Deliverable 1.1: Data Management Plan

E. Apostolidis, S. Papadopoulos, V. Mezaris, L. Nixon, R. Garcia, A. Scharl, G. Innerwinkler, G. Rudinger, J. Spangenberg, R. Bouwmeester, T. Koch, D. Teyssou

28/04/2016

Work Package 1: Project and Innovation Management

InVID - In Video Veritas: Verification of Social Media Video Content for the News Industry

Innovation Action

Horizon 2020, Research and Innovation Programme

Grant Agreement Number 687786

Dissemination level	*Public*
Contractual date of delivery	*30/04/2016*
Actual date of delivery	*28/04/2016*
Deliverable number	*D1.1*
Deliverable name	*Data Management Plan*
File	*D1.1*_v1.0
Nature	*Report*
Status & version	*Final, V1.0*
Number of pages	*32*
WP contributing to the deliverable	*WP1, and ALL WPs*
Task responsible	*CERTH*
Other contributors	*MODUL, UdL, WLT, APA-IT, DW, AFP*
Author(s)	*Evlampios Apostolidis, Symeon Papadopoulos, Vasileios Mezaris (CERTH), Lyndon Nixon (MODUL), Roberto Garcia (UdL), Arno Scharl (WLT), Gerald Innerwinkler, Gerhard Rudinger (APA-IT), Jochen Spangenberg, Ruben Bouwmeester, Tim Koch (DW), Denis Teyssou (AFP)*
Quality Assessor	*Max Göbel, WLT*
EC Project Officer	*Miguel Montarelo-Navajo*
Keywords	*InVID Data Management Plan*

Abstract:

This deliverable outlines the Data Management Plan of the InVID project. Based on the guidelines of the European Commission for the definition of a Data Management Plan, it lists the datasets that will be collected, processed or generated by the project, identifying whether and how they will be exploited or made accessible for re-use, and how they will be curated and preserved. The Data Management Plan of the InVID project is a working document that evolves during the lifespan of the project, and an updated version of this plan, improved by integrating findings about these datasets and possibly also additional datasets, as the project progresses, will be submitted to the European Commission in Month 21 of the project (September 2017).

Table of contents

1 Introduction

This deliverable presents the Data Management Plan of the InVID project. In particular, it describes in detail the adopted management policy for the datasets that will be collected, processed or generated by the project. The utilized approach: (a) ensures that any sensitive data are kept safe, (b) identifies whether and how the data will be exploited or made publicly accessible so as to maximize their reuse potential, and (c) indicates how these data will be curated and preserved, in accordance with the activities described in Task T1.3 Quality, data and knowledge management.

The European Commission (EC) has defined a number of guidelines / requirements for maximizing the reuse potential of scientific data, via making them easily discoverable, intelligible, usable beyond the original purpose for which they were collected and interoperable to specific quality standards. Using as a basis these guidelines we apply the methodology that is outlined in Section 2. According to this approach, for each dataset we specify: (a) its name (based on a standardized referencing approach), (b) its description, (c) the utilized standards and metadata, (d) the applicable data sharing policy and (e) the intended actions for its archiving and preservation. Further explanation regarding the information that needs to be considered and reported for each one of these features is given in Sections 2.1 to 2.5. Subsequently, based on this methodology, Section 3 lists and describes the datasets of the InVID project in a per-workpackage-basis (Sections 3.1 to 3.7). The concluding Section 4 briefly summarizes the information reported in the deliverable.

The InVID Data Management Plan is a working document that evolves during the lifespan of the project. For this reason an updated version of the Data Management Plan, enhanced by exploiting the findings and the decisions made as the project proceeds, will be produced and delivered as part of deliverable D1.3 titled "Updated Data, quality and knowledge management plan", which will be submitted to the EC in Month 21 of the project (September 2017).

1.1 History of the document

Table 1: History of the document

Date	Version	Name	Comment
11/02/2016	V0.1	E. Apostolidis, V. Mezaris, CERTH	Skeleton of the deliverable
17/02/2016	V0.2	S. Papadopoulos, CERTH	Addition of a first list of WP3 datasets
25/02/2016	V0.3	R. Garcia, UdL	Addition of WP4 dataset
10/03/2016	V0.4	G. Innerwinkler, G. Rudinger, APA-IT	Addition of WP7 datasets

Date	Version	Name	Comment
11/03/2016	V0.5	D. Teyssou, AFP	Addition of WP8 Market Study and WP3 TVLogos datasets
11/03/2016	V0.6	L. Nixon, MODUL	Addition of two WP2 datasets
18/03/2016	V0.7	J. Spangenberg, R. Bouwmeester, T. Koch, DW	Addition of WP6 dataset
22/03/2016	V0.8	A. Scharl, WLT	Addition of WP5 dataset
04/04/2016	V0.9	E. Apostolidis, V. Mezaris, CERTH	Complete draft version
06/04/2016	V0.10	E. Apostolidis, V. Mezaris, CERTH	Complete version submitted for Quality Assurance
13/04/2016	V0.11	E. Apostolidis, V. Mezaris, CERTH	After QA version of the deliverable; input from partners requested
28/04/2016	V1.0	E. Apostolidis, S. Papadopoulos, V. Mezaris, CERTH	Final document after Quality Assurance, submitted to the EC

2 Applied methodology

The applied methodology for drafting this initial Data Management Plan of the project was based on the guidelines of the EC[1] and the DMPonline[2] tool which can be used for implementing such a plan in a structured manner via a series of questions that need to be clarified for each dataset of the project. According to these guidelines, the Data Management Plan of the InVID project addresses the points below on a per dataset basis, reflecting the current status within the consortium about the data that will be produced:

- Dataset reference and name
- Dataset description
- Standards and metadata
- Data sharing
- Archiving and preservation (including storage and backup)

A more detailed description of the information that is considered and reported for each one of these subjects, is provided in the following subsections.

2.1 Dataset reference and name

For convenient reference on the data that will be collected and/or generated in the project we had to define a naming pattern. A referencing approach that contains information about the WP that owns/uses the dataset, the serial number of the dataset and the title of the dataset is the following: *InVID_Data_"WPNo."_"DatasetNo."_"DatasetTitle"*. According to this pattern, an example dataset reference name could be *InVID_Data_WP1_1_UserGeneratedContent*.

2.2 Dataset description

The description of the dataset that will be collected and/or generated includes information regarding the origin (in case of data collection), nature and scale of the data, as well as details related to the potential users of the data. In later editions of this document, this section will also clarify whether these data have been used in InVID to support a scientific publication (as a general rule, we expect most of the InVID datasets to indeed support one or more scientific publications). Information on the existence of similar data and the possibilities

[1] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://dmponline.dcc.ac.uk

for integration and reuse, if any, is also provided. Last but not least, concerning the nature of the data, potential negative effects on persons that are dealing with these data due to mentally traumatic and/or frustrating content will also be highlighted in this section (at present, this does not apply to any of the datasets listed in this document).

2.3 Standards and metadata

This section outlines how the data will be collected and/or generated and which community data standards (if any) will be used at this stage. Moreover it provides information on how the data will be organized during the project, mentioning for example naming conventions, version control and folder structures. For a detailed overview of the used standards the following questions were considered:

- How will the data be created?
- What standards or methodologies will be used?
- Which structuring and naming approach will be applied for folders and files?
- How different versions of a dataset will be easily identifiable?

In addition this section reports the types of metadata that will be created to describe the data and aid their discovery. Information about how this information will be created/captured and where it will be stored is also reported. The aspects bellow were examined for determining the necessary ways and types of generating and using metadata:

- How these metadata are going to be captured/created?
- Can any of this information be created automatically?
- What metadata standards will be used and why?

2.4 Data sharing

This point describes how the collected and/or generated data will be shared. For this, it reports on access procedures and embargo periods (if any), and lists technical mechanisms and software/tools for dissemination and exploitation/re-use of these data. Moreover it determines whether access will be widely open or restricted to specific groups (e.g. due to participant confidentiality, consent agreements or Intellectual Property Rights (IPR)), while it outlines any expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties. In case a dataset cannot be shared, the reasons for this are mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related). Last but not least, identification of the repository where data will be stored, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.) is also performed. The questions bellow were studied for concluding to the most appropriate sharing policy for each dataset of the project:

- How these data are going to be available to others?
- With whom will the data be shared, and under what conditions?
- Are any restrictions on data sharing required (e.g. limits on who can use the data, when and for what purpose)?
- What restrictions are needed and why?
- What actions will be taken to overcome or minimise restrictions?
- Where (i.e. in which repository) will the data be deposited?

2.5 Archiving and preservation

The established data archiving and preservation policy defines the procedures that will be put in place for long-term preservation of the data. In particular it indicates how long the data will be preserved and what is their approximated end volume. It also outlines the plans for preparing and documenting data for sharing and archiving. In case of not using an established repository, the Data Management Plan demonstrates the resources and systems that will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

A set of questions that were considered for defining the archiving and preservation policy for the datasets of the project is given bellow:

- What is the long-term preservation plan for the dataset (e.g. deposit in a data repository)?
- Are any additional resources needed to deliver our plan?
- Is there sufficient storage and equipment or additional may be needed?

3 Datasets in InVID

This section lists the datasets that will be created or collected for the needs of the InVID project, grouping them in a per-workpackage basis. Based on the methodology presented in Section 2, each dataset is defined by: (a) its name, (b) its description, (c) the used standards and accompanying metadata, (d) the applied data sharing policy, and (e) the adopted mechanisms for its archiving and preservation.

As a key component for the creation and management of these datasets, data privacy issues will be closely monitored from the beginning of the project, and the project's Data Protection Officer (Mr. Max Göbel from WLT) as well as, where necessary, the external Ethics Board with be consulted on this, to ensure that the collection, use and sharing of the data will not raise ethical concerns.

As a general statement about the adopted data collection and management policy for the datasets reported in the following subsections, we would like to declare that InVID is a scientific project. Therefore, any use of third-parties copyrighted material within its scope is meant to be made for scientific purposes and under the exception set forth in article 5.3.a of the Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. In order to set the licensing needs of the project, should it become a commercial one, as well as any personal data issues that need to be addressed, each WP will consider any copyright, personal data and/or contractual limitations that applies to the media, software and/or data involved in their study. These limitations will be studied in order to provide recommendations on any agreements with the main services/platforms where User Generated Video (UGV) is found and/or with owners of such content that may be deemed necessary for the InVID tools to be able to treat such contents and deliver their verification and licensing outputs to the media industry.

3.1 WP2 Datasets

Dataset name	**InVID_Data_WP2_1_TRECVID**
Dataset description	This dataset is provided by NIST[3] to the participants of the TRECVID SIN[4] and MED[5] tasks. It will be used for developing technologies for video annotation with visual concept and event labels. The dataset is divided in two main parts.

[3] http://www.nist.gov/

Dataset name	**InVID_Data_WP2_1_TRECVID**
	The first part consists of approx. 18500 videos (354 GB, 1400 hours) under a Creative Commons (CC) license, in MPEG-4/H.264 format, and it is typically partitioned into training (approx. 11200 videos, 10 seconds to 6,4 minutes long; 210 GB, 800 hours total) and testing set (approx. 7300 videos, 10 seconds to 4,1 minutes long; 144 GB, 600 hours total) for video concept detection methods. The total number of concepts is 346, and the annotation of each of these videos is based on a pair of XML and TXT files; the XML file contains information about the shot segments of the video and the TXT file includes the shot-level concept-based annotation of the video via a number of positive and negative concept labels. Finally, a TXT file with metadata describing sets of relations between these concepts in the form of "concept A implies concept B" and "concept A excludes concept B", is also available. The second part is a collection of approx. 63000 videos (736 GB, 2520 hours) in MPEG-4/H.264 format, created by the Linguistic Data Consortium[6] and NIST. It is used for the development of video event detection techniques and is divided in three subsets: (a) a training set with 3000 (50 GB, 80 hours) positive or near-miss videos, and 5000 (51 GB, 200 hours) background (i.e., negative) videos, (b) a validation set of 23000 videos (272 GB, 960 hours), and (c) an evaluation set of 32000 videos (363 GB, 1280 hours). The number of considered events is 20, and the ground truth for this collection is stored in CSV files. These files provide the event-based annotations of the videos by defining the list of positive or near-miss videos for each visual event.
Standards and metadata	The videos of this static dataset are in MPEG-4/H.264 format, while their annotations and metadata are in TXT, XML and CSV files. The generated results after processing this dataset (extracted features, if any; automatic annotation results) will be stored in XML, JSON and MPEG-7 formats. They will be accompanied by a document (a word or pdf file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify the funder and research discipline, and (c) appropriate key words to help users to locate the data.
Data	This is a dataset created and provided to us by NIST, under specific conditions

[4] http://www-nlpir.nist.gov/projects/tv2015/#sin

[5] http://www-nlpir.nist.gov/projects/tv2015/#med

[6] https://www.ldc.upenn.edu

Dataset name	**InVID_Data_WP2_1_TRECVID**
sharing	that are linked with the TRECVID benchmarking activity. Sharing of the dataset is regulated by NIST, and we will comply with their requirements. We are not allowed to further share this dataset with third parties. We can, however, and will share the results of our processing of the dataset (automatic annotation results in XML, JSON and MPEG-7 formats) via the free-of-charge OpenAIRE[7] or Zenodo[8] platforms, under the express conditions that the data is used solely for the purposes of evaluating concept detection algorithms and may not be copied and re-used for any other purpose.
Archiving and preservation	The original dataset and the analysis results will be stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made. Moreover, as stated above, a set of processing outcomes of this dataset will be also made available on the free-of-charge OpenAIRE or Zenodo platforms.

Dataset name	**InVID_Data_WP2_2_ImageNet**
Dataset description	This dataset contains images of the online ImageNet[9] collection, which is organized and managed by the Stanford and Princeton Universities. It will be used for building and training Deep Convolutional Neural Networks (DCNNs) for video concept detection. In particular, ImageNet is an image dataset organized according to the WordNet[10] hierarchy (currently only the nouns); for each node of the hierarchy, related images (often several hundreds or thousands of them) are provided. The current dataset is the one released in fall 2011 and is an updated version of the initial collection[11]. It contains approx. 15 million images in high resolution JPEG format, which are clustered in categories that correspond to 22000 distinct concepts of the WordNet structure.

[7] https://www.openaire.eu
[8] https://zenodo.org
[9] http://image-net.org/index
[10] http://wordnet.princeton.edu
[11] http://www.image-net.org/papers/imagenet_cvpr09.pdf

Dataset name	**InVID_Data_WP2_2_ImageNet**
	Images of each concept are quality-controlled and human-annotated.
Standards and metadata	This static dataset is composed by images that are mainly in high resolution JPEG format. The created metadata after analyzing these images can be: (a) local features extracted from these images, that are stored in BIN of TXT files, and (b) the output of the trained DCNNs (i.e., the classification decision), which is stored in TXT files. These data will be accompanied by a document (a word file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify the funder and discipline of the research, and (c) appropriate key words to help internal users to locate the data.
Data sharing	The ImageNet dataset is freely available for non-commercial research and/or educational use, by following the procedure and adopting the terms of use that are described in the ImageNet website[12].
Archiving and preservation	The original dataset and the results of processing it will be stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made. The archiving and preservation of this dataset are performed by the Stanford and Princeton Universities; InVID will have no involvement in this process.

Dataset name	**InVID_Data_WP2_3_TopicDetection**
Dataset description	This dataset is intended for the benchmark evaluation of the topic detection results produced in the InVID project. For a baseline, we will have one set of documents which contains 24 hours of collected news articles from English international media, together with a ground truth annotation of topics which emerge in this collection. For topic detection from Twitter streams we will have another set of documents in the dataset, which is a collection of Twitter content (from the Streaming API) over a 24 hour period.

[12] http://image-net.org/download-faq

Dataset name	**InVID_Data_WP2_3_TopicDetection**
Standards and metadata	This static dataset will be an index of JSON serialised documents, where each document captures the textual content and metadata (e.g. date-time published) for one news article or tweet, according to the webLyzard document model. The ground truth will be stored in a file as a description of the newsworthy topics which occur in the dataset.
Data sharing	This dataset will be generated from the documents crawled in a 24hr period by the webLyzard platform. The resulting data will be made available to third parties under the express conditions that the data is used solely for the purposes of evaluating topic detection algorithms and may not be copied and re-used for any other purpose.
Archiving and preservation	The dataset will be stored persistently (i.e. guaranteed until project's end and planned to be kept also after the project for an undefined period of time) on a MODUL University server (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches), and on request can be made available for download.

Dataset name	**InVID_Data_WP2_4_SocialMediaRetrieval**
Dataset description	This dataset is intended for the benchmark evaluation of the social media retrieval produced in the InVID project. It will consist of a set of social media postings collected from different social networks as a result of different general queries on named entities who are in the news at that time, e.g. the name of a celebrity, or a geographical location. A ground truth annotation will tag which posts in the dataset are directly related to a news story about the named entity.
Standards and metadata	This static dataset will be an index of JSON serialised documents, where each document captures the textual content and metadata (e.g. date-time published) for one social media posting, according to the webLyzard document model and extended with the ground truth annotation with the news story the posting is directly related to.
Data sharing	This dataset will be generated from the documents queried in a 24-hour period by the webLyzard platform. The resulting data will be made available to third

Dataset name	**InVID_Data_WP2_4_SocialMediaRetrieval**
	parties under the express conditions that the data is used solely for the purposes of evaluating social media retrieval and may not be copied and re-used for any other purpose.
Archiving and preservation	The dataset will be stored persistently (i.e. guaranteed until project's end and planned to be kept also after the project for an undefined period of time) on a MODUL University server (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches), and on request can be made available for download.

3.2 WP3 Datasets

Dataset name	**InVID_Data_WP3_1_WildWebTamperedImages**
Dataset description	This dataset was collected by CERTH within the REVEAL project. It will be used for testing the existing image forensics capabilities offered by TUNGSTEN. Its description is available on: http://mklab.iti.gr/project/wild-web-tampered-image-dataset The dataset contains 80 cases of forgeries, all confirmed from multiple reliable sources and with the help of the original photographs, where available. For each forgery, the dataset contains all instances that we could find on the Web using the Google and TinEye reverse image search services. The downloaded files went through a hash comparison to filter out exact file duplicates. After this step, the entire collection contains 13,577 unique images. By further removing images that were considered inappropriate for the task of evaluating image tampering detection algorithms, the remaining images are 10,870. In addition, the dataset contains manually created masks corresponding to the tampered area (ground truth).
Standards and metadata	The root folder of this static dataset contains two subfolders: WildWeb and UnsplicedSources. The former contains 90 subfolders, each containing one subcase. The naming convention is, in all cases, the name of the case, followed by a number, if multiple subcases exist. Within each such folder are the images, plus two subdirectories. The first subdirectory, called Mask contains all the mask files for the subcase, in the form of PNG images, with

Dataset name	**InVID_Data_WP3_1_WildWebTamperedImages**
	white (255) corresponding to the tampered region and black (0) to the rest of the image pixels. The second subdirectory, called Crops – PostSplices, contains all cropped and re-spliced versions of the subcase.
Data sharing	Due to copyright considerations, the dataset is not publicly available. However, for research purposes, the dataset creator may share the dataset following an electronic request by interested parties.
Archiving and preservation	The original dataset and the results of processing it will be stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made.

Dataset name	**InVID_Data_WP3_2_InVidFakeVideos**
Dataset description	This dataset will be collected for testing a number of verification approaches. It will be composed by a set of videos that have been found to be fake (or misleading). For each video the dataset will contain: the source (link where the video was found), metadata about the video (both embedded in the video file and available from the platform hosting the video), contextual information (e.g. website(s) or social media posts where the video appeared). In addition, we consider including in the dataset annotations that journalists produce during the verification process.
Standards and metadata	A simple and lightweight annotation scheme will be defined to accommodate the needs of this corpus. The serialization format will most likely be JSON to enable easy parsing, extensibility and ease of storage and retrieval. The dataset will be versioned by the WP3 leader (CERTH).
Data sharing	Since the corpus will be collected by the InVID consortium, we will consider making it publicly available. However, since part of the data will come from third party platforms (e.g. YouTube, Twitter, etc.), we will first need to investigate the legal constraints and issues that may arise from such an action.
Archiving	The original dataset and the results of processing it will be stored on the file

Dataset name	**InVID_Data_WP3_2_InVidFakeVideos**
and preservation	servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made.

Dataset name	**InVID_Data_WP3_3_VisualGeometryGroupDatasets**
Dataset description	This refers to two datasets from the Visual Geometry Group, namely the Oxford buildings (http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/) and the Paris dataset (http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/). These datasets have been extensively used to test similarity-based search approaches and hence are considered as one of the benchmarks to use for assessing the InVID near-duplicate search solution. The Oxford Buildings Dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. The Paris Dataset consists of 6412 images collected from Flickr by searching for particular Paris landmarks.
Standards and metadata	Each of these two static datasets consists of a set of image files (from Flickr) and ground truth in custom text format.
Data sharing	The datasets are available from the dedicated pages of the Visual Geometry Group, and hence no further sharing is foreseen within InVID.
Archiving and preservation	The dataset are stored and maintained by the Visual Geometry Group on a dedicated dataset page: http://www.robots.ox.ac.uk/~vgg/data/

Dataset name	**InVID_Data_WP3_4_InriaDatasets**
Dataset description	This refers to two datasets available from INRIA, namely the Holidays and Copydays datasets. These are expected to be useful for evaluating the near-duplicate detection solution of InVID. The Holidays dataset is a set of images which mainly contains some of the creators' personal holiday photos. The remaining ones were taken on purpose to test the robustness to various attacks: rotations, viewpoint and illumination changes, blurring, etc. The dataset includes a very large variety of scene types (natural, man-made, water and fire effects, etc.) and images are in high resolution. The dataset contains 500 image groups, each of which represents a distinct scene or object. The first image of each group is the query image and the correct retrieval results are the other images of the group. The Copydays dataset is a set of images which is exclusively composed of the creators' personal holiday photos. Each image has suffered three kinds of artificial attacks: JPEG, cropping and "strong". The motivation is to evaluate the behavior of indexing algorithms for most common image copies. More information is available on: https://lear.inrialpes.fr/~jegou/data.php.
Standards and metadata	This static dataset contains: (a) the images themselves, (b) the set of descriptors extracted from these images, (c) a set of descriptors produced, with the same extractor and descriptor, for a distinct dataset (Flickr60K), (d) two sets of clusters used to quantize the descriptors (again obtained from Flickr60K), (e) some pre-processed feature files for one million images, that were used by the dataset creators to perform the evaluation on a large scale.
Data sharing	The datasets are available from the dedicated page of INRIA and hence no further sharing is foreseen within InVID.
Archiving and preservation	The datasets are stored and maintained by INRIA on a dedicated dataset page: https://lear.inrialpes.fr/~jegou/data.php.

Dataset name	**InVID_Data_WP3_5_CCWEBVIDEO**
Dataset description	The dataset is called CC_WEB_VIDEO, named by the initials of City University of Hong Kong and Carnegie Mellon University, and which was collected from

Dataset name	**InVID_Data_WP3_5_CCWEBVIDEO**
	the web video sharing web site YouTube and video search engines Google Video and Yahoo! Video. It will be used for evaluating the near-duplicate detection solution of InVID. This static dataset was collected by considering 24 queries designed to retrieve the most viewed and top favorite videos from YouTube. Each text query was issued to YouTube, Google Video, and Yahoo! Video respectively. The videos were collected in November, 2006. Videos with time duration over 10 minutes were removed from the dataset. The final data set consists of 12,790 videos. More information is available on: http://vireo.cs.cityu.edu.hk/webvideo/.
Standards and metadata	Links to the videos, metadata and ground truth information are stored in simple text files, which are further described in the dataset page.
Data sharing	The dataset is available from the dedicated page of City University Hong Kong, and hence no further sharing is foreseen within InVID.
Archiving and preservation	The dataset is stored and maintained by City University Hong Kong on a dedicated page: http://vireo.cs.cityu.edu.hk/webvideo/.

Dataset name	**InVID_Data_WP3_6_MediaevalVerifyingMultimediaUse**
Dataset description	This is a dataset consisting of tweets spreading both fake and real images and videos. It has been used as a benchmark in the Verifying Multimedia Use task in Mediaeval 2015. It is expected to be of interest for testing contextual verification approaches. The dataset was collected in a semi-automatic way, by first manually collecting a set of known cases of images and videos and then in any automatic way collecting tweets that shared those images/videos. Data cleaning has also been done using manual inspection.
Standards and metadata	The dataset comprises a set of tweet ids associated with basic metadata and ground truth information. All information is serialized in simple tab-separated text files.
Data	The dataset is available on:

Dataset name	**InVID_Data_WP3_6_MediaevalVerifyingMultimediaUse**
sharing	https://github.com/MKLab-ITI/image-verification-corpus
Archiving and preservation	The dataset will continue to be maintained on GitHub.

Dataset name	**InVID_Data_WP3_7_YFCC100M**
Dataset description	This is a dataset consisting of 99 million CC-licensed Flickr images and one million videos. It is currently the largest publicly available multimedia dataset. We primarily foresee its usage for the purpose of evaluating location detection approaches (relevant for T3.3), since a large percentage of the images and videos are geo-located. In addition, the dataset has been extensively used within the Placing Task of Mediaeval. More details on the dataset are available on the following article from Communications of the ACM: http://cacm.acm.org/magazines/2016/2/197425-yfcc100m/fulltext
Standards and metadata	This static dataset comprises the metadata of the images in tab-separated text file format. Furthermore, some extensions of the dataset available from http://mmcommons.org include the original images, visual features extracted from the images and audio features extracted from the videos.
Data sharing	The dataset is available through the Yahoo Research WebScope program, while several extensions to the dataset are available at http://mmcommons.org. Hence, no further sharing is foreseen within InVID.
Archiving and preservation	The dataset is stored and maintained by Yahoo Research through their WebScope program: https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67 Furthermore, the Lawrence Livermore National Laboratory hosts several extensions of the dataset on: https://multimediacommons.wordpress.com/features/

Dataset name	**InVID_Data_WP3_8_TVChannelsLogos**
Dataset description	This dataset will be built for the needs of task T3.2, which is related to the collection of logos of TVs and user-generated channels on video platforms, along with the name and a description of the channel, a DBPedia URI if available, and tags. We intend to use this dataset to assess the performance of methods recognizing automatically logos in videos.
Standards and metadata	The dataset will be stored in a schemaless database and exposed as a web service to display relevant information on the channel's logos in the InVID verification platform. Moreover, a spreadsheet that will be versioned by AFP will be used as index of this dataset, storing for each logo its name, a short description and (potentially) a number of indicative images.
Data sharing	As part of the dissemination and exploitation strategy, we will consider exposing publicly the dataset as an API and/or a web tool.
Archiving and preservation	The dataset will be stored and maintained on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made.

3.3 WP4 Datasets

Dataset name	**InVID_Data_WP4_1_UGCRegisteredProviders**
Dataset description	This dataset will register User Generated Content (UGC) creators collected from social networks and other UGC online sources (such as YouTube, Twitter or Facebook). These creators will be registered, after obtaining their informed consent, whenever one of their digital media items is selected because a potential user is interested in reusing it. Consequently, just preselected users will be gathered and no crawling of social networks or UGC sources will be performed. The dataset will keep the username and the source social network, plus all the reuse policies defined by the creator. In case that there are agreements between the creator and the reusers, these will be also stored in the database, associated to the creator and the licensed UGC. Moreover, a set of security measures will be defined (which will be reported in the corresponding project deliverable D4.2 "Framework and Workflows for UGC

Dataset name	**InVID_Data_WP4_1_UGCRegisteredProviders**
	Copyright Management") and applied in order to ensure that the aforementioned data within the project is not used for improper or unauthorized purposes. Finally, registered users will be offered the option to opt-out of the service. In this case, additional personal data collected during registration will be erased. However, links to original users in social networks, content and policies will be kept if they are required to contextualize existing agreements by the user opting-out.
Standards and metadata	This dataset will be based on Resource Description Framework (RDF) metadata and use different Web Ontologies to structure the data, including for example FOAF, SIOC, Schema.org, Media Ontology and Copyright Ontology. It will be stored in a database capable of storing semantic data based on RDF. Specific RDF properties for time intervals and instants will be used to track the evolution of the dataset, for instance keeping track of when a particular agreement between a creator and a reuser was established.
Data sharing	This dataset will be generated as a result of the InVID platform operation when the Rights Module is involved and is specific to its operation. As stated in its description, this dataset will basically contain UGV creators reuse policies and bilateral agreements between them and the reusers, which we expect that they will prefer not to fully expose in public. Consequently, this dataset won't be shared outside InVID.
Archiving and preservation	This dataset will be preserved at the same location where the Rights Management module is deployed, i.e. a server hosted at the premises of Universitat de Lleida. It will be protected by preventing unauthorized access to the server and ensuring that security software is up-to-date. Moreover, backup provisions will be made.

3.4 WP5 Datasets

Dataset name	**InVID_Data_WP5_1_News-Media**
Dataset description	This dataset is intended as a generic, domain-independent basis for building the initial system prototype (T5.2) including the multimodal analytics dashboard (T5.3), and help to assess the achieved progress on document annotation and

Dataset name	**InVID_Data_WP5_1_News-Media**
	topic detection. It will be continuously updated through WLT's crawling architecture, and by accessing RSS feeds embedded in the crawled Web content. Specific InVID content feeds from social media will later complement the dataset, to be analyzed individually or in combination.
Standards and metadata	The dataset will be a continuously updated index of JSON serialised documents, where each document captures the textual content and metadata (e.g. date-time published) for one news article or tweet, according to the webLyzard document model.
Data sharing	The resulting data will be made available as part of the InVID dashboard under the express conditions that the data is used solely for the purposes of evaluating individual technical components as well as the overall system (T5.4), and may not be copied and re-used for any other purpose.
Archiving and preservation	The dataset will be stored persistently on a webLyzard server, during and beyond the project, and will be downloadable (with certain restrictions) via the multimodal analytics dashboard (T5.3).

3.5 WP6 Datasets

Dataset name	**InVID_Data_WP6_1 _Industrial Requirements**
Dataset description	This dataset will contain all data on related UGC verification tools and initiatives and the ones focusing on video verification in particular, as well as the interviews that have been reported in the deliverable D6.1, entitled "InVID Initial Industrial Requirements". By its nature it will also list all requirements that have been derived from the market analysis as well as the interviews with key persons active in the field that have been conducted. The dataset is meant to list all relevant activities in the research fields InVID tackles in order to identify the advantages and shortcomings of already existing solutions and to collect a complete list of what needs to be developed in InVID to make it a commercially successful video verification platform.
Standards and	This dataset is designed to analyse the industrial requirements. The latter will be collected in a shared spreadsheet and can be stored in a repository or

Dataset name	**InVID_Data_WP6_1 _Industrial Requirements**
metadata	database if required. The spreadsheet will be versioned by the WP6 leader (CONDAT).
Data sharing	The dataset will be made available for project partners only. Nevertheless, D6.1 and its updates are public deliverables that can be downloaded from the project website.
Archiving and preservation	The spreadsheet will be maintained by the WP6 leader (CONDAT). Updates of the industrial requirements will be created in the course of the project.

3.6 WP7 Datasets

Dataset name	**InVID_Data_WP7_1_UGVideo1**
Dataset description	This dataset will include UGV and their relevant metadata that are created by the utilized mobile applications for capturing these videos (e.g. data about the creator/registered user of the video, details about the used device, geolocation data and so on). The owners of these videos will be requested to sign up to the platform and agree to the usage terms, thus providing their informed consent for the collection and processing of their data. The users will also have an option to “opt-out” by notifying the local newspapers representative. Moreover, a set of security measures will be defined (which will be reported in the corresponding project deliverable D7.1 "Activities and outcome of the Pilots, first report") and applied in order to ensure that the aforementioned data is not used for improper or unauthorized purposes.
Standards and metadata	The videos will be stored in their native format that is defined by the mobile phone type. The metadata provided by the mobile application (user-id, date and time of video taken, location if agreed by the user) are distributed according to the possibilities of the appropriate device (either embedded in the video file itself or in a sidecar file (XML) managed by the mobile application).
Data sharing	The videos of this dataset (which is a static dataset as videos will not be updated) that will be selected by the editors will be shared via the websites of local newspapers, mentioning also the credit (as provided by the user) and

Dataset name	**InVID_Data_WP7_1_UGVideo1**
	usually the location of the video. Both fake and validated videos will be shared (after been anonymised) within the project consortium in order to be used for further tests and evaluations. So, no sensitive information will be shared, something that will be clearly indicated upon signing up to the platform and agreeing to the usage terms.
Archiving and preservation	UGV will be stored in the data-center of APA-IT on high-availability object store hosted in two data centers (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches). Videos will be deleted after a time to be agreed on with the newspapers. Videos identified as fake videos, and validated videos will be stored for a longer period, something that has to be agreed on with the consortium.

Dataset name	**InVID_Data_WP7_2_CommunityManagement**
Dataset description	This dataset will contain all data needed to manage the selected online-user-groups of newspapers for the pilot tests. These data include email addresses of the users, usernames, date and time of agreeing to the usage terms, users' feedback, usage statistics and device-information, assignments to groups (e.g. members of firebrigades, local sportsclubs and similar). The users will also have an option to “opt-out” by notifying the local newspapers representative. The involved persons in these tests will be requested to sign up to the platform and agree to the usage terms, thus providing their informed consent for the collection and processing of their data. Moreover, a set of security measures will be defined (which will be reported in the corresponding project deliverable D7.1 "Activities and outcome of the Pilots, first report") and applied in order to ensure that the aforementioned data is not used for improper or unauthorized purposes.
Standards and metadata	These data will be stored in an SQL-database, and changes will be logged accordingly without versioning.
Data sharing	Data will be shared as aggregated data only within the consortium. This dataset will show which user-groups were involved in the pilot tests, how

Dataset name	**InVID_Data_WP7_2_CommunityManagement**
	actively they participated and similar statistics. Details on specific users are owned by the publishers who manage their user-base and are of no importance for the project's results itself, something that will be clearly indicated upon signing up to the platform and agreeing to the usage terms.
Archiving and preservation	The relational database will be run on servers in the data-center of APA-IT (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches), and backup provisions will be made.

3.7 WP8 Datasets

Dataset name	**InVID_Data_WP8_1_MarketStudy**
Dataset description	This dataset will include all data collected regarding the market of UGV verification. It will include company names, UGV publishers (such as broadcast TVs) and their websites, online video platforms, technology companies dealing with forensic verification or contextual verification on social networks, market figures, contact names and company information, which will be gathered mainly from the web.
Standards and metadata	As this dataset is designed to support the efforts for exploitation of the InVID consortium, it will be initially collected as a shared spreadsheet and later will be included in an SQL database if needed. The spreadsheet will be versioned by the WP8 leader (AFP).
Data sharing	Being collected by InVID partners for exploitation purposes, we will maintain internally this dataset, although some findings about new tools, companies, or publishers will be shared on our website and social networks accounts as part of our dissemination policy.
Archiving and preservation	The spreadsheet will be maintained by the WP8 leader (AFP). A backup procedure will be set up for the preservation of the data.

Dataset name	**InVID_Data_WP8_2_InVidDeliverables**
Dataset description	This dataset will be composed of the project deliverables that have to be prepared and submitted to the EC during the project's lifespan, according to the contractual obligations of the InVID consortium.
Standards and metadata	These documents will be stored in PDF format. For each deliverable we will provide: (a) the list of authors, (b) a brief description of its content (i.e. its abstract), (c) the related WP of the project, and (d) the contractual date for their submission to the EC. This dataset will be extended whenever new deliverables are submitted to the EC. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The public project deliverables will be made publicly available after their submission to the EC, via the project website.
Archiving and preservation	This dataset will be maintained on the project wiki and the relevant webpage of the project website[13], both hosted by a CERTH server which is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. This webpage will grant open access to the PDF file of each listed public deliverable.

Dataset name	**InVID_Data_WP8_3_InVidPublications**
Dataset description	This dataset will contain manuscripts reporting the conducted scientific work in InVID, which have been accepted for publication in peer-reviewed journals and conferences. All these publications will inlcude a statement with acknowledgement to the InVID project, while their content may vary from the description of specific analysis techniques, to established evaluation datasets and individual components or parts of the InVID platform.
Standards	Most commonly, these documents will be stored in PDF format. Each

[13] http://www.invid-project.eu/deliverables

Dataset name	**InVID_Data_WP8_3_InVidPublications**
and metadata	document will be also accompanied by: (a) details about the venue (e.g. conference, workshop or benchmarking activity) or journal where it was published, (b) a short description with the abstract of the publications, and (c) the LaTeX-related BIB file with its citation. This dataset will be extended whenever new submitted works are accepted for publication in conferences or journals. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	This dataset will be publicly available, following the guidelines of the EC[14] for open access to scientific publications and research data in Horizon2020.
Archiving and preservation	Self-archiving (also known as "green" open access) will be applied for ensuring open access to these publications. According to this archiving policy the author(s) of the publication will archive (deposit) the published article or the final peer-reviewed manuscript in online repositories, such as personal webpage(s), the project website[15] and the free-of-charge OpenAIRE[16] or Zenodo[17] repositories, after its publication. Nevertheless, the employed archiving policy will also be fully aligned with restrictions concerning embargo periods that may be defined by the publishers of these publications, making the latter publicly available in certain repositories only after their embargo period has elapsed.

Dataset name	**InVID_Data_WP8_4_InVidPresentations**
Dataset description	This dataset will consist of presentations prepared for reporting InVID-related scientific work or progress made, in a variety of different events, such as conferences, workshops, meetings, exhibitions, interviews and so on.
Standards	Most commonly these presentations will be in PPT or PDF format. Information

[14] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[15] http://www.invid-project.eu/publications

[16] https://www.openaire.eu

[17] https://zenodo.org

Dataset name	**InVID_Data_WP8_4_InVidPresentations**
and metadata	related to: (a) the authors, (b) the presenter, (c) the venue and (d) the date of the presentation will be also stored in plain text. This dataset will be extended whenever new InVID presentations are prepared and publicly released. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The project presentations will be made publicly available after their presentation at the venue/event they were prepared for.
Archiving and preservation	The project presentations will be publicly available for view and download via the SlideShare channel of the project[18], while links to the presentations of this channel will be also added on the relevant webpage of the project website[19], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches.

Dataset name	**InVID_Data_WP8_5_InVidSoftwareDemosAndTutorials**
Dataset description	This dataset will collect information regarding the developed and utilized InVID technologies. Public video demonstations, tutorials with instructions of use, documentations as well as links to publicly-released online instances of these technologies will be also included.
Standards and metadata	A variety of different formats will be used for storing the necessary information. In particular, video demonstrations can be (but not limited to) MP4, AVI or WEBM files, software tutorials and documenations can be written in PDF format, online documentations of tools and services can be presented in plain text, and presentations can be stored in PPT or PDF format. This dataset will be extended whenever new content related to the InVID developed technologies (e.g. video/web demos, tutorials, documentation) is prepared and publicly released. A simple log file of the performed updates of the dataset will

[18] http://www.slideshare.net/InVID_EU

[19] http://www.invid-project.eu/presentations

Dataset name	**InVID_Data_WP8_5_InVidSoftwareDemosAndTutorials**
	be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	Information related to the developed InVID technologies, including video demonstrations, documentations, presentatons and tutorials with instructions of use, will be publicly available supporting the dissemination of the project's activities and the exploitation of the project's outcomes. However, confidentiality control will be applied on each piece of information in order to avoid the release of inappropriate information that could have a negative impact to the project's progress and developments.
Archiving and preservation	Data related to the developed InVID technologies, tools and applications will be archived and made publicly available through the relevant webpage of the project website[20], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Moreover, the created video demos and tutorials will be also available for view via the YouTube channel of the InVID project[21].

Dataset name	**InVID_Data_WP8_6_InVidNewsletters**
Dataset description	This dataset will comprise the released newsletters for disseminating the activities and the progress made in the InVID project.
Standards and metadata	The newsletters will be prepared and stored in PDF format, while information regarding their release date will be provided. This dataset will be extended whenever new project newsletters are publicly released. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The newsletters of the project will be publicly available online right after their official release.

[20] http://www.invid-project.eu/tools-services

[21] https://www.youtube.com/channel/UCFp4OyFkV7cwQsDLCFRyBJQ

Dataset name	**InVID_Data_WP8_6_InVidNewsletters**
Archiving and preservation	An online archive with open access to the released newsletters of the project will be maintained at the relevant webpage of the project website[22], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches.

[22] http://www.invid-project.eu/newsletters

4 Summary

The initial Data Management Plan by the members of the consortium of the InVID project was presented in this deliverable. This plan involves every dataset that will be collected, processed or generated during the lifespan of the project. Aligned with the guidelines of the European Commision, the aim of the Data Management Plan is to ensure the safety of data, to enhance data accessibility, exploitability and reuse potential, as well as to support their long-term preservation. The applied methodology for defining the DMP of the InVID project was presented in Section 2, while detailed explanations about what will be considered for the reported datasets were provided in Sections 2.1 to 2.5. The entire list of datasets was presented in Section 3, where each subsection (see Sections 3.1 to 3.7) groups the datasets of each workpackage of the project. An updated version of the Data Management Plan integrating newer findings of the project in relation to datasets and their management will be described in D1.3 "Updated Data, quality and knowledge management plan", which is due in Month 21 of the project (September 2017).