Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Integrating Open Data through Innovation Accelerator Platforms***
Deliverable No.	***D8.4***
Delivery Date	***31 August 2016***
Authors	***Izabela Moise (ETHZ), Nino Antulov-Fantulin (ETHZ)***

SoBigData receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D8.4
Deliverable Title	Integrating Open Data through Innovation Accelerator Platforms
Contractual Delivery Date	31 August 2016
Actual Delivery Date	22 August 2016
Author(s)	Izabela Moise (ETHZ), Nino Antulov-Fantulin(ETHZ)
Editor(s)	Nino Antulov-Fantulin (ETHZ)
Reviewer(s)	Gerhard Gossen (LUH), Thomas Risse (LUH), Valerio Grossi (CNR)
Contributor(s)	Evangelos Pournaras (ETHZ)
Work Package No.	WP8
Work Package Title	WP8 - JRA1_Big Data Ecosystem
Work Package Leader	LUH
Work Package Participants	ETHZ, CNR
Dissemination	Public
Nature	Report
Version / Revision	V1.0
Draft / Final	Final
Total No. Pages (including cover)	17
Keywords	Open data, Integration, Participatory platforms

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.
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 document has been produced with the funding of the European Commission. 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.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
API	Application programming interface
iOS	iPhone OS is a mobile operating system created and developed by Apple Inc.
JSON	JavaScript Object Notation is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate
SQL	Structured Query Language is a special-purpose programming language designed for managing data held in a relational database management system

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable describes the integration of open data through Innovation Accelerator Platforms. The title of deliverable D8.4. was “Living Archive Platform” with the corresponding task T5.4. “Integrating Open Data through the Living Archive Platform”. However, the Living Archive platform and its functions became obsolete since the time of writing the description of task T5.4. and its development had stopped due to the emergence of novel search engine options such as: https://www.google.com/publicdata/, https://webscope.sandbox.yahoo.com/, https://scholar.google.com/ and others.

For the above-stated reasons, the Living Archive platform has been replaced with the more general Innovation Accelerator Platforms such as Nervousnet platform, which is built and designed as an open, decentralized, participatory platform which aligns better with the scopes of the SoBigData project. Consequently, the title “Integrating Open Data through the Living Archive Platform” is changed to “Integrating Open Data through Innovation Accelerator Platforms”. The Nervousnet system with respect to the Living Archive platform provides extra functionalities and addresses broader scopes such as:

- crawling is replaced with distributed data acquisition,
- searching is replaced with more general data analytics engine,
- scientific data from Living Archive is replaced with more general open data: sensory data, social data (messages), proximity data, spatio-temporal data and enviromental data,
- filtering is replaced with querying with additional privacy-preserving options.

The deliverable D8.4. describes the new Nervousnet platform that is being used as Innovation Accelerator Platform, the arhitecture of the system and the integration of open data through this platform to the SoBigData infrastructure. Nervousnet uses the sensor networks that make up the Internet of Things, including those in smartphones, to measure the world around us and to build a collective “data commons”. Nervousnet now enables anyone to measure and analyse aspects of the world in real time.

The Nervousnet app allows users to activate or deactivate about ten smartphone sensors that measure, among others, acceleration, light and noise. A range of other functions are being shaped by the core research and development team at ETH. Nervousnet is run as a ‘citizen web’, built and managed by its users. Inspired by Wikipedia and OpenStreetMap, people can interact with Nervousnet in three ways. They can contribute data, analyse the crowd sourced data sets and share code and ideas. Anyone can create data-driven services and products using a generic programming interface. Nervousnet uses distributed data storage and distributed control, so that it is resilient to attacks and centralized manipulation attempts, easy to scale up and tolerant to faults. Because data encryption is not enough, a secure personal-data store is needed to allow each user to determine which data to share with whom and for what purpose.

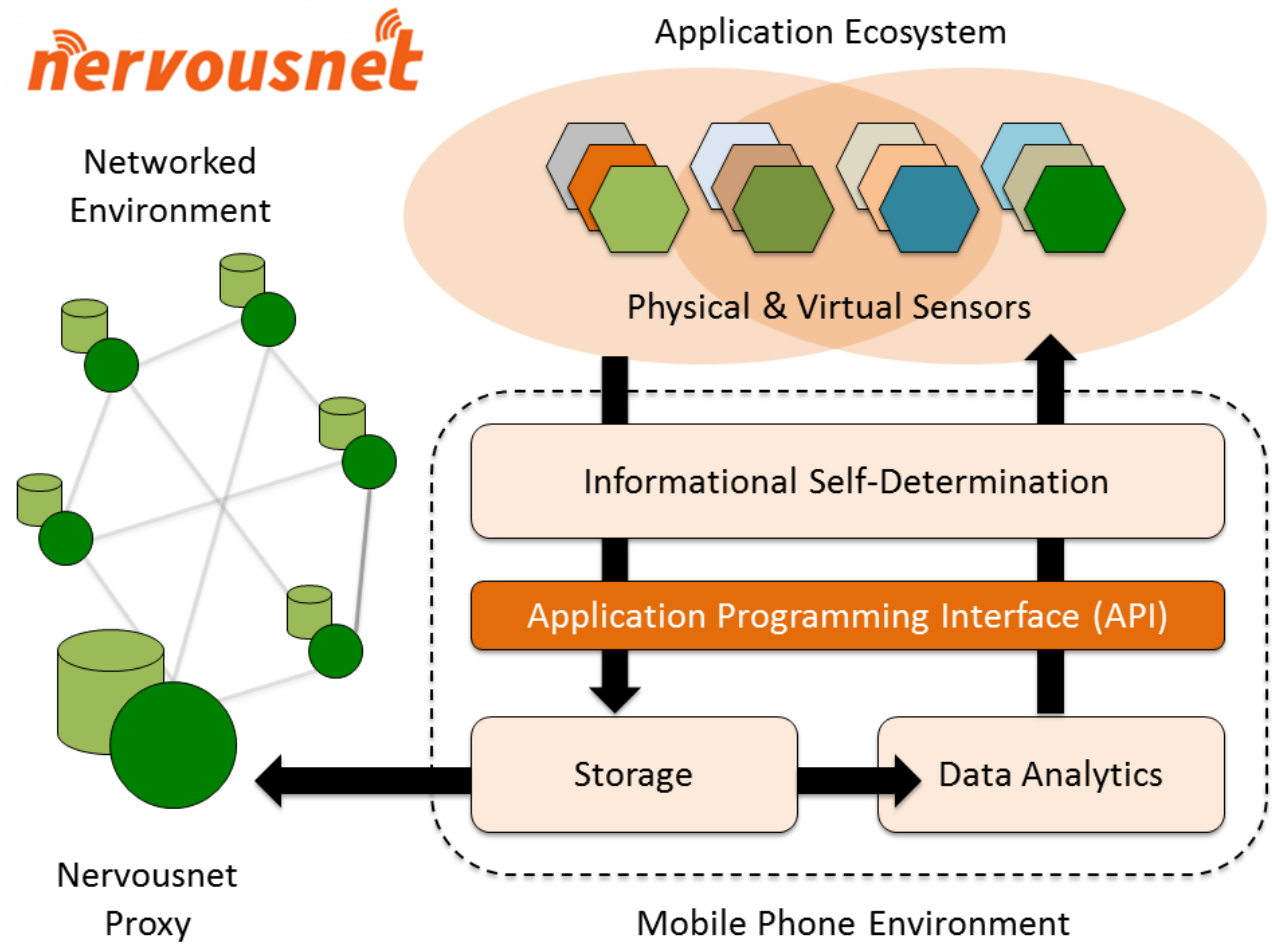
Due to the above-stated functionalities, the Nervousnet platform brings wider contributions to the SoBigData project, while being better aligned to the scope of the project: participatory and privacy-preserving social mining.

EXECUTIVE SUMMARY

This deliverable D8.4. describes the integration of open data through Innovation Accelerator Platforms in the task T8.4. Innovation Accelerator platforms (described in WP5) enable the acquisition of open data sets which will be integrated in the SoBigData infrastructure through servers provided by the Nervousnet system. Open data include scientific data, sensory data, social data, proximity data, spatial-temporal data, environmental data. The open data will be integrated within the e-Infrastructure resource catalogue described in D10.6.

1 INTRODUCTION

The Nervousnet system is a large-scale distributed research platform that provides real-time social sensing services as a public good. Existing Big Data systems threaten social cohesion as they are designed to be closed, proprietary, privacy-intrusive and discriminatory. In contrast, the Nervousnet system is an open, privacy-preserving and participatory platform designed to be collectively built by citizens and for citizens. The system is enabled by Internet of Things technologies and aims at seamlessly interconnecting a large number of different pervasive devices, e.g. mobile phones, smart sensors, etc. For this purpose, several universal state-of-the-art protocols and communication means are introduced. A novel social sensing paradigm shift is engineered: Users are provided with freedom and incentives to share, collect and, at the same time, protect data of their digital environment in real-time. In this way, social sensing turns into a knowledge extraction service of public good. The social sensing services of the Nervousnet system can be publicly used for building novel innovative applications. Whether you would like to detect an earthquake, perform a secure evacuation or discover the hot spots of a visited city, the Nervousnet makes this possible by collectively sensing social activity of participatory citizens.



2 ARCHITECTURE OF NERVOURSNET

The system consists of two software parts: (i) the NERVOUSNET BACKEND and (ii) the NERVOUSNET VIRTUAL SENSORS.

2.1 NERVOUSNET BACKEND

The Nervousnet backend is a distributed middleware software for social sensing. The current stable version of the NervousNet Backend includes a centralized single proxy server. Decentralized versions is under development and specifications are still being developed. The goal of the implementation is to gave a fully decentralized process for collecting and managing that data.

It also exposes the APIs on top of which various applications can be built, the so called “virtual sensors”. The Nervousnet backend is implemented in two mobile platforms, Android and iOS. It engineers a novel feed loop:

- Nervousnet stores and manages sensor data. Local data is locally stored, within memory constraints and all data are stored on the server, if the privacy settings defined by the users, allow this.
- New data can be generated by performing data analytics on the sensor data, both on the local and global analytics engine.
- New data can be stored and managed by the Nervousnet backend, if allowed by the user’s privacy settings.

2.1.1 STORAGE ENGINE

The storage engine is responsible for the storage and management of sensor data locally in the phone. Data are managed in the following temporal form: time, sensor type, value(s).
Data are structured in the JSON format and sent directly to servers, without being serialized.
Two storage implementations are provided for benchmark measurements and comparisons:

- A red black tree implementation for the Android and
- SQLite DB implementation for iOS. The storage engine exposes an API with which virtual sensors can stream data back to the Nervousnet backend for a unified seamless storage and management.

2.1.2 LOCAL ANALYTICS ENGINE

The local analytics engine performs lightweight statistics and machine learning operations on the data of the storage engine. It is the core component that bridges the Nervousnet backend with the virtual sensors. The local analytics engines expose an extensible and evolving API for accessing the supported operations. New novel sensors can be entirely built by composing a pipeline of API calls to the local analytics engine. Moreover, if this pipeline is generic and relevant for other virtual sensors, it can be integrated in the local analytics engine as part of the Nervousnet backend.

2.1.3 PROXY

The Nervousnet proxy is a locally deployable autonomous software for remote data collection. The proxy opens up new possibilities for an autonomous and participatory social sensing without relying on a centralized Big Data system. You can configure one or more proxies for one or more groups of users for full control of the data collection process.

The Nervousnet backend is capable of pushing sensor data to the proxy in an efficient and smart way as data transfers occur in large chunks when there is an available wireless connection. Currently this component is implemented by a single server, future work will develop this as a peer-to-peer servers interaction.

2.1.4 PRIVACY REGULATOR

The privacy regulator in provides control of the data collection process in respect to privacy. The system provides three levels of privacy control:

- self-determination of the sensor data logged in the phone,
- self-determination of the sensor data shared with the proxy and
- self-determination of the data collection frequency for each involved sensor.

2.2 VIRTUAL SENSORS

Virtual sensors [1] sense an environment with one or more data streams and outputs a new data stream. A virtual sensor consists of an AGGREGATOR that processes the input data stream and a FILTER that regulates the availability of the output data stream. Most virtual sensors are expected to run as independent applications supported by the Nervousnet backend, though generic virtual sensors can be integrated in the Nervousnet backend, e.g. the real-time visualizer.

3 THE VIRTUAL SENSR MODEL

Figure 1 illustrates the model of virtual sensors. This model can be realized as a generic programming interface, with which open participatory platforms for privacy-preserving ubiquitous social mining can be software engineered.

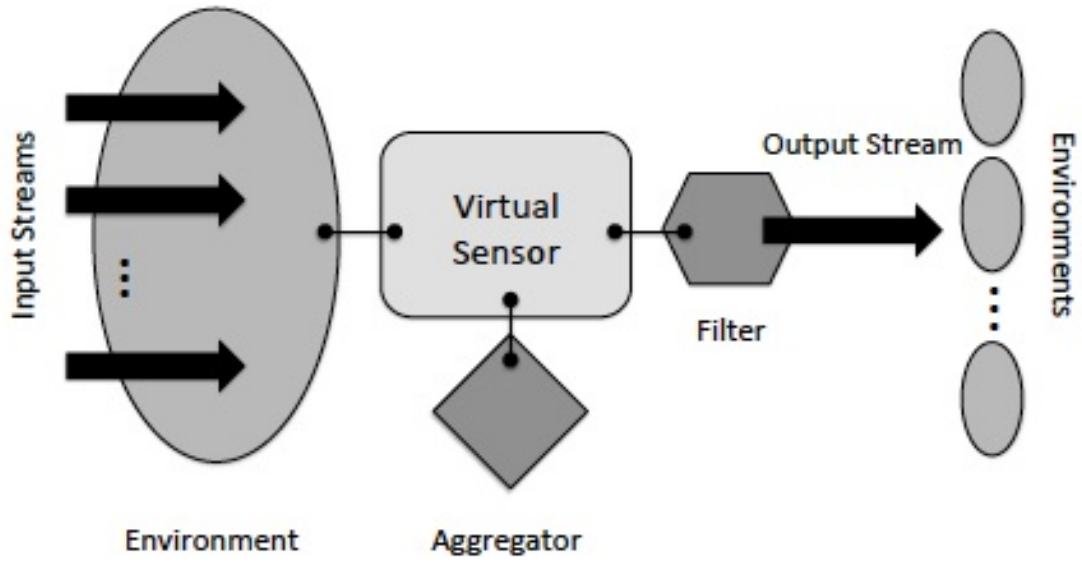


Figure 1 Virtual sensor model

The core component of this model is the virtual sensor. A virtual sensor is defined by its environment, an aggregator, a filter and its output stream. The environment of a sensor is a set of input streams of data generated from physical or virtual sensors. The environment defines the context within which the virtual sensor operates to generate its output stream. The aggregator processes the input streams from the environment of the sensor in real-time and transforms them to the output stream of this sensor. An aggregator can be part of different types of sensors. For example, an aggregator may perform summation of input streams with numerical values, with each input stream having a given weight. The values of the weights may vary depending on the type of the sensor in which the aggregator is applied. Similarly, different aggregators may operate within the same type of sensor, for instance, a sensor that computes the error of the input streams in its environment can be realized with aggregators that compute the absolute, relative or root mean square error.

The output stream of a virtual sensor is a type of real-time data signal generated by the aggregator of the virtual sensor. An output stream can be part of one or more other environments.

The filter controls the availability of the output stream to all other environments in real-time. In other words, the filter introduces privacy-by-design within the virtual sensor model. A filter can be realized by a scheduling algorithm or even by a user interface through which users have full control on which sensor information they make available. The information flow of this model is designed to be recursive: It starts from an environment sensed by a virtual sensor. The output stream of this virtual sensor can form new enhanced environments for further sensing. This recursive design in the information flow of the model

enables a highly modular, compositional and extensible environment for building data driven ubiquitous platforms for social mining.

Figure 2 illustrates the design of the Nervousnet system according to the model of virtual sensors. The first observation is that all software components of the system are elements of the virtual sensor model. Data are collected from different ubiquitous environments with both physical and virtual sensors. The current implementation focuses on mobile phones such as Android and iOS systems, however, an extension to the physical sensors of the Arduino platform is ongoing work.

Smart phones provide access to various physical sensors, such as accelerometer, humidity, battery, temperature, etc. An aggregator of a virtual sensor can control the frequency of data sampling in the output stream of the virtual sensor. Moreover, the output streams generated by these virtual sensors on a user's device come with a self-determining privacy control in two levels. The first privacy level provides user control for storing the data of the output streams in the phone. Users can select to log or not data from each sensor, but they can also schedule logging at certain time periods. This privacy control functionality is implemented in the filter of the Android and iOS virtual sensors. However, a user may desire different privacy control for storing data locally and sharing data with other users. This specialized privacy functionality can be engineered as a virtual sensor, the sharing sensor as shown in Figure 1, whose aggregator controls the streams shared in the network environment. This is the second privacy level introduced in the Planetary Nervous System. Both privacy levels are designed via the same model of virtual sensors. Figure 2 illustrates an example of a user interface that implements the two privacy levels of control.

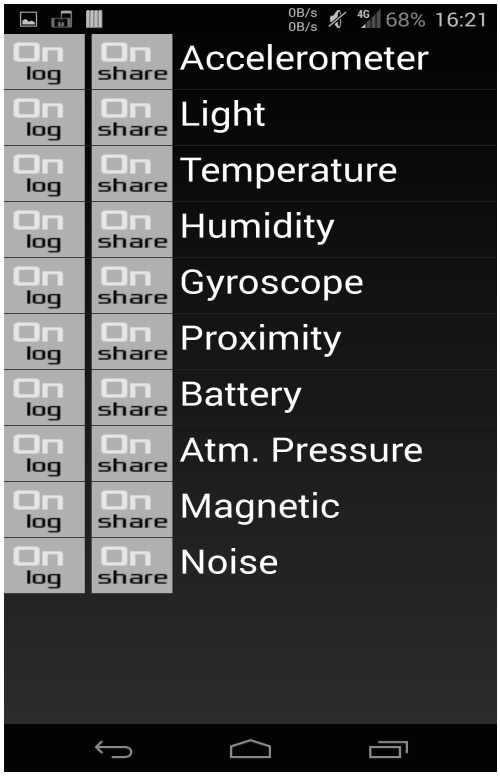


Figure 2 An implementation of a user interface with the two privacy control levels in the Nervousnet.

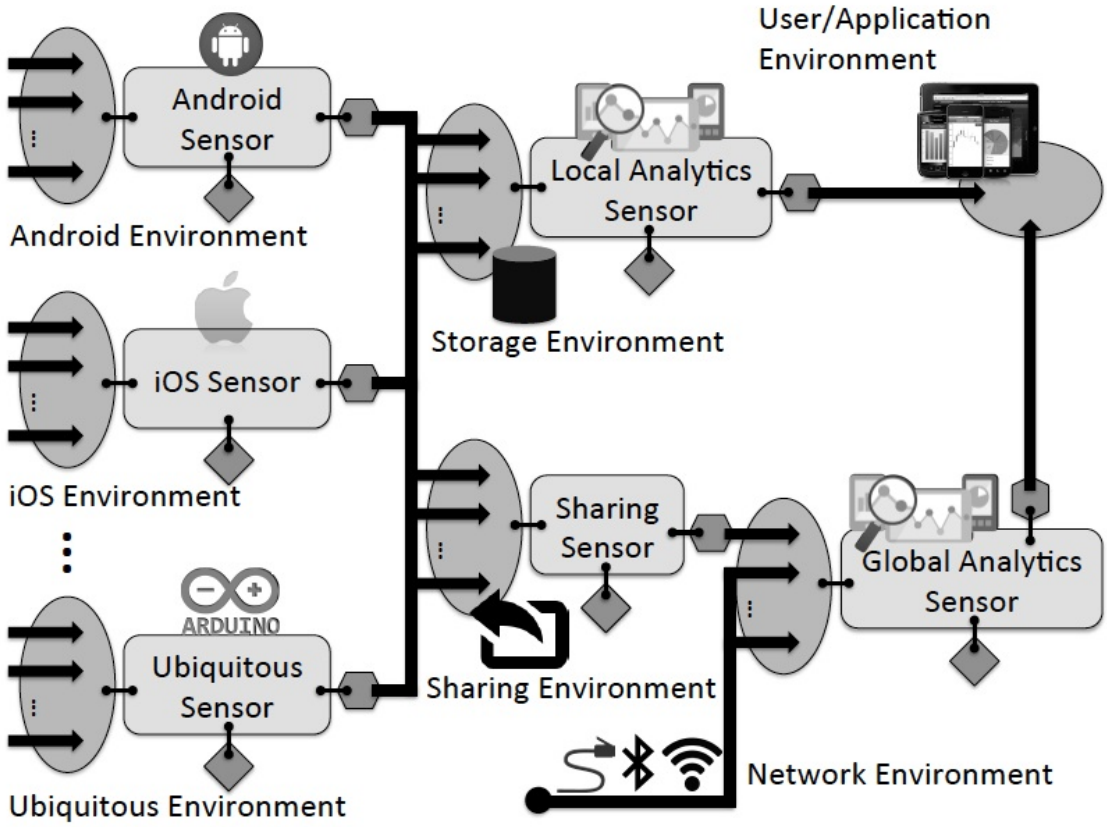


Figure 3 The Nervousnet system designed according to the model of virtual sensors.

Data sensed from sensors are stored in the local storage environment shown in Figure 3. The storage environment implements an efficient method for serializing structured data with the Protocol Buffer library6 . Data are stored locally stored by relying on the SQLite storage engine, and they are indexed in Red-Black trees for fast retrievals based on range queries that define a period of time. The data stored in phones act as a data pool over which lightweight data analytics are performed. Such analytics are implemented in the aggregator of a local analytics sensor and include aggregation functions such as summation, average, maximum, minimum, standard deviation, but also data mining algorithms such as clustering. The aggregator interface of the local analytics sensor defines a toolkit for real-time operations performed in time series data, with which application developers can further build other virtual sensors.

Local analytics are performed over the data of a sensor type for a defined period of time. The purpose of the local analytics sensor is twofold:

- It provides data for an engaging, interactive and gamifying visual experience to users in order to understand and explore in real-time their own social environment and activity.
- (ii) It provides intuition for users and developers to build their own applications with virtual sensors. While local analytics provide interesting information about single users, collective information about the status of the participatory community cannot be captured in real-time via the local analytics sensor only.

System-wide analytics is the objective of a global analytics sensor, currently work-in-progress in the Nervousnet system. A global analytics sensor is ambitious and challenging to realize as computations should be performed in a decentralized manner. Distributed privacy-preserving aggregation services, such as DIAS, the Dynamic Intelligent Aggregation Service [2] and OpenPDS [3], can realize a global analytics sensor.

4 INTEGRATION OF OPEN DATA THROUGH NERVOUSNET

Innovation Accelerator platforms (described in WP5) enable the acquisition of open data sets which will be integrated in the SoBigData infrastructure through servers provided by the Nervousnet system. Open data include scientific data, sensory data, social data, proximity data, spatial-temporal data, environmental data. In the following, we describe the integration of open data through the Nervousnet platform.

The Integration server is hosted at ETH Zurich and the code is made available at: https://github.com/nervousnet/nervousnet-proxy
The integration of the Nervousnet server functionality in the SoBigData infrastructure consists of the following steps:

1. Install and setup the mysql server:
a. Download mysql http://dev.mysql.com/downloads/mysql/ for your system
b. Install mysql and run it
c. Create an account for a user "user" having password **"**password"
d. Create a database db in which you wish your data to be stored.

2. Configuring and running the proxy on the SoBigData infrastructure:
a. Download the entire https://github.com/nervousnet/nervousnet-proxy/tree/master/proxy/proxy/build folder.
b. Modify the config.xml file, by setting the attributed sqlUsername**,** sqlPassword**,** sqlDatabase to user**,** password and db**.**
c. Start the proxy from the console by running BASH PROXY-START.SH

Nervousnet 1.0 backend relies on the following libraries:

- Protobuf
- Protobuf Swift
- JQuery
- Highcharts
- SQLiteDB

For the client side (mobile), it is only required to add the SoBigData proxy server configuration (IP and port number). Then the SoBigData server will collect all the open data from Nervousnet mobile phones, which have configured the SoBigData server as their proxy. And all the Nervousnet functionalities (described in WP5) can be used in the SoBigData platform.

The integration through the Nervousnet will enable in (WP5) the following:

- acquisition of sensory data, social data, proximity, spatial-temporal data, environmental data
- querying the sobigdata servers (in Nervousnet) with filters (privacy preserving)
- monitoring and visualizing the open data
- search functionality provided through the data analytics engine

5 FUTURE WORK

As described in the previous section, the current implementation of the integration is a “light” one, as the Nervousnet server remains hosted at ETH Zurich and is accessible via web services. As future directions, we aim at addressing the shortcomings currently existing in Nervousnet: decentralized version of the proxy server, with peer-to-peer communication; serializing mechanism of data before it is sent to the server; achieving a deeper integration of the Nervousnet in the SoBigData platform.

REFERENCES

[1] E. Pournaras, I. Moise and D. Helbing, "Privacy-Preserving Ubiquitous Social Mining via Modular and Compositional Virtual Sensors," 2015 IEEE 29th International Conference on Advanced Information Networking and Applications, Gwangiu, 2015, pp. 332-338.

[2] E. Pournaras, M. Warnier, and F. M. Brazier, “A generic and adaptive aggregation service for large-scale decentralized networks”, Complex Adaptive Systems Modeling, vol. 1, no. 19, 2013.

[3] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, “openPDS: Protecting the privacy of metadata through safeanswers”, PloS one, vol. 9, no. 7, p. e98790, 2014.