Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D8.2

Social Mining Services and Application Integration

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (48 months)
ENDING DATE	31/12/2023
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP8 JRA1 – Social Mining and Big Data Resource Integration
WORK PACKAGE LEADER	LUH
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, UT, IMT, LUH, SNS, AALTO, ETHZ, CNRS, CEU, URV, BSC, UPF, UvA
DELIVERABLE NUMBER	D8.2
DELIVERABLE TITLE	Social Mining Services and application integration
AUTHOR(S)	Giulio Rossetti (CNR), Jurek Leonhardt (LHU)
CONTRIBUTOR(S)	
EDITOR(S)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
REVIEWER(S)	Carlos Castillo (UPF), Jesus A. Manjon Panigua (URV)
CONTRACTUAL DELIVERY DATE	31/03/2020
ACTUAL DELIVERY DATE	12/06/2020
VERSION	V2.0
TYPE	Websites, patents filling, etc
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	16
KEYWORDS	Services, Applications, Algorithm census

EXECUTIVE SUMMARY

The deliverable provides the current census at Month 3 of the methods and applications present in the SoBigData++ online catalogue. Such a description has to be considered as a first snapshot of the resources available within the consortium while an ongoing and up to date view will be provided by the online catalogue, accessible through the project RI.

Furthermore, this document discusses the preliminary plan and guidelines for the development and integration of enhanced services, libraries and applications within the SoBigData RI

The document is organized as follows:

- **Section 1:** provides an introduction to the aim of the deliverable and its relation with the other work packages;
- **Section 2:** reports on the status of the methods available within the consortium and provides results of the census at Month 3. We recall that the list of the available methods is an ongoing catalogue that will be updated through the project lifetime;
- **Section 3:** provides the preliminary guidelines for the development and integration of enhanced services, libraries and applications within the SoBigData RI.

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

This document contains information on SoBigData++ core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData++ Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The content of this publication is the sole responsibility of the SoBigData++ Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

GLOSSARY

EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
RI	Research Infrastructure
VA	Virtual Access
TA	Transnational Access
SNA	Social Network Analysis
ViA	Visual Analytics
WA	Web Analytics
TSSM	Text and Social Media Mining
SD	Social Data
HMA	Human Mobility Analytics

TABLE OF CONTENTS

1 Relevance to SoBigData++

1.1 Purpose of this document

The deliverable outlines a description of the current census of methods available within the consortium, as evicted from the RI catalogue. Moreover, it collects preliminary guidelines on how to integrate and develop enhanced services, applications and libraries to be included within the SoBigData RI.

1.2 Relevance to project objectives

This document shows a first census of the methods available in the consortium and set the principles that will guide the integration of novel enhanced services, libraries and applications.

1.3 Relation to other work packages

Work package 8 is part of “social mining research infrastructure building”, one of three axes the SoBigData++ work plan comprises. It is therefore strongly connected to the other work packages within the same axis, namely WP9 (“SoBigData e-Infrastructure and supercomputing network”) and WP10 (“Exploratories”). They are aimed at building the project core and infrastructure as well as advance research in social mining.

Additionally, WP8 is connected to work packages in the “community building” axis, such as WP2 (“Critical Data Literacy, Ethics and Legal Framework”) and WP4 (“Training”), as they go hand in hand with the creation of the platform and infrastructure. Finally, WP8 maintains connections to the work packages in the “user accessibility” axis, WP6 (“Transnational Access”) and WP7 (“Virtual Access”), as those deal with providing access to the integrated resources.

This deliverable, along with D8.1, is intended to report on the current status of the resources that are available for exploitation in the connected WPs, as well as to describe the preliminary plans devised to enrich and reshape such collection.

1.4 Structure of the document

The document is organized into 3 main sections: Section 2 provide a first overview of the methods already available in the SoBigData++ project catalog. It is important to recall that this document represents only a first census of the available resources in the consortium at Month 3. The list and description of methods and applications will be continuously updated through the catalog hosted on the RI. Finally, Section 3 provides preliminary guidelines that will allow the partners to plan the integration and development of enhanced services, libraries and application within the SoBigData RI.

2 Census of algorithmic resources

2.1 Census of the methods of the consortium

This section proposes a first census of the methods resources available in the consortium. This represents only an initial analysis of the resources available in the project. The list of methods is ongoing and up-to-date through time. At the end of this first census, 31 March 2020, the consortium shows:

Resources: 81 methods, six thematic clusters covered with the following distribution:

- [HMA] Human Mobility Analytics: 24 methods
- [SD] Social Data: 4 methods
- [SNA] Social Network Analysis: 15 methods
- [TSMM] Text and Social Media Mining: 33 methods
- [SDS] Sport Data Science: 1 method
- [WA] Web Analytics: 4 methods

Accessibility: a method can be accessed by virtual and/or transnational modalities, such information is integrated within the metadata.

Focusing on the distribution among VA and TA access as specified for the dataset in the catalog we observe the following results:

- Virtual: 1 method
- Transnational: 10 methods
- Both: 70 methods

Finally, Appendix A reports the list of all methods available in the RI catalog at 31 March 2020. For each method we report some relevant information such as the accessibility, availability and related thematic cluster.

3 Resource Integration and Development Guidelines

3.1 Integration goals

This section provides preliminary guidelines to aid the partners in the process of planning the development and integration of enhanced services, libraries and applications in the SoBigData research infrastructure. In essence, these services and libraries will be organized in thematic areas (or sub-areas). We initially define the thematic areas analogously to the tasks in the work package; they may be modified or extended later.

3.1.1 Guidelines

In this subsection we present the guidelines for partners to follow during the development and integration of methods. We begin by reshaping the aforementioned thematic areas using as guidelines the following tasks of the work package:

- T8.3: Text and Social Media Mining [TSMM]
- T8.4: Complex Network Analysis Mining [SNA]
- T8.5: Human Mobility Analytics [HMA]
- T8.6: Web Analytics [WA]
- T8.7: Visual Analytics [ViA]
- T8.8: Privacy Enhancing Technology and Discrimination Prevention [PETDP]
- T8.9: Explainable AI [XAI]
- T8.10: Scalable Machine Learning [SML]

In the future, these areas may be extended, split into sub-areas or otherwise modified if need be. In the following, we describe specific guidelines for methods and services in the SoBigData research infrastructure:

- Each method to be integrated is expected to be classified as part of one of the thematic areas;
- Methods that are part of the same thematic area should be combined into libraries in order to improve the ease of use and overall usability;
- Similarly, both existing (in the catalog) and new services alike should provide a unified interface. This enables them to be usable online directly from within the SoBigData portal;
- Finally, it is possible to provide stand-alone web applications to serve single methods in appropriate cases.

3.1.2 Next steps

To ensure a uniform and homogeneous integration of novel and existing tools, we will schedule monthly meetings for each Task that will be dedicated to discuss implementation details and advancements status. The first task-related meetings will be dedicated to identify the specific purposes of the libraries/applications to be developed/integrated as well as to identify technological requirements, and to define task related software standards (e.g., target programming language, desired input/output formats, general software structure, etc.). Finally, following the same scheduling, monthly WP meetings among task leaders will be dedicated to report on advancement status, to underline specific issues and coordination among tasks.

4 Conclusions

This deliverable reports an initial census of the algorithmic resources available in the SoBigData++ RI catalog. Such a list (as well as the datasets one, see deliverable D8.1) will be continuously updated throughout the project lifetime. Moreover, this document discusses the goals and rationale behind the integration of novel enhanced services, libraries and applications, thus proposing preliminary guidelines aimed at simplifying such activities as well as the cooperation among the members of the consortium.

Appendix A. The complete algorithm census (as available at 31 March 2020)

Human Mobility Analytics

Name	Thematic Cluster	Availability	Accessibility
Borders	HMA	On-Site	Both
Carpooling	HMA	On-Line	Both
Carpooling Network Analysis	HMA	On-Line	Both
Data-driven ranking of soccer teams	HMA	On-Site	Both
Diary-based Trajectory Generator	HMA	On-Site	Both
Exploration of time use by citizens based on their movement tracks	HMA	On-Line	Both
GeoTopics - A method and system to explore urban activity	HMA	On-Line	Both
Human Mobility Data Privacy Risk Estimator	HMA	On-Site	Both
Injury forecaster for soccer players	HMA	On-Site	Both
Matlas - TClustering	HMA	On-Line	Both
Matlas - Trajectory Builder	HMA	On-Line	Both
Mobility Profile	HMA	On-Line	Both

Modelling Scientific Migration	HMA	On-Site	Trans National Access
MyWay - Trajectory Prediction	HMA	On-Line	Both
Nowcasting migration stocks and flows	HMA	On-Site	Trans National Access
Nowcasting well-being with human mobility data	HMA	On-Line	Trans National Access
Origin Destination Matrix Computation	HMA	On-Line	Both
PlayeRank evaluation framework	HMA	On-Line	Both
Prediction of next career moves from scientific profiles	HMA	On-Site	Trans National Access
Privacy Risk on Trajectories	HMA	On-Line	Both
Soccer teams ranking simulator	HMA	On-Site	Both
Sociometer	HMA	On-Line	Both
TripBuilder	HMA	On-Line	Virtual Access
Urban Mobility Atlas	HMA	On-Line	Both

Social Data

Name	Thematic Cluster	Availability	Accessibility
Economic Integration Model	SD	On-Site	Trans National Access
LORE	SD	On-Line	Both
Privacy Risk on Sociometer	SD	On-Line	Both
TARS based prediction for Next Basket	SD	On-Line	Both

Social Network Analysis

Name	Thematic Cluster	Availablity	Accessibility
DEMON	SNA	On-Line	Both
DebtRank Systemic Risk Estimation Method	SNA	On-Line	Both
Egonetworks	SNA	On-Line	Both
Estimating Collective Wellbeing	SNA	On-Line	Both
F1-Communities	SNA	On-Line	Both
KDDMultiGraph	SNA	On-Line	Both
Leader Detect	SNA	On-Line	Both
MaxAndSam Network Reconstruction Method	SNA	On-Line	Both
Maximum entropy network reconstruction	SNA	On-Line	Both

NDlib	SNA	On-Line	Both
NDlib-rest	SNA	On-Line	Both
SCube	SNA	On-Site	Both
Statistical validation	SNA	On-Line	Both
TILES	SNA	On-Line	Both
Tail granger causality network construction	SNA	On-Line	Both

Text and Social Media Mining

Name	Thematic Cluster	Availability	Accessibility
Annie Plus Measurements	TTSMM	On-Line	Both
Cymrie Welsh Named Entity Recognizer	TTSMM	On-Line	Both
DecarboNet Environmental Annotator	TTSMM	On-Line	Both
DecarboNet German Environmental Annotator	TTSMM	On-Line	Both
Dictionary creator	TTSMM	On-Line	Both
Digital DNA fingerprinting	TTSMM	On-Site	Trans National Access
Distance Calculator	TTSMM	On-Line	Both
English Named Entity Recognizer	TTSMM	On-Line	Both
English Named Entity Recognizer for Tweets	TTSMM	On-Line	Both

English Part of Speech and Morphology Anaylizer	TTSMM	On-Line	Both
English Tweet Tokenizer	TTSMM	On-Line	Both
Epidemic Sentiment Analysis	TTSMM	On-Site	Trans National Access
French Named Entity Recognizer	TTSMM	On-Line	Both
French Named Entity Recognizer for Tweets	TTSMM	On-Line	Both
GSP - Geo-Semantic-Parsing	TTSMM	On-Site	Trans National Access
Generic Opinion Mining English	TTSMM	On-Line	Both
German Named Entity Recognizer	TTSMM	On-Line	Both
German Named Entity Recognizer for Tweets	TTSMM	On-Line	Both
Language Identification for Tweets	TTSMM	On-Line	Both
MARLENA	TTSMM	On-Line	Both
Measurement Expression Annotator	TTSMM	On-Line	Both
Noun Phrase Chunker	TTSMM	On-Line	Both
Open NLP Dutch Pipeline	TTSMM	On-Line	Both
Open NLP English Pipeline	TTSMM	On-Line	Both
Open Nlp German Pipeline	TTSMM	On-Line	Both
Part Of Speech Tagger for Tweets	TTSMM	On-Line	Both

Polarized User and Topic Tracking	TTSMM	On-Line	Both
Summa Text Summarization (En)	TTSMM	On-Line	Both
Summa Text Summarization (Es)	TTSMM	On-Line	Both
Superdiversity and Sentiment	TTSMM	On-Site	Trans National Access
The Brexit Analyzer Pipeline	TTSMM	On-Line	Both
Twitter Opinion Mining English	TTSMM	On-Line	Both
Twitter preprocessor	TTSMM	On-Site	Trans National Access

Sport Data Science

Name	Thematic Cluster	Availability	Accessibility
Influence of missing RR-Intervals caused by motion artifacts on HRV features estimations	SDS	On-Line	Both

Web Analytics

Name	Thematic Cluster	Availability	Accessibility
ArchiveSpark	WA	On-Line	Both
QuickRank	WA	On-Line	Both
Web Archive Collection Extractor	WA	On-Line	Both
WikiData Geo Mapper	WA	On-Line	Both