Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D8.1

Social Data resources and Social media observatory report

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	CNR, LUH
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, UT, IMT, LUH, SNS, AALTO, ETHZ, CNRS, CEU, URV, BSC, UPF, UvA
DELIVERABLE NUMBER	D8.1
DELIVERABLE TITLE	Social Data resources and Social media observatory report
AUTHOR(S)	Giulio Rossetti (CNR), Guglielmo Cola (CNR), Valerio Grossi (CNR)
CONTRIBUTOR(S)	Maurizio Tesconi (CNR), Roberto Trasarti (CNR)
EDITOR(S)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
REVIEWER(S)	Nadia Tonello (BSC)
CONTRACTUAL DELIVERY DATE	31/03/2020
ACTUAL DELIVERY DATE	12/06/2020
VERSION	V2.0
TYPE	ORDP: Open Research Data Pilot
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	23
KEYWORDS	Data Management, datasets, dataset census, social media observatory

EXECUTIVE SUMMARY

The deliverable provides the current census at Month 3 of the datasets present in the SoBigData++ online catalogue. Such a description has to be considered as a first snapshots of the data 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. The census includes statistics, metadata and a pointer to the sharing policies, archiving technologies as well as the preservation provisions and lifespans for the collected data.

Furthermore, this document presents the idea of Social Media Observatory. To this extent, are briefly introduced and discussed a few tools already available within the consortium around which the Social Media Observatory will be built.

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 data management plan and provides results of the current dataset census at Month 3. We recall that the list of the available datasets is an ongoing catalogue that will be updated through the project lifetime;
- **Section 3:** provides the description of what will become the Social Media Observatory, briefly discussing existing tools and application already developed within the consortium that can be identified as prototypical backbones of such service.

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 datasets available within the consortium, as evicted from the RI catalogue. The description includes statistics and a pointer to the metadata, sharing policies and archiving technologies as well as to the preservation provisions and lifespan. Moreover, it provides a prototypical description of what will become the Social Media Observatory, discussing a few already available methods and applications that will become its initial backbone.

1.2 Relevance to project objectives

This document shows a first census of the datasets available in the consortium and an initial characterization of the Social Media Observatory.

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 deliverable D8.2, 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 data management in the SoBigData++ project. It is important to recall that this document represent only a first census of the datasets available in the consortium at Month 3. The list and description of the dataset will be continuously updated through the catalogue hosted on the RI. Finally, Section 3 provides a preliminary description of what will become the Social Media Observatory: to such extent, in this section will be briefly described a few tools and applications already available within the consortium that will act as prototypical backbones for such a service.

2 Data Management and census of the datasets

2.1 Data Management

The purpose of the data management plan is to define policies that will guide the partners in the collection, description, preservation and sharing of their data sets for VA and TA.

Research on social mining relies on massive datasets of digital traces from human activities. Many big datasets were already made available within the SoBigData RI: such resources include, and are not limited to, transaction micro-data from diverse retailers, networks crawled from several online social networks, query logs from search engines and e-commerce, GPS tracks from personal navigation devices, survey data about customer satisfaction, large Web archives and data from location-aware networks. All partners will continue to make such data available, as well as including new ones, by adopting all the policies already defined within the SoBigData project. The access under VA and TA will concern both existing and newly collected datasets. Datasets will be made available so to enforce the requirements expressed by their sharing policies (public, restricted and private availability). In particular, the access through VA will be granted for all those datasets whose sharing policies allow open diffusion; conversely, for all the datasets whose access is restricted due to licensing constraints (e.g., restricted and private availability), access will be provided exclusively through TA.

In accordance to the requirements identified during the SoBigData project – as expressed by its task T10.1 “e-infrastructure interoperability” and T10.2 “Integration to the e-infrastructure” – the metadata collected for each dataset (along with their sharing policies) are, and will continue to be, the ones discussed in D8.1 of the same project[1]. Symmetrically, for details on what concern archiving technologies as well as the data preservation provisions and lifespan the reference deliverable from SoBigData is D10.1 “Best practices and guidelines towards interoperability”[2]. The first version of the Data Management Plan (DMP) was defined in the SoBigData H2020 Project. The DMP is updated by the SoBigData++ project annually. In particular, the DMP will be updated to 31st December 2020. Further details about data management can be found in Appendix A.

2.1.1 Census of datasets of the consortium

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

[1] https://goo.gl/kjcBZS
[2] https://goo.gl/YrOi1P

Resources: 90 datasets, five thematic clusters covered with the following distribution:

- [HMA] Human Mobility Analytics: 20 datasets
- [SD] Social Data: 11 datasets
- [SNA] Social Network Analysis: 19 datasets
- [TSMM] Text and Social Media Mining: 30 datasets
- [ViA] Visual Analytics: 1 dataset
- [WA] Web Analytics: 9 datasets

Accessibility: a dataset can be accessed by virtual and/or transnational modalities. This information is integrated within the metadata and can assume three different values:

- Public: for public data, e.g., open data;
- Restricted: for data available under specific restrictions, e.g., NDA;
- Private: for data that cannot be accessed by the user directly but only through APIs, services ore views that provide only aggregated information or analysis.

It is worth to notice that multiple choices can be selected for a dataset. As an example, a dataset can be marked as “Virtual/Public” and “Transnational/Public” if it is freely accessible through both VA and TA. Focusing on the distribution among VA and TA access as specified for the dataset in the catalogue we observe the following results:

- Virtual/Public: 6 datasets
- Transnational: 35 datasets
- Both: 48 datasets

Finally, Appendix A reports the list of all datasets available in the RI catalogue at 31 March 2020. For each dataset we report some relevant information such as the accessibility, or the manifestation type that shows if a dataset is a replica, i.e., if it has been pre-processed or transformed in some way, or it is original, I.e., it is has been taken as it has been generated.

3 Social media observatory

3.1 Definition, ambition and goals

The rapid growth of social networking platforms produced a great interest in assessing how massive real-time social data can be used as a mine of information in numerous domains. In this context, we aim to develop the Social Media Observatory, which will consist of a set of tools that facilitate the creation of social media listening campaigns ("social sensing") and high-level interpretation of collected data.

In social sensing, also known as crowdsensing, the users of social media platforms actually represent the "human sensors" of data collection campaigns, as they spontaneously produce a massive amount of information related to disparate topics [1]. In particular, social networking applications like Facebook, Instagram and Twitter have been the source of information for many crowdsensing systems over the last ten years.

The tools of the Social Media Observatory will be available to researchers through Transnational Access. It will be possible to target data collection campaigns towards specific topics of interest, such as political elections, natural and man-made disasters, the spread of epidemics, etc. More specifically, the tools offered by this Observatory will enable easy setup of targeted data collection on social platforms by specifying keywords (e.g., "presidential election", "covid19"), user accounts, or geographic areas on interest. In addition, the Observatory will include a set of postprocessing tools to enrich data with high-level information which is not directly available on the raw data. An example is represented by fake and bot account detection tools [2], which will be implemented by leveraging digital DNA-based techniques. These tools could be exploited to investigate the sources of infodemic phenomena, such as the spread of misinformation and disinformation surrounding events like the recent Covid-19 pandemic. As for data enrichment, the Observatory will include tools to classify the type of sentiment and hate speech level, the extraction of geographic information obtained through geoparsing [3], and finally the detection of possible "links" between two or more user accounts by means of user identity linkage techniques.

3.2 Examples of preexisting tools and applications

The main preexisting tool that will be exploited to implement the Social Media Observatory is Twitter Monitor, which was developed by CNR [4] as part of the SoBigData project. Twitter Monitor is a crowdsensing tool written in PHP, designed to access Twitter streams through the Twitter Streaming API. It offers an easy-to-use web interface mainly developed in AJAX, HTML5, CSS3, and Javascript. Through this interface, users can perform a number of complex operations which are automatically translated into specific API requests issued to Twitter. Twitter Monitor is able to manage concurrent monitors: it is possible to launch parallel listening sessions (i.e., more than one Twitter crawler at the same time) using different parameters and collecting different sets of data. Twitter Monitor also offers a set of functionalities aimed at minimizing the loss of data due to network or local machine failures. In particular, Twitter Monitor is automatically capable of detecting and recovering from simple error conditions, such as a closed or disconnected Twitter streams.

It is also capable of detecting more serious issues, such as Twitter refusing to open new streaming connections, and automatically sends targeted alerts to system administrators.

The Social Media Observatory will be built on top of Twitter Monitor, as the latter represents the main tool to start social sensing campaigns. In order to facilitate data filtering and enrichment, the use of different libraries will be integrated in the Observatory. One preexisting example of such libraries is represented by the Digital DNA Toolbox (digitaldna package), developed by CNR. This tool exploits a novel approach to modeling online user behavior as digital DNA-inspired sequences. Sequences are then analyzed by means of standard DNA analysis techniques to discriminate between genuine and spambot accounts. Thanks to this technique, researchers using the Observatory will be able to quickly filter out possibly irrelevant data (noise) produced by bots and fake accounts, as well as investigate the impact of fake accounts on the spread of disinformation related to a topic of interest.

4 Conclusions

This deliverable reports an initial census of the dataset available in the SoBigData++ RI catalogue. Moreover, it discusses the ambitions and goals of the Social Media Observatory, providing descriptions and statistics of a subset of the applications and services that will act as its initial backbones. The list of available datasets (as well as the algorithmic resources one, see deliverable D8.2) will be continuously updated throughout the project lifetime.

References

[1] Avvenuti, M., Bellomo, S., Cresci, S., La Polla, M. N., & Tesconi, M. (2017, April). Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In Proceedings of the 26th international conference on World Wide Web companion (pp. 1413-1421).

[2] Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019, June). RTbust: exploiting temporal patterns for botnet detection on twitter. In Proceedings of the 10th ACM Conference on Web Science (pp. 183-192).

[3] Avvenuti, M., Cresci, S., Nizzoli, L., & Tesconi, M. (2018, June). GSP (Geo-Semantic-Parsing): Geoparsing and Geotagging with machine learning on top of linked data. In European Semantic Web Conference (pp. 17-32). Springer, Cham.

[4] Cresci, S., Minutoli, S., Nizzoli, L., Tardelli, S., & Tesconi, M. (2019, January). Enriching digital libraries with crowdsensed data. In Italian Research Conference on Digital Libraries (pp. 144-158). Springer, Cham.

Appendix A. Data Management Plan details

The first version of the Data Management Plan was defined in the SoBigData H2020 Project in 2015. The DMP is updated during the SoBigData++ project starting from January 2020. The DPM (and all the associated resources) is updated annually and is available at the following link https://data.d4science.net/5i2k (registration is required to download the file). The SoBigData Resource Catalogue represents a primary place for users to be informed on what is available. A heterogeneous community of researchers powers SoBigData RI to study and analyze social data. The list of datasets reported in Appendix B shows different kinds of data collected and created to support the research activities developed inside WP8 and WP10. Figure 2.1.1 reports the taxonomy and the data formats of the datasets currently (March 2020) available in the catalogue.

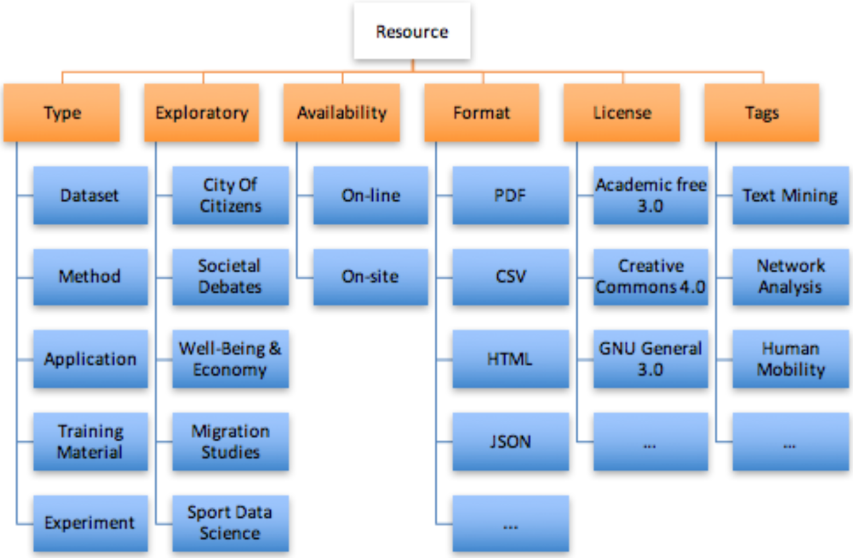


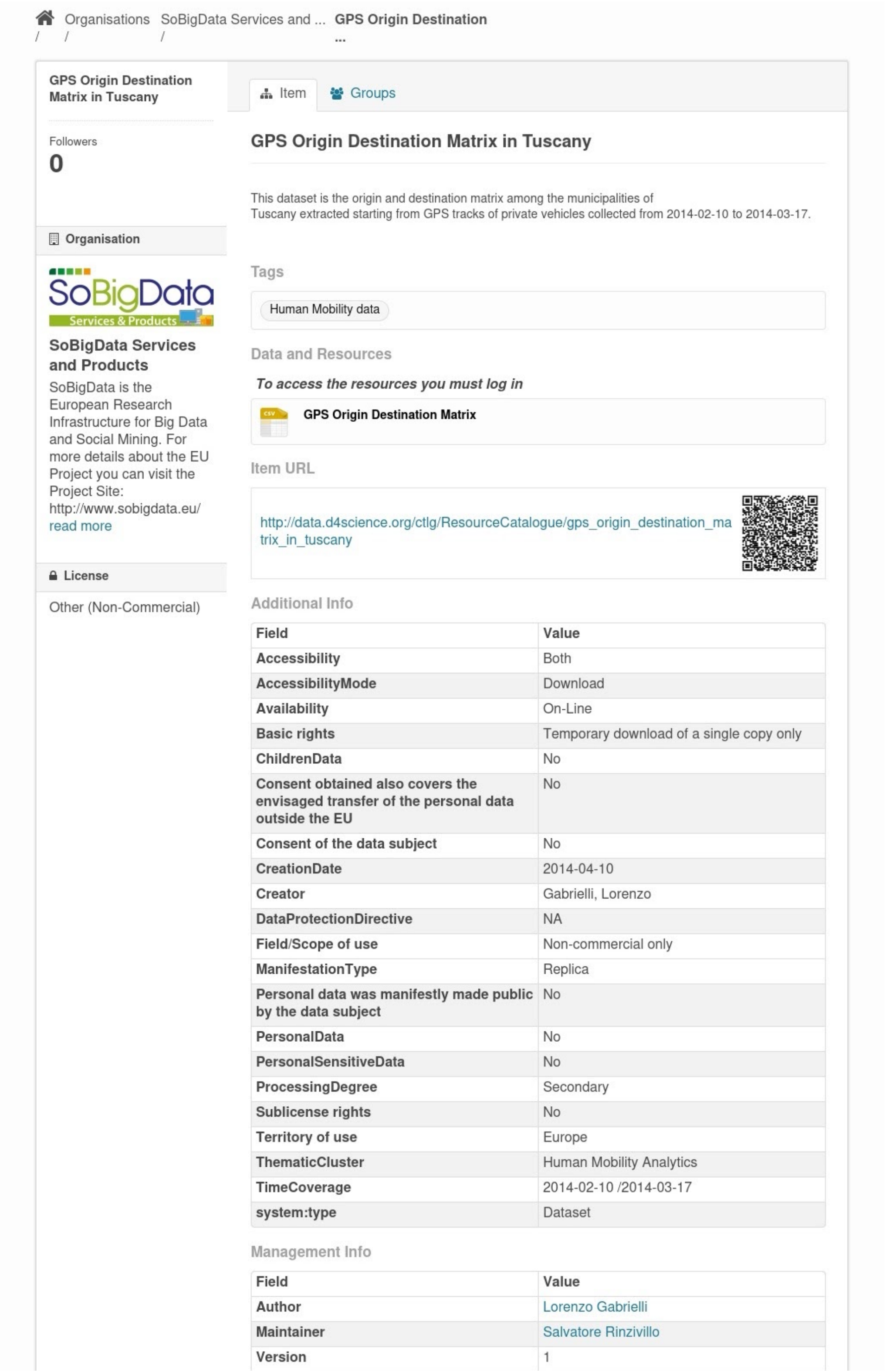
Figure 2.1.1 Taxonomy of the resources in the SoBigData RI Catalogue

SoBigData RI promotes open science and its development under FAIR (Findable, Accessible, Interoperable, and Reusable) and FACT (Fair, Accurate, Confidential, and Transparent) principles. At the moment, SoBigData RI counts two ways to provide data access inside the RI. Three fields define the accessibility rules for a dataset: accessibility defines how the access to the resource is regulated (Virtual Access, TransNational Access, or Both); availability outlines how the availability to the resource is offered (on-line by e-infrastructure facilities, on-site by visiting the institution that is the data controller of the dataset); accessibility mode describes the nature of the dataset and how access to the resource is implemented. On-line access is used for connecting to the servers that provide data (e.g., a DBMS or API Access). We plan the integration of two types of datasets. Publicly Available Datasets are made available by private and public entities and included in the RI resources. Restricted data, i.e., datasets that will be made available prevalently on-site through Transnational Access due to the restriction imposed by data owners. Furthermore, dataset metadata include a unique reference and an assessment of their nature, scale, and available metadata (such as related scientific publications, privacy issues, data governance policies, licensing, or similar resources). The preservation and re-using procedures describe how the partners store the data, which technology is used, and how long the data is available.

Access through VA will be granted for all those datasets whose policies allow open diffusion; conversely, access will be provided only through TA for all the data sets whose access is restricted due to licensing restrictions. Moreover, for some datasets, to avoid Term of Usage (ToS) infringements, access will be offered in the form of data crawlers which can be used both in VA and TA to obtain data directly from the original source and for a specific and time-limited experiment (e.g., Twitter data).

Without any registration, SoBigData Catalogue enables the user to discover datasets and access the dataset's metadata. Figure 2.1.2 reports an example of the metadata available for a dataset. For the datasets freely available, a direct link for the download is provided. The download of an item requires the registration and the log-in of the user. For the dataset with different access policies, the instructions for reaching the data are provided. As reported in Figure 2.1.2 each dataset has associated a maintainer, who is then responsible for the dataset. In the first months of the SoBigData++ project, a complete audit of all the datasets has been performed. We removed all the datasets where the maintainer was no more available, and we updated and integrated the metadata required for managing this kind of situation.

At the moment each dataset includes the territory of use, the period of time during which the dataset may be used; and the technical measures and organisational conditions (FAIR ecosystem link) for Dataset Re-Use Safeguards and Retention Period, i.e., the dataset should be set to private after this period.



Organisations / SoBigData Services and ... / **GPS Origin Destination ...**

GPS Origin Destination Matrix in Tuscany

Followers

0

Organisation

SoBigData Services and Products

SoBigData is the European Research Infrastructure for Big Data and Social Mining. For more details about the EU Project you can visit the Project Site: http://www.sobigdata.eu/ read more

License

Other (Non-Commercial)

GPS Origin Destination Matrix in Tuscany

This dataset is the origin and destination matrix among the municipalities of Tuscany extracted starting from GPS tracks of private vehicles collected from 2014-02-10 to 2014-03-17.

Tags

Human Mobility data

Data and Resources

To access the resources you must log in

GPS Origin Destination Matrix

Item URL

http://data.d4science.org/ctlg/ResourceCatalogue/gps_origin_destination_matrix_in_tuscany

Additional Info

Field	Value
Accessibility	Both
AccessibilityMode	Download
Availability	On-Line
Basic rights	Temporary download of a single copy only
ChildrenData	No
Consent obtained also covers the envisaged transfer of the personal data outside the EU	No
Consent of the data subject	No
CreationDate	2014-04-10
Creator	Gabrielli, Lorenzo
DataProtectionDirective	NA
Field/Scope of use	Non-commercial only
ManifestationType	Replica
Personal data was manifestly made public by the data subject	No
PersonalData	No
PersonalSensitiveData	No
ProcessingDegree	Secondary
Sublicense rights	No
Territory of use	Europe
ThematicCluster	Human Mobility Analytics
TimeCoverage	2014-02-10 /2014-03-17
system:type	Dataset

Management Info

Field	Value
Author	Lorenzo Gabrielli
Maintainer	Salvatore Rinzivillo
Version	1

Figure 2.1.2 Public metadata for a dataset

Appendix B. The complete dataset census (as available at 31 march 2020)

Human Mobility Analytics Datasets

Dataset Name	Manifestation	Thematic Cluster	Processing Degree	Availability	Accessibility
CDR Data - Rome	Original	HMA	Primary	On-Site	Trans National Access
CDR data Tuscany	Original	HMA	Primary	On-Site	Trans National Access
Call Data Record District of Pisa 2013 October	Replica	HMA	Primary	On-Site	Trans National Access
Call Data Record Pisa 2012	Replica	HMA	Primary	On-Site	Trans National Access
Call Data Record Tuscan cities 2014	Replica	HMA	Primary	On-Site	Trans National Access
Car sharing dataset	Virtual	HMA	Primary	On-Site	Trans National Access
City-to-city migration	Replica	HMA	Primary	On-Site	Both
Flickr and Wikipedia Tourism Trajectories	Virtual	HMA	Secondary	On-Site	Both
GPS Tracks - Calabria Italy 2012	Replica	HMA	Primary	On-Site	Trans National Access
GPS Tracks - Mestre Italy 2010	Replica	HMA	Primary	On-Site	Trans National Access
GPS Tracks - Milan Italy - Simulated	Virtual	HMA	Secondary	On-Line	Trans National Access
GPS Tracks - Tuscany 2011	Replica	HMA	Primary	On-Site	Trans National Access
GPS Tracks Pisa - Italy 2010	Replica	HMA	Primary	On-Site	Trans National Access

GPS Tracks Tuscany by volunteers	Original	HMA	Primary	On-Site	Trans National Access
GeoLife GPS trajectories dataset	Original	HMA	Primary	On-Line	Both
Open data from NervousNet	Original	HMA	Primary	On-Line	Both
Scientific Publications Dataset	Replica	HMA	Primary	On-Line	Both
Soccer Events	Replica	HMA	Secondary	On-Line	Virtual Access
Soccer Team Performance	Replica	HMA	Secondary	On-Site	Trans National Access

Social Data Datasets

Dataset Name	Manifestation	Thematic Cluster	Processing Degree	Availability	Accessibility
Churn Dataset	Virtual	SD	Primary	On-Line	Both
Compas	Virtual	SD	Primary	On-Line	Both
Dataset Adult	Virtual	SD	Primary	On-Line	Both
Food consumption data at the canteens of University of Pisa	Virtual	SD	Primary	On-Site	Both
German Credit	Virtual	SD	Primary	On-Line	Both
Medical Dataset	Virtual	SD	Primary	On-Line	Both
Retail Market Data	Replica	SD	Secondary	On-Site	Trans National Access
Retail market dataset	Replica	SD	Primary	On-Site	Trans National Access
UCR Time Series Classification Archive	Original	SD	Primary	On-Line	Both

Yeast	Virtual	SD	Primary	On-Line	Both
e-MID dataset	Original	SD	Primary	On-Site	Restricted

Social Network Analysis Datasets

Dataset Name	Manifestation	Thematic Cluster	Processing Degree	Availability	Accessibility
Aalto-Twitter	Original	SNA	Primary	On-Site	Trans National Access
Amazon Network	Original	SNA	Primary	On-Line	Both
Congress Network	Virtual	SNA	Primary	On-Line	Both
DBLP Network	Original	SNA	Primary	On-Line	Both
Disease Twitter Dataset	Virtual	SNA	Primary	On-Site	Trans National Access
Estonian public sector electronic services and service providers and consumers	Replica	SNA	Secondary	On-Site	Both
European Banks Asset Class exposures	Virtual	SNA	Secondary	On-Line	Both
FED data	Replica	SNA	Primary	On-Site	Both
Facebook - New Orleans regional network	Virtual	SNA	Primary	On-Line	Virtual Access
Facebook EuroSys 2009	Virtual	SNA	Primary	On-Line	Virtual Access
Facebook Wallpost	Virtual	SNA	Primary	On-Line	Both
Formal network of Estonian companies and board members	Replica	SNA	Secondary	On-Site	Trans National Access

NYSE transactions	Replica	SNA	Primary	On-Site	Trans National Access
Russell 3000 stock prices	Replica	SNA	Secondary	On-Site	Trans National Access
Social Network dataset - LiveJournal	Original	SNA	Primary	On-Line	Both
Twitter Dataset 2013-2014	Virtual	SNA	Primary	On-Site	Trans National Access
WEIBO interactions	Virtual	SNA	Secondary	On-Line	Both
Bond yield_equity log-returns_CDS spreads	Replica	SNA	Primary	On-Site	Trans National Access
e-MID interbank transactions	Replica	SNA	Primary	On-Site	Trans National Access

Text and Social Media Mining Datasets

Dataset Name	Manifestation	Thematic Cluster	Processing Degree	Availability	Accessibility
Aalto-Foursquare	Original	TSMM	Secondary	On-Site	Trans National Access
Amazon reviews	Original	TSMM	Primary	On-Line	Both
Articles and comments of major Estonian newspapers	Replica	TSMM	Secondary	On-Site	Trans National Access
Brexit Tweets Linked Domains	Virtual	TSMM	Secondary	On-Line	Both
Brexit Twitter User Vote Intent	Virtual	TSMM	Secondary	On-Site	Trans National Access
Broad Twitter Corpus	Virtual	TSMM	Secondary	On-Line	Both
Emergency Tweets 2009 L'Aquila earthquake	Original	TSMM	Primary	On-Line	Both

Emergency Tweets 2011 Christchurch earthquake	Original	TSMM	Primary	On-Line	Both
Emergency Tweets 2012 Emilia earthquake	Original	TSMM	Primary	On-Line	Both
Emergency Tweets 2013 Milan blackout	Original	TSMM	Primary	On-Line	Both
Emergency Tweets 2013 Sardinia flood	Original	TSMM	Primary	On-Line	Both
Emergency Tweets 2014 Genoa flood	Original	TSMM	Primary	On-Line	Both
Emergency Tweets 2016 Amatrice earthquake	Original	TSMM	Primary	On-Line	Both
GERDAQ Dataset	Original	TSMM	Primary	On-Line	Virtual Access
Geo-annotated tweets ENG-ITA	Original	TSMM	Primary	On-Line	Both
IMDB Network	Original	TSMM	Primary	On-Line	Virtual Access
ISTAT Census zone Tuscany	Original	TSMM	Primary	On-Line	Both
Official administrative information of Tuscany	Original	TSMM	Primary	On-Line	Both
Sheffield NERD Tweet Corpus	Virtual	TSMM	Secondary	On-Line	Both
Soccer Data Challenge dataset	Virtual	TSMM	Primary	On-Line	Both
The Italian Music Dataset	Virtual	TSMM	Primary	On-Line	Both
Twitter Dumps	Virtual	TSMM	Primary	On-Site	Trans National Access

Twitter fake followers	Original	TSMM	Primary	On-Site	Both
Twitter social bots	Original	TSMM	Primary	On-Site	Both
UK General Election Vote Intent	Virtual	TSMM	Secondary	On-Site	Trans National Access
UK election abuse data	Virtual	TSMM	Secondary	On-Line	Both
WIRE dataset	Virtual	TSMM	Primary	On-Line	Both
Wikinews dataset	Virtual	TSMM	Primary	On-Line	Both
Wikipedia Word Embeddings	Virtual	TSMM	Secondary	On-Line	Both
Word Sense Evolution Testset	Virtual	TSMM	Primary	On-Line	Both

Visual Analytics Datasets

Dataset Name	**Manifestation**	**Thematic Cluster**	**Processing Degree**	**Availability**	**Accessibility**
Public data set of spatio-temporal match events in soccer competitions	Virtual	ViA	Primary	On-Line	Both

Web Analytics Datasets

Dataset Name	**Manifestation**	**Thematic Cluster**	**Processing Degree**	**Availability**	**Accessibility**
.ee Web archive	Original	WA	Primary	On-Site	Trans National Access
ClueWeb09	Replica	WA	Primary	On-Site	Trans National Access

ClueWeb12	Replica	WA	Primary	On-Site	Trans National Access
CoPhIR	Virtual	WA	Secondary	On-Site	Virtual Access
DE webarchive	Replica	WA	Primary	On-Line	Trans National Access
German Academic Web	Original	WA	Primary	On-Site	Trans National Access
Global Peace Index data	Replica	WA	Primary	On-Site	Trans National Access
MSN Search query log	Virtual	WA	Primary	On-Site	Trans National Access
Wyscout soccer-logs dataset	Replica	WA	Primary	On-Site	Both