteria that each dataset must comply to be used in the project, while Section 2.2 shows the workflow that each dataset must go through in the acquisition process.

2.1. Dataset acceptance criteria

ASGARD deliverable D1.1 "Data Protection Policy Report" addresses the subject of data protection within the ASGARD project and describes the actions put in place to comply with EU ethical, data protection and privacy principles (EP). It includes the establishment of a SELP Unit (Societal Ethical Legal and Privacy) that takes the responsibility for dataset evaluation and classification according the perceived potential risks against the principles mentioned above.

To measure those risks the SELP Unit pays attention to four main aspects on each dataset:

1. Type of data: identification of the type(s) of data contained in the dataset; for example, simulated data, aggregated/statistical data, de-identified/anonymised, or (sensitive) personal data;
2. Provenance: open source data, open public data, or identification of the Data Controller;[1]
3. Collection process: identification of how the data gathering has been performed;
4. Other legal issues: evaluation of other legal issues (e.g., intellectual propriety rights).

The criteria for acceptance and SELP approval for datasets can be summarized in the following points:

- Public datasets used for testing purposes;
- Public datasets used during the Hackathons;
- Small-sized public datasets (even if only used for training, or not used at all);
- Synthesized datasets, created for the project (e.g. the fake conversations created for the previous Hackathon);
- Any other datasets on a "per request" basis.

[1] According European Commission definition, the data controller determines the purposes for which and the means by which personal data is processed.

Admitted exceptions, that require further detailed evaluation, are:

- Datasets subject to some kind of access control;
- Datasets requiring acceptance of "terms and conditions".

As a result of that evaluation process, each dataset is classified according to the following colour code:

- **Green** datasets: low data protection and privacy (DP) risks, these datasets can be used in the ASGARD project;
- **Yellow** datasets: medium DP risks, these datasets can probably be used after going through the formal ASGARD Ethical Review Process or satisfying other legal requirements (e.g., Intellectual property or copyright information);
- **Red** datasets: high DP risks, these datasets CANNOT be used in ASGARD.

2.2. Dataset Request Process

Whenever ASGARD partners identify the need for a dataset the following steps must be followed:

1. Check with T5.2 team (responsible for the dataset management, as well as for this deliverable D5.2) if there is any dataset already available in the ASGARD Datasets collection that fulfils the identified need.
2. To reuse an existent dataset, the color code (Green, Yellow, Red and white) must be taken in consideration. It indicates the SELP approval status according to:
    a. A Red dataset cannot be used. It has not the required SELP approval;
    b. A Green dataset can be used after an ethical self-assessment process;
    c. A Yellow dataset has some restrictions. Contact SELP Unit to verify legal and ethical issues;
    d. A White dataset was not yet classified. Contact SELP Unit to verify legal and ethical issues.
3. To add an existing dataset to the ASGARD dataset repository, from other organizations or from publicly available sources the SELP Unit must be contacted to verify ethical and legal issues.
4. To add a new dataset that will be created collecting new data, partners must contact the SELP unit in order to:
    a. to verify legal and ethical issues;
    b. to implement technical and legal safeguards to fulfill all the SELP requirements.

Figure 1 presents the workflow to add a new dataset to the collection. Section 5 describes how the datasets are maintained and the process of requesting the approval of a new dataset.

Figure 1 – New dataset acquisition workflow

3. Collected datasets

The goal of the process of dataset collection is to support developers and LEAs during development, training and testing of ASGARD tools, when addressing the ASGARD Use Cases (as described in ASGARD deliverable D3.1) or attempting to fulfil the identified User Requirements and Specifications. Therefore, the process for dataset acquisition uses those Use Cases as a starting point to identify the collection initial needs. Taking this in consideration, ASGARD partners were invited to contribute with datasets designed/created in previous projects and open source/public datasets.

3.1. Addressed use cases

The following use cases are addressed by the received contributions from partners:

- **Audio analysis** – includes activities such as audio events detection, environment recognition, acoustic-phonetic studies, gender detection, language detection, speech recognition and speaker identification;
- **Image analysis** – includes activities such as face detection, face recognition, artefact detection and recognition, image-based localization or place classification, image OCR (characters recognition), as well as image enhancement (e.g. image Super-Resolution);
- **Video analysis** – includes activities such as people and vehicle identification and tracking, background scenery identification and classification and summarizing videos through most significant frames extraction;
- **Forensics/data storage** – includes activities such as data recovery from raw images of devices, information recovery from deleted or damaged file systems, system activity identification through raw images of devices;
- **Text analysis** – includes activities such as natural language processing, semantic classification, keyword or named entities mentions identification, language identification and author profiling;
- **Posts & social media** – includes activities such as sentiment analysis, terms extractor, users profiling, topic prediction and relations graphic visualization;
- **People information** – includes activities such as finding missing people and transnational migration analysis;
- **Darkweb** – includes activities such as data collection from darkweb and searching the darkweb for arbitrary terms.

3.2. Dataset information

For each shared dataset the following information was recorded:

- **Use Case(s) to be used** – The use case where the dataset can be used;
- **Dataset name** – Name of the dataset;
- **Dataset description** – A brief description of the dataset that should be enough to decide if the dataset could fulfil the purpose of partners looking for datasets;
- **Data Controller** – Partner assuming the role of data controller for the dataset;
- **Data Processors** – Partner assuming the role of data processors for the dataset;
- **Recommended by** – Partner that have recommended the dataset;
- **Point of contact in ASGARD** – The contact point to address any subject related with the dataset;
- **Source** – The source of the dataset;
- **Dataset properties** – Technical description of the dataset;
- **Data quality** – Image or video quality;
- **Format** – data formats;
- **Size** – Size of the dataset;
- **"Can it be shared within the consortium?"** – Indication if this dataset is sharable with the other partners or usage is restricted for the partner that is asking for dataset approval;
- **Pre-processing / modifications to make available** – Actions that must be taken before partners are able to use the dataset;
- **Restriction / Constraints** – Already known restrictions or constraints for the dataset usage;
- **Licences and Costs** – Licenses and costs for usage of the dataset;
- **Documentation of the dataset (link)** – Links for additional documentation related with the datasets (e.g, sources website);
- **Annotations** – Additional information not included in the previous points.

3.3. Identified datasets

During ASGARD task T5.2, 185 datasets from different sources covering a wide range of use cases (see Section 3.1) were identified. Each dataset has received a unique ID (format: DSxxx), used to identify it within the project. ANNEX III presents the full list of datasets, including its ID, SELP Unit approval status, availability for ASGARD, name, source, use case and description. Table 3 shows a sample with a single dataset.

SELP Unit approved	Available for ASGARD?	ID	Dataset	Source	Use Case(s) to be used	Dataset description
Yes	Yes	DS007	YouTube Faces DB	internet	Face Recognition	YouTube Faces Database, a database of face videos designed for studying the problem of unconstrained face recognition in videos.

Table 3 – datasets list sample

4. Created datasets

From the set of collected datasets described in Section 3, some identified gaps were not possible to cover with available datasets, even after a new procurement process. The solution was to generate new datasets or customize some of the already existent using the project resources. The content of these generated datasets can be summarized in the following categories:

- Device images containing a selection of existent datasets;
- Social Media content with specific content;
- CCTV videos showing specific actions;
- Language corpus to be used in speech recognition tools.

Table 4 shows the list of created datasets. Sections 4.1 to 4.9 describes the process of creation of the datasets.

Dataset	Description
DS032	Dataset implemented to test digital forensics triage tools. Consists of images of different Windows OS with limited activity on web browsing, file sharing and social media usage and a range of file type datasets.
DS043	Social network records synthesized through a python script and stored in a csv file.
DS160	Dataset created reusing a sample of an existing dataset (DS048) containing a list of tweets in a csv file. In the DS160, Names, Usernames and Tweets were modified.
DS169	Dataset created reusing a sample of an existing dataset (DS048) containing a list of tweets in a csv file. In the DS169, only a link was modified from the existing tweets;
DS172	Disk image created for the 1st Hackathon, containing approved datasets
DS180	Video of internal CCTV cameras of INOV, researchers acting accordingly to a previously defined script.
DS181	Videos recorded during the Mallorca hackathon.
DS182	Video recorded by INOV. It shows an INOV researcher talking directly to a camera, mentioning words associated with terrorism and religion.
DS185	Portuguese corpus including about 20 hours of annotate audio, to be used in speech recognition tools.

Table 4 – Created datasets

4.1. Creation of dataset DS032

The dataset DS032 addresses the “Computer and Storage” dataset and was constructed to test digital forensics triage tools. It consists of different Windows Operating Systems devices images, with limited activity on web browsing, file sharing and social media usage and a range of file type datasets.

It aims to provide independent comparative performance testing of commercial digital forensics triage tools, in representative user scenarios. The content consists in 100GB of forensic images of hard drives containing a wide range of file types, mostly sourced from open corpora.

Relies on the use of licensed software - e.g., Windows operating system.

4.2. Creation of dataset DS043

During the technical validation process, it was identified that ASGARD testers will need a model for the classification of texts and images in social media. Social media data is used because it contains geo-temporal labels together with texts and images. This information could be potentially relevant for Law Enforcement Agencies (LEA). Due to legal limitations of processing "Personal Information", it is not possible to use real data. The solution was to follow a different approach, creating a similar dataset that somehow can simulate a reasonable evolution of social media records.

A python script was created to generate a csv file containing the following fields as seen in Table 5Table 5 - List of the generated fields:

Field Name	Description	Example values
date	Timestamp of the record	'2018-01-01 06:34:19'
LAT	Latitude	53.34297274174823
LONG	Longitude	-6.256036314629904
AUTHOR	Label	user_1606051
TEXT	A sentence taken from module NLTK. is a python module for natural language processing	Art Lund; a fine big actor with a great head of blond hair and a good voice; impersonates Enright.
TOPIC	One of the Brown corpora of NLTK. We have chosen Brown because the topic classification of sentences is easy to understand and it is also open source so we can legally use it.	reviews[2]
IMAGE	File name of the CIFAR10 dataset contained in KERAS.	13143.jpeg
IMAGE_CLASS	One of the classes of the CIFAR10 . CIFAR10 is an open source dataset of labelled images.	horse

Table 5 - List of the generated fields

The above fields are created following these steps:

1. Date: the timestamp is created by adding N records per each day of simulation. Each day, the records are generated from a Gaussian distribution with a mean of 12 am Figure 2). The reason for this is to simulate an increase of activity during working hours.

[2] "reviews"is one of the topics of the the Brown corpora of NLTK. The others topics are: news, editorial, religion, hobbies, lore, belles_lettres, government, learned, fiction, mystery, science_fiction, adventure, romance and humor

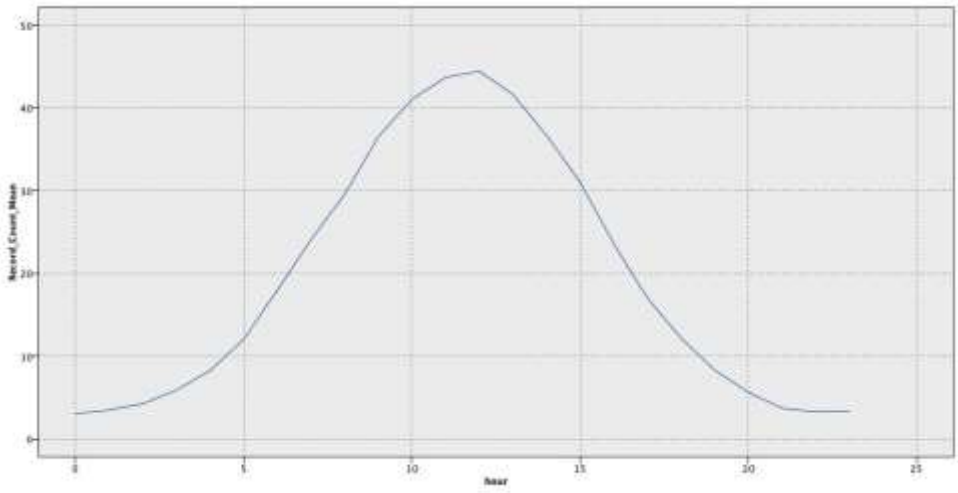

Figure 2 - Hourly frequency of messages as result of data generation

2. Space: Latitude and longitude are obtained from two Gaussian distributions. The means of these two Gaussians are the coordinates of the centre of a city. The standard deviations decrease in working hours in order to simulate the movement of people towards the city centre.

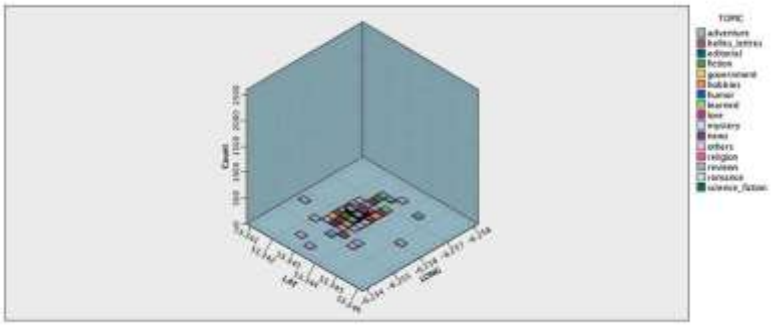

Figure 3 -Example of 3D histogram of the volumes of messages during night hours in a given day

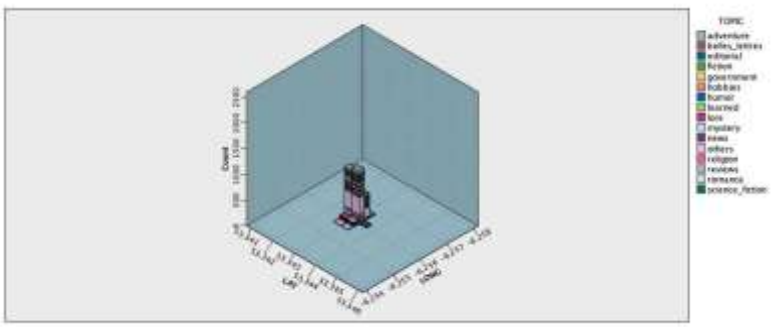

Figure 4 - Example of 3D histogram of volumes of messages in working hours for a given day

3. Topic: the topic is selected from a uniform distribution between 0 and the total number of topics. But at every iteration (so after the selection of one topic), the topic can generate other messages of the same type. The number of messages to add will be decided by a power law distribution of per of -2. This choice is made in order to simulate messages that are created independently of each other but in some cases correlation between them can occur. This choice follows the work of Zapperi.

   See Figure 5 to view the final result of the simulation.

4. Text: is randomly selected from the Brown corpus in NLTK. The text is selected with the relative topic.

5. Images: they come from the CIFAR10 dataset of KERAS. They are added only if in the text there is an explicit reference to the class label of the image; in that case they are selected randomly.

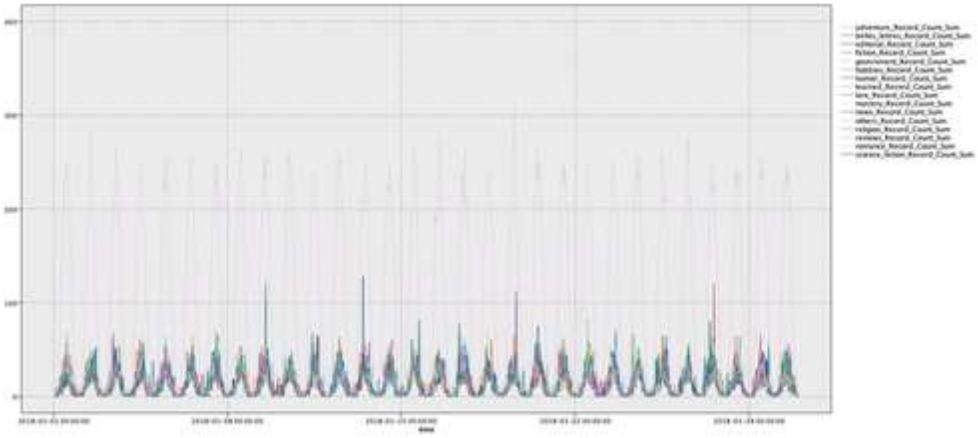

Figure 5 - Volume of messages for each topic along time

4.3. Creation of dataset DS160

The dataset DS160 was derived from DS048 to fit with a particular case for hackathon purpose. The dataset DS048 consists of about 17,000 tweets, scraped from more than 100 Pro-ISIS fans all over the world. The dataset includes: name, username, description, location, number of followers at the time the tweet was downloaded, number of statuses by the user when the tweet was downloaded, date and timestamp of the tweet, the tweet itself.

For the dataset DS160 construction, Names, Usernames and Tweets on 111 tweets were manually modified from the original (DS048). The goal was to replace names and usernames, as well as tweets' contents, to fit in an investigation case prepared for the hackathon.

4.4. Creation of dataset DS169

The dataset DS169 was derived from DS048 and DS152 to fit with a particular use case.

As described in 4.3, the dataset DS048 consists of about 17,000 PRO-ISIS tweets. The dataset includes: name, username, description, location, number of followers at the time the tweet was downloaded, number of statuses by the user when the tweet was downloaded, date and timestamp of the tweet, the tweet itself.

The dataset DS152 consists on a Jihadist attack threat video where a known ISIS member threatening Spain with a Jihadist attack.

For the creation of the new dataset DS169, some tweets were modified replacing the existing link with a new one pointing to the video in DS152.

4.5. Creation of dataset DS172

The dataset DS172 addresses the use case of "recovering deleted/damaged file systems and files". It was created for the first hackathon consists on a disk image created with several partitions encrypted, some deleted and partially deleted content/files. The disk contains Skype messages, synthetized emails, pictures, text files and PDF files.

4.6. Creation of dataset DS180

The dataset DS180 consists of a simulated CCTV video (mp4 format), where ASGARD members are acting according to a previously defined script inside of a room.

The camera was positioned on the ceiling, pointing down to capture the participants passing by a limited area. The participants act as they are not aware that the camera is there, never looking at it. Captured footages try to simulate images from indoor CCTV, including:

- Images without employees;
- Images with a single employee walking on both directions;
- Images with more than one employee walking on the same directions
- Images with more than one employee walking on opposite directions greeting each other
- Images with more than one employee walking on opposite directions without interaction.

To allow this video usage, participants have signed the informed consent.

4.7. Creation of dataset DS181

The dataset DS181 consists of an mp4 video recorded during the Mallorca hackathon for a particular case during that hackathon. It aims to simulate the video captured from a CCTV camera pointing to a crowded walkthrough.

Like in the DS180 case (Section 4.6), the camera was positioned on the ceiling, pointing down to capture the participants passing by a limited area. The participants act as they are not aware that the camera is there, never looking at it.

To allow this video usage, participants have signed the informed consent.

4.8. Creation of dataset DS182

The dataset DS182 consists of an mp4 video, where a single ASGARD member sends a video message.

The participant is seated in front of a table and the camera is positioned right in front of him, capturing the video on a close-up of his face. The participant talks directly to the camera mentioning words associated with terrorism and religion. It lasts for about 1 minute.

The message starts with an intro where the participant describes the context of that video, making it clear that

it is part of a training dataset. Even so the rest of the message was written aiming to avoid that it could be misused in the wrong context. It includes the mentions to terrorism, religion, war, governmental, etc. but avoiding hate expressions.

To allow this video usage, participant has signed the informed consent.

4.9. Creation of dataset DS185

The accuracy of any speech recognition tool is directly related with its training process and with the quality and comprehensiveness of the datasets used in it. A well succeeded training process needs a large number of hours with annotated audios where the variety of speakers and expressions are a key factor. Whoever needs to train a speech recognition tool, developed from scratch or reusing available open-source tools (e.g., Kaldi[3] and CMUSphinx[4]), has several options for training datasets in the most common languages (e.g., English). The same is not true for Portuguese.

This dataset includes about 20 hours of annotated audio, in the Portuguese language as it is spoken in continental Portugal. The audio is stored in ".wav" files (mono, 32bit, 44100Hz) and the text in plain ".txt" files. Sentences are divided in separated audio files and, for each audio file, there is a text file with its content.

The steps included in this dataset creation are:

1. Identify public available audio with its content available in text format;
2. Evaluate audio quality and text accuracy. Some differences between audio and text are admitted but the more differences the more processing effort;
3. Split audio and text in sentences, ensuring that for each audio file there is a text file exactly matching with the audio content. This step includes human intervention to cut the audio as well as select and, if necessary, correct text.

The source of audios selected for this data set creation was the librivox[5], a free public domain audiobooks platform that makes available readings from books that are in Public Domain made by volunteers. The read texts are available in Project Gutenberg[6] , a provider of free electronic books.

It was possible to find 20 hours of audio read by 9 volunteers. To make it possible to process this amount of audio in a reasonable amount of time, a dedicated tool with an intuitive graphic interface was developed (see Figure 6). Within this tool user process the audio and text through the following steps:

1. Select input folder that should include 2 subfolders with audio and text files respectively
2. Select the audio file to process
3. Select the text file to process – that should include the read text
4. Start processing
    a. The processing tool will start playing the sentences according the automatically suggested beginning and ending

[3] Kaldi-asr.org
[4] https://cmusphinx.github.io
[5] Librivox.org
[6] http://www.gutenberg.org/

b. The content of the text file is shown in a text box

c. User can add, remove and edit the text

d. User can adjust the beginning and ending of audio of each sentence

e. User can select the select a block of text to be associated to the played audio

f. After user confirm the played sentence transcription, two files are generated with the played audio – wav file mono, 32bit, 44100Hz – and the selected text. These two files have the same name with different extensions

g. A new sentence is played as describe above in "a."

At the end of the audio processing the generated audio/text pairs of files are stored in a new folder called exported and can be used as input to the speech recognition tools training.

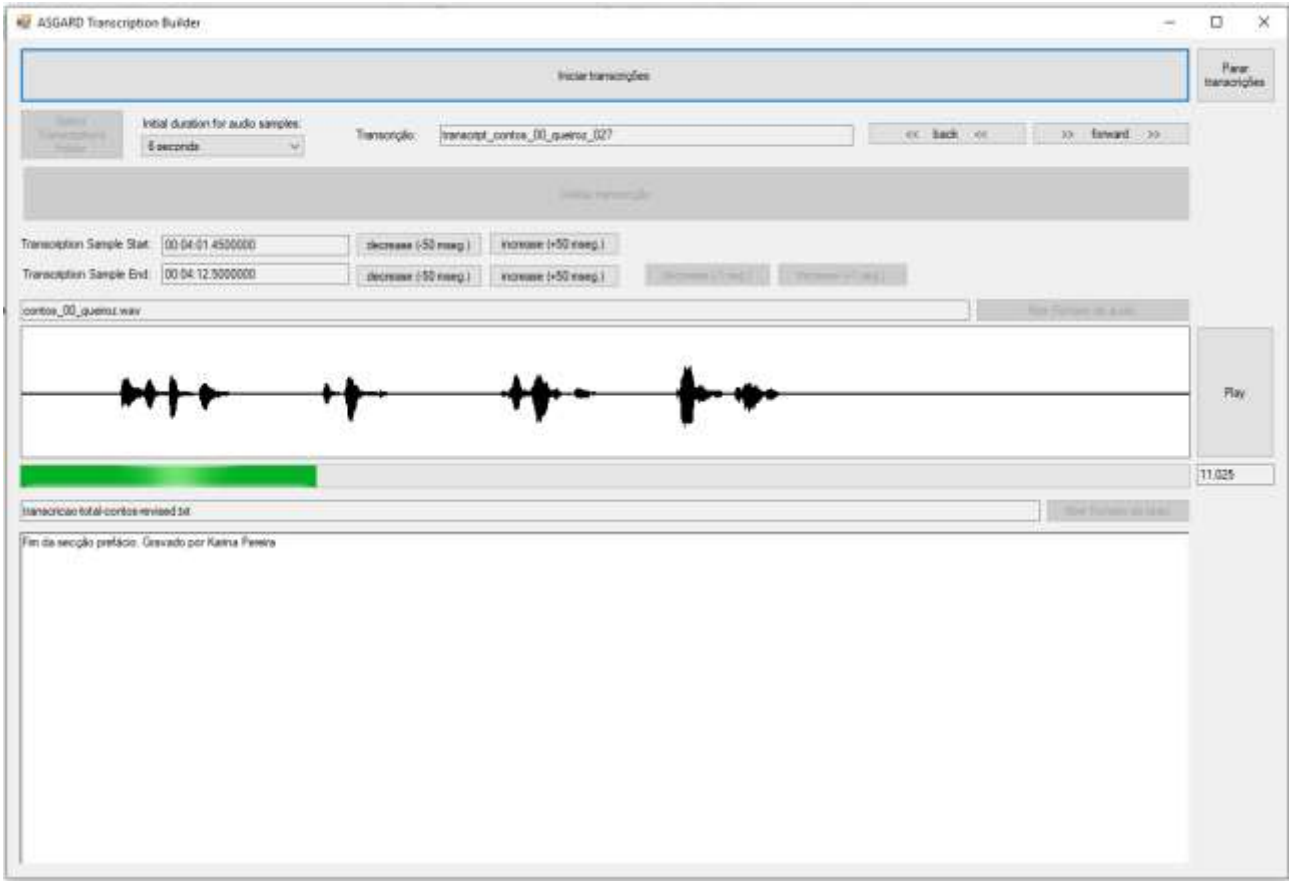

Figure 6 – Print screen of the audio processing tool

5. Datasets availability and maintenance

The ASGARD project live-cycle includes several developments milestone and evaluation moments through hackathons and "Capture the Flag" (CTF) exercises. Since its initial stage until the end of the project, the datasets collection aims to support developers and LEAs in development, training and testing of ASGARD tools. For this to be possible, a mechanism was put in place to ensure 4 main goals: an easy to consult dataset catalogue, the availability of the datasets collection, support for the defined workflow for new datasets acquisition (Figure 1 in Section 2.2) and get efficiency during all stages of the process. This mechanism was supported on some collaborative tools used in the project like "mattermost" and "gitlab", as well as the "box" and "nextcloud" repositories. It also includes a dedicated instance of "MongoDB" - a document-oriented database – accessed through a "REST API". The bullet points in the next paragraph details the process.

- **Easy to consult dataset catalogue** – An internal catalogue of the datasets was created and maintained in two supports: an Excel file stored in the ASGARD Box repository and the MongoDB database. Both contains all metadata referred within Section 3.2 plus a unique ID for each dataset and the result of SELP Unit approval;
  The catalogue in the database can be accessed through a web interface (see Figure 7), able of filtering information by metadata values (e.g. the unique ID or SELP approval status);
  All tools developed in ASGARD should be validated with SELP approved datasets included in this catalogue (see ANNEX III);

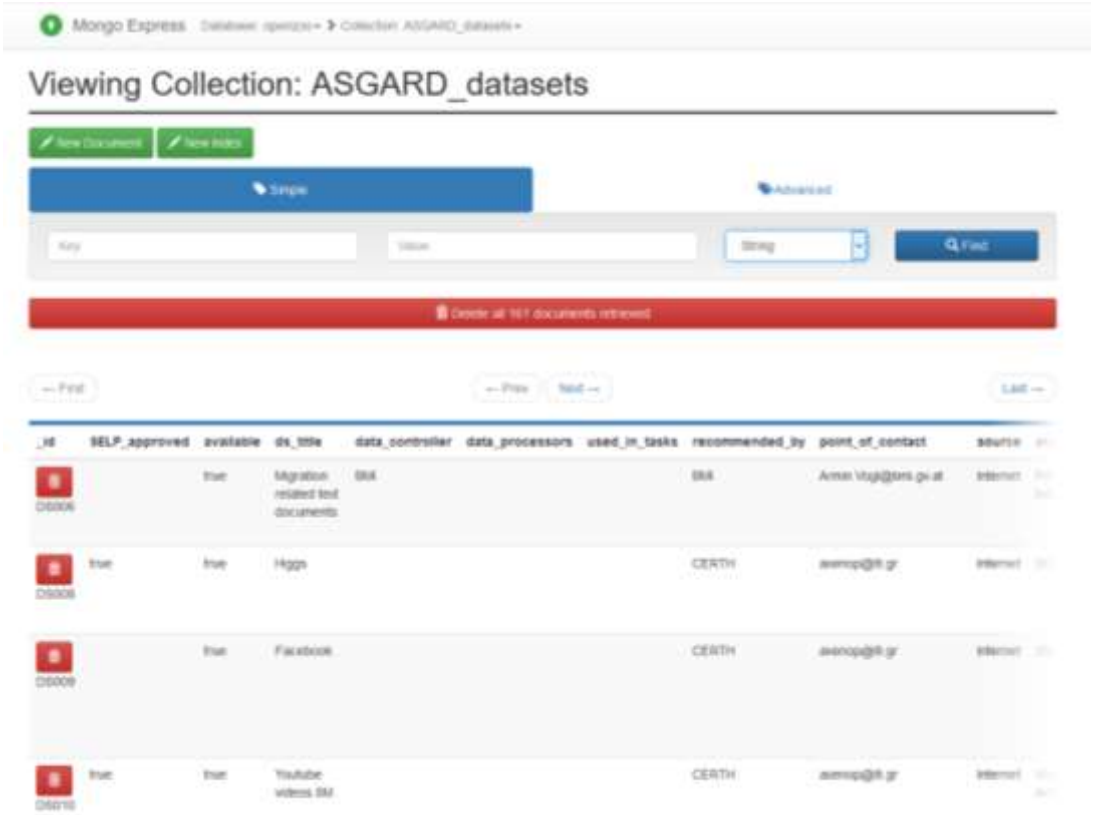

Figure 7 – web interface to access datasets catalogue

- **Availability of the datasets collection** – The datasets were made available in the ASGARD "nextcloud" repository. The datasets were stored in a folder named with its unique ID (see Figure 8);

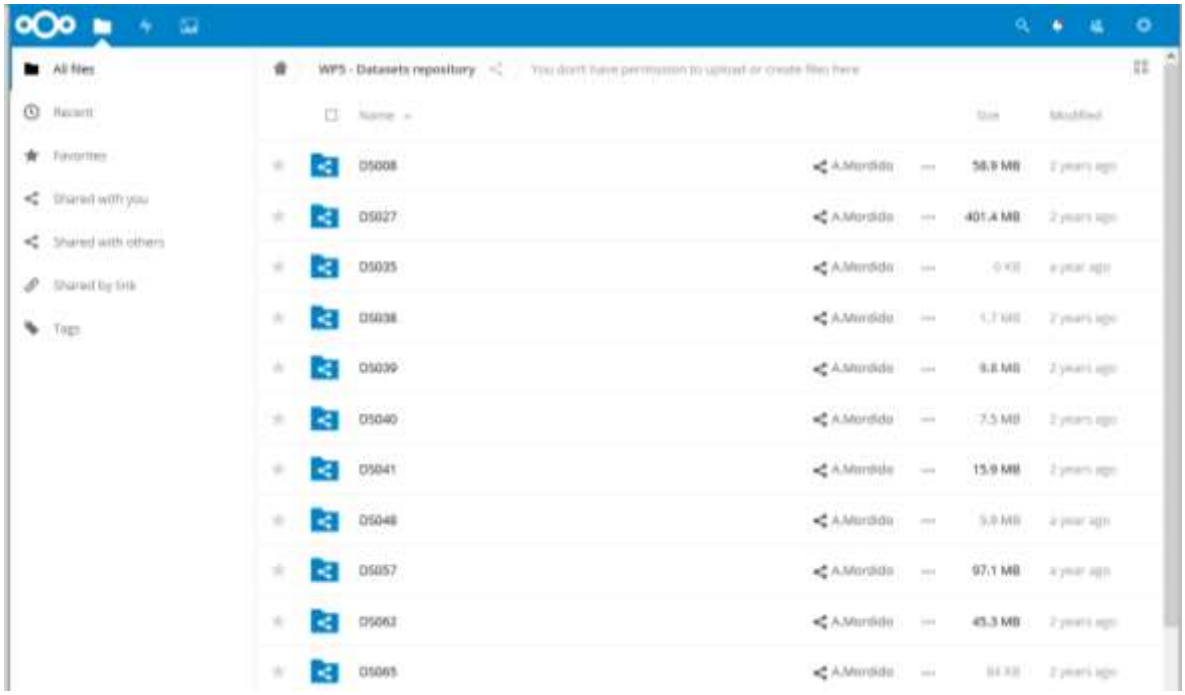

Figure 8 – ASGARD Nextcloud Repository with datasets

- **New datasets acquisition according with the data-flow presented Figure 1, Section 2.2 –** Whenever a dataset is needed, ASGARD developers or LEAs should follow the steps described below:

  1. Consult the catalogue and check if there is any useful dataset already available in the ASGARD Datasets collection.
  2. To reuse an existent dataset, evaluate its SELP approval status according the color on the excel file and the value in the column "SELP Unit approved".
  3. To add an existing dataset, from other institutions or from other publicly available source, the SELP Unit must be contacted to verify ethical and legal issues
  4. To add a new dataset that will be created collecting new data, partners must contact the SELP unit to Contact WP12-Selp unit in order to:
     a. to verify legal and ethical issues
     b. to implement technical and legal safeguards requirements
  5. The result of the SELP approval process must be updated in the catalogue

- **Efficiency during all stages of the process –** Any request to the SELP Unit should be made creating a new issue in the "WP5-datasets" project on ASGARD gitlab. If needed, the SELP Unit will request additional information adding new comments to the gitlab issue. The issue is closed after a final assize about the SELP approval. Mattermost is also available as an additional communication channel, instead of using emails. Parts in the process can use it to solve problems during the validation process. Nevertheless, all information about the dataset and reasons supporting the SELP evaluation should be provided on the gitlab issue.

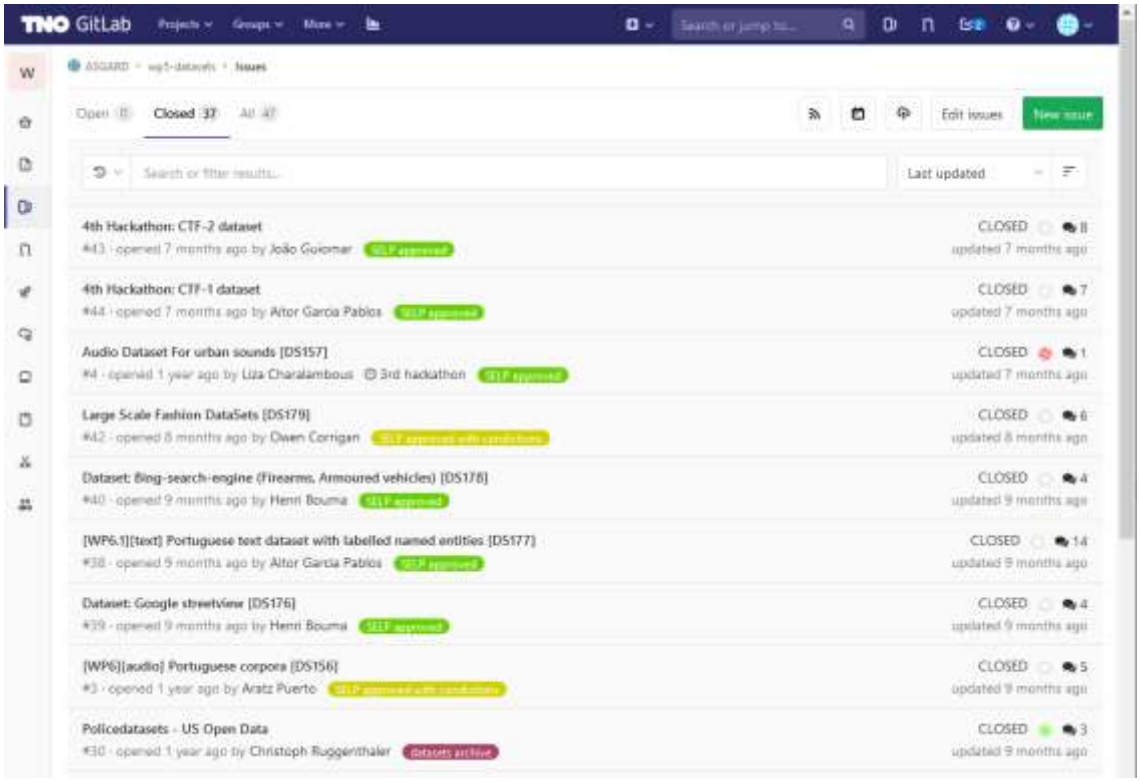

Figure 9 – Issues list on ASGARD gitlab

6. Datasets used in hackathons

During the preparation process for each hackathon, partners involved were asked about the datasets that can be useful for the planned use cases. Sections 6.1, 6.2 and 6.3 present the list of datsets used in the hackathons conducted so far.

6.1. Datasets selected for usage during the 1st and 2nd Hackathons

SELP Unit approved	ID	Dataset	Source	Use Case(s) to be used	Dataset description
Yes	DS067	jihad-related videos	internet	visual analysis	videos from Jihadology.net
Yes	DS068	non-jihad videos Columbia Consumers Video dataset (CCV)	internet	visual analysis	Columbia Consumers Video public dataset (CCV)
Yes	DS069	Dabiq, Inspire, Rumiyh	internet and Pydio	visual and text analysis	Islamic State's magazines with english text and pictures; this dataset also contains different articles split in different files
Yes	DS072	jihadist videos	internet	audio analysis	Jihadist videos with english audio
Yes	DS078	Skype logs	Box	text analysis, social media anlaysis	Skype messages jihad-related. Adapted to fit the Generic Use Case for the 1st Hackathon.
Yes	DS079	Audio conversations	Pydio	audio analysis	Audio recordings of slot conversations with jihad-related content from DS078; VoIP filter
yes	DS172	GUC disk image	nextcloud	computer and storage	Disk image created for the 1st Hackathon - content SELP approved and selected for the Generic Use Case

Table 6 – List of datasets selected for usage during 1st and 2nd hackathons

6.2. Datasets selected for usage during the 3rd Hackathon

SELP Unit approved	ID	Dataset	Use Case(s) to be used	Dataset description
yes	DS007	YouTube Faces DB	Face Recognition	YouTube Faces Database, a database of face videos designed for studying the problem of unconstrained face recognition in videos.
yes	DS035	Standford I2V	visual analysis	a state-of-the-art query by image video dataset
yes	DS010	Youtube videos 8M	visual analysis	8million youtube videos with annotations
yes	DS067	jihad-related videos	visual analysis	videos from Jihadology.net
yes	DS068	non-jihad videos Columbia Consumers Video dataset (CCV)	visual analysis	Columbia Consumers Video public dataset (CCV)

yes	DS074	Europe's wanted Fugitives	visual analysis	images of the Europe's most wanted ugitives
yes	DS075	Labeled Faces in the Wild (LFW)	visual analysis	images of public figures collected from the Internet
yes	DS155	YouTube videos - traffic cams	image and video analysis	Traffic Cams and scenes from the 2014 oscars awards
necessary a data sharing agreementw with GUCI	DS170	Jihadist Propaganda Videos	video	Harddisc with recorded/stored propaganda videos
yes	DS152	Jihadist attack threat video	Face detection	A video of a known ISIS member threatening Spain with a Jihadist attack. From: https://www.almasdarnews.com/article/disturbing-video-isis-militants-spain-celebrate-barcelona-terror-attack/
yes	DS154	Town Center Database	Facial recognition	Dataset composed of videos of busy streets to be used for facial recognition in the scope of ASGARD
yes	DS161	UA-DETRAC labelled traffic cameras	video analysis	The dataset consists of 10 hours of videos captured with a Cannon EOS 550D camera at 24 different locations at Beijing and Tianjin in China. The videos are recorded at 25 frames per seconds (fps), with resolution of 960×540 pixels. There are more than 140 thousand frames in the UA-DETRAC dataset and 8250 vehicles that are manually annotated, leading to a total of 1.21 million labeled bounding boxes of objects.
yes	DS159	London Traffic Cams	video	different types of datasets (public transport) -Open Data, examples http://jamcams.tfl.gov.uk/00002.00878.jpg http://jamcams.tfl.gov.uk/00001.09642.mp4?time=636452153679172956
yes	DS072	jihadist videos	audio analysis	Jihadist videos with english audio
yes	DS073	10 second sound clips	audio analysis	2 million ten-second YouTube excerpts labeled with a vocabulary of 527 sound event categories, with at least 100 examples for each category.
yes	DS017	LibriSpeech ASR corpus	audio	Large-scale (1000 hours) corpus of read English speech
yes	DS079	Audio conversations	audio analysis	Audio recordings of slot conversations with jihad-related content from DS078; VoIP filter
SELP has approved ONLY the gender recognition voice datasets.	DS149	Gender Recognition by Voice	audio analysis	This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz (human vocal range).
yes	DS157	Urban Sounds	urban sounds	2 datasets: URBANSOUNS dataset contains 1302 labeled sound recordings. Each recording is labeled with the start and end times of sound events from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music. Each recording may contain multiple sound events, but for each file only events from a single class are labeled. URBANSOUNDS8K contains 8732 labeled sound excerpts (<=4s) of urban sounds from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music.

yes	DS158	Mivia Audio Events Dataset	environmental sounds	a total of 6000 events for surveillance applications, namely glass breaking, gun shots and screams. The 6000 events are divided into a training set (composed of 4200 events) and a test set (composed of 1800 events). The data set is designed to provide each audio event at 6 different values of signal-to-noise ratio (namely 5dB, 10dB, 15dB, 20dB, 25dB and 30dB) and overimposed to different combinations of environmental sounds in order to simulate their occurrence in different ambiences.
yes	DS008	Higgs	Social Network analysis	he Higgs dataset has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered.
yes	DS166	Broad Twitter Corpus dataset	social media	The dataset of tweets collected over stratified times, places and social uses. The goal is to represent a broad range of activities, giving a dataset more representative of the language used in this hardest of social media formats to process. Further, the BTC is annotated for named entities. The entities and the crowd annotations are all provided with the corpus, as well as (where possible) the raw twitter JSON.
yes	DS048	Pro-IS tweets	Posts and social media	Pro-IS tweets. The dataset includes: name, username, description, location, number of followers at the time the tweet was downloaded, number of statuses by the user when the tweet was downloaded, date and timestamp of the tweet, the tweet itself
yes	DS128	Stormfront	Social Network analysis	Posts from Stormfront Forum. Stormfront is a white nationalist, white supremacist and neo-Nazi Internet forum, and the Web's first major racial hate site.
yes	DS045	Words by lone offenders	Posts and social media	Manifestos and words written by lone offenders before they commited an attack. On english
yes	DS078	Skype logs	text analysis, social media anlaysis	Skype messages jihad-related. Adapted to fit the Generic Use Case for the 1st Hackathon.
yes	DS169	Fake synthesized tweets on DS048 (second sub-dataset)	social network analysis	syntethic data from DS048. Only a link was modified on the existing tweets; the inserted link was for the video approved in DS152
yes	DS160	DS048 synthesized for GUC	text analysis	syntethic data from DS048. Names, Usernames and Tweets are modified from the original
yes	DS069	Dabiq, Inspire, Rumiyh	visual and text analysis	Islamic State's magazines with english text and pictures; this dataset also contains different articles split in different files
yes	DS070	jihadi-related blogs (some already removed from internet)	visual and text analysis	text and images from blogs publicly available online, but currently closed down
yes	DS071	Vox-Pol	visual and text analysis	Vox-Pol journal data
yes	DS065	VAST Challenge 2010 MC1	text analysis	Text files containing intelligence reports, communication interceptions, blog and email records, providing evidences for the existence of criminal networks.
yes	DS039	VAST Challenge 2014 MC2	location	32MB of GPS car tracking data plus some additional info
yes	DS027	Computer Forensic Tool Testing Dataset (CFReDS)	computer and storage	These reference data sets (CFReDS) provide to an investigator documented sets of simulated digital evidence for examination. Since CFReDS would have documented contents, such as target search strings seeded in known locations of CFReDS, investigators could compare the results of searches for the target strings with the known placement of the strings. The CFReDS site is a repository of images.

yes	DS172	GUC disk image	computer and storage	Disk image prepared for the 1st hackathon
yes	DS038	VAST Challenge 2014 MC1	people information	844 News articles (.txt), 35 Resumes (.docx) and ~1170 Email headers (single .csv) plus some additional info
yes	DS040	VAST Challenge 2011 MC3	people information	4474 News reports
yes	DS041	VAST Challenge 2006	people information	~1200 News articles plus some additional info
yes	DS085	Webite of the German Federal Office for Migration and Refuggees	people information	Repository of information on how to migrate, enter in or return to Germany
yes	DS087	World Bank data	people information	At the World Bank, the Development Data Group coordinates statistical and data work and maintains a number of macro, financial and sector databases. Working closely with the Bank's regions and Global Practices, the group is guided by professional standards in the collection, compilation and dissemination of data to ensure that all data users can have confidence in the quality and integrity of the data produced. World Bank databases are essential tools for supporting critical management decisions and providing key statistical information for Bank operational activities. The application of internationally accepted standards and norms results in a consistent, reliable source of information.
yes	DS089	Website of the Centre on Religion and Geopolitics	people information	The Centre on Religion & Geopolitics (CRG) is an international affairs think-tank. It presents informed analysis on the global interaction of religion, geopolitics, and conflict.
yes	DS090	Country Reports on Human Rights Practices for 2016	people information	The annual Country Reports on Human Rights Practices – the Human Rights Reports – cover internationally recognized individual, civil, political, and worker rights, as set forth in the Universal Declaration of Human Rights and other international agreements. The U.S. Department of State submits reports on all countries receiving assistance and all United Nations member states to the U.S. Congress in accordance with the Foreign Assistance Act of 1961 and the Trade Act of 1974.
yes	DS091	Website of EU Institute for Security Studies	people information	Repository for many security-related topics: EU foreign policies, global governance, secuirty and defense, transnational challenges.
yes	DS092	Latest reports from DTM	people information	Latest reprots from the Displacement Tracking Matrix (DTM) system. DTM tracks and monitors the displacement and population mobility. It is designed to regurlarly and systematically capture, process and disseminate information to provide a better understanding of the movements and evolving needs of displaced populations, whether on site or en route.

yes	DS094	EXODI	people information; geo location	EXODI is an interactive web map built upon testimonies of 1,000 migrants from sub-Saharan Africa that were collected in nearly three years of activity (2014-2016) by the operators and volunteers of Medici per i Diritti Umani/Doctors for Human Rights (MEDU).They are part of those 730 thousand men, women and children landed on Italian shores in the last 15 years, of which more than half in the last 32 months. The map describes in the simplest and detailed way the Migratory Routes from Sub-Saharan Countries to Italy, the difficulties, the violence, the tragedy and hopes encountered during the trip by the protagonists. EXODI is not only a map showing the stages and paths, as well as a report with data and statistics, but above all, a testimony that describes life stories. It is an interactive and in progress web map that will be periodically updated with new testimonies gathered from all those who will share the story of their own journey. Through updated data EXODI aims also to describe the physical and mental consequences of the journey on the health of an entire generation of young Africans; a journey in which, as a witness said, "you are no longer considered as a human being".
yes	DS095	ecoi.net	people information	ecoi.net, the country of origin information system of the Austrian Red Cross, gathers, structures and processes publicly available country of origin information with a focus on the needs of asylum lawyers, refugee counsels and persons deciding on claims for asylum and other forms of international protection. Comprehensive country of origin information not only encompasses many dimensions of human rights but also other aspects not usually covered by human rights reports. This includes information on the living situation in a given country, presentation of ethnic groups and cultural traditions and assessments of the possible development of a security situation.
yes	DS096	EMM News Explorer	people information	The NewsExplorer uses JRC developed technology to automatically generate daily news summaries, allowing users to see: - the major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages; - The list of most mentioned names and find further automatically derived information (eg. variant name spellings, titles and phrases, list of the most recent articles and list of related persons and organisations.
yes	DS097	EMHRN's Migration and Asylum Blog	people information	News on migration and asylum from around the region
yes	DS098	NewsBrief Media Monitor	people information	Europe Media Monitor (EMM) lets you easily see, explore and understand current news reported by the world's online media. Monitoring thousands of news sources in over 70 languages, the system uses advanced information extraction techniques to automatically determine what is being reported in the news, where things are happening, who is involved and what they said. It provides a unique and independent viewpoint of what is being reported in the world right now. The EMM App allows you to track what is being said by people and organizations, follow news on a given topic (more than 2000 predefined topics) and see what are the biggest stories that are happening right now in the world in a given language.
yes	DS099	Countries Data	people information	For each data set the latest value for the country is displayed. If historical values are available the drill down icon is displayed. The rank column displays the ranking of the selected country within all countries. Datasets from: electricity exports/imports, CO2 emissions, pollution, climate, etc
yes	DS100	Blog of Forced Migration Current Awareness	people information; text analysis	A service highlighting web research and information relating to refugees, asylum-seekers, internally displaced persons (IDPs), and other forced migrants

yes	DS101	Migratory routes map	people information; geo location	Detections of illegal border-crossings statistics download (updated monthly)
yes	DS102	Migration Geo-Portal	people information; geo location	The Migration Geo-Portal aspires to promote a better understanding of migratory trends towards Europe through in-depth data analysis and visualisation. Our work focuses specifically on migrant arrivals and fatal incidences during the sea journeys to Italy, Greece, and Spain. We update the Migration Geoportal every two months giving insight into the most recent developments in the Mediterranean diaspora.
yes	DS103	Giz	people information; geo location	GIZ provides numerous information on its projects. It thus supports the international efforts to improve the effectiveness of international cooperation by publishing clear, timely, easily accessible and detailed information. Against this background, GIZ publishes detailed project data as well as project presentations on all ongoing projects. On the website, you will find a large number of aggregated data on GIZ's projects and programs as well as detailed information on ongoing projects in the partner countries and regions. The data are updated daily. Worldwide projects.
yes	DS104	Global Detention Project	people information	The GDP's activities inlcude: (1) providing policy-makers, civil society actors, and human rights institutions with a source of accurate information and analysis about detention and other immigration control regimes, with a particular focus on the impact these policies have on the health, human rights, and well being of undocumented migrants, asylum seekers, and refugees; (2) developing and maintaining a measurable and regularly updated database that can be used to assess the evolution of detention practices, provide an evidentiary base for advocating reforms, and serve as a framework for comparative analysis; (3) working with academics and practitioners to develop policy relevant scholarship about detention systems; and (4) collaborating with advocacy organisations to document policies and practices through the launching of a interactive online database, the Global Immigration Detention Observatory.
yes	DS106	Blog: What drives human migration?	people information; text analysis; visual analysis	Blog posts about migration and refugees
yes	DS107	FEWS NET	people information	FEWS NET, the Famine Early Warning Systems Network, is a leading provider of early warning and analysis on acute food insecurity. Created in 1985 by the US Agency for International Development (USAID) after devastating famines in East and West Africa, FEWS NET provides objective, evidence-based analysis to help government decision-makers and relief agencies plan for and respond to humanitarian crises. Our products, published here on our website, include: - monthly reports and maps detailing current and projected food insecurity - timely alerts on emerging or likely crises - specialized reports on weather and climate, markets and trade, agricultural production, livelihoods, nutrition, and food assistance
yes	DS108	UNOSAT	people information	UNOSAT provides timely and high-quality geo-spatial information. UNOSAT develops solutions on integrating field collected data with remote sensing imagery and GIS data through web-mapping and information sharing mechanisms, including remote monitoring of development projects and sharing of geographic data using web-services. UNOSAT delivers integrated satellite-based solutions for human security, peace and socio-economic development, in keeping with the mandate given to UNITAR by the UN General Assembly since 1963. UNOSAT's goal is to make satellite solutions and geographic information easily accessible to the UN family and to experts worldwide who work at reducing the impact of crises and disasters and help nations plan for sustainable development.

yes	DS109	Aljazeera news	people information	Timeline: The rise of Yemen's Houthi rebels. A look at how Shia rebels changed the balance of power, eventually prompting Saudi-led military intervention.
yes	DS110	International Conference on Migration in Africa	people information	The International Conference on Migration in Africa (ICMA) is a forum that connects research on migration with a focus on the African continent.
yes	DS111	News about Syria	people information; visual analysis; text analysis	This website is from a group of Syrian human rights activists. They noticed the lack of bodies which document abuses against civilians inside Syria, therefore they decided to establish this project, it is specialized to document the violations which have committed by all sides in the ongoing conflict in Syria against civilians in a professional way. The organization is documenting all kinds of violations, and working on the accounting for the groups which committed these violations by the international community. The organization is completely independent, and does not follow any political or military bodies, whether inside or outside Syria.
yes	DS112	Iraq bease fire	people information	News, reports, pictures of violation occurring in Iraq
yes	DS113	Database of Documents on Peace and Security	people information; text analysis	The "Database on Peace and Security" (DFS) by the Institute for Religion and Peace (IRF) provides the full text of church documents and similar institutions. Information about religions in Egipt, Syria, Israel, Jordan, Lebanon
yes	DS114	Historical Data Diagrams per Year	people information	Historical Data Diagrams per Year in many countries: HIV/AIDS rates, unemployment rates, ...
yes	DS115	News website: online Focus	people information; text analysis	News about the crisis in the Arab world
yes	DS116	Wikipedia: List of govern systems by state/country	people information	List of govern systems by state/country
yes	DS118	KCMD Data Catalogue	people information	The Knowledge Centre on Migration and Demography (KCMD) Data Catalogue is a table of data sources relevant to Migration and Demography policies. Each data source is listed with its summary description, the link to its web site and other metadata. The catalogue will include official EU and international statistics, as well as important data sets at Member State level. This catalogue comprises 120 datasets: emmigration, immigration, population change, aylum applications, etc
yes	DS120	Emigration country website - current country information for emigrants	people information	Data about emigrating to slightly endangered countries in face of natural catastrophes
yes	DS121	Global Migration Futures	people information	Publications helping understand and prepare for future changes in international migration. Scenarios for: North Africa, Europe, Horn of Africa and Yemen, and the Pacific
yes	DS122	African Refugees in Israel	people information; text analysis	Diverse information about refugees, migrants, infiltrators, rights of African refugees, etc.
yes	DS123	Operational Portal: Refugee situations	people information	The Refugees Operational Portal is a Partners coordination tool for Refugee situations provided by UNHCR; Mediterranean Situation
yes	DS126	RSCAS Research Project Reports	people information	Reports on migration, asylum, fundamental rights, etc
yes	DS127	ACAPS	people information	ACAPS's information products and insight can be used by humanitarians to make better decisions, and our training and methodology work supports others to develop better assessments and analysis.

yes	DS164	London Police Records	people information	These 3 datasets provide a complete snapshot of crime, outcome, and stop and search data, as held by the Home Office from late 2014 through mid 2017 for London, both the greater metro and the city.
yes	DS163	Synthetic Data from a financial payment system		BankSim is an agent-based simulator of bank payments based on a sample of aggregated transactional data provided by a bank in Spain. The main purpose of BankSim is the generation of synthetic data that can be used for fraud detection research.
yes	DS168	EU Data Portal	aggregated dat a	The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries.

Table 7– List of datasets selected for usage during 3rd hackathon

6.3. Datasets selected for usage during the 4th and 5th Hackathons

SELP Unit approved	ID	Dataset	Use Case(s) to be used	Dataset description
yes	DS096	EMM News Explorer	people information	The NewsExplorer uses JRC developed technology to automatically generate daily news summaries, allowing users to see: - the major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages; - The list of most mentioned names and find further automatically derived information (eg. variant name spellings, titles and phrases, list of the most recent articles and list of related persons and organisations.
Yes	DS017	LibriSpeech ASR corpus	audio	Large-scale (1000 hours) corpus of read English speech
Yes	DS038	rob	people information	844 News articles (.txt), 35 Resumes (.docx) and ~1170 Email headers (single .csv) plus some additional info
yes	DS045	Words by lone offenders	Posts and social media	Manifestos and words written by lone offenders before they commited an attack. On english
Yes	DS065	VAST Challenge 2010 MC1	text/audio analysis	Text files containing intelligence reports, communication interceptions, blog and email records, providing evidences for the existence of criminal networks.
Yes	DS157	Urban Sounds	urban sounds	2 datasets: URBANSOUNS dataset contains 1302 labeled sound recordings. Each recording is labeled with the start and end times of sound events from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music. Each recording may contain multiple sound events, but for each file only events from a single class are labeled. URBANSOUNDS8K contains 8732 labeled sound excerpts (<=4s) of urban sounds from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music.
yes	DS057	FlickrLogos-27	visual analysis	annotated logo dataset downloaded from Flickr and contains more than four thousand classes in total

Yes	DS058	BelgaLogos	visual analysis	set of images covering all aspects of life and current affairs: politics and economics, finance and social affairs, sports, culture and personalities. Two different groundtruth are provided: a global groundtruth and a local groundtruth
Yes	DS074	Europe's wanted Fugitives	visual analysis	images of the Europe's most wanted ugitives
Yes	DS068	non-jihad videos Columbia Consumers Video dataset (CCV)	visual analysis	Columbia Consumers Video public dataset (CCV)
Yes	DS007	YouTube Faces DB	Face Recognition	YouTube Faces Database, a database of face videos designed for studying the problem of unconstrained face recognition in videos.
Yes	DS067	jihad-related videos	visual analysis	videos from Jihadology.net
Yes	DS075	Labeled Faces in the Wild (LFW) Dataset	visual analysis	images of public figures collected from the Internet
Yes	DS128	Stormfront	Social Network analysis	Posts from Stormfront Forum. Stormfront is a white nationalist, white supremacist and neo-Nazi Internet forum, and the Web's first major racial hate site.
yes	DS152	Jihadist attack threat video	Face detection	A video of a known ISIS member threatening Spain with a Jihadist attack. From: https://www.almasdarnews.com/article/disturbing-video-isis-militants-spain-celebrate-barcelona-terror-attack/
Yes	DS154	Town Center Database	Facial recognition	Dataset composed of videos of busy streets to be used for facial recognition in the scope of ASGARD
yes	DS169	Fake synthesized tweets on DS048 (second sub-dataset)	Posts and social media	syntethic data from DS048. Only a link was modified on the existing tweets; the inserted link was for the video approved in DS152
yes	DS180	INOV simulated cctv video	video	video of internal CCTV cameras of INOV, researchers are acting and simulating, researchers have singed the informed consent, it can be used only in the hackathon sections

Table 8 – List of datasets selected for usage during 4th and 5th hackathons

7. Conclusion

7.1. Summary

The purpose of this document is to report the outcomes of T5.2, describing the datasets acquisition and creation process. Throughout T5.2 activity, the most effective way has been sought to, for the one hand, fulfil the partners requests demanding the needed datasets for developing, testing and evaluating the ASGARD tools and by the another hand, ensure that datasets within ASGARD project comply with EU ethical, data protection and privacy principles (EP). As described before, this was ensured with a close collaboration between T5.2 team, the ASGARD SELP Unit and all partners requesting datasets. The description on how it was made is included in the following sections:

- In Section 2 we have shown how it was possible to create a process for datasets acquisition, taking in consideration EU ethical, data protection and privacy principles.
- In Section 3 we have described the dataset collection and covered use cases.
- In Section 4 we have listed the created datasets to fulfil not covered needs.
- In Section 5 we have shown the maintenance process of the dataset collection taking in consideration the defined 4 goals: an easy to consult dataset catalogue, the availability of the datasets collection and support for the defined workflow for new datasets acquisition (Figure 1 in Section 2.2) as well as get efficiency during all stages of the process.
- In Section 6 we have presented the list of datasets used in the hackathons.

7.2. Evaluation

The mechanism put in place to respond to developers and LEAs demands have proved to work both on keep the datasets available and to procure and create datasets to address development and evaluation needs. It was also possible to ensure that datasets within the ASGARD project comply with EU ethical, data protection and privacy principles.

To achieve these goals, without considerable delays, it was crucial to establish an efficient communication process between the requesting partners, T5.2 team and SELP Unit and to encourage the reuse and sharing of the approved datasets through an easy to consult catalogue.

7.3. Future work

The datasets maintenance will continue until the end of the project to support developers and LEAs on their tools development and evaluation tasks. This will include the work to ensure the dataset's catalogue and collection availability, as well as the evaluation, on a per-request basis, of not evaluated yet datasets (not coloured in the list presented in ANNEX III).

ANNEX I.GLOSSARY AND ACRONYMS

Term	Definition / Description
LEA	Law Enforcement Agency
GDPR	General Data Protection Regulation (EU) 2016/679
SELP	Societal Ethical Legal and Privacy
EP	ethical, data protection and privacy principles
DoA	ASGARD document containing the description of work during the project
Data Controller	According European Commission (EC), the data controller determines the purposes for which and the means by which personal data is processed.
CCTV	Close circuit television
CTF	Capture the flag

Table 9 - Glossary and Acronyms

ANNEX II. REFERENCES

The table below shows the most significant references used and/or cited to prepare this document:

Reference	Source
Zapperi	Zapperi, Stefano, Kent Bækgaard Lauritsen, and H. Eugene Stanley. "Self-organized branching processes: mean-field theory for avalanches." *Physical review letters* 75.22 (1995): 4071.
Brown	www.nltk.org/book/ch02.html
CIFAR10	www.cs.toronto.edu/~kriz/cifar.html keras.io/datasets/
KERAS	keras.io
NLTK	www.nltk.org
European Commission definition for Data Controller	https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controller-processor/what-data-controller-or-data-processor_en

ANNEX III. Datasets collection list

SELP Unit approved	Available for ASGARD?	ID	Dataset	Source	Use Case(s) to be used	Dataset description
Yes	Yes	DS007	YouTube Faces DB	internet	Face Recognition	YouTube Faces Database, a database of face videos designed for studying the problem of unconstrained face recognition in videos.
Yes	Yes	DS008	Higgs	Internet	Posts & Social Media	The Higgs dataset has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered.
Yes	Yes	DS010	Youtube videos 8M	Internet	video analysis	8million youtube videos with annotations
Yes	Yes	DS017	LibriSpeech ASR corpus	Johns Hopkins University	audio analysis	Large-scale (1000 hours) corpus of read English speech
Yes	Yes	DS027	Computer Forensic Tool Testing Dataset (CFReDS)	DHS/NIST	computer and storage	These reference data sets (CFReDS) provide to an investigator documented sets of simulated digital evidence for examination. Since CFReDS would have documented contents, such as target search strings seeded in known locations of CFReDS, investigators could compare the results of searches for the target strings with the known placement of the strings. The CFReDS site is a repository of images.
Yes	Yes	DS035	Standford I2V	Stanford Digital Repository	Image/video analysis	a state-of-the-art query by image video dataset
Yes	Yes	DS036	TRECVID	TRECVID	Image/video analysis	a well known benchmarking dataset for image/video retrieval. Most videos have some metadata provided by the donor available e.g., title, keywords, and description.
Yes	Yes	DS038	VAST Challenge 2014 MC1	internet	people information	844 News articles (.txt), 35 Resumes (.docx) and ~1170 Email headers (single .csv) plus some additional info
Yes	Yes	DS039	VAST Challenge 2014 MC2	internet	Geo location	32MB of GPS car tracking data plus some additional info
Yes	Yes	DS040	VAST Challenge 2011 MC3	internet	people information	4474 News reports
Yes	Yes	DS041	VAST Challenge 2006	internet	people information	~1200 News articles plus some additional info

yes	Yes	DS045	Words by lone offenders	internet	Posts & Social Media	Manifestos and words written by lone offenders before they commited an attack. On english
yes	Yes	DS048	Pro-IS tweets	internet	Posts & Social Media	Pro-IS tweets. The dataset includes: name, username, description, location, number of followers at the time the tweet was downloaded, number of statuses by the user when the tweet was downloaded, date and timestamp of the tweet, the tweet itself
Yes	Yes	DS056	FlickrLogos-32	internet	image analysis	photos showing brand logos with binary masks and bounding boxes that mark the position of the logo in each image
Yes	Yes	DS057	FlickrLogos-27	internet	image analysis	annotated logo dataset downloaded from Flickr and contains more than four thousand classes in total
Yes	Yes	DS058	BelgaLogos	internet	image analysis	set of images covering all aspects of life and current affairs: politics and economics, finance and social affairs, sports, culture and personalities. Two different groundtruth are provided: a global groundtruth and a local groundtruth
Yes	Yes	DS059	Logos-160	internet	image analysis	logo image database for brand recognition, with a total of 100 brands, 160 logo categories, 73414 images, and 130,608 logo objects annotated with a bounding box and category label.
Yes	Yes	DS060	METU	internet	image analysis	large dataset (the largest publicly available logo dataset as of 2014), which is composed of more than 900K real logos belonging to real companies worldwide. The dataset also includes query sets of varying difficulties, allowing Trademark Retrieval researchers to benchmark their methods against other methods to progress the field.
Yes	Yes	DS062	CoNLL-2003	internet	text analysis	Data for language-independent named entity recognition. We will concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. This data can be used for developing a named-entity recognition system.
Yes	Yes	DS065	VAST Challenge 2010 MC1	internet	text/audio analysis	Text files containing intelligence reports, communication interceptions, blog and email records, providing evidences for the existence of criminal networks.
Yes	Yes	DS067	jihad-related videos	internet	videos analysis	videos from Jihadology.net
Yes	Yes	DS068	non-jihad videos Columbia Consumers Video dataset (CCV)	internet	video analysis	Columbia Consumers Video public dataset (CCV)
Yes	Yes	DS069	Dabiq, Inspire, Rumiyh	internet and Pydio	image and text analysis	Islamic State's magazines with english text and pictures; this dataset also contains different articles split in different files

Yes	Yes	DS070	jihadi-related blogs (some already removed from internet)	Pydio	image and text analysis	text and images from blogs publicly available online, but currently closed down
Yes	Yes	DS071	Vox-Pol	Pydio	image and text analysis	Vox-Pol journal data
Yes	Yes	DS072	jihadist videos	internet	audio analysis	Jihadist videos with english audio
Yes	Yes	DS073	10 second sound clips	internet	audio analysis	2 million ten-second YouTube excerpts labeled with a vocabulary of 527 sound event categories, with at least 100 examples for each category.
Yes	Yes	DS074	Europe's wanted Fugitives	EUROPOL/ENFAST	image analysis	images of the Europe's most wanted ugitives
Yes	Yes	DS075	Labeled Faces in the Wild (LFW) Dataset	Labeled Faces in the Wild (LFW) Dataset	image analysis	images of public figures collected from the Internet
Yes	Yes	DS076	Color FERET Dataset	Color FERET Dataset	image analysis	a database of facial imagery was collected between December 1993 and August 1996. The database is used to develop, test, and evaluate face recognition algorithms.
Yes	Yes	DS078	Skype logs	Box	Posts & Social Media	Skype messages jihad-related. Adapted to fit the Generic Use Case for the 1st Hackathon.
Yes	Yes	DS079	Audio conversations	Pydio	audio analysis	Audio recordings of slot conversations with jihad-related content from DS078; VoIP filter
Yes	Yes	DS085	Webite of the German Federal Office for Migration and Refuggees	internet	people information	Repository of information on how to migrate, enter in or return to Germany
Yes	Yes	DS087	World Bank data	internet	people information	At the World Bank, the Development Data Group coordinates statistical and data work and maintains a number of macro, financial and sector databases. Working closely with the Bank's regions and Global Practices, the group is guided by professional standards in the collection, compilation and dissemination of data to ensure that all data users can have confidence in the quality and integrity of the data produced. World Bank databases are essential tools for supporting critical management decisions and providing key statistical information for Bank operational activities. The application of internationally accepted standards and norms results in a consistent, reliable source of information.

Yes	Yes	DS089	Website of the Centre on Religion and Geopolitics	internet	people information	The Centre on Religion & Geopolitics (CRG) is an international affairs think-tank. It presents informed analysis on the global interaction of religion, geopolitics, and conflict.
Yes	Yes	DS090	Country Reports on Human Rights Practices for 2016	internet	people information	The annual Country Reports on Human Rights Practices – the Human Rights Reports – cover internationally recognized individual, civil, political, and worker rights, as set forth in the Universal Declaration of Human Rights and other international agreements. The U.S. Department of State submits reports on all countries receiving assistance and all United Nations member states to the U.S. Congress in accordance with the Foreign Assistance Act of 1961 and the Trade Act of 1974.
Yes	Yes	DS091	Website of EU Institute for Security Studies	internet	people information	Repository for many security-related topics: EU foreign policies, global governance, secuirty and defense, transnational challenges.
Yes	Yes	DS092	Latest reports from DTM	internet	people information	Latest reprots from the Displacement Tracking Matrix (DTM) system. DTM tracks and monitors the displacement and population mobility. It is designed to regurlarly and systematically capture, process and disseminate information to provide a better understanding of the movements and evolving needs of displaced populations, whether on site or en route.
Yes	Yes	DS094	EXODI	internet	people information; geo location	EXODI is an interactive web map built upon testimonies of 1,000 migrants from sub-Saharan Africa that were collected in nearly three years of activity (2014-2016) by the operators and volunteers of Medici per i Diritti Umani/Doctors for Human Rights (MEDU).They are part of those 730 thousand men, women and children landed on Italian shores in the last 15 years, of which more than half in the last 32 months. The map describes in the simplest and detailed way the Migratory Routes from Sub-Saharan Countries to Italy, the difficulties, the violence, the tragedy and hopes encountered during the trip by the protagonists. EXODI is not only a map showing the stages and paths, as well as a report with data and statistics, but above all, a testimony that describes life stories. It is an interactive and in progress web map that will be periodically updated with new testimonies gathered from all those who will share the story of their own journey. Through updated data EXODI aims also to describe the physical and mental consequences of the journey on the health of an entire generation of young Africans; a journey in which, as a witness said, "you are no longer considered as a human being".
Yes	Yes	DS095	ecoi.net	internet	people information	ecoi.net, the country of origin information system of the Austrian Red Cross, gathers, structures and processes publicly available country of origin information with a focus on the needs of asylum lawyers, refugee counsels and persons deciding on claims for asylum and other forms of international protection. Comprehensive country of origin information not only encompasses many dimensions of human rights but also other aspects not usually covered by human rights reports. This includes information on the living situation in a given country, presentation of ethnic groups and cultural traditions and assessments of the possible development of a security situation.

yes	Yes	DS096	EMM News Explorer	internet	people information	The NewsExplorer uses JRC developed technology to automatically generate daily news summaries, allowing users to see: - the major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages; - The list of most mentioned names and find further automatically derived information (eg. variant name spellings, titles and phrases, list of the most recent articles and list of related persons and organisations.
yes	Yes	DS097	EMHRN's Migration and Asylum Blog	internet	people information	News on migration and asylum from around the region
yes	Yes	DS098	NewsBrief Media Monitor	internet	people information	Europe Media Monitor (EMM) lets you easily see, explore and understand current news reported by the world's online media. Monitoring thousands of news sources in over 70 languages, the system uses advanced information extraction techniques to automatically determine what is being reported in the news, where things are happening, who is involved and what they said. It provides a unique and independent viewpoint of what is being reported in the world right now. The EMM App allows you to track what is being said by people and organizations, follow news on a given topic (more than 2000 predefined topics) and see what are the biggest stories that are happening right now in the world in a given language.
yes	Yes	DS099	Countries Data	internet	people information	For each data set the latest value for the country is displayed. If historical values are available the drill down icon is displayed. The rank column displays the ranking of the selected country within all countries. Datasets from: electricity exports/imports, CO2 emissions, pollution, climate, etc
yes	Yes	DS100	Blog of Forced Migration Current Awareness	internet	people information; text analysis	A service highlighting web research and information relating to refugees, asylum-seekers, internally displaced persons (IDPs), and other forced migrants
yes	Yes	DS101	Migratory routes map	internet	people information; geo location	Detections of illegal border-crossings statistics download (updated monthly)
yes	Yes	DS102	Migration Geo-Portal	internet	people information; geo location	The Migration Geo-Portal aspires to promote a better understanding of migratory trends towards Europe through in-depth data analysis and visualisation. Our work focuses specifically on migrant arrivals and fatal incidences during the sea journeys to Italy, Greece, and Spain. We update the Migration Geoportal every two months giving insight into the most recent developments in the Mediterranean diaspora.

yes	Yes	DS103	giz	internet	people information; geo location	GIZ provides numerous information on its projects. It thus supports the international efforts to improve the effectiveness of international cooperation by publishing clear, timely, easily accessible and detailed information. Against this background, GIZ publishes detailed project data as well as project presentations on all ongoing projects. On the website, you will find a large number of aggregated data on GIZ's projects and programs as well as detailed information on ongoing projects in the partner countries and regions. The data are updated daily. Worldwide projects.
Yes	Yes	DS104	Global Detention Project	internet	people information	The GDP's activities inlcude: (1) providing policy-makers, civil society actors, and human rights institutions with a source of accurate information and analysis about detention and other immigration control regimes, with a particular focus on the impact these policies have on the health, human rights, and well being of undocumented migrants, asylum seekers, and refugees; (2) developing and maintaining a measurable and regularly updated database that can be used to assess the evolution of detention practices, provide an evidentiary base for advocating reforms, and serve as a framework for comparative analysis; (3) working with academics and practitioners to develop policy relevant scholarship about detention systems; and (4) collaborating with advocacy organisations to document policies and practices through the launching of a interactive online database, the Global Immigration Detention Observatory.
Yes	Yes	DS106	Blog: What drives human migration?	internet	people information; text analysis; image analysis	Blog posts about migration and refugees
Yes	Yes	DS107	FEWS NET	internet	people information	FEWS NET, the Famine Early Warning Systems Network, is a leading provider of early warning and analysis on acute food insecurity. Created in 1985 by the US Agency for International Development (USAID) after devastating famines in East and West Africa, FEWS NET provides objective, evidence-based analysis to help government decision-makers and relief agencies plan for and respond to humanitarian crises. Our products, published here on our website, include: - monthly reports and maps detailing current and projected food insecurity - timely alerts on emerging or likely crises - specialized reports on weather and climate, markets and trade, agricultural production, livelihoods, nutrition, and food assistance

Yes	Yes	DS108	UNOSAT	internet	people information	UNOSAT provides timely and high-quality geo-spatial information. UNOSAT develops solutions on integrating field collected data with remote sensing imagery and GIS data through web-mapping and information sharing mechanisms, including remote monitoring of development projects and sharing of geographic data using web-services. UNOSAT delivers integrated satellite-based solutions for human security, peace and socio-economic development, in keeping with the mandate given to UNITAR by the UN General Assembly since 1963. UNOSAT's goal is to make satellite solutions and geographic information easily accessible to the UN family and to experts worldwide who work at reducing the impact of crises and disasters and help nations plan for sustainable development.
yes	Yes	DS109	Aljazeera news	internet	people information	Timeline: The rise of Yemen's Houthi rebels. A look at how Shia rebels changed the balance of power, eventually prompting Saudi-led military intervention.
Yes	Yes	DS110	International Conference on Migration in Africa	internet	people information	The International Conference on Migration in Africa (ICMA) is a forum that connects research on migration with a focus on the African continent.
Yes	Yes	DS111	News about Syria	internet	people information; visual analysis; text analysis	This website is from a group of Syrian human rights activists. They noticed the lack of bodies which document abuses against civilians inside Syria, therefore they decided to establish this project, it is specialized to document the violations which have committed by all sides in the ongoing conflict in Syria against civilians in a professional way. The organization is documenting all kinds of violations, and working on the accounting for the groups which committed these violations by the international community. The organization is completely independent, and does not follow any political or military bodies, whether inside or outside Syria.
Yes	Yes	DS112	Iraq bease fire	internet	people information	News, reports, pictures of violation occurring in Iraq
Yes	Yes	DS113	Database of Documents on Peace and Security	internet	people information; text analysis	The "Database on Peace and Security" (DFS) by the Institute for Religion and Peace (IRF) provides the full text of church documents and similar institutions. Information about religions in Egipt, Syria, Israel, Jordan, Lebanon
Yes	Yes	DS114	Historical Data Diagrams per Year	internet	people information	Historical Data Diagrams per Year in many countries: HIV/AIDS rates, unemployment rates, ...
Yes	Yes	DS115	News website: online Focus	internet	people information; text analysis	News about the crisis in the Arab world
Yes	Yes	DS116	Wikipedia: List of govern systems by state/country	internet	people information	List of govern systems by state/country

Yes	Yes	DS118	KCMD Data Catalogue	internet	people information	The Knowledge Centre on Migration and Demography (KCMD) Data Catalogue is a table of data sources relevant to Migration and Demography policies. Each data source is listed with its summary description, the link to its web site and other metadata. The catalogue will include official EU and international statistics, as well as important data sets at Member State level. This catalogue comprises 120 datasets: emmigration, immigration, population change, aylum applications, etc
Yes	Yes	DS120	Emigration country website - current country information for emigrants	internet	people information	Data about emigrating to slightly endangered countries in face of natural catastrophes
Yes	Yes	DS121	Global Migration Futures	internet	people information	Publications helping understand and prepare for future changes in international migration. Scenarios for: North Africa, Europe, Horn of Africa and Yemen, and the Pacific
Yes	Yes	DS122	African Refugees in Israel	internet	people information; text analysis	Diverse information about refugees, migrants, infiltrators, rights of African refugees, etc.
Yes	Yes	DS123	Operational Portal: Refugee situations	internet	people information	The Refugees Operational Portal is a Partners coordination tool for Refugee situations provided by UNHCR; Mediterranean Situation
Yes	Yes	DS126	RSCAS Research Project Reports	internet	people information	Reports on migration, asylum, fundamental rights, etc
Yes	Yes	DS127	ACAPS	internet	people information	ACAPS's information products and insight can be used by humanitarians to make better decisions, and our training and methodology work supports others to develop better assessments and analysis.
Yes	Yes	DS128	Stormfront	internet	Posts & Social Media	Posts from Stormfront Forum. Stormfront is a white nationalist, white supremacist and neo-Nazi Internet forum, and the Web's first major racial hate site.
yes	Yes	DS151	GALE Phase 3 Arabic Broadcast News Speech Part 1	LDC	audio analysis	GALE Phase 3 Arabic Broadcast News Speech Part 1 was developed by the Linguistic Data Consortium (LDC) and is comprised of approximately 132 hours of Arabic broadcast news speech collected in 2007 by the Linguistic Data Consortium (LDC), MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 3 of the DARPA GALE (Global Autonomous Language Exploitation) program.
yes	Yes	DS152	Jihadist attack threat video	Youtube	Face detection	A video of a known ISIS member threatening Spain with a Jihadist attack. From: https://www.almasdarnews.com/article/disturbing-video-isis-militants-spain-celebrate-barcelona-terror-attack/
Yes	Yes	DS154	Town Center Database	Internet	Facial recognition	Dataset composed of videos of busy streets to be used for facial recognition in the scope of ASGARD
yes	yes	DS155	YouTube videos - traffic cams	YouTube	image and video analysis	Traffic Cams and scenes from the 2014 oscars awards

Yes	yes	DS157	Urban Sounds	Internet	environmental sounds	2 datasets: URBANSOUNS dataset contains 1302 labeled sound recordings. Each recording is labeled with the start and end times of sound events from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music. Each recording may contain multiple sound events, but for each file only events from a single class are labeled. URBANSOUNDS8K contains 8732 labeled sound excerpts (<=4s) of urban sounds from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music.
Yes	Yes	DS158	Mivia Audio Events Dataset	Internet	environmental sounds	a total of 6000 events for surveillance applications, namely glass breaking, gun shots and screams. The 6000 events are divided into a training set (composed of 4200 events) and a test set (composed of 1800 events). The data set is designed to provide each audio event at 6 different values of signal-to-noise ratio (namely 5dB, 10dB, 15dB, 20dB, 25dB and 30dB) and overimposed to different combinations of environmental sounds in order to simulate their occurrence in different ambiences.
Yes	Yes	DS159	London Traffic Cams	Internet	Video analysis	different types of datasets (public transport) -Open Data, examples http://jamcams.tfl.gov.uk/00002.00878.jpg http://jamcams.tfl.gov.uk/00001.09642.mp4?time=636452153679172956
yes	yes	DS160	DS048 synthesized for GUC	DS048	text analysis	syntethic data from DS048. Names, Usernames and Tweets are modified from the original
yes	yes	DS161	UA-DETRAC labelled traffic cameras	internet	video analysis	The dataset consists of 10 hours of videos captured with a Cannon EOS 550D camera at 24 different locations at Beijing and Tianjin in China. The videos are recorded at 25 frames per seconds (fps), with resolution of 960×540 pixels. There are more than 140 thousand frames in the UA-DETRAC dataset and 8250 vehicles that are manually annotated, leading to a total of 1.21 million labeled bounding boxes of objects.
yes	yes	DS162	DS128 syntethic dataset for hate speech labelling (using only the texts)	DS128	text analysis	Create a new simulated/fake dataset extracting texts from the posts and assign to each a label indicating whether they contain hate-speech or not.
yes	yes	DS163	Synthetic Data from a financial payment system	internet		BankSim is an agent-based simulator of bank payments based on a sample of aggregated transactional data provided by a bank in Spain. The main purpose of BankSim is the generation of synthetic data that can be used for fraud detection research.
yes	yes	DS164	London Police Records	internet		These 3 datasets provide a complete snapshot of crime, outcome, and stop and search data, as held by the Home Office from late 2014 through mid 2017 for London, both the greater metro and the city.
yes	yes	DS166	Broad Twitter Corpus dataset	internet	Posts & Social Media	The dataset of tweets collected over stratified times, places and social uses. The goal is to represent a broad range of activities, giving a dataset more representative of the language used in this hardest of social media formats to process. Further, the BTC is annotated for named entities. The entities and the crowd annotations are all provided with the corpus, as well as (where possible) the raw twitter JSON.

yes	yes	DS168	EU Data Portal	internet	aggregated da ta	The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries.
yes	yes	DS169	Fake synthesized tweets on DS048 (second sub-dataset)	DS048	Posts & Social Media	syntethic data from DS048. Only a link was modified on the existing tweets; the inserted link was for the video approved in DS152
yes	yes	DS172	GUC disk image	nextcloud	computer and storage	Disk image created for the 1st Hackathon - content SELP approved and selected for the Generic Use Case
yes	yes	DS174	Interpol - Wanted Persons	internet	Image analysis	Interpol Wanted persons public dataset
yes	yes	DS175	Interpol - Missing Persons	internet	Image analysis	Interpol Missing persons public dataset
yes	yes	DS176	Google streetview dataset		Image analysis	images from Google street view were used from 32 European cities
yes	yes	DS177	Linguateca - Portuguese text dataset	internet	Text analysis	The dataset contains texts in Portuguese labelled with named entities and other linguistic information. Texts from newspaper and scientific, literary and transcribed spoken texts
yes	yes	DS178	Bing-search-engine (Firearms, Armoured vehicles)		Image analysis	images form BING-search engine of various types of firearms and various types of armoured vehicles
yes	yes	DS180	INOV simulated cctv video		Video analysis	video of internal CCTV cameras of INOV, researchers are acting and simulating, researchers have singed the informed consent, it can be used only in the hackathon sections
yes	yes	DS181	MALLORA video recorded		video/ image analysis	videos recorded during the Mallorca hackathon. Reserachers participants have signed the informed consent
yes	yes	DS183	AIT dummy txt files		Txt, e-mail	Lores ipsum self generated data: 100 txt files
yes	yes	DS184	AIT dummy pdf files		PDF	Lorem ipsum self generated data: 5 pdf files
Please contact SELP Unit	Yes	DS030	i-LIDS	CAST	visual analysis	video footage, annotated, for specific event detection and tracking scenarios, developed to test video analytics products (scenarios include perimeter intrusion, doorways surveillance, parked vehicle, abandoned baggage and person tracking) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/143875/ilids-user-guide.pdf

Please contact SELP Unit	Yes	DS031	FRL2011	CAST	visual analysis	Video and stills to test automated face recognition technology. The library comprises of still images, simulated surveillance footage, and CCTV footage taken from 10 camera positions at a major station, and is an extract from a larger dataset.
Please contact SELP Unit	Yes	DS084	Website of the Country of Origin Information Unit of the Austrian Federal Office for Immigration and Asylum	internet	people information	In the database of the Austrian Federal Office for Immigration and Asylum (in cooperation with the Red Cross/ACCORD - www.ecoi.net) you will find a comprehensive collection of all relevant and up-to-date documents regarding country of origin information.
Please contact SELP Unit	Yes	DS086	Base Maps	internet	people information; geo location	Base Map is a map of a country or area showing typical geographical information, including borders, administrative divisions, major cities, roads and disputed areas. The map can be used in PowerPoint to add information on a situation using icons and other PowerPoint objects like arrows and text boxes.
Please contact SELP Unit	Yes	DS093	Knowledge Centre on Migration and Demography - Dynamic Data Hub	internet	people information; geo location	Through interactive mapping, the Dynamic Data Hub gives direct access to single datasets to visualise migration and demography data and trends. Its restricted version, for Commission and EEAS internal use only, gives access to multiple datasets at a time and allows for some more in-depth analysis.
Please contact SELP Unit	Yes	DS105	Germany Trade and Invest Website	internet	people information; text analysis	Magazines and information about: markets in Germany, international markets, local and international business activities of German forms worldwide, ..
Please contact SELP Unit	Yes	DS119	Migration Information Source	internet	people information; text analysis; visual analysis	The Migration Information Source provides fresh thought, authoritative data, and global analysis of international migration and refugee trends.
Please contact SELP Unit	Yes	DS124	ReliefWeb: crisis information at your fingerprints	internet	people information	ReliefWeb provides reliable and timely information, enabling humanitarian workers to make informed decisions and to plan effective response. They collect and deliver key information, including the latest reports, maps and infographics and videos from trusted sources.
Please contact SELP Unit	Yes	DS125	ReliefWeb	internet	people information	ReliefWeb provides reliable and timely information, enabling humanitarian workers to make informed decisions and to plan effective response. They collect and deliver key information, including the latest reports, maps and infographics and videos from trusted sources. ReliefWeb editors identify and select the content that is most relevant to global humanitarian workers.
Please contact SELP Unit	Yes	DS149	Gender Recognition by Voice	internet	audio analysis	This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz (human vocal range).

Please contact SELP Unit	Yes	DS150	SAVAS	Internal	audio analysis	Corpus composed of 100 hours of clean and 120h of non-clean read speech (with transcriptions) in Spanish.
Please contact SELP Unit	Yes	DS153	Crowd video	Internet	Facial recognition	Video of a busy street to be used for facial recognition in the scope of ASGARD
Please contact SELP Unit	yes	DS156	Portuguese corpora	Internet	voices recognition	The Fundamental Portuguese Corpus is a corpus of spoken language, collected between 1970 and 1974, composed of 1800 recordings (500 hours) made in Continental Portugal and the Islands. Of these 1800 conversations, a sample was selected and transcribed.
Please contact SELP Unit	yes	DS165	SCface - Surveillance Cameras Face Database	internet	video/image analysis	SCface is a database of static images of human faces. Images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. Database contains 4160 static images (in visible and infrared spectrum) of 130 subjects. Images from different quality cameras mimic the real-world conditions and enable robust face recognition algorithms testing, emphasizing different law enforcement and surveillance use case scenarios.
Please contact SELP Unit	yes	DS167	Mobile phone records of German politician Malte Spitz	internet	mobile phone	This dataset contains records of all communications of a mobile phone serviced by Deutsche Telekom. Whenever the phone made a call or a connection, a timestamped record of the phone's location, service used, source, and destination was saved by the service provider. The phone's location was most likely estimated by the strength of the signal received at the cell towers.
Please contact SELP Unit	yes	DS170	Jihadist Propaganda Videos	GUCI	Video analysis	Hard disc with recorded/stored propaganda videos
Please contact SELP Unit	yes	DS171	Mobile phone records of Czech Ph.D. student Michal Ficek	internet	mobile phone	Mobile phone records of Czech Ph.D. student Michal Ficek collected in 2010-2011
Please contact SELP Unit	yes	DS173	DARK WEB datasets anonymised	internet	Darkweb	Anonymized data for the AlphaBay online anonymous marketplace (2014-2017). This dataset does not make available any textual information (item name, description, or feedback text). All handles (user id, item id) have been anonymised.
Please contact SELP Unit	yes	DS179	DeepFashion dataset	internet	Image analysis	Large-scale clothes database, which has several appealing properties: First, DeepFashion contains over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos. Second, DeepFashion is annotated with rich information of clothing items. Each image in this dataset is labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks. Third, DeepFashion contains over 300,000 cross-pose/cross-domain image pairs.
NO	Yes	DS011	2D MOT 2015	Internet	Tracking evaluation	a total of 22 sequences, of which half is for training and half for testing. the test data contains over 10 minutes of footage and 61440 annotated bounding boxes
No	Yes	DS061	GWERN	internet	text and visual analysis	Dataset with many dumps of marketplaces (text, images)
No	Yes	DS077	CASIA WebFace Dataset	CASIA WebFace Dataset	Image analysis	a large scale dataset containing 10,575 subjects and 494,414 images

NO	Yes	DS117	Website of the State Border Guard Service of Ukraine	internet	people information	The State Border Guard Service of Ukraine is charged with the tasks of ensuring inviolability of state borders and protection of sovereign rights of Ukraine within its exclusive (maritime) economic zone.
Not evaluated, Please contact SELP Unit	Yes	DS006	Migration related text documents	BMI - internet	People – Forecasting	pdf documents extracted from Websites including pictures including source and lately dates
Not evaluated, Please contact SELP Unit	Yes	DS009	Facebook	Internet	Posts & Social Media	The is the directed network of a small subset of posts to other user's wall on Facebook. The nodes of the network are Facebook users, and each directed edge represents one post, linking the users writing a post to the users whose wall the post is written on. Since users may write multiple posts on a wall, the network allows multiple edges connecting a single node pair. Since users may write on their own wall, the network contains loops.
Not evaluated, Please contact SELP Unit	Yes	DS012	.ISO image from a laptop	Internal	Forensic/ Computer & Storage	An .iso image as an exact image from a laptop. This laptop contained traces of web surfing habits and some documents or video files with content related with terrorism activities.
Not evaluated, Please contact SELP Unit	Yes	DS013	YFCC100M	Flickr	images analysis	100 Million images with their metadata, many parties including the UvA adding additional derived descriptions.
Not evaluated, Please contact SELP Unit	Yes	DS014	Audio Event Detection	Internal	audio analysis	Datasets for the detection of Audio events related to screaming, gunshooting, glass breaking and general noise
Not evaluated, Please contact SELP Unit	Yes	DS022	Corrupt NTFS MFT test data set	NICC	recovering deleted/damaged file systems and files	datadump contains corrupt MS Windows NTFS MFT file system meta-data
Not evaluated, Please contact SELP Unit	Yes	DS024	Image recovery dataset SKL/NFC	SKL/NFC	computer and storage	The file SKL_FAT32_16KiB.rar contains the disk image SKL FAT32 16KiB.001 depicting a digital storage device with the approximately size of 2GB. The disk image ins a partition that is formatted with FAT32 with cluster size of 16KiB. The disk image contains allocated and deleted complete and partial images. The disk image may only be used in research and education purpose. Commercial use is not allowed. Some of the images used are taken from Mathworks Matlab. When used the origin of the disk image must be referenced.
Not evaluated, Please contact SELP Unit	Yes	DS025	NPS - 2009 - CANON2	internet	computer and storage	Multiple images of a 32MB SD card shot in a Canon PowerShot SD800IS

Not evaluated, Please contact SELP Unit	Yes	DS026	Many disk images, RAM dumps, pictures, files, text	internet	computer and storage	From here you can view the available data: - ipod images - files jpeg, zip, html, txt, doc, mp3. pdf, exe, etc - emails - outlook pst files - disk images - pictures - videos - pcap files - logs - RAM dumps from laptops, PS3 - SD images - Smartphone images - SIM card images - tablet images - USB images - XBOX images http://datasets.fbreitinger.de/datasets/
Not evaluated, Please contact SELP Unit	Yes	DS028	Forensics Challenge File Image Layout (DFRWS)	NICC	computer and storage	THE dataset is a 50MB raw file. It has no file system, but it contains JPEG, ZIP, HTML, Text, and Microsoft Office files and fragments.
Not evaluated, Please contact SELP Unit	Yes	DS029	overview of forensic datasets	internet	computer and storage	large-scale corpora of forensically interesting information that are available for those involved in forensic research.
Not evaluated, Please contact SELP Unit	Yes	DS032	Digital forensics tool testing dataset	CAST	computer and storage	Dataset constructed to test digital forensics triage tools. Consists of images of different Windows OS with limited activity on web browsing, file sharing and social media usage and a range of file type datasets.
Not evaluated, Please contact SELP Unit	Yes	DS034	RSS feeds from different sources	various news portals	people information	RSS structured documents enhanced with locations, persons, organisations
Not evaluated, Please contact SELP Unit	Yes	DS042	Criminal Information Database (data model)	Internal	People information, Illicit Markets	A data model of a criminal information system so the data can be sinthesized
Not evaluated, Please contact SELP Unit	Yes	DS043	IBM Connection	IBM	Posts & Social Media	Social network records within the company
Not evaluated, Please contact SELP Unit	Yes	DS044	Twitter Dataset	Twitter	Posts & Social Media	twitter records collected daily
Not evaluated, Please contact SELP Unit	Yes	DS046	Blogs	internet	Posts & Social Media	Blogs written on different languages (english, swedish, spanish, french and russian)

Not evaluated, Please contact SELP Unit	Yes	DS047	Halummu	internet	Posts & Social Media	Pro-IS blog with lots of material, including all issues of dabiq
Not evaluated, Please contact SELP Unit	Yes	DS049	Flashback	internet	Posts & Social Media	Swedish xenophopic discussion forum
Not evaluated, Please contact SELP Unit	Yes	DS050	Pascal-VOC	internet	image analysis	a set of images annotated with ground truth; Ground truth information in each annotated image includes a bounding box for the objects of interest and might also include pixel segmentation masks or polygonal boundaries
Not evaluated, Please contact SELP Unit	Yes	DS051	ILSVRC / ImageNet	internet	image analysis	150,000 photographs hand labeled with the presence or absence of 1000 object categories
Not evaluated, Please contact SELP Unit	Yes	DS052	ILSVRC / ImageNet	internet	image analysis	images annotated with ground truth, i.e. bounding boxes for categories in the image have been labeled
Not evaluated, Please contact SELP Unit	Yes	DS053	MS-COCO	internet	image analysis	images with annotations; COCO currently has three annotation types: object instances, object keypoints, and image captions.
Not evaluated, Please contact SELP Unit	Yes	DS054	YFCC100M	internet	Image/video analysis	dataset contains a list of photos and videos.
Not evaluated, Please contact SELP Unit	Yes	DS055	Open Images	internet	image analysis	dataset of ~9 million URLs to images that have been annotated with labels spanning over 6000 categories.
Not evaluated, Please contact SELP Unit	Yes	DS063	Digital Forensics Tool testing images	internet	computer and storage	The following are file system and disk images for testing digital (computer) forensic analysis and acquisition tools.
Not evaluated, Please contact SELP Unit	Yes	DS064	Computer forensics certification	internet	computer and storage	This is a data set that is used by digital forensics analysts to get certified.
Not evaluated, Please contact SELP Unit	Yes	DS066	SPOKE	CAST	audio analysis	Database of voices intended for use in testing speaker recognition technology. Consists of 2,191 recordings from 100 speakers (65 male and 35 female, aged 17 to 65, average age 32) recorded on various recording devices.

Not evaluated, Please contact SELP Unit	Yes	DS080	Arabic Speech Corpus	internet	audio analysis	Corpus recorded in south Levantine Arabic (Damascian accent). Contains 1813 spoken utterances. Contains also extra 18 minutes of fully annotated corpus.
Not evaluated, Please contact SELP Unit	Yes	DS081	Arabic Speech Corpus for Isolated words	internet	audio analysis	Arabic speech corpus for isolated words containing 9992 utterances of 20 words spoken by 50 native male Arabic speakers.
Not evaluated, Please contact SELP Unit	Yes	DS082	Statista	internet	people information	Statistics from unemployment rate in Ukraine from 2007 to 2017
Not evaluated, Please contact SELP Unit	Yes	DS083	ReliefWeb Afghanistan	internet	people information; video analysis	ReliefWeb is the leading humanitarian information source on global crises and disasters. It is a specialized digital service of the UN Office for the Coordination of Humanitarian Affairs (OCHA). We provide reliable and timely information, enabling humanitarian workers to make informed decisions and to plan effective response. We collect and deliver key information, including the latest reports, maps and infographics and videos from trusted sources. This dataset contains humanitarian information about Afghanistan.
Not evaluated, Please contact SELP Unit	Yes	DS088	ReliefWeb countries	internet	people information	ReliefWeb is the leading humanitarian information source on global crises and disasters. It is a specialized digital service of the UN Office for the Coordination of Humanitarian Affairs (OCHA). We provide reliable and timely information, enabling humanitarian workers to make informed decisions and to plan effective response. We collect and deliver key information, including the latest reports, maps and infographics and videos from trusted sources. In this website we are able to browse by country for overviews, news, analysis and maps on crises and disasters. A red dot beside the country name denotes an ongoing crisis or disaster.
Not evaluated, Please contact SELP Unit	Yes	DS129	Xbox One partitions	internet	Computer and storage	5 Xbox One partitions
Not evaluated, Please contact SELP Unit	Yes	DS130	BOSS - Break Our Steganographic System	internet	image analysis	The dataset is part of a steganographic challenge. The goal of the player is to figure out, which images contain a hidden message and which images don't.

Not evaluated, Please contact SELP Unit	Yes	DS131	Pictures - University of Florence Image Communication Laboratory	internet	image analysis	Several datasets: - MICC-F220: this dataset is composed by 220 images; 110 are tampered and 110 originals. - MICC-F2000: this dataset is composed by 2000 images; 700 are tampered and 1300 originals. - MICC-F8multi: 8 tampered images with realistic multiple cloning. - MICC-F600: this dataset is composed by 440 original images, 160 tampered images and 160 ground truth images
Not evaluated, Please contact SELP Unit	Yes	DS132	The Database of Faces	internet	image analysis	The ORL Database of Faces'), contains a set of face images taken between April 1992 and April 1994 at the lab. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department. There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). A preview image of the Database of Faces is available.
Not evaluated, Please contact SELP Unit	Yes	DS133	TrustFoto - Columbia Image Splicing Detection Evaluation Dataset	internet	image analysis	1845 image blocks with a fixed size of 128 pixels x 128 pixels. The image blocks are extracted from images in the CalPhotos collection, with a small number of additional images captured by digital cameras. The dataset includes about the same number of authentic and spliced image blocks, which are further divided into different subcategories (smooth vs. textured, arbitrary object boundary vs. straight boundary).
Not evaluated, Please contact SELP Unit	Yes	DS134	TrustFoto - Columbia Uncompressed Image Splicing Detection Evaluation Dataset	internet	image analysis	There are 2 directories in this dataset: 4cam_auth & 4cam_splc. 4cam_auth containts authentic images, and 4cam_splc contains spliced images. By the term 'authentic', we mean an image that is taken using just one camera. In 4cam_auth, there are 183 images, and in 4cam_splc, there are 180. The image sizes range from 757x568 to 1152x768 and are uncompressed, in either TIFF or BMP formats. The spliced images are created using the authentic images, without any post processing. Full EXIF information is included in authentic images. The images are mostly indoor scenes: labs, desks, books ...etc. Only 27 images, or 15%, are taken outdoors on a cloudy day (which makes the outdoor illumination similar to indoor conditions).
Not evaluated, Please contact SELP Unit	Yes	DS135	TrustFoto - Columbia Photographic Images and Photorealistic Computer Graphics Dataset	internet	image analysis	The dataset is composed of four component image sets, i.e., the Photorealistic Computer Graphics Set, the Personal Photographic Image Set, the Google Image Set, and the Recaptured Computer Graphics Set. This dataset, available from the Trustfoto website, will be for those who work on the photographic images (PIM) versus photorealistic computer graphics (PRCG) classification problem, which is a subproblem of the passive-blind image authentication research

Not evaluated, Please contact SELP Unit	Yes	DS136	Dresden Image Database	internet	image analysis	The 'Dresden Image Database' website provides an interface to two different sources of images: - the 'Dresden Image Database', which includes the original set of images as taken by the maintainer of the website and his colleagues, and - an additional collection of flickr(r) images, taken by camera models that have also been used to compile the 'Dresden Image Database'.
Not evaluated, Please contact SELP Unit	Yes	DS137	Detection and Localization of Region-Level Video Forgery with Spatio-Temporal Coherence Analysis	internet	image analysis	Two datasets: The first data set contains 11 test video sequences, which is numbered 1-11, are captured using a Canon IXUS 750 digital video camera. The second data set contains 7 test video sequences, which is numbered 12-18, are obtained from the PETS 2009
Not evaluated, Please contact SELP Unit	Yes	DS138	NRCS Photo Gallery e USDA Natural Resources	internet	Image/video analysis	11 videos with different sizes, from ~2GB to ~280GB
Not evaluated, Please contact SELP Unit	Yes	DS139	e-mails Digital Corpora	internet	text analysis	12 emails
Not evaluated, Please contact SELP Unit	Yes	DS140	Outlook PST file - DFRWS 2009 Forensics Rodeo	internet	text analysis	Dataset from a Challenge: PST File in a Forensics Challenge: DFRWS 2009 Rodeo
Not evaluated, Please contact SELP Unit	Yes	DS141	DFRWS 2006	internet	Computer and storage	Datasets from a challenge: The dataset is a 50MB raw file. It has no file system, but it contains JPEG, ZIP, HTML, Text, and Microsoft Office files and fragments.
Not evaluated, Please contact SELP Unit	Yes	DS142	DFRWS 2008 Rodeo: Laptop memory image, Thumb drive	internet	Computer and storage	Dataset from a Challenge: Memory capture of a computer system; 128 MB USB thumb drive and Canon digital camera
Not evaluated, Please contact SELP Unit	Yes	DS143	Nitroba University Harassment Scenario - Digital Corpora	internet	Computer and storage	Dataset from a Scenario: PCAP file
Not evaluated, Please contact SELP Unit	Yes	DS144	Computer Forensic Tool Testing (CFTT) - NIST	internet	Computer and storage	11 hard disk images

Not evaluated, Please contact SELP Unit	Yes	DS145	Disk Images - Digital Corpora	internet	Computer and storage	**nps-2009-canon2** — A set of images taken on with a Canon digital camera that can be used to test basic file recovery, fragmented file recovery, and file carving. **nps-2009-casper-rw** — An ext3 file system from a bootable USB token that had an installation of Ubuntu 8.10. The operating system was used to browse several US Government websites. **nps-2009-hfsjtest1** — A test image of a journaled HFS system in which the data from a previous version of a file can only be recovered from the HFS journal **nps-2009-ntfs1** — A test image of an NTFS file system including unfragmented and highly fragmented files stored in raw, compressed, and encrypted directories. The decryption key is provided. **nps-2009-ubnist1** — The FAT32 file system from which the nps-2009-capser-rw disk image was extracted. **nps-2009-domexusers** — This is a disk image of a Windows XP SP3 system that has two users, domexuser1 and domexuser2, who communicate with a third user (domexuser3) via IM and email. Two versions of this disk image will be provided: **nps-2009-domexusers** – The full system, distributed as an encrypted disk image. **nps-2009-domexusers-redacted** – The full system with the Microsoft Windows executables redacted so that they cannot be executed. **nps-2010-emails** — is a test disk image consists of 30 different email addresses, each one stored in a different document with a different coding scheme. **nps-2014-usb-nondeterministic** – this is a series of disk images that were made from a USB storage device that produced different data each time it was read. The original submission ZIP file and narrative are presented, as well as E01 files that were created by extracting the raw files from the ZIP image and re-encoding them.
Not evaluated, Please contact SELP Unit	Yes	DS146	iPod images	internet	Computer and storage	10 iPod images
Not evaluated, Please contact SELP Unit	Yes	DS147	DFRWS 2009 Forensics Challenge	internet	Computer and storage	Dataset from a Scenario: - 3 network traces in pcap format - 1 PS3 physical memory dump - 2 filesystem images
Not evaluated, Please contact SELP Unit	Yes	DS148	The Art of Memory Forensics	internet	Computer and storage	Dataset from a book: - memory images - evidence files
Not evaluated, Please contact SELP Unit	yes	DS182	INOV video recorded		Video analysis	video recorded by INOV. It shows na INOV researcher talking directly to a camera, mentioning words associated with terrorism and religion

ANNEX IV. Script for DS043 dataset generation

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import datetime as dt
import nltk

def create_dir(directory):
    wpath=os.path.dirname(os.path.abspath(__file__))
    new_path=os.path.join(wpath, directory)
    if not os.path.exists(new_path):
        os.makedirs(new_path)
        return new_path
    else:
        return new_path

def create_input_image():
    '''
    :return:  folder with sample images frp, the cfer dataset. call this function only one time
    '''
    from keras.datasets import cifar10
    from PIL import Image
    import datetime as dt

    labels=['airplane','automobile','bird','cat','deer','dog','frog','horse','ship', 'truck']
    (x_train, y_train), (x_test, y_test) = cifar10.load_data()
    ipath=create_dir('Images')
    trn_path= create_dir(os.path.join('Images', 'train'))
    tst_path= create_dir(os.path.join('Images', 'test'))
    val_path= create_dir(os.path.join('Images', 'valid'))
    for i in range(len(y_train)):
        A=x_train[i]
        lbs= labels[int(y_train[i])]
        c_path= create_dir(os.path.join('Images', 'train', lbs ))
        im = Image.fromarray(A)
        filename=os.path.join(trn_path, lbs,'trn_'+str(i)+'.jpeg')
        im.save(filename)
    for i in range(len(y_test)):
        r=np.random.uniform(0,1)
        lbs= labels[int(y_test[i])]
        if r <0.5:
            A=x_test[i]
            c_path= create_dir(os.path.join('Images', 'test', lbs))
```

```
            im = Image.fromarray(A)
            filename=os.path.join(tst_path, lbs ,'tst_'+str(i)+'.jpeg')
            im.save(filename)
        else:
            A=x_test[i]
            c_path= create_dir(os.path.join('Images', 'valid', lbs))
            im = Image.fromarray(A)
            filename=os.path.join(val_path, lbs  ,'tst_'+str(i)+'.jpeg')
            im.save(filename)
    print('DOne')


def plot_space_dis(x1,y1):
    import matplotlib
    import pylab
    import numpy as np
    import matplotlib.pyplot as plt
    from mpl_toolkits.mplot3d.axes3d import Axes3D
    from matplotlib import cm
    xAmplitudes = x1
    yAmplitudes = y1

    x = np.array(xAmplitudes)   #turn x,y data into numpy arrays
    y = np.array(yAmplitudes)

    fig = plt.figure()          #create a canvas, tell matplotlib it's 3d
    ax = fig.add_subplot(111, projection='3d')

    #make histogram stuff - set bins - I choose 20x20 because I have a lot of data
    hist, xedges, yedges = np.histogram2d(x, y, bins=(20,20))
    xpos, ypos = np.meshgrid(xedges[:-1]+xedges[1:], yedges[:-1]+yedges[1:])

    xpos = xpos.flatten()/2.
    ypos = ypos.flatten()/2.
    zpos = np.zeros_like (xpos)

    dx = xedges [1] - xedges [0]
    dy = yedges [1] - yedges [0]
    dz = hist.flatten()

    cmap = cm.get_cmap('jet') # Get desired colormap - you can change this!
    max_height = np.max(dz)   # get range of colorbars so we can normalize
    min_height = np.min(dz)
    # scale each z to [0,1], and get their rgb values
    rgba = [cmap((k-min_height)/max_height) for k in dz]

    ax.bar3d(xpos, ypos, zpos, dx, dy, dz, color=rgba, zsort='average')
    plt.title("X vs. Y Amplitudes for ____ Data")
    plt.xlabel("My X data source")
```

```
    plt.ylabel("My Y data source")
    plt.savefig("Your_title_goes_here")

def sigma(t):
    h=t.hour
    if h>12: 24-h
    s=1.0/(h+1)
    return s


def plotting():
    sel=df[df.date.apply(lambda s: s.hour) == 12]
    sel=sel[sel.date.apply(lambda s: s.day) == 12]
    print(sel)
    plot_space_dis(sel.LAT.tolist(),sel.LONG.tolist())
    sel1=df[df.date.apply(lambda s: s.hour) == 0]
    sel1=sel1[sel1.date.apply(lambda s: s.day) == 12]
    plot_space_dis(sel1.LAT.tolist(),sel1.LONG.tolist())
    plt.show()


def get_sentence(d):
    N=len(list(d.keys()))
    id=np.random.randint(N)
    return  d[id]


def get_sentence4topic(t, chat, Brown):
    if t=='others':
        s= get_sentence(chat)
    else:
        s= get_sentence(Brown[t])
    return s



def get_chattext():
    from nltk.corpus import webtext
    chat={}
    i=0
    ids=[fileid for fileid in webtext.fileids()]
    for id in ids :
        Sents=webtext.sents(id)
        for Sent in Sents:
            s=" ".join(Sent)
            chat[i]=s
            i+=1
    return chat
```

```
def randomize_topic(d_topic):
    r=np.random.uniform(0,1)
    if r < 0.5:
        return d_topic[0]
    else:
        return  d_topic[int(np.random.randint(1,len(topics)))]


def avalanches(l,a,N) :
    if l!='others':
        return  int(rndm(1, N, a, size=1))
    else:
        return 1


def rndm(a, b, g, size=1):
    """Power-law gen for pdf(x)\propto x^{g-1} for a<=x<=b"""
    r = np.random.random(size=size)
    ag, bg = a**g, b**g
    return (ag + (bg - ag)*r)**(1./g)


def collect_sentence():
    chat= get_chattext()
    topics=['others']
    brown_categories=nltk.corpus.brown.categories()
    topics.extend(brown_categories)
    d_topic=dict([(i,topics[i]) for i in range(len(topics))])
    Brown=dict()
    for c in brown_categories:
        Brown[c]=dict()
        sents=nltk.corpus.brown.sents(categories=c)
        i=0
        for sent in sents:
            Brown[c][i]=" ".join(sent)
            i+=1
    return topics, d_topic, chat,  Brown


def add_image(text, image_classes):
    r=None
    text=str(text)
    for c in  image_classes:
        if text.lower().find(c)>-1:
            r=c
            break
    return r

def add_image_file(t):
```

```
    if t != None:
        fnames=os.listdir(os.path.join('Images','test', t))
        i = np.random.randint(len(fnames))
        f=fnames[i]
    else:
        f=None
    return f




topics, d_topic, chat,  Brown=   collect_sentence()
image_classes= os.listdir(os.path.join('Images','test'))   #for thise created from
create_input_image(   use ['airplane','automobile','bird','cat','deer','dog','frog','horse','ship',
'truck']
print(image_classes)
columns=['date','LAT','LONG','TOPIC']
n_days=30
daily_records=5000
d=dict()
lat= 53.342998628  # Dublin citu centre coordinates
long= -6.256165642
dfres=pd.DataFrame(columns= columns  )
j=0
dr=dict()
for day in range(n_days):
    for line in range(daily_records):
        hour=np.int64(np.abs(np.random.normal(12,4)))
        if hour> 23: hour=23
        min=np.random.randint(0,60)
        sec=np.random.randint(0,60)
        t= dt.datetime(2018,1,1,hour,min,sec)+dt.timedelta(day,0)
        d['date']=pd.Timestamp(year=t.year, month=t.month, day=t.day, hour=t.hour,
minute=t.minute, second=t.second)
        d['LAT']=  lat + 0.001*np.random.normal(0,sigma(t))
        d['LONG']= long + 0.001*np.random.normal(0,sigma(t))
        d['TOPIC']= randomize_topic(d_topic)
        dr[j]=pd.Series(d)
        j+=1
        s=avalanches( d['TOPIC'],-2,100  )
        a=0.000001
        if s> 1:
            d1=dict()
            for l in range(s):
                t1 =t +dt.timedelta(0, np.random.randint(1,15))
                d1['date']=  pd.Timestamp(year=t1.year, month=t1.month, day=t1.day,
hour=t1.hour, minute=t1.minute, second=t1.second)
                d1['LAT']=d['LAT'] + +  a*np.random.uniform(1,2)
```

```
            d1['LONG']=d['LONG'] + +   a*np.random.uniform(1,2)
            d1['TOPIC']=d['TOPIC']
            dr[j]=pd.Series(d1)
            j+=1


dfres=pd.DataFrame(dr ).transpose()
#dfres=dfres.T

dfres['AUTHOR']=dfres.TOPIC.apply(lambda t :
'user_'+str(int(np.ceil(np.random.lognormal(topics.index(t),1)))))
dfres['TEXT']=dfres.TOPIC.apply(lambda t : get_sentence4topic(t, chat, Brown))
dfres.TEXT=dfres.TEXT.str.replace(',', ';').str.replace('"', ' ').str.replace("'", " ")
dfres['IMAGE_CLASS']=dfres.TEXT.apply(lambda t : add_image(t, image_classes))
dfres['IMAGE']=dfres.IMAGE_CLASS.apply(lambda t : add_image_file(t))

dfres.to_csv('real_time.csv')
```