Co-funded by the Horizon 2020
programme of the European Union

h2020mirror.eu

MIRROR

Migration-Related Risks caused by
misconceptions of Opportunities and Requirements

Grant Agreement No. GA832921

Deliverable D1.2

Work-package	WP1: Project Management and Coordination
Deliverable	D1.2: Data and Knowledge Management Plan
Deliverable Leader	LUH
Quality Assessor	RUG
Dissemination level	Public
Delivery date in Annex I	November, 30, 2019
Actual delivery date	November, 30, 2019
Revisions	3
Status	Final
Keywords	Datasets, Knowledge assets

Disclaimer

List of Authors

Partner Acronym	Authors
LUH	Erick Elejalde, Claudia Niederee
CERTH	Alexandros Pournaras, Vasileios Mezaris
UNIVIE	Jakob-Moritz Eberl

Table of Contents

Executive Summary **5**

1 Introduction **6**

2 Data Management **7**

2.1 Applied Methodology 7

2.1.1 Dataset reference and name 7

2.1.2 Dataset description 7

2.1.3 Standards and metadata 7

2.1.4 Data sharing 8

2.1.5 Archiving and preservation 9

2.2 Existing Dataset in MIRROR 9

2.2.1 WP5 Datasets 10

2.2.2 WP6 Datasets 11

2.2.3 WP8 Datasets 14

2.3 Protection concerns 15

3 Knowledge Management **16**

3.1 Knowledge exchange tools 16

3.2 Publication notification 17

3.3 Knowledge management plan definitions 17

3.4 Knowledge assets in MIRROR 18

4 Conclusions **22**

Executive summary

The Data and Knowledge Management Plan of the MIRROR project is outlined in this deliverable, including the methodology, activities and measures that will be employed for realizing this plan throughout the project's life. It starts by introducing the adopted policies for monitoring and assuring the safety of the activities related to data management during research, software development, project deliverable preparation and the overall progress of the project based on the defined time-plan. Then, it proceeds with the documentation of rules and definitions related to the management of knowledge and intellectual property, specifying factors related to their protection and assessibility. The Data and Knowledge Management Plan of the MIRROR project is a working document that evolves during the lifespan of the project, and can be updated or improved by integrating findings concerning the data manipulation and created knowledge as the project progresses. An up to date version of the document will be always available in the shared repository created for the project (https://git.l3s.uni-hannover.de/mirror/)

In particular, in the section dedicated to the data, we describe in detail the adopted management policy for the datasets that will be collected, processed or generated by the project. The utilized approach: (a) identifies which data and how they will be exploited or made accessible (to other partners in the consortium) so as to maximize their reuse potential, (b) specifies how these data will be curated and preserved, to support their reuse, and (c) identifies any data that should not be made publicly available and measures to be taken for their safe-keeping.

Following, we present our management plan concerning the extracted scientific knowledge and the created intellectual property assets by the project consortium. The production of these assets will be performed ensuring that no ethical requirements are being violated. The knowledge management plan of the project is outlined by discussing a set of rules and definitions related to the management of the provided or produced knowledge and intellectual property assets by the members of the consortium, and describing an established procedure for publication notification among the MIRROR partners.

1 Introduction

This deliverable documents the Data and Knowledge Management Plan of the MIRROR project and introduces the procedures and materials that are necessary for realizing this plan throughout the project's life, in accordance with the activities described in T1.1 Planning, Scheduling and Risk Management.

For handling of data, the European Commission (EC) has defined a number of guidelines / requirements for maximizing scientific-data's reuse potential, via making them easily discoverable, intelligible, usable beyond the original purpose for which they were collected and interoperable to specific quality standards. In MIRROR we incorporate these guidelines as a basis for our Data Management Plan. 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 the above points is given in sections below. Given the sensitive political and personal nature of some of the dataset used and generated in the MIRROR project, special care will be given to data protection policies (point (d) above).

With regards to the knowledge in MIRROR, the consortium partners are willing to share their expertise to make the project a success. Knowledge created during the project will be distributed within the consortium and (where adequate beyond the consortium) to enable a targeted and coordinated development towards the project goals, which also requires active knowledge exchange between the project partners and between the work-packages. As an entry point to project related knowledge, a project web portal will be created. This portal will provide also links to a restricted area for trusted exchange of insights and documents between the MIRROR members. The restricted area is based on a Wiki system, and complemented by a shared repository system (both based on a GitLab server). Multiple access protection levels have been set to allow appropriate access to researchers, WP leaders, committees and European Commission officers and reviewers. All intermediate results (i.e., deliverables, milestones and WP progress reports) are being documented in this area. All meeting preparations, administrative and technical management, discussion groups and deliverables drafts will also be managed there. Formal issues regarding the management of knowledge have been addressed in the MIRROR Consortium Agreement (CA) in Sections 8 (Section: Results) and 9 (Section: Access Rights). This includes - at consortium level - the coordination of rights for new knowledge generated by the project, i.e., for the project results, including information, and whether or not they can be protected. These conventions (which are discussed and extended in the sections below) aim to ensure the protection of confidential information and patenting of important knowledge that will be created during the project's lifetime.

2 Data Management

2.1 Applied Methodology

The applied methodology for drafting the Data Management Plan of the project was based on the guidelines of the EC (https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf) and the DMP online tool (https://dmponline.dcc.ac.uk), 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 MIRROR addresses the points below on a dataset-by-dataset basis, reflecting the 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.

2.1.1 Dataset reference and name

For convenient referencing of 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:
"*MIRROR_Data_WP(#)_(Dataset#)_(DatasetTitle)*".

According to this pattern, an example dataset reference name could be:
"*MIRROR_Data_WP6_1_WorldNewsAboutEurope*".

2.1.2 Dataset description

This field describes the dataset that will be collected and/or generated. It 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. Moreover, the description clarifies whether these datasets (are expected to) support a scientific publication, while information on the existence (or not) of similar data and the possibilities for integration and reuse is provided.

2.1.3 Standards and metadata

This section outlines how the data was collected and/or generated, and which community data standards (if any) are 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 should be considered:

- How have been the data created?
- What standards or methodologies are being 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 must be created to describe the data and aid their discovery. How this information will be created/captured and where it will be stored is also reported. The aspects below should be 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?

2.1.4 Data sharing

This section 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 and privacy-related considerations, intellectual property and commercial interests). 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 below should be considered 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 be the data 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 minimize restrictions?
- Where (i.e. in which repository) will be the data stored?

2.1.5 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 approximate 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 describes the resources and systems that will be put in place to enable the data to be curated and used effectively beyond the lifetime of the project. A set of questions that should be considered for defining the archiving and preservation policy for the datasets of the project is given below:

- What is the long-term preservation plan for the dataset (e.g. deposit in a data repository)?
- Are there sufficient resources, including storage and other equipment, to carry out this plan, or are any additional resources needed?

2.2 Existing Dataset in MIRROR

This section lists the datasets that have been created or collected so far for the needs of the MIRROR project. This datasets are specified based on the methodology presented in Section the previous section. This means that each dataset is defined by: (a) its name, (b) 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 new datasets are collected their specification will be added to this list and a up-to-date version of it will be kept in the shared repository (https://git.l3s.uni-hannover.de/mirror/).

2.2.1 WP5 Datasets

Dataset name	MIRROR_Data_WP5_1_VideoConceptAnnotation
Dataset description	**YouTube-8M**: This is a large-scale video dataset consisting of more than 6 million videos annotated with one or more concepts from 3862 categories (3.4 concepts per video on average). The videos have been pre-processed to extract visual and audio features at video-, frame and segment-level granularity (at 1-second resolution). The visual features were extracted using Inception-V3 image annotation model, trained on ImageNet, while the audio features were extracted using a VGG-inspired acoustic model on a preliminary version of YouTube-8M. Both the visual and audio features were PCA-ed and quantized. The overall dataset is divided into a training, validation and testing partitions of 3,888,919, 1,112,356, and 1,133,323 videos, respectively. Link: https://research.google.com/youtube8m/ **ImageNet (ILSVRC-2012)**: This dataset consists of over 15 millions labelled high-resolution images with around 22,000 distinct concepts of the WordNet structure. Image annotation is quality-controlled and human-generated. It is organised and managed by the Stanford and Princeton Universities. ILSVRC-2012 is a popular subset of ImageNet providing approx. 1.28 million training and 50,000 validation images, categorized to 1,000 distinct concepts. ImageNet: http://www.image-net.org/ ILSVRC-2012: http://www.image-net.org/challenges/LSVRC/2012/ **TrecVid SIN**: This dataset is provided by the US National Institute for Standards and Technology (NIST) to the participants of the TRECVID SIN evaluation activity. It is used for developing technologies for video annotation with visual concept labels. It consists of approx. 18,500 videos (354 GB, 1,400 hours) under a Creative Commons (CC) license, in MPEG- 4/H.264 format, and it is partitioned into training (approx. 11,200 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. Link: https://trecvid.nist.gov/index.html **SentiBank**: this dataset contains a Visual Sentiment Ontology (VSO) consisting of more than 3,000 adjective noun pairs (ANPs), such as “beautiful flowers” or “sad eyes”, and a set of 1,200 trained visual concept detectors created using a set of 500,000 Flickr images. Link: http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html **CIFAR-10**: It is one of the most widely used datasets for machine learning research. It consists of 60,000 32x32 colour images annotated with one of ten different concepts. The dataset is divided to 50,000 training and 10,000 testing subsets. Link: https://www.cs.toronto.edu/~kriz/cifar.html **Places**: This dataset is provided by MIT. It contains over 10 million annotated images and more than 400 unique scene categories. Link: http://places2.csail.mit.edu/index.html
Standards and metadata	This is a dataset collection of third-party pre-existing datasets that were developed, are maintained and are distributed to the scientific community by third-parties (outside the MIRROR consortium). The methodology, standards and metadata used in these datasets are those defined by their respective owners. In MIRROR, we only use these datasets for developing concept detection methods, and for training concept detectors, in accordance with the licenses that accompany each of these datasets.

Data sharing	These are third-party datasets; data sharing is controlled by each dataset's respective owners (who are external to MIRROR).
Archiving and preservation	These are third-party datasets; their archiving and preservation are performed by each dataset's respective owners (who are external to MIRROR).

Dataset name	MIRROR_Data_WP5_2_VideoCaptioning
Dataset description	**MSVD**: The dataset consists of 2089 video clips accompanied by 85,000 English descriptions. Each video clip depicts a single action or event. The clips were taken from videos on YouTube. Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/ **MSR-VTT**: This dataset was created for the MSR Video to Language Challenge by Microsoft. It consists of 13,000 videos, 10,000 of which are used for training and 3,000 for testing. The videos are taken from YouTube and each video is annotated with 20 natural sentences. Link: http://ms-multimedia-challenge.com/2017/dataset **ActivityNet Captions**: This dataset was created by Stanford University. The dataset consists of 20,000 videos taken from YouTube and 100,000 annotated phrases in total. The videos were preprocessed by a C3D network and the extracted network features are provided in the project page. Link:https://cs.stanford.edu/people/ranjaykrishna/densevid/ **MPII-MD**: This dataset was created by Max-Planck-Institut für Informatik (MPII). The dataset consists of 68,337 movie clips taken from 94 Hollywood movies. Each clip is annotated by one sentence resulting in a total of 653,467 words. Link: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/mpii-movie-description-dataset/
Standards and metadata	This is a dataset collection of third-party pre-existing datasets that were developed, are maintained and are distributed to the scientific community by third-parties (outside the MIRROR consortium). The methodology, standards and metadata used in these datasets are those defined by their respective owners. In MIRROR, we only use these datasets for developing video transcription methods, in accordance with the licenses that accompany each of these datasets.
Data sharing	These are third-party datasets; data sharing is controlled by each dataset's respective owners (who are external to MIRROR).
Archiving and preservation	These are third-party datasets; their archiving and preservation are performed by each dataset's respective owners (who are external to MIRROR).

2.2.2 WP6 Datasets

Dataset name	MIRROR_Data_WP6_1_WorldNews
Dataset description	This dataset contains news from the English news outlets of 85 countries for the period of May-July, 2019. This collection contains outlets from five major geographical regions EU, Middle-East, Central-Asia, Africa, and US-Canada, containing 198, 94, 163, 59, 429 English news outlets. This dataset contains total 225,210 news articles published over the three months.

Standards and metadata	We start from a large collection of 7064 news outlets compiled by Sail Labs Technology. Out of this, we find 2859 English news outlets. We map these news sources to their host countries based on their top-level domain in the Domain Name System of the Internet. For example, news sources with .af domain name are assigned to Afghanistan. However, news sources with generic domain names such as .net or .com cannot be directly mapped to any specific country. This constituted an issue for US news channels, since most of them use such domain names. To deal with this case, we collected all the English news channels having these domains (1644) and manually identified 408 US news channels. The rest of the generic news sources were discarded. This dataset contains total 225,210 news articles published over the period of above-mentioned three months. We create a folder for each month containing five subfolders for the five different regions (EU, Central-Asia, Middle-East, Africa, US-Canada). Each of these region folders contains one folder for each of its countries. For example, Central-Asia contains following seven folders — (a). Afghanistan (b). Armenia (c). Azerbaijan (d). Kyrgyzstan (e). Pakistan (f). Tajikistan (g). Uzbekistan. The news articles of a country are stored in .xml format under its corresponding directory. Each news article has the following metadata: Episode id (epiid), Date (date), News Channel (channel), Language (language), URL link (WebSiteUrl), Positive and Negative Sentiment (posSentiment and negSentiment), Polarity (polarity), Topic (topic), Entities (entities), Publication Date (pubDate), News content (paragraph).
Data sharing	The raw dataset is originally collected by Sail Labs Technology. L3S signed a data transfer agreement with SAIL to collect and use the data for the MIRROR project. Under this agreement, LUH will not share this data.
Archiving and preservation	The lifetime of this dataset is the project period. It will be stored in the L3S' MIRROR server up to that point.

Dataset name	MIRROR_Data_WP6_2_WorldNewsAboutEurope
Dataset description	This dataset contains Europe-related news from the English news outlets of 85 countries for the period of May-July, 2019. This is a subset of *MIRROR_Data_WP6_1_WorldNews*. This collection contains outlets from five major geographical regions EU, Central-Asia, Africa, Middle-East, and US-Canada containing 198, 94, 163, 59, 429 English news outlets respectively. The dataset contains 61169 news.
Standards and metadata	We have three files corresponding to the three months of collected news. Files are named under following convention: [MONTH]_2019_europe.csv. The metadata is the same listed for *MIRROR_Data_WP6_1_WorldNews*
Data sharing	As with the *MIRROR_Data_WP6_1_WorldNews* dataset, under the agreement signed between L3S and SAIL Labs, this collection cannot be made public or shared outside the LUH.
Archiving and preservation	The lifetime of this dataset is the project period. It will be stored in the L3S' MIRROR server up to that point.

Dataset name	MIRROR_Data_WP6_3_GeoNames
Dataset description	This dataset contains the geographical information about 11 million locations around the world. The information includes the location's name, the country's name, the latitude and longitude etc.

Standards and metadata	The GeoNames geographical database is available for download free of charge under a creative commons attribution license. It contains over 25 million geographical names and consists of over 11 million unique features whereof 4.8 million populated places and 13 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. GeoNames is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. All lat/long coordinates are in WGS84 (World Geodetic System 1984). Users may manually edit, correct and add new names using a user friendly wiki interface.
Data sharing	The data is accessible free of charge through a number of webservices and a daily database export.
Archiving and preservation	The download dataset and the results of processing it will be stored on the file server of MIRROR (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	MIRROR_Data_WP6_4_TwitterMigration
Dataset description	Set of Twitter users, mostly in the EU and Middle East countries, who are active in migration and refugee topics.
Standards and metadata	The dataset was collected using snowball sampling method. Specifically, the collection consists of the following steps: (1) Identifying the core users: by manually scanning for influential users in the EU and Middle East countries on migration and refugee topics, starting from the followees of @Refugees, @RefugeesOlympic, @TeamRefugees. This core set is iteratively grown and refined by scanning their followees. This is due to the assumption that "influencers are followed by other influencers". (2) Collecting followees and followers of the core users using Twitter APIs. (3) Collecting tweets of all the collected users using Twitter APIs. The versions of the dataset will be named by adding version number and created date (in yyyymmdd format) to the end of its name. For example, "MIRROR_Data_WP6_4_Twitter_Migration_1.0_20191127" is the version 1.0 of the dataset which is created on Nov 27th, 2019
Data sharing	Due the Twitter's policy, only the id of tweets in the dataset will be publicly shareable.
Archiving and preservation	The dataset will be stored in the LUH's servers dedicated to the MIRROR project.

2.2.3 WP8 Datasets

Dataset name	MIRROR_Data_WP8_1_ExpertAndFieldInterviews
Dataset description	Workpackage 8 (WP8) will conduct several interviews with experts and irregular migrants at different locations in Europe as well as in MENA countries. The data will first be collected in the form of raw audio recordings (formats may include – but are not restricted to – mp3, m4a and WMA). Interviews will then have to be (partially) transcribed and translated. A qualitative summary of responses, including selected quotes in English may be prepared. Finally, translated interview transcripts may be coded following a qualitative research approach (formats may include – but are not restricted to – WORD, pdf and CSV). The raw data will – at any point – only be available to members of WP8. Completely anonymized/pseudonymized transcribed interviews or quotes may be made public in the process of dissemination. In the case that expert interviewees have formally and explicitly consented to it, non-anonymized quotes from their interviews may as well be made public in the process of dissemination. All collected data are expected to support deliverables as well as resulting scientific publications. In compliance with the consent form for experts, all data concerning expert interviews will be pseudonymized if explicitly asked for by interviewees, where personal identifiers will be kept in a separate file within a UNIVIE secured database. In these cases, all identifiers will be kept secure and separated from other data and never shared publicly. Should expert interviewees consent on the non-anonymized use of their data, their data will not be pseudonymized. However, following the consent forms for interviews with irregular migrants, we will not keep any personal identifiers in our database concerning interviews with irregular migrants. The lists of participants' contacts, that may be used for recruiting of the interview-partners and focus groups will not be shared at any point with anybody, not even with the core team and will be destroyed after usage. In compliance with the GDPR, we will not keep any personal identifiers of irregular migrants in our database. Contacts of participants will not be needed in the further course of the research, and WP8 will not re-use them for any further purpose. It will be not possible to deanonymize data collect from interviews with irregular migrants. Considering the specific research question of this Workpackage, the collected data is unique. Due to the high sensitivity of the data, reuse beyond this project and beyond what will have been published in a pseudonymized fashion within the dissemination process is not being considered.
Standards and metadata	Conventions concerning "version control" and "meta data" do not apply to data generated within WP8.
Data sharing	Due to participant confidentiality and consent agreements, none of the raw (audio) data will be shared beyond the core team of WP8. Anonymized (for irregular migrants), pseudonymized (for some experts), or non-anonymized (if experts give explicit consent) summaries of interviews, anonymized/pseudonymized/non-anonymized quotes and anonymized/pseudonymized/non-anonymized coded data may be shared within the project and will be part of published deliverables.

Archiving and preservation	Due to participant confidentiality and consent agreements, none of the raw (audio) data will be shared beyond the core team of WP8. Anonymized (for irregular migrants), pseudonymized (for some experts), or non-anonymized (if experts give explicit consent) summaries of interviews, anonymized/pseudonymized/non-anonymized quotes and anonymized/pseudonymized/non-anonymized coded data may be shared within the project and will be part of published deliverables. The data, such as CSV exports, anonymized transcripts, summarized responses of interviews, CSV exports of coded interviews will be stored within a secured MIRROR data repository at UNIVIE for three more years after the end of MIRROR to allow for adequate dissemination. UNIVIE must authorize any access data stored within that data repository. Due to high sensitivity of the data no access can be given to any raw data at any point.

2.3 Protection concerns

All partners in MIRROR understand that we are dealing with sensitive information that may impact vulnerable migrant populations. That is why we place research ethics at the core of our work. During the development of the project, several research tools and methods, such as in-depth interviews and participant observation (e.g., WP8), are used. This level of human involvement produces a set of ethical and legal issues, which the consortium is well aware of and determined to address. Details on how MIRROR plans to undertake these tasks are described in the project's Description of Actions (DoA). In this section, we focus on some measures concerning the manipulation and storage of digital datasets.

We think it is important to mention again that MIRROR will comply with applicable data protection legislation. All partners in the consortium will comply with the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) and other applicable national laws. Generally, MIRROR has different strategies to ensure not merely compliance with the legal requirements but to make sure that the privacy and data protection rights of the participants are fully protected.

In particular, during the collection of personal data, MIRROR will ensure that the amount of information is kept at the minimum necessary to achieve the project's objectives and will, as far as possible, work with anonymized data.

With respect to field work, interviewers will be supervised regularly by PIs, and the interviews will be recorded digitally. After being anonymized, recordings will be stored in safe digital environments, and deleted from recording devices.

In the same vein, most of the data collection from social and mass media will be centralized in one partner that already has the infrastructure for this task. This contributes to preventing duplication of collection effort and ensures homogeneity across datasets (including issues regarding to protection concerns).

To distribute the data collected, we will create Peer-to-Peer agreements. These will formalize the transfer procedures and safe handling of all the information once the receiving partner stores it. A template of the bilateral agreement is available in the project's shared space (https://git.l3s.uni-hannover.de/mirror/).

3 Knowledge Management

A set of established rules related to the management of knowledge and intellectual property in the MIRROR project was documented in the Consortium Agreement (CA). These conventions are discussed and extended in the following section.

A well-designed Intellectual Property Rights (IPR) management strategy will contribute to the smooth functioning of the consortium and to the commitment of individual project partners to the project tasks. In the design of this IPR strategy the consortium will follow a systematic approach. IPR issues within the project will be structured along three dimensions: (a) type of asset considered (i.e., content/data or technology), (b) point of asset creation (i.e., pre-existing assets versus assets created during the project), and (c) the type of intended use (i.e., use by consortium partner during the project or after the project, commercial use outside the project, non-commercial use outside the project).

For most of the technology created as part of the MIRROR project, it is the goal to keep its future non-commercial use free of charge (see also next section). Different licensing models, such as LGPL, Apache License, 2.0, etc., will be evaluated during the project to identify the most adequate one for fostering the wide use and the further extension and refinement of MIRROR components and solutions.

The possibility of commercial exploitation of technologies developed within European projects and the creation of new services and products on top of such technologies is one of the major reasons for larger companies and SMEs to join European projects. Thus, some technologies can be chosen for IPR that enable systematic commercial exploitation by the companies. In general, Foreground IP shall be owned by the project partner carrying out the work leading to such Foreground IP. If any Foreground IP is created jointly by at least two project partners and it is not possible to distinguish between the contributions of each of the project partners, such work will be jointly owned by the contributing project partners. All the details concerning the exposure to jointly owned Foreground IP, joint inventions and joint patent applications have been addressed in the Consortium Agreement.

Preexisting know-how will be provided by each partner in order to contribute to the success of the MIRROR project. In principle, the IPRs for preexisting technology stay with their original owners. In order to ensure a smooth execution of the project, the project partners agree to grant each other royalty-free access rights to their Background and Foreground IP for the execution of the project. Further regulations for the use of these technologies have been defined in the Consortium Agreement (CA).

Open-sourcing the software developed in the project will be considered in conjunction with the MIRROR exploitation plan. We will pursue open source release of pieces of software only to the extent that this does not undermine in any way the exploitation potential and the exploitation efforts undertaken or planned by the consortium as a whole and by each individual partner. Scientific results will follow the “green model” for free online access. We will publish our results in prestigious venues and pre-print versions will be shared in dedicated author communities (e.g., arXiv), EU-supported repositories such as Zenodo, as well as in our institutional online repositories and the project’s website.

3.1 Knowledge exchange tools

Within the scope of the project, we have made available a set of tools that allow partners to exchange and collaborate in the creation of knowledge. Here, we describe some of the most used such tools.

Given the nature of the project, one crucial and visible outcome is the MIRROR digital platform. This plat-

form will be developed internally, and all partners will be involved in a way or another through the continuous integration process. For this process, we prepared instances of Jenkins and Artifactory. For all programming code, scripts generated by different work packages, we have configured a private GitLab instance. Based on Git, Gitlab allows for code versioning in collaborative development environment, adding useful features like Wiki pages for additional documentation and issue tracking system, among others.

The datasets are stored on a dedicated storage attached to a small cluster, assembled specifically for the MIRROR project. Access to the cluster and the data is granted according to the diverse needs of the partners. Additional distributed storage across the cluster nodes allows for meaningful usage of modern technologies, which can digest very large datasets, e.g., Cassandra, Elasticsearch, Janusgraph, Spark. These storage and computational frameworks will play crucial role when building the MIRROR digital platform.

For research articles and other dissemination documents, we have created a dedicated section in the project's Wiki. This space will be used to share pre-publication drafts of the papers in advance. These can be then appropriately vetted by the interested partners according to the procedures defined in the CA (see Section 3.2).

3.2 Publication notification

The members of the MIRROR consortium agreed on and put in place a simple, lightweight publications notification procedure with the help of a dedicated wiki page so that partners wishing to publish or otherwise disclose project activities and results notify the other partners in good time to enable IP protection to be obtained if necessary. According to this process, prior notice of any planned publication shall be given to the other partners. Any objection to the expected publication shall be made in accordance with the Grant Agreement in writing to the Project Coordinator and the partner(s) proposing the dissemination. If no objection is made within the time limit stated in the CA, the publication is permitted.

3.3 Knowledge management plan definitions

For the management plan regarding the extracted knowledge and the created Intellectual Property (IP) assets generated during the project, the MIRROR consortium has defined the following set of definitions:

- **Foreground** or project results, including information, whether or not they can be protected, which are generated by the project.
- **Ownership** of foreground resulting from the project.
- **Protection** of valuable foreground by its owner(s) through filing of patent applications where possible, or other IPR protection measures.
- **Background** information that is held by Parties prior to their accession to the Grant Agreement, and which is needed to carry out the Project or for using the Foreground.
- **Access rights** to another participant's Foreground or Background, as well as joint exploitation activities, will be agreed upon between the partners in separate agreements.

Formal issues regarding the management of knowledge have been addressed in the MIRROR Consortium Agreement (CA). This includes - at consortium level - the coordination of rights for new knowledge generated by the project, i.e., for the project results, including information, and whether or not they can be

protected. Specific details on the implementation of this plan (asset-by-asset) will be included in the final version of the "Exploitation Plan" (deliverable D11.5).

Based on this group of definitions, we crafted an initial record of knowledge assets that will be created in the different work packages of the project, so that attention can be given to the protection of the most valuable IP by means of patents or other methods (e.g., trade secrets), while ensuring that no ethics requirements are being violated in any case. The identified assets are reported on a per work package basis in the following Section.

3.4 Knowledge assets in MIRROR

This section reports the foreground assets that are expected to be generated by the project, grouping them on a per work package basis. As presented in this subsections, preexisting (Background) expertise and new (Foreground) knowledge will be blended by the project consortium for the needs of the project. Further specifications of ownership, protection, and access rights of these assets will be resolved during the development of the project, based on the conditions of the Consortium Agreement (CA).

Work-package 2

- **Migration Related Semantic Concepts (MRSC):** This asset will offer a hierarchical classification of the semantic concepts that will be used in the project. These classes should serve as labels for items collected from different sources, creating the foundations for further analysis. It will also provide a common language among partners and other users of the system, facilitating the scrutiny of the system's output at different levels.

Work-package 3

- **Societal Frame of Migration:** this asset aims to provide a specific "social frame", to circumscribe and describe how some relevant deviations are occurring "from certain perceptions abroad" and "the reality of the EU". The construction of this social frame is essential to identify the most relevant "messages" to launch to potential migrants and to figure out the information channels considered and trusted by these. Also, this frame allows engaging more directly in the migrants' decision-making process by providing information, which would help migrants properly assess the likely rewards of risky behavior, rather than just focusing on the risks themselves.

- **Practical Guidance and Policy Recommendations:** this asset is intended for partners, technology developers, border security authorities, FRONTEX, stakeholders, and the European Commission. It regards possible ways of resolving legal, ethical, and societal issues and risks that may arise in the course of using systems like MIRROR on a daily basis.

Work-package 4

- **Named Entity and Concept Detection Component:** this asset will enrich text-documents (and textual-meta-data) with migration-related-semantic-concepts (MRSCs). Based upon an infrastructure for pattern-based detection of Named-Entities as well as statistical tagging of entities, the existing mechanisms and models will be extended and adapted to the target relevant domains and use-cases.

- **Sentiment/Polarity/Stance Detection Component:** this asset will provide annotation of sentiment polarity and stance of textual documents. It will be based on SAIL's existing SentiStrength sentiment analysis system, which will be extended to be applicable to migration-specific-concepts.

- **Topic Detection and Topic Evolution Component:** this asset's function is mining and tracking topics discussed among migrants and/or about MRSCs across media channels. It will include novel methods for deriving topics from multi-modal, multilingual and heterogeneous data sources as well as for joint modelling of topics and networks/communities.

- **Assessment of the Impact of Editorial Information and Perceived Information Quality:** Producing and spreading junk news and fake news is perhaps the most well-known way to influence the public. However, the editorial material does not have to be direct lies or untruthful to influence the readers and in some cases also the public opinion. This asset includes techniques and methods for analysing what kind of editorial messages related to migration are spread in relevant media sources. It will also explain how perceived information quality and the choice of editorial material correspond and might be used for opinion manipulation.

Work-package 5

- **Visual Media Annotation and Sentiment Analysis Component:** This asset will generate image and video concept annotations using MRSCs, sentiment concepts and other generic concept pools. The annotations will be used for the media summarization and to provide a deeper understanding of the media content, for search and retrieval.

- **Visual Media Collection Summarization Component:** this asset will create a concise, yet descriptive, summary of media items retrieved by the content collection component of the MIRROR platform. This condensed representation will facilitate search and navigation in large media collections.

- **Automatic Speech-Recognition Component:** this asset will adapt existing models for automatic transcription of audio-content to migration domains.

Work-package 6

- **Cross-media Network Construction Component:** this asset will construct network(s) for different use cases identified during the end-user requirement analysis. These networks will serve as a starting point for other advanced analyses.

- **Bias Detection and Reduction:** this asset will provide appropriate tools and frameworks for detecting, examining, investigating, and reducing bias when seeking, receiving, consuming, and working with media data. This asset will focus on two types of biases: (i) social biases introduced by data seeking and retrieving methods or originally introduced by social selection and alignment of data sources, and (ii) the biases introduced by data processing and analysing methods.

- **Evolution of Networks and Communities Component:** this asset will deal with the evolution of migrants' networks and their communities over time. It will include a wide range of statistical observations on (i) the networks' growth and shrinkage, (ii) the communities' formation and characteristics, and (iii) interactions among the communities. These will give insights into how migrants connect and interact and the factors which affect their behaviour and perception.

- **Information Diffusion and Manipulation:** this asset will show how information is diffused and manipulated within and across migrants' networks. At network level, it will examine and model the rate and phases of the diffusion: how these deviate with regards to the networks, information types, and the origins of the information. At individual level, it will identify users playing different roles in the diffusion process.

Work-package 7

- **MIRROR Information Model:** this asset will define an information model which will enable the representation of the different data sources and the combination of the results from the automated media analysis and from empirical studies into a coherent model used for the implementation of the MIRROR framework.
- **MIRROR Framework:** this asset will integrate the analysis methods produced by the technical partners into a coherent framework.

Work-package 8

- **Literature Review and Expert Interviews:** this asset will include a comprehensive review of the academic literature in communication, political science and migrations studies. It will bring together existing research on perceptions of irregular migrants and the role of media for perceptions of migration decisions. The review will look at driving (push and pull) factors for irregular migration as well as media use and media perceptions from the perspective of irregular migrants. This review aims at a better understanding of the driving factors behind forced migration as well as the role of social media in this process. This initial analysis will pave the way for structuring subsequent interviews with five experts in Vienna (e.g. UNHCR, Criminal Intelligence Service Austria, Task-force Human Trafficking).
- **Interview Records:** this asset will compile transcriptions of interview records. This include comparative and semi-structured interviews with irregular migrants and experts at key locations (e.g. Turkey, Jordan, Greece etc.).
- **Analysis and Reporting:** this asset will provide analysis of interviews' transcripts. This will provide an assessment of whether knowledge about and (mis-)perceptions of the EU and the migration process are influenced by means of digital communication, and at what point during the migration process knowledge and perceptions have been shaped due to incoming information.

Work-package 9

- **Threat Analysis Connecting to Existing Work:** this asset will perform threat analysis with special focus on threats resulting from misperception of Europe. The threat analysis will not be done from scratch but will strongly connect to existing European Risk Assessment/Threat Analysis Frameworks for Border Control (e.g. FRONTEX Common Integrated Risk Analysis Model (CIRAM), Eastern Partnership Risk Analysis Network (EAP-RAN)). the asset will answer how risk indicators are constructed, and further translated and negotiated in border police practice.
- **Recommendations Based on MIRROR Results:** this asset will provide recommendations to end-users on how to integrate MIRROR results into their own Risk Assessment methodologies. Given the characteristics of the MIRROR project the type of output provided will fall into two categories: risk indicators based on sentiment analysis and actionable insights (e.g. geographical concentration of migrants,

smuggling routes etc.). While the core stages for developing the methodology will be the same, each of the two categories will be treated separately (e.g. each having its own sets of indicators) as they are used in different stages of the security process.

- **Toolkit of Actionable Insights:** this asset will provide a toolkit of actionable insight including different fields such as information policies, law, ethics, and technology for border control as well as for policy makers. The toolkit of actionable insights will be a collection of promising practices and other relevant instruments (e.g. methodologies, auditing checklists, codes of conduct) on the use of social media in addressing security challenges caused by migration.

Work-package 10

- **Consolidation of Lessons Learnt:** this asset will collect lessons learnt from the pilot run and will compile suggestions for modifications in the beta system. It will also include guidelines to be incorporated in the MIRROR toolkit (WP9).This report will help to understand how regions differ from a migration perspective. Perception of different countries will be evaluated in a comparative study where existing intelligence and knowledge is used as benchmark.

Work-package 11

- **Exploitation Plan:** this asset will list the activities of consortium partners to exploit the project results for future research or commercialization. The project's stakeholders are divided into academic, industrial partners, other project partners such as civil-society organizations, and border and security authorities. Industry partners will describe their future markets and the project's contribution to market evolution.

  IPR Rights: This asset will also manage the secure maintenance of the project's background IPR contributed by the consortium members. It includes secure documentation of terms and conditions of licenses and status.

4 Conclusions

The initial Data and Knowledge Management Plan defined by the members of the MIRROR consortium was documented in this deliverable.

This plan includes guidelines to handle datasets that will be collected, processed, or generated during the lifetime of the project. The present document represents the dataset-related status and planning at month 6 of the MIRROR project; as such, it may change to some extent during the remaining lifetime of the project, by, for example, the use or generation of new datasets that are not currently foreseen. To account for such changes, which are typical and expected within a 36-month project, we anticipate that one or more updated versions of the Data Management Plan will be produced as the project progresses and per the project's needs. Though not being formal deliverables of the project, these updates to the Data and Knowledge Management Plan will be made available via the project's website. This plan also includes the prescribed actions for assuring smooth collaboration and knowledge sharing between partners and with the community. This involves a set of different aspects of the project, namely the innovation management, the research activities, the software development, the preparation of the project deliverables, and the overall progress of the project according to the defined time plan.