MixedEmotions

Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D1.9 Data Management Plan

Project ref. no	**H2020 644632**
Project acronym	MixedEmotions
Start date of project (dur.)	01 April 2015 (24 Months)
Document due Date	30 June 2015
Responsible for deliverable	Paradigma Tecnológico
Reply to	jruiz@paradigmatecnologico.com
Document status	Final

Project reference no.	**H2020 644632**
Project working name	MixedEmotions
Project full name	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets
Security (distribution level)	PU
Contractual delivery date	30 June 2015
Deliverable number	D1.9
Deliverable name	Data Management Plan
Type	Report
Version	1.0
WP / Task responsible	WP1
Contributors	José Ruiz (PT), All
EC Project Officer	Susan Fraser

Table of Contents

1 INTRODUCTION AND SCOPE 4

2 DATASET DESCRIPTION FOR DATA LIFECYCLE MANAGEMENT 4

3 DATASET IDENTIFICATION AND LISTING 4

3.1 DEUTSCHE WELLE CONTENT 4
3.2 TWITTER CONTENT 5
3.3 TWITTER GRAPH 5
3.4 FACEBOOK CONTENT 5
3.5 WEBSITES CONTENT 6
3.6 TAGGED TEXT 6
3.7 SINDICETECH KNOWLEDGE GRAPH 6

4 CONCLUSIONS 7

1 Introduction and scope

This Data Management Plan (DMP) describes the data management life cycle for all data sets that will be collected, processed or generated by the MixedEmotions project. It outlines how research data will be handled during the project, and even after it is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved. As the DMP is not a fixed document, it will evolve and gain more precision and substance during the lifespan of the project; therefore the first versions will be necessarily incomplete.

2 Dataset description for data lifecycle management

This initial version of the DMP will describe each available dataset using the fields below. To allow for more context and a better understanding of the purpose of the different datasets, they are listed and categorized according to the consortium partner that will collect the data. In future versions of this DMP, when the data is more complete, a more detailed categorization system will be used.

- **Dataset reference and name**: dataset identifier
- **Dataset description:** short dataset profile, summary and origin
- **Standards and metadata:** formats used
- **Data sharing:** access policies including restrictions on use
- **Archiving and preservation:** storage and backup provisions
- **Responsible partner:** partner in charge of collecting and maintaining the data

3 Dataset identification and listing

3.1 *Deutsche Welle content*

- **Data set reference and name**: DW texts
- **Data set description:** Texts obtained from Deutsche Welle API regarding selected brands
- **Standards and metadata:** Text, brand, date, language

- **Data sharing:** No sharing. That data is already available from DW.
- **Archiving and preservation:** Preserved in a "sources" index in the platform elasticSearch.
- **Responsible partner:** DW

3.2 *Twitter content*

- **Data set reference and name:** Tweets
- **Data set description:** Tweets extracted from Twitter regarding selected brands
- **Standards and metadata:** Text, brand, date, language, account.
- **Data sharing:** None. There are legal issues sharing this data.
- **Archiving and preservation:** Preserved in a "sources" index in the platform elasticSearch.
- **Responsible partner:** BUT

3.3 *Twitter graph*

- **Data set reference and name:** Twitter graph
- **Data set description:** Relationships for Twitter accounts. That would be followers and followings of accounts that tweeted about our selected brands.
- **Standards and metadata:** RDF.
- **Data sharing:** No sharing. There are legal issues sharing this data.
- **Archiving and preservation:** In a graph database that could be Elasticsearch with the Siren plugin.
- **Responsible partner:** UPM

3.4 *Facebook content*

- **Data set reference and name:** Facebook content
- **Data set description:** A dataset of publicly available user accounts content as provided by SODATO (Copenhagen Business School). SODATO stores the public facebook wall data into a MS SQL Server db and can export a variety of csv files.
- **Standards and metadata:** tbd
- **Data sharing:** Open access.

- **Archiving and preservation:** In a graph database that could be Elasticsearch with the Siren plugin.
- **Responsible partner:** NUIG

3.5 *Websites content*

- **Data set reference and name:** Websites content
- **Data set description:** In case DW text is not enough, web text from some sites should be extracted.
- **Standards and metadata:** Text, brand, date, language, source.
- **Data sharing:** No sharing. There are legal issues sharing this data.
- **Archiving and preservation:** Preserved in a "sources" index in the platform elasticSearch.
- **Responsible partner:** PT

3.6 *Tagged Text*

- **Data set reference and name:** Tagged Text
- **Data set description:** Once text is processed (splitted and emotion, polarity and terms are added) the results are saved to be the base of the analytics.
- **Standards and metadata:** Sentence, brand, date, language, account, original_text, emotions, polarity, concepts, topics, source, media.
- **Data sharing:** No sharing, for commercial reasons.
- **Archiving and preservation:** Preserved in a "results" index in the platform elasticSearch.
- **Responsible partner:** PT

3.7 *SindiceTech Knowledge Graph*

- **Data set reference and name:** Knowledge graph
- **Data set description:** Basis for the MixedEmotions knowledge graph
- **Standards and metadata:** RDF Dumps available,
- **Data sharing:** ST will provide both low level data dumps (RDF) and virtual machines preloaded with the data.

- **Archiving and preservation:** ST will not per se preserve the data as they are integrating sources which are preserved already. The main work will be of integration and cleanup of the data coming from Wikidata and DBpedia along with the integration of support tools
- **Responsible partner:** ST

4 Conclusions

It is too early in the project to have a complete data set identification. Some of the data that will need to be collected is still not clear enough to be detailed with the required level of specification, and others will surely be identified later in the project, so this first version of the Data Management Plan should be taken as a work in progress, still incomplete.

As new data sets to be collected are clearly identified by the consortium partners, the Data Management Plan will be updated accordingly.