D1.2 DATA MANAGEMENT PLAN

Deliverable No.:	1.2
Deliverable Title:	Data Management Plan
Project Acronym:	Fandango
Project Full Title:	FAke News discovery and propagation from big Data and artificial inteliGence Operations
Grant Agreement No.:	780355
Work Package No.:	1
Work Package Name:	Project Management
Responsible Author(s):	Massimo Magaldi, Silvia Boi
Date:	31.07.2018
Status:	v1.0 – Draft
Deliverable type:	ORDP
Distribution:	PUBLIC

REVISION HISTORY

VERSION	DATE	MODIFIED BY	COMMENTS
V0.1	31.06.2018	Angelo Manfredi	First draft
V0.2	19.07.2018	Massimo Magaldi	Second draft
V0.3	23.07.2018	Luca Bevilacqua	Quality check
V0.9	31.07.2018	Silvia Boi	Draft and request for evaluation of the opt-out option

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

ABBREVIATION	DESCRIPTION
H2020	Horizon 2020
EC	European Commisiion
WP	Work Package
EU	European Union

EXECUTIVE SUMMARY

Fake News are a hot issue in Europe and worldwide, particularly in relation to Political and Social Challenges. The state of the art is still lacking a systematic approach to address the aggressive emergence of fake news and post-truth evalutaions of facts and circumstances.

FANDANGO aims at contributing to semi-automatic approaches that can aid humans in the evaluation of the trusthworthiness of news (potentially fake ones).

To achieve this goal, FANDANGO will use several approaches, collecting a large volume of data and providing sophisticated machine learning approaches to aid in the investigation and validation purposes. To do so different data sources will be used leveraging an open software stack that will include a wide range of big data oriented technologies.

This document was supposed to be the FANDANGO Data Management Plan (DMP), and as such it was meant to describe in particular how research data will be produced, collected or processed by the FANDANGO project (and also procedures and decisions to allow data to be *FAIR*).

For this reason this deliverable was also structured following the official guidelines set forth in Horizon 2020 for similar documents.

However, while studying and working to write it, one circumstance began to stand out and become prominent to our managerial attention.

The circumstance is the following: while FANDANGO will have no interest to process personal data as such, it will need to process massively data about the emerging news, especially data related to potentially untrustworthy (fake) news.

In so doing, some processing steps may offer the potential opportunity to infer personal data.

The easiest example showing such a risk is to be able to assess that a personal account on Twitter is usually writing untrustworthy news and being in the nedd to keep this information for future decision making about other news published by the same person.

At the present stage of research we are still not in the position to evaluate and put in place specific measures to counter this type of risk and sharing data for scientific purposes, in this condition, would exacerbate the risk.

For this reason we are:

- asking for the opt-out option,
- submitting this deliverable still in a draft version, with only the information collected at the stage when the opt out option was considered.

Ethical aspects remaining after the opt-out choice are dealt with in D9.1 and D9.2 documents. In particular the information presented in D9.2 beras some resemblance to the information provided by this document.

1. Data summary

1.1. What is the purpose of the data collection/generation and its relation to the objectives of the project?

In short, FANDANGO will collect data about potential news whose trustworthiness (i.e. are they real news or fake news ?) is uncertain and will generate assessment scores about the probability of each of being genuine or fake.

If we call S_i such fakeness scores (where i = 1, 2 … N and N >> 1) Fandango will generate the S_i fakeness scores by combining partial scores $S_{i,j}$, where j = 1, 2, 3, 4 are the output of specific software modules, studied and developed as part of FANDANGO original results (in WP4, tasks 4.1 to 4.5).

Each module will perform an assessment on the basis of a specific criterion, thus computing a partial fakeness score.

Fandango will generate the S_i fakeness scores by optimally combining partial scores $S_{i,j}$.

This process is depicted in the following figure:

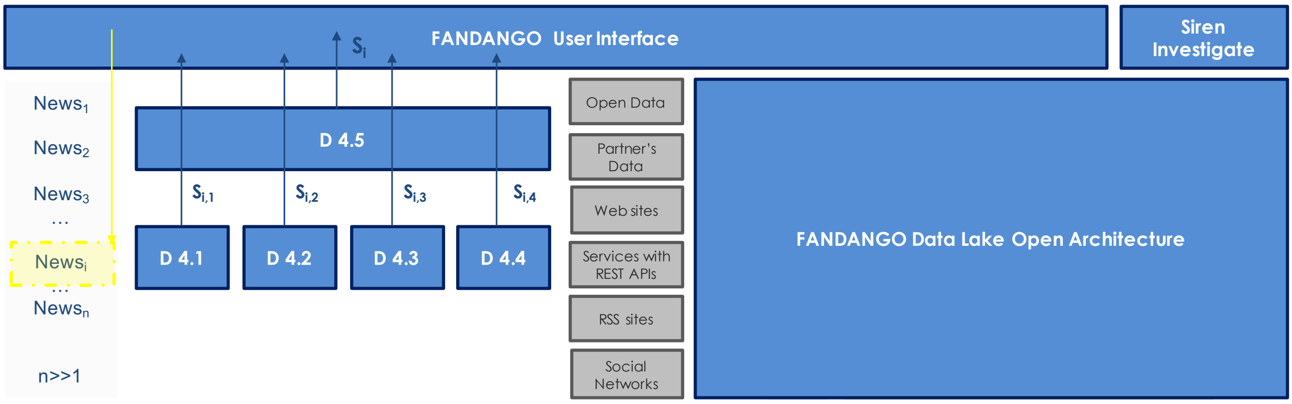

Figure 1

Final users of the FANDANGO platform are working at more clearly specifying their needs and priorities, that will be addressed while performing the research work complying to the approved FANDANGO DoA.

Preliminary user results show that our user champions would rather see FANDANGO as a set of tools prescriprions instead of one completely integrated software solution. The tools deemed most useful are

- news verification tool
- photo/video verification tool
- alert system

In addition to that our users stated that Fandango should never make a final decision about the trustworthiness of information, but rather help the journalists in doing that (This is the reason why FANDANGO results are outlined in the diagram above not only in terms of a general S_i fakeness score (where i = 1, 2 … N and N >> 1) but also in terms of partial scores $S_{i,j}$ (j = 1, 2, 3, 4).

Please notice that FANDANGO is dealing with a big data class problem since the sheer number of potential news to be examined (n) may be very large and each piece of potential news will be analyzed by considering all its text and multimedia content (multimedia content will be in potentially large) and potentially many past texts or contant fragment related to past news (real or fake ones).

This is the reason why to achieve this result FANDANGO will need to leverage an integrated big data platform based on Open Source middleware.

It is worth noticing that FANDANGO will target users will be professional journalists that need to evaluate in a short time the genuinity of a potential news.

Being the target user a professional it is very likely that the Si fakeness scores as well as the Si,j partial scores will be treated as a decision making help, for a final decision that will still rely on human judgement.

1.2. What research data do we collect and for what purpose?

FANDANGO will collect data about potential fake news, for the sole purpose of evaluating the effectiveness of algorithms and check their ability to *partially* automate the process of deciding about each news being fake or not.

The achieve this main objective several minor objectives will need to be achieved:

- ingest cross-domain and cross-lingual data sources of different nature to the FANDANGO platform
- provide state of the art algorithms for fake news related feature extraction (i.e. computing the $S_{i,j}$ partial fakeness scores)
- provide an higher level fake news evaluation (i.e. computing the S_i fakeness scores)
- back-track the propagation of potential fake news, determining the original sources and the diffusion points for source scoring regarding fake news distribution

Such algorithms will analyse not only text content but also images and videos and in structured and unstructured formats.

FANDANGO will leverage a Data Lake architecture to manage all relevant data found in relevant data sources (a preliminary selection of data sources is identified in DoA Section 1.3.4.4, but will be enlarged during the project).

FANDANGO partners are aware of the fact that the Open Research Data Pilot applies primarily to the data needed to validate the results presented in scientific publications and that other data can be provided on a voluntary basis.

1.3. WHAT RESEARCH DATA DO WE GENERATE AND FOR WHAT PURPOSE?

In FANDANGO all collected data will be processed/analysed by a set of software modules to extract markers and cues in order to reveal fake or misleading news.

As already stated different (four) analysis modules will be in the FANDANGO toolset:

1. The Spatio-temporal analytics and out of context fakeness markers module will be responsible for analyzing news posts and finding duplicate or near duplicate posts in the past or referring to other geographic/physical locations or contexts. In fact, a common case of fake news is the re-posting of a real past piece of news that it is no longer relevant or is removed from its original context. Such spatio-temporal or out-of-context correlations can generate strong fakeness markers (i.e. generating $S_{1,i}$).
2. The Multilingual text analytics for misleading messages detection module will handle multilingual content and score it the text as potentially misleading or not. To establish such scoring ability it will digest data from the public web as well as existing and well updated knowledge bases such as YAGO, DBPedia, Geonames etc.) to identify contradictions and potentially intentional errors. (i.e. generating $S_{2,i}$).
3. The Copy-move detection on audio-visual content module will detect the manipulation of images and videos to modify their visual content. This module will leverage deep learning architectures to identify such content and the pool of near duplicate content and visuals that were used as sources for creating the fake object. Synthetic data and publicly available big image datasets will be used to train the models. Moreover, state of the art audio analysis algorithms will be deployed to detect modified or voice-over attacks in news videos. (i.e. generating $S_{3,i}$).
4. The Source credibility scoring, profiling and social graph analytics module will profile the sources of news and apply graph analytics to detect paths and nodes that tend to produce fake news and spread them widely on the public web. (i.e. generating $S_{4,i}$).

To fuse the output of the above-mentioned modules a machine learnable approach will be used for overall fake news scoring (i.e. generating S_i)

A machine learnable score function that will learn how to weight and what data to use from the data lake to decide about the fakeness or not of a news post. The task will apply existing and successful predictive analytics deep learning architectures in order to be able to score news posts incrementally and update the score as new data populate the data lake, thus being able to provide hints from the early beginning of the appearance of a post.

Finally, for the visualisation and analysis of fake news, FANDANGO will provide a set of front end web applications and investigative intelligence tools with focus on identifying case studies. However, these tools are suitable for any kind of fake news discovery application. Siren platform is commercial product that delivers a unique investigative experience to solve real world data driven problems, enabling Analysts, Investigators and Data Scientists. It uniquely allows you to identify relationships across multiple data sets, accessible via search, dashboard analytics, knowledge graphs and real-time alerts, providing journalists and investigative agents to get contextual information and elaborate on their analysis.

1.4. What types and formats of data will the project generate/collect?

The FANDANGO project will leverage a Data Lake architecture that will store all available data types found in the identified data sources, i.e. free text, structured and unstructured data, images, videos and audio data from.

The data types we will be handling are plain text, images, videos, JSON unstructured files, structured data from open data databases.

1.5. Will you re-use any existing data and how?

We will reuse existing data mainly for the machine learning training set, some other data may be kept to refine evaluations of trustworthiness of specific news ("ground truth").

As an example, CERTH will be reusing existing publicly available datasets that are found in many publications and provide a reference for comparison with other algorithms.

A list of possible datasets we will be using is the following:

- Moments (http://moments.csail.mit.edu/)
- Imagenet (www.image-net.org/)
- MIT Places (http://places.csail.mit.edu/
- 20bn-Something (https://20bn.com/datasets/something-something)
- Coverage (https://github.com/wenbihan/coverage)
- MS-COCO (http://cocodataset.org/#home)
- COMOFOD (http://www.vcl.fer.hr/comofod/)
- EUREGIO Image forensics challenge (http://euregiommsec.info/image-forensics-challenge/)
- Image manipulation dataset (https://www5.cs.fau.de/research/data/image-manipulation/)
- NIST media forensics challenge (https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018)
- SULFA (http://sulfa.cs.surrey.ac.uk/)
- REWIND (https://sites.google.com/site/rewindpolimi/downloads/datasets)

As another example UPM will use the existing data stored in the FANDANGO platform for the graph analysis tasks. In addition, UPM will employ the data provided by the ingestion process (crawler to make the proper data wrangling and data transformation processes in order to have the data in the correct format for both Machine learning and Deep learning procedures.

1.6. What is the origin of the data?

The main goal of FANDANGO project can be pursued by aggregating data, from different Data sources in a suitable Data Lake.

1.7. What is the expected size of the data?

FANDANGO will deal with Big data size for ingested and homogenized data, a very different size for generated data (e.g. source thrustability and fakeness scoring).

1.8. To whom might it be useful ('data utility')?

Mainly to other Research Projects.

2. FAIR DATA

All considerations about FAIRness of data will be postponed because of the opt-out request.

3. DATA SECURITY

All considerations about long term preservation and curation of data will be postponed because of the opt-out request.

4. ETHICAL ASPECTS

Ethical aspects remaining after the opt-out choice are dealt with in D9.1 and D9.2 documents.