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	***SoBigData e- Infrastructure release plan 2***
Deliverable No.	***D10.3***
Delivery Date	***30 April 2017***
Authors	***Leonardo Candela (CNR), Paolo Manghi (CNR), Pasquale Pagano (CNR)***

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.	D10.1
Deliverable Title	SoBigData e- Infrastructure release plan 2
Contractual Delivery Date	30 April 2017
Actual Delivery Date	14 November 2017
Author(s)	Leonardo Candela (CNR), Paolo Manghi (CNR), Pasquale Pagano (CNR)
Editor(s)	Paolo Manghi (CNR)
Reviewer(s)	Valerio Grossi (CNR)
Contributor(s)	
Work Package No.	WP10
Work Package Title	JRA3_SoBigData e-Infrastructure
Work Package Leader	CNR
Work Package Participants	USFD, UNIPI, FRH, UT, IMT LUCCA, LUH, KCL, SNS, AALTO, ETHZ
Dissemination	Public
Nature	Report
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	30
Keywords	Release planning; D4Science; SoBigData e-Infrastructure;

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
LtR	Learning to Rank
NLP	Natural Language Processing
Research Infrastructure	Facilities, resources and services that are used by a research community to conduct research and foster innovation in their fields. Include: major scientific equipment (or sets of instruments), knowledge-based resources such as collections, archives and scientific data, e-infrastructures, such as data and computing systems and communication networks and any other tools that are essential to achieve excellence in research and innovation. They may be 'single-sited', 'virtual' and 'distributed'.
RI	Research Infrastructure
SNA	Social Network Analysis
VA	Virtual Access
Virtual Access	**Open and free access** through communication networks to resources needed for research, without selecting the researchers to whom access is provided.
Virtual Research Environment	Innovative, web-based, community-oriented, comprehensive, flexible, and secure working environments conceived to serve the needs of modern science.
VRE	Virtual Research Environment
WP	Work Package

TABLE OF CONTENT

TABLE OF FIGURES

DELIVERABLE SUMMARY

This deliverable refines the plan as described in deliverable D10.2, characterizing the release and development of the SoBigData e-Infrastructure in the second year. This is the second of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable describes the current status of the e-infrastructure and focuses on the plan leading to the second release of the SoBigData e-Infrastructure at M24.

EXECUTIVE SUMMARY

SoBigData WP10 is called to support the development of the SoBigData e-Infrastructure in close collaboration with other work packages that are respectively called (a) to operate the infrastructure to provide virtual access to the integrated resources (WP7), (b) integrate existing and newly collected datasets in the infrastructure (WP8), and (c) integrate existing tools and methods for mining social data in the infrastructure (WP9). In particular, WP10 puts in place actions comprising: (i) studies and definition of best practices/policies for the harmonization of federated resources available at the local infrastructure sites; (ii) support for adaptation of existing resources to the identified best practices; and (iii) realization of VREs supporting scientists in benefitting from the integration of the federated resources and infrastructures.

This deliverable describes the development plan characterising the release and development of the SoBigData e-Infrastructure. This is the second of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable focuses on the plan leading to the second release of the SoBigData e-Infrastructure at M24.

1 INTRODUCTION

SoBigData is a research infrastructure (RI) for ethic-sensitive scientific discoveries and advanced applications of social data mining to the various dimensions of social life, as recorded by "big data". It is planned to serve the wide cross-disciplinary community of data scientists involved in social mining, i.e., researchers studying all aspects of societal complexity from a data- and model-driven perspective, including data and text miners, visual analytics researchers, socio-economic scientists, network scientists, political scientists, humanities researchers, and more.

In order to serve its "designated community"[1], the project is setting up an e-Infrastructure providing "virtual access" to the resources of interest, namely datasets and methods for social mining. In particular, SoBigData implements an e-Infrastructure that is "open" and "aggregative" by design, i.e. it is conceived to aggregate into a unifying resource space resources coming from many and heterogeneous providers. In order to do this, it will rely on the D4Science e-Infrastructure [1]. D4Science offers a common ground for hosting the domain specific resources and dynamically building and operating *Virtual Research Environments* (VREs) offering specific and web-based working environments to target communities [2]. In order to be able to serve the needs of the social mining community, it is of paramount importance to invest effort in adapting and extending the resources currently owned by the SoBigData community thus to make them benefitting from the e-Infrastructure capacity. Adaptation and extension of existing resources goes in the direction to integrate them into a unifying space. Depending from the "level of integration" that is achieved for each resource it will be possible to support and guarantee a diverse level of management ranging from a simple discovery to their repurposing to better serve the needs arising in a specific VRE.

This deliverable describes the SoBigData e-Infrastructure current status of operation and presents the developments for its second year of operation.

[1] This term is borrowed from the archival community here it is used to refer to an identified group of potential consumers that should be able to understand the particular set of information and benefit from them. The group may consist of multiple communities and may change over time.

2 D4SCIENCE: SYSTEM ARCHITECTURE AND OPERATION

The SoBigData e-Infrastructure is powered by an existing e-Infrastructure – D4Science – that plays the role of "integration infrastructure" collecting resources and services from the infrastructures and resource providers participating in SoBigdata and enabling, where possible, their interoperation. Figure 1 gives a conceptual view of the SoBigData e-Infrastructure.

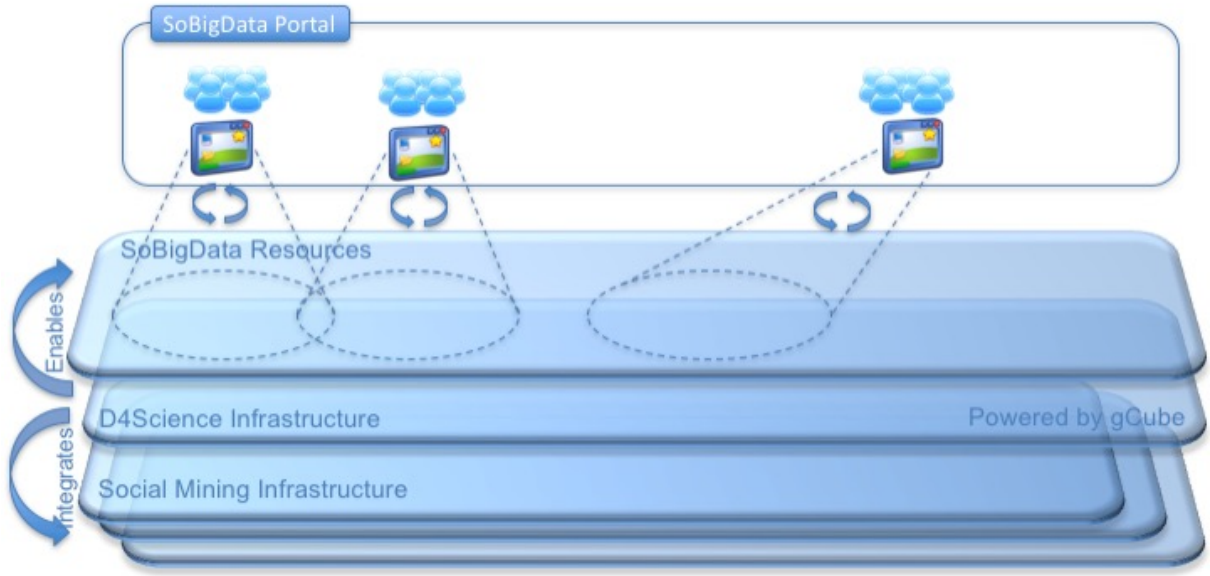

Figure 1. SoBigData Infrastructure

D4Science[2] [1] is an operational Hybrid Data Infrastructure that (a) supports the creation and management of Virtual Research Environments by dynamically acquiring the resources (data, tools, services, computing) from a resource space, (b) is designed to integrate resources from other e-Infrastructures and (commercial) vendors/providers by using a "system of systems" strategy, (c) is offering unified access to the integrated resources by abstracting from the underlying e-Infrastructures, (d) is designed to maximize resources exploitation and minimize operational costs, and (e) relies on the gCube open source[3] technology that offers a feature rich set of services for data management and collaborative work. Over time D4Science supported a set of projects and initiatives promoting the development of virtual research environments in various application domains [1] including iMarine (fisheries management and marine living resources conservation), EUBrazilOpenBio (biodiversity), ENVRI (environmental science), and OpenAIRE-Connect (scholarly communication).

The D4Science platform proves to be, by practice, extremely apt for those scenarios where sub-systems (e.g. local infrastructures) characterised by highly heterogeneous settings (e.g. services, programming platforms and frameworks, formats, protocols) are attempting to integrate their resources in a uniform environment, where end-users can hope to access all such resources in a coherent manner. This is due to the following main features of D4Science: (i) the ability to integrate methods at different degrees of effort, (ii) the ability to integrate web applications, (iii) the ability to integrate/add access rights to any resource (e.g. datasets, methods, applications), and (iv) the ability to dynamically group methods and applications into Virtual Research Environments. VREs are web applications where users can, under given access rights, use the methods and applications included in the VRE to process or generate research data. VREs can

[2] https://www.d4science.org/
[3] https://www.gcube-system.org

therefore be created to address specific needs of groups of users, for a given period of time, by providing them with the methods and applications they need as well as social tools (Facebook-like posting and messaging) and a shared file system to facilitate collaboration.

3 THE SOBIGDATA E-INFRASTRUCTURE

Because of the highly aggregative nature of the SoBigdata e-Infrastructure where key resources (social mining data and methods) are coming from four national infrastructures, the D4Science platform proved to be the most reasonable choice. Alternatives would have been service-oriented frameworks whose constraints would have raised expectations and required coding efforts that are too high for scientific communities to succeed in integrating their methods and applications. D4Science allows incremental and evolving deployment of the e-infrastructure following agile principles in the integration of the methods and the applications. Thanks to this flexibility the project could deliver the first release of the SoBigData e-infrastructure at M12 (August 2016), start basic integration immediately and elaborate deeper level of integration on project's course.

3.1 SOBIGDATA E-INFRASTRUCTURE FIRST RELEASE

The first version of the SoBigData e-Infrastructure comprises the following basic components, whose dependencies are illustrated in Figure 2:

- The ***SoBigData Virtual Organisation***, i.e. an organizational structure and a set of basic services created and operated in the context of D4Science to serve the needs of SoBigData. This Virtual Organisation realizes the actual operational context for realizing and operating the SoBigData e-Infrastructure and its resources in autonomy with respect to the other communities and initiatives supported by D4Science;
- The ***SoBigData Resource Catalogue***, i.e. a core service where all the resources contributing to form the SoBigData e-Infrastructure are expected to be registered thus to make it possible for clients to discover them and be informed on their characteristics for, e.g. properly using them. This catalogue is expected to serve both (a) human users willing to know the offering of the e-Infrastructure in terms of datasets and services / methods and (b) other services willing to dynamically discover resources to consume / interact with to deliver their services;
- The ***SoBigData Gateway***, i.e. a Liferay[4] based web-portal customized to interface with the D4Science infrastructure and equipped with gCube portlets. This portal will act as the "one stop shop" for the entire SoBigData e-Infrastructure. Through it users will have access to the resources and Virtual Research Environments created to serve the needs of the SoBigData community and scenarios;

The SoBigData e-infrastructure lives as a Virtual Organization over the D4Science platform, which is an operation system enabling the construction of multiple e-infrastructures. Beyond SoBigData, D4Science is today supporting the following e-infrastructures: i-Marine, BlueBRIDGE[5], and PARTHENOS. D4Science is operated by means of a set of "enabling services" which can offer the abstraction of an e-infrastructure (Virtual Organizations), of its available resources (catalogue), of its VREs (set of resources ruled by common policies), and of its users and usability (Authorization, Authentication, and Accounting service). E-infrastructures can be deployed over D4Science as one VO with a corresponding set of registered resources, whose information model can be custom to the e-infrastructure (e.g. datasets and methods for SoBigData). A powerful, flexible and configurable access policy framework based on the notion of Virtual Research Environment, allows resource owners to decide which users can access and use their resources in a given VRE scope; i.e.

[4] http://www.liferay.com/
[5] BlueBRIDGE project portal, http://www.bluebridge-vres.eu/

VREs are a flexible way to package a set of resources and make them available to users authorized to access the VRE.

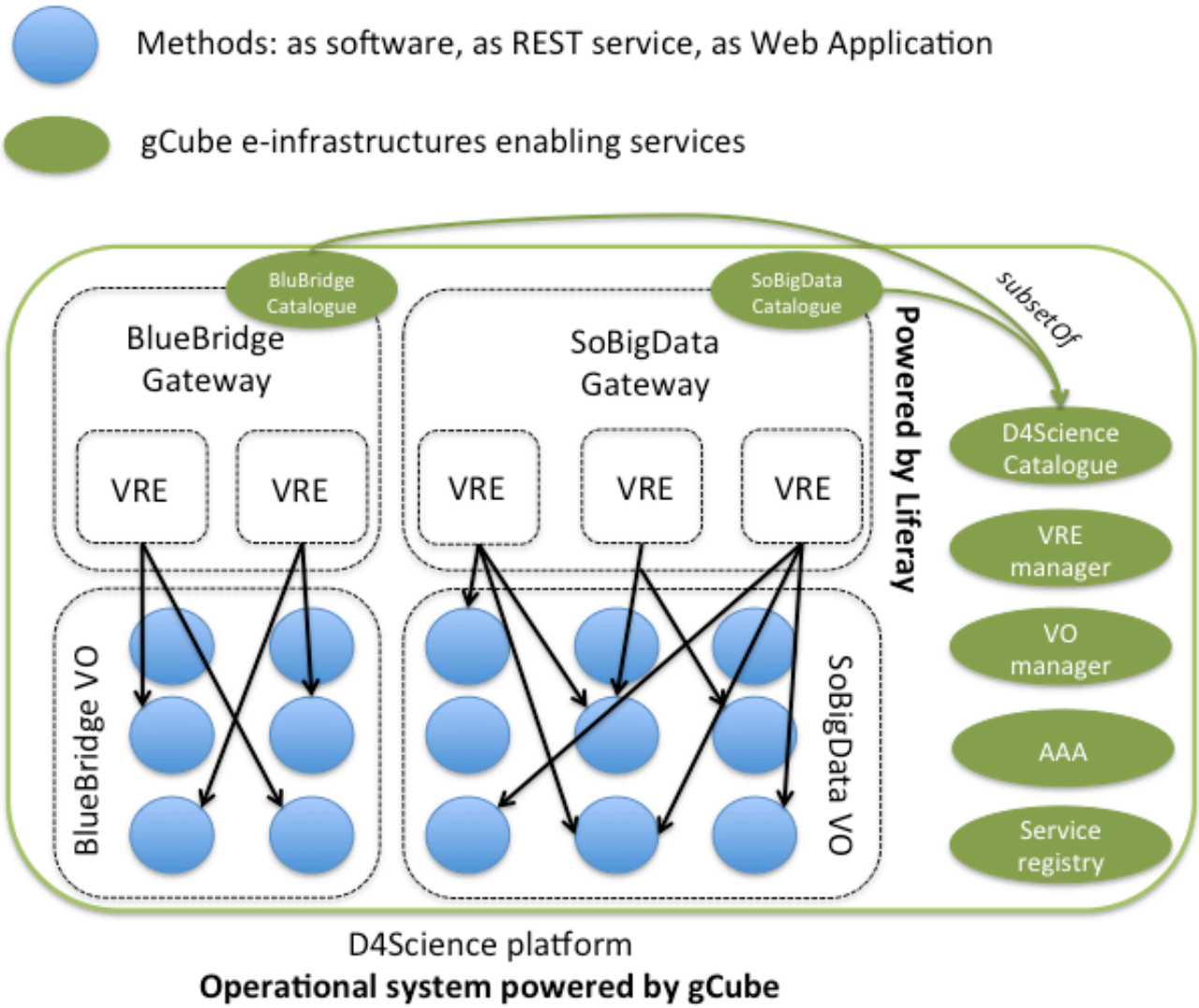

Figure 2. D4Science and e-infrastructures: how the SoBigData e-infra coexists with other e-infras (e.g. BlueBRIDGE) over the same D4Science platform

3.2 SOBIGDATA RESOURCES INTEGRATED SO FAR

SoBigData e-Infrastructure is a RI whose primary objective is to support social mining practitioners by providing them with seamless access to datasets and methods of interest. Thus it is of paramount importance to integrate these domain specific resources in a unifying resource space. The integration of datasets and methods in infrastructure release actually depend on a series of factors including:

- The "level of integration" expected, i.e. SoBigData has developed a series of guidelines and best practices to make existing resources interoperable in the context of the e-Infrastructure [3]. Depending on the degree of guidelines and best practices implemented the resource will result integrated and interoperable with the rest to some extent. In the course of the project, the level of integration will be properly tuned depending on the needs and expectations raised by the community;

- Development teams' velocity differences, i.e. the integration of existing resources is expected to be driven by resource owners and service providers. These actors are in the best position to reconsider their products, to identify the major limitations with respect to the guidelines for interoperability, and to plan concrete steps and timelines leading to enhanced versions of their resources;
- A negotiation between users' expectations/desiderata and effort needed to satisfy them, i.e. the availability of a resource within the e-Infrastructure range from the simple discovery to the actual reuse. Diverse users might have different expectations with respect to the actual reuse of a resource and reconsidering the implementation of a given resource thus to serve the expected reuse scenario has a cost for the resource provider;
- Dependencies among resources in a user scenario, i.e. when combining a series of existing resources in a certain exploitation scenario the workflows assume that the resources are integrated to a well expected level/degree. Thus the willingness to implement a certain exploitation scenario poses specific requirements on diverse resources and on how these resources have to be integrated.

These factors and the willingness to provide users with a working e-Infrastructure as quickly as possible call for a planning activity that is almost continue and where it is possible to correct the course as the project progresses. Because of this, SoBigData developers and practitioners rely on the open source software RedMine[6]. RedMine is an "issue tracking system", i.e. a service for managing and maintaining lists of "issues" or "tickets". Each ticket focuses on a specific activity/problem and describes the current state of the art and plan including a due date. Tickets have an issuer and a responsible person who must keep it up to date to reflect its status of completion. They are used to exchange comments and favour the resolution among the issuers, the responsible, and a set of watchers who are interested and may help in the process. RedMine is integrated in the SoBigData gateway as an application shared among all users with roles of responsibility in the project: task leaders, WP leaders, coordinators, project managers, developers; the application is available at

https://sobigdata.d4science.org/group/sobigdata.eu/activity-tracker

By using this environment the teams called to develop and deliver the subsequent versions of the SoBigData e-Infrastructure are able to plan, track, and monitor:

- Software development tasks (actually complex workflows consisting of multiple and interrelated tasks) needed to gradually adapt each target resource to the operational environment by implementing the project guidelines and best practices [3];
- Software deployment tasks needed to plan the actual action of bringing a certain version of the resource into effective action in the context of the SoBigData e-Infrastructure.

By using this knowledge base every project member is informed of the state of the art and the planned activities with their detailed time plan and can intervene to identify potential issues and propose corrective actions (e.g. reconsider a deadline, reconsider an implementation decision) to be agreed with the proper team.

In the following sections we described the resources that have been so far integrated in the SoBigData Infrastructure, classifying them in applications, methods, and datasets.

[6] http://www.redmine.org/

3.2.1 APPLICATIONS

Scientists can integrate Web Applications, intended as web accessible applications, within a given VRE. The effort required is that of:

1. Writing code to integrate the web graphical interface in a VRE by means of Liferay portlets;
2. Register the application resource to the SoBigData catalogue.

Figure 3 illustrates how provider and consumer of the resource can share an application via SoBigData VREs.

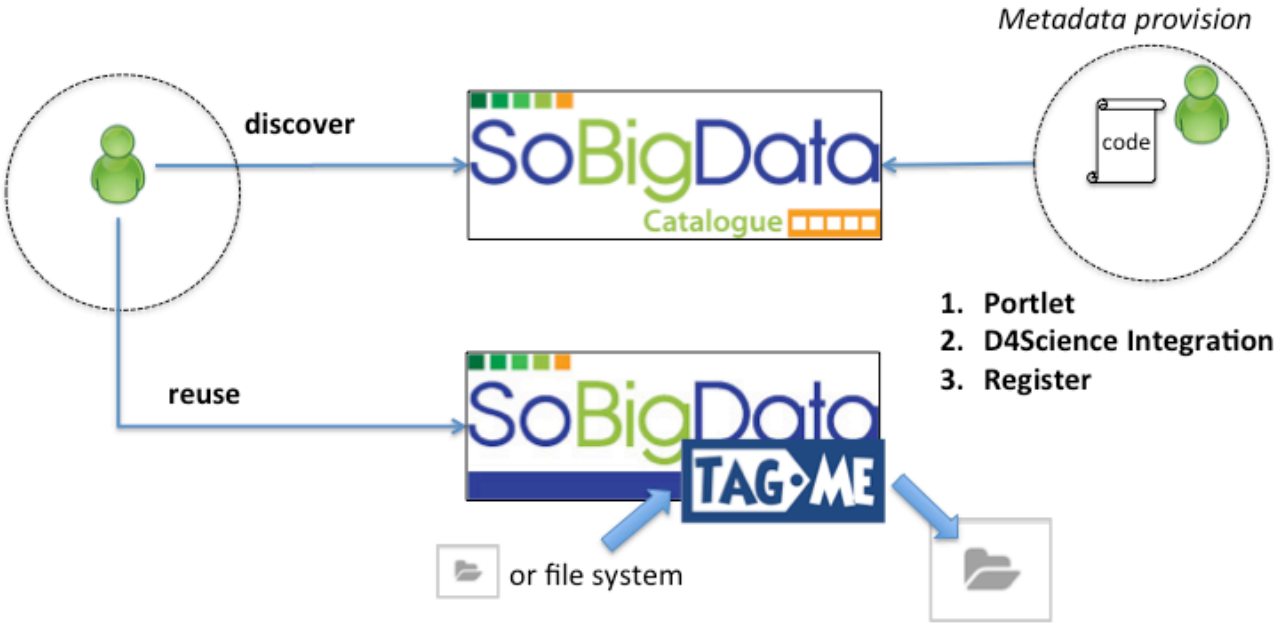

Figure 3. Registration and re-use of applications

In the following we describe the list of applications currently integrated in the e-infrastructure.

3.2.1.1 GATECLOUD

GATECloud[7] is a unique, cloud-based infrastructure for large-scale, data-intensive NLP and text mining research. Important infrastructural issues are dealt with by the platform, completely transparently for the researcher: NLP algorithm distribution, load balancing, efficient data upload and storage, deployment on the virtual machines, security and fault tolerance. Another unique feature of this specialized NLP Platform-as-a-Service is its support for researchers who want to develop and run their own text mining/NLP pipelines on big data.

It offers a growing number of NLP and text mining services for multiple European languages. Currently only around 30% of the algorithms and tools from the GATE infrastructure have been made available through GATECloud, and the number will continue to grow during the project lifetime. GATECloud offers a web interface for data upload, social media data collection, programming-less execution of the GATECloud text analytics services, and results download. This new improved version of the GATE Cloud platform is now available via the SoBigData Lab VRE.

SoBigData catalogue reference: http://data.d4science.org/ctlg/ResourceCatalogue/gate_cloud

[7] https://gatecloud.net/

3.2.1.2 TAGME

TAGME [6] identifies meaningful sequences of terms in an unstructured text and links them to the Wikipedia page describing the mentioned entity, in a fast and effective way. This annotation process has implications which go far beyond the enrichment of the text with explanatory links because it concerns with the contextualization and, in some way, the understanding of the text.

SoBigData catalogue reference: http://data.d4science.org/ctlg/ResourceCatalogue/tagme

Figure 4. TagMe VRE: TagMe UI

3.2.1.3 SMAPH QUERY ENTITY LINKER

The SMAPH system links queries to the entities it mentions, disambiguating mentions if needed. Entities are Wikipedia pages. This problem is known as "entity recognition and disambiguation in queries". For example, the query "armstrong moon landing" should point to Neil Armstrong and Moon Landing, while the query "armstrong trumpet" should point to Louis Armstrong and Trumpet. This system won the Entity Recognition and Disambiguation Challenge 2014 (short-text track).

SoBigData catalogue reference:
http://data.d4science.org/ctlg/ResourceCatalogue/smaph_system_for_query_entity_linking

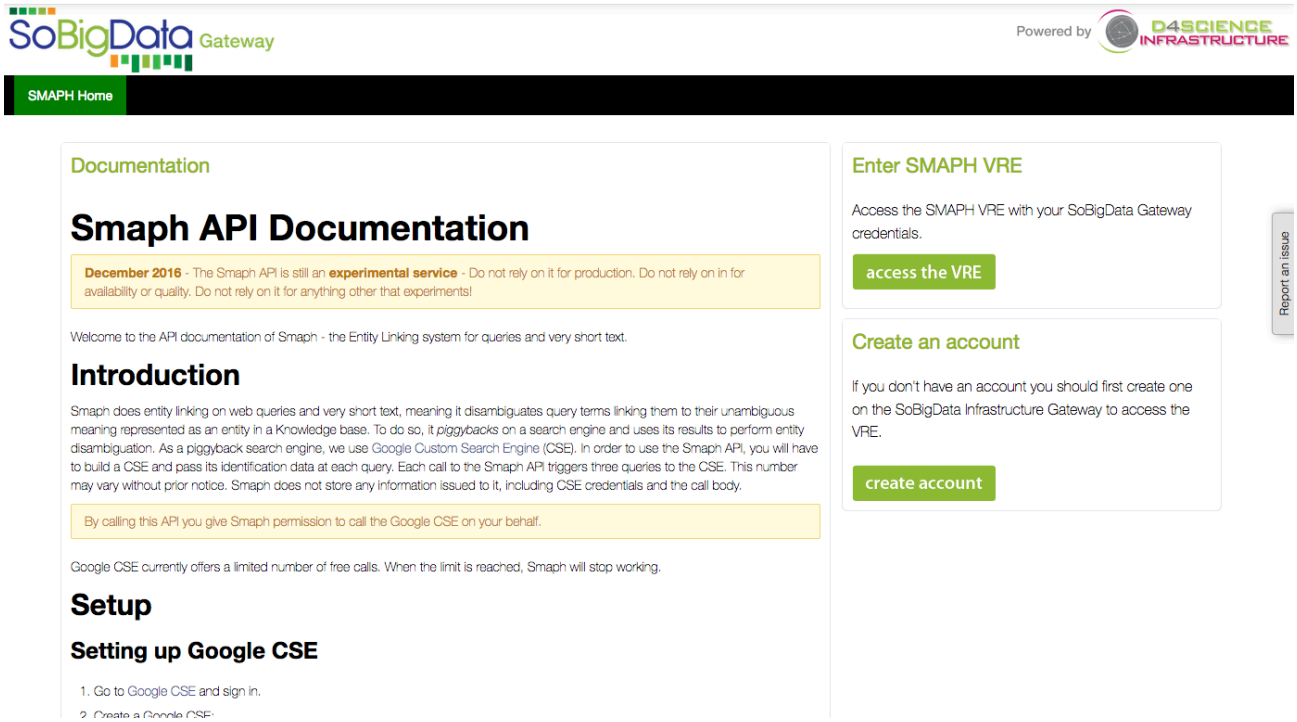

Figure 5. SMAPH VRE

3.2.1.4 TWITTER MONITOR

The Twitter Monitor features an interactive Web application designed to access the Twitter stream by exploiting the public Twitter Streaming APIs. The application 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. In addition to offering an interactive Web interface in order to ease all the operations related to Twitter crawling, the Twitter Monitor also offers a set of functionalities aimed at minimizing the loss of data due to network or local machine problems.

SoBigData catalogue reference:
http://data.d4science.org/ctlg/ResourceCatalogue/twitter_monitor

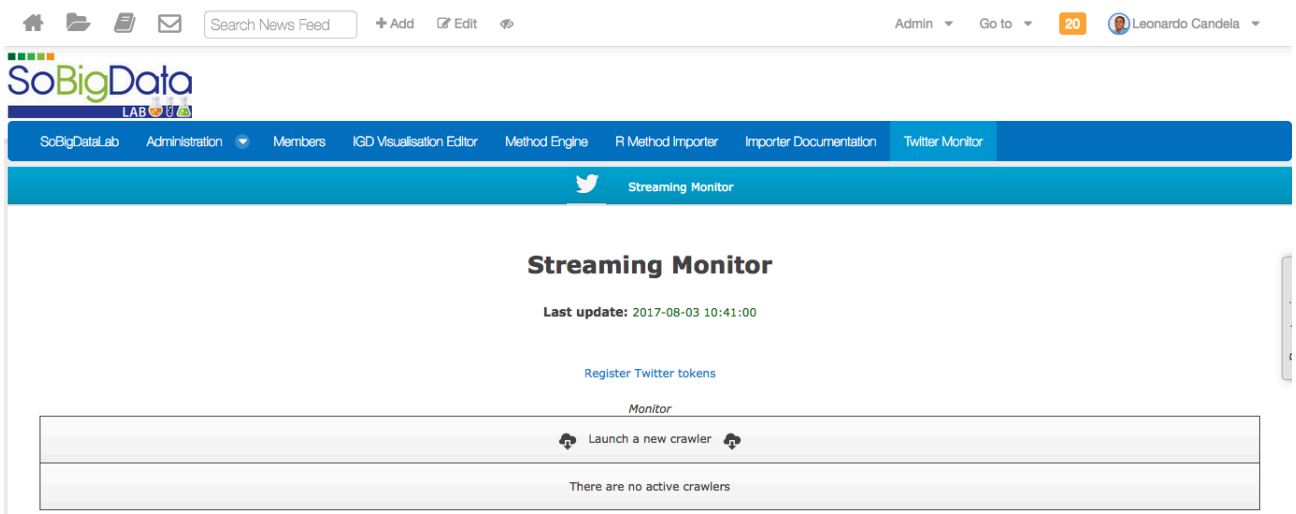

Figure 6. SoBigData Lab VRE: Twitter Monitor

3.2.1.5 IGD VISUALISATION EDITOR

The IGD Visualization Editor is a local infrastructure that offers services for the visualization and especially the visual analysis of research data. It is developed and maintained by the Fraunhofer Group. The services offered by the platform are primarily interactive visualization techniques. As a starting point, the repository contains implementation of generic techniques, yet the platform should be capable to attract both users and contributors alike.

SoBigData catalogue reference:

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

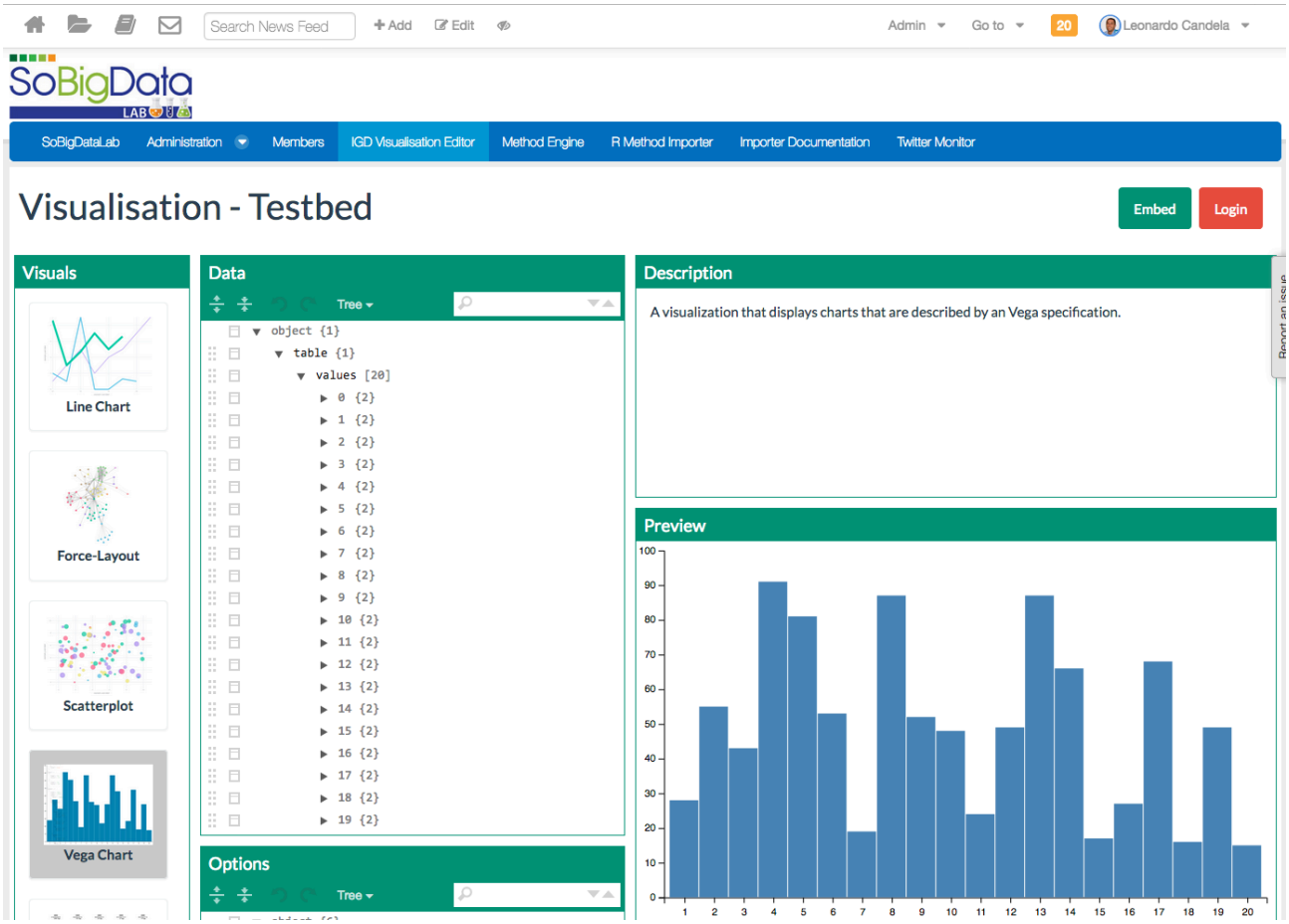

Figure 7. SoBigData Lab VRE: IGD Visualisation Editor

3.2.2 METHODS

Scientists can integrate methods according to three main strategies:

- Methods as software: software of the method is deposited on web accessible repositories, e.g. GitHub, and registered in the e-infrastructure catalogue with metadata, including the link; scientists can discovery the method via the catalogue and download it for installation in their local machines;
- Methods as a services: the method is available as a WPS web service; we have two cases supported:

1. The method is software installed in a D4Science WPS engine (called Data Miner) and made available for execution (possibly benefiting from parallel execution over D4Science clusters);
2. The method is a remote web service, which is invoked by a WPS mediator deployed on the D4Science platform.

In both cases, end-users can search and invoke any methods by means of the same user interfaces (filling or uploading input parameters).

Figure 8 and Figure 9 illustrate how provider and consumer of a method can share it via SoBigData VREs.

Figure 8. Registration and re-use of methods as software

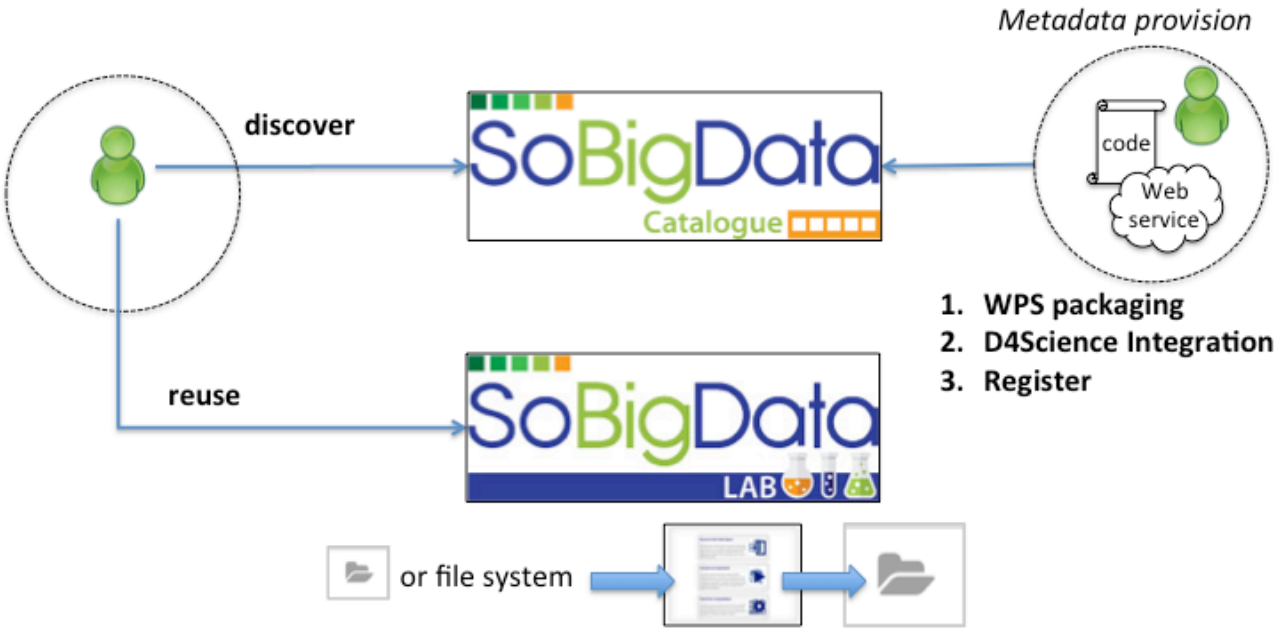

Figure 9. Registration and re-use of methods as web (REST) service or code

In the following we describe the methods currently integrated in the e-infrastructure.

3.2.2.1 QUICK RANK METHODS

Quick Rank[8] [4] is an efficient Learning to Rank (LtR) toolkit providing several C++ implementations of specific algorithms. In particular, the toolkit offers an implementation of the following algorithms:

- *GBRT*: J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232, 2001.
- *LamdaMART*: Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. Information Retrieval, 2010.
- *Oblivious GBRT / LamdaMART*: Inspired to I. Segalovich. Machine learning in search quality at yandex. Invited Talk, SIGIR, 2010.
- *CoordinateAscent*: Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. Information Retrieval 10(3), pages 257–274, 2007.

SoBigData catalogue reference: http://data.d4science.org/ctlg/ResourceCatalogue/quickrank

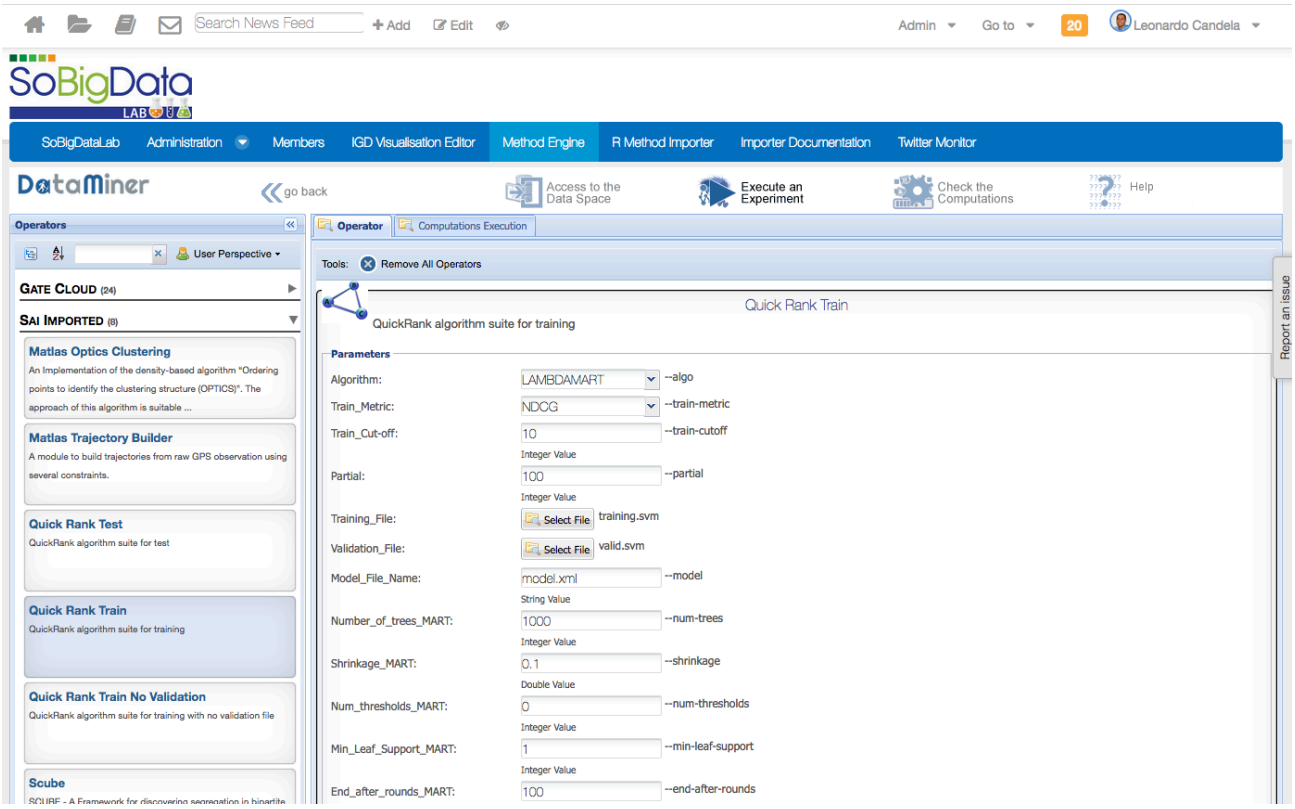

Figure 10. SoBigData Lab VRE: Quick Rank methods by DataMiner

3.2.2.2 GATECLOUD METHODS

Several methods (24) made available as REST web services via the GateCloud infrastructure have been integrated in SoBigData and made available via the SoBigData DataMiner Engine. These are visible at:

[8] http://quickrank.isti.cnr.it/

https://sobigdata.d4science.org/group/sobigdatalab/method-engine

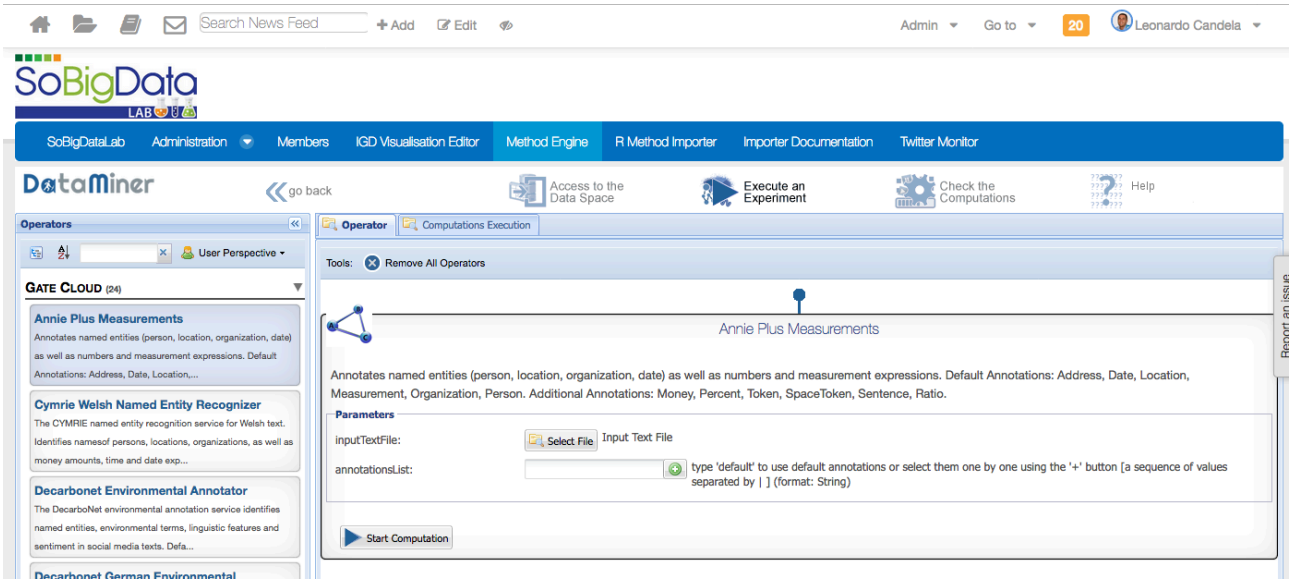

Figure 11. SoBigData Lab VRE: GATECloud methods by DataMiner

3.2.2.3 M-ATLAS METHODS

Two methods from M-ATLAS have been integrated:

Trajectory builder: A module to build trajectories from raw GPS observation using several constraints.

SoBigData catalogue reference:

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

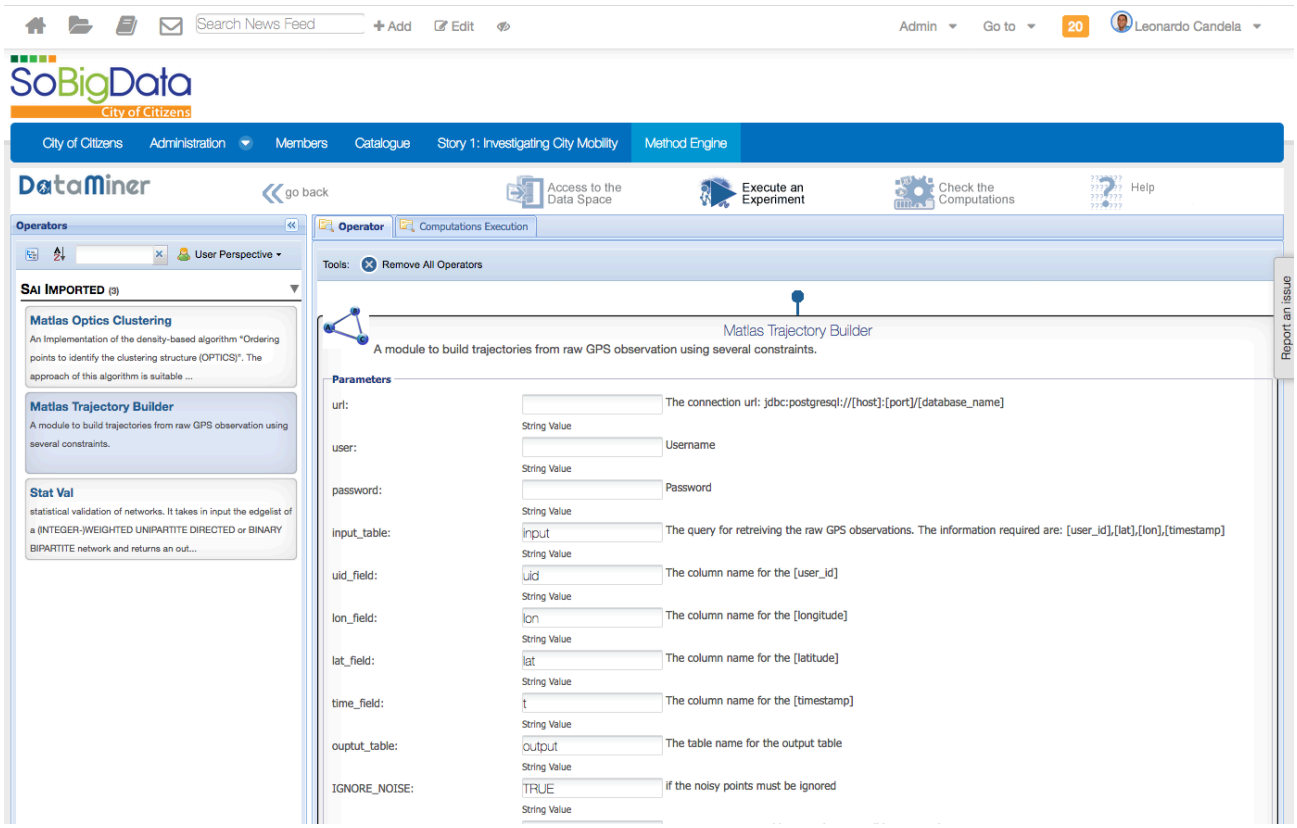

Figure 12. City of Citizens VRE: M-Atlas Trajectory Builder by DataMiner

Optics Clustering: An Implementation of the density-based algorithm "Ordering points to identify the clustering structure (OPTICS)" part of the M-ATLAS package. OPTICS is an algorithm for finding density-based clusters in spatial data. Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. The approach of this algorithm is suitable for complex data such as patio-temporal Trajectories (see Trajectory Builder). The algorithm is equipped with a set of distance functions which are able to adapt to different analytical objectives.

SoBigData catalogue reference:

http://data.d4science.org/ctlg/ResourceCatalogue/matlas_-_optics_algorithm

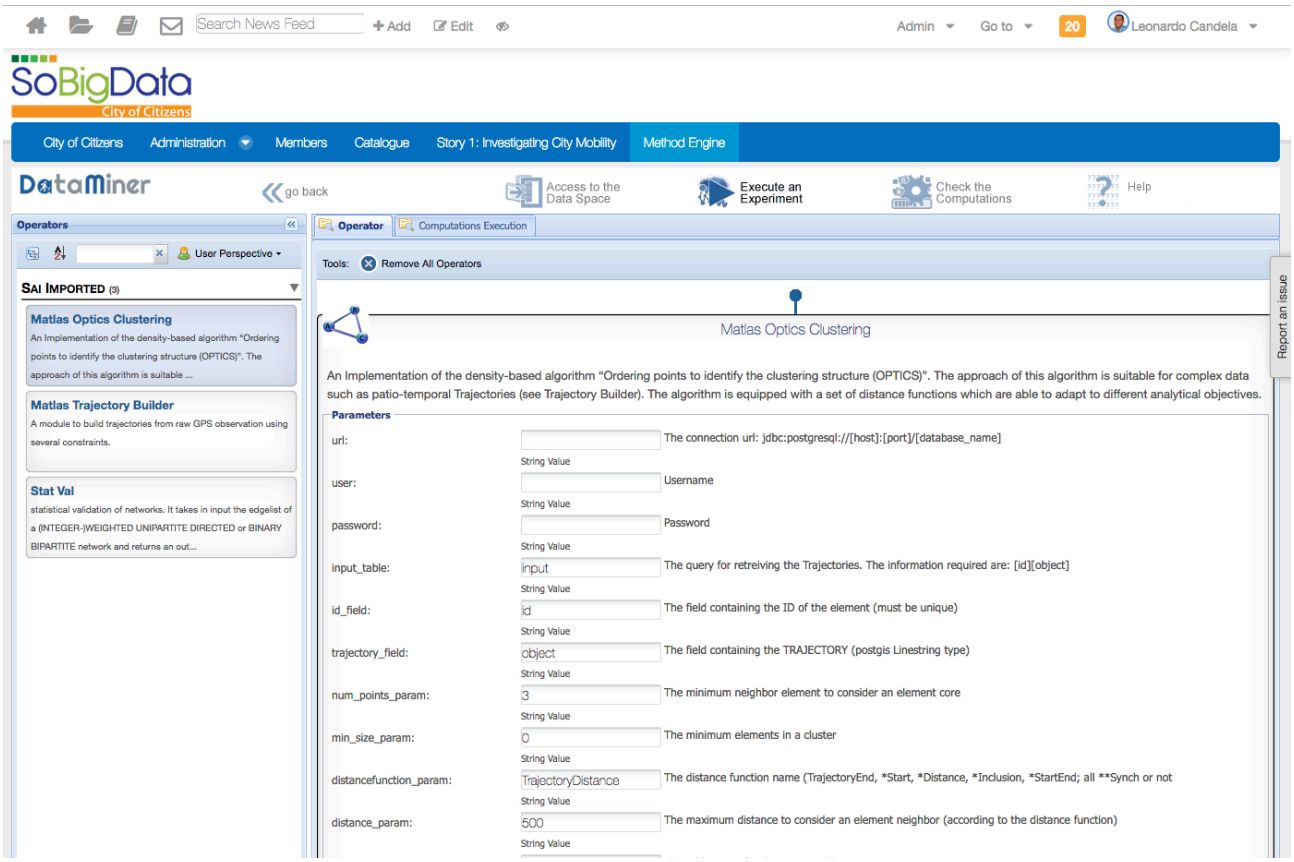

Figure 13. City of Citizens VRE: M-Atlas Optimistic Cluster by DataMiner

3.2.2.4 STATVAL: STATISTICAL VALIDATION OF NETWORKS

StatVal[9] is an approach offering a statistical method for filtering a network to its backbone structure. It takes in input the edgelist of a (INTEGER-)WEIGHTED UNIPARTITE DIRECTED or BINARY BIPARTITE network and returns an output network that is filtered by statistical comparison of the given network with a NULL model which has the same number of nodes and the same degree/strength sequence

SoBigData catalogue reference:

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

[9] http://mathfinance.sns.it/statistical_validation/

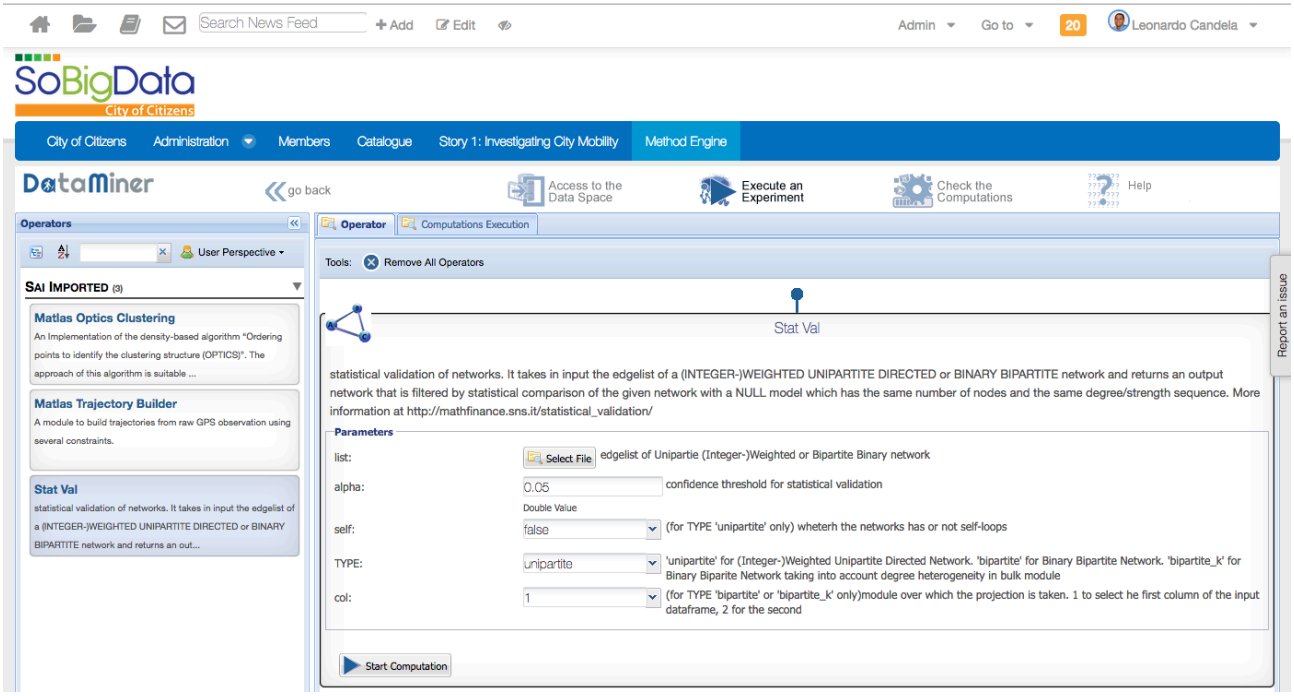

Figure 14. City of Citizens VRE: StatVal by DataMiner

3.2.3 DATASETS

A preliminary list of datasets worth to be integrated in the SoBigData e-Infrastructure has been identified and described by D8.1 Data Management report [7]. The list of datasets is expected to grow during the project lifetime. The picture emerging from the census concluded in November 2015 identified a total of 63 datasets covering five of the six thematic clusters:

- Human Mobility Analytics: 15 datasets;
- Social Data: 6 datasets;
- Social Network Analysis: 15 datasets;
- Text and Social Media Mining: 21 datasets;
- Web Analytics: 6 datasets.

No dataset for Visual Analytics has been identified yet.

Regarding accessibility, the survey identified that there are 22 datasets suitable for virtual access (10 public and 12 with restricted access) while there are 38 datasets suitable for transnational access and 27 are actually private.

After the first year of operation the infrastructure counts 47 datasets registered in the catalogue, hence available for discovery by scientists.

An up to date list of available datasets can be obtained by the Catalogue https://sobigdata.d4science.org/catalogue

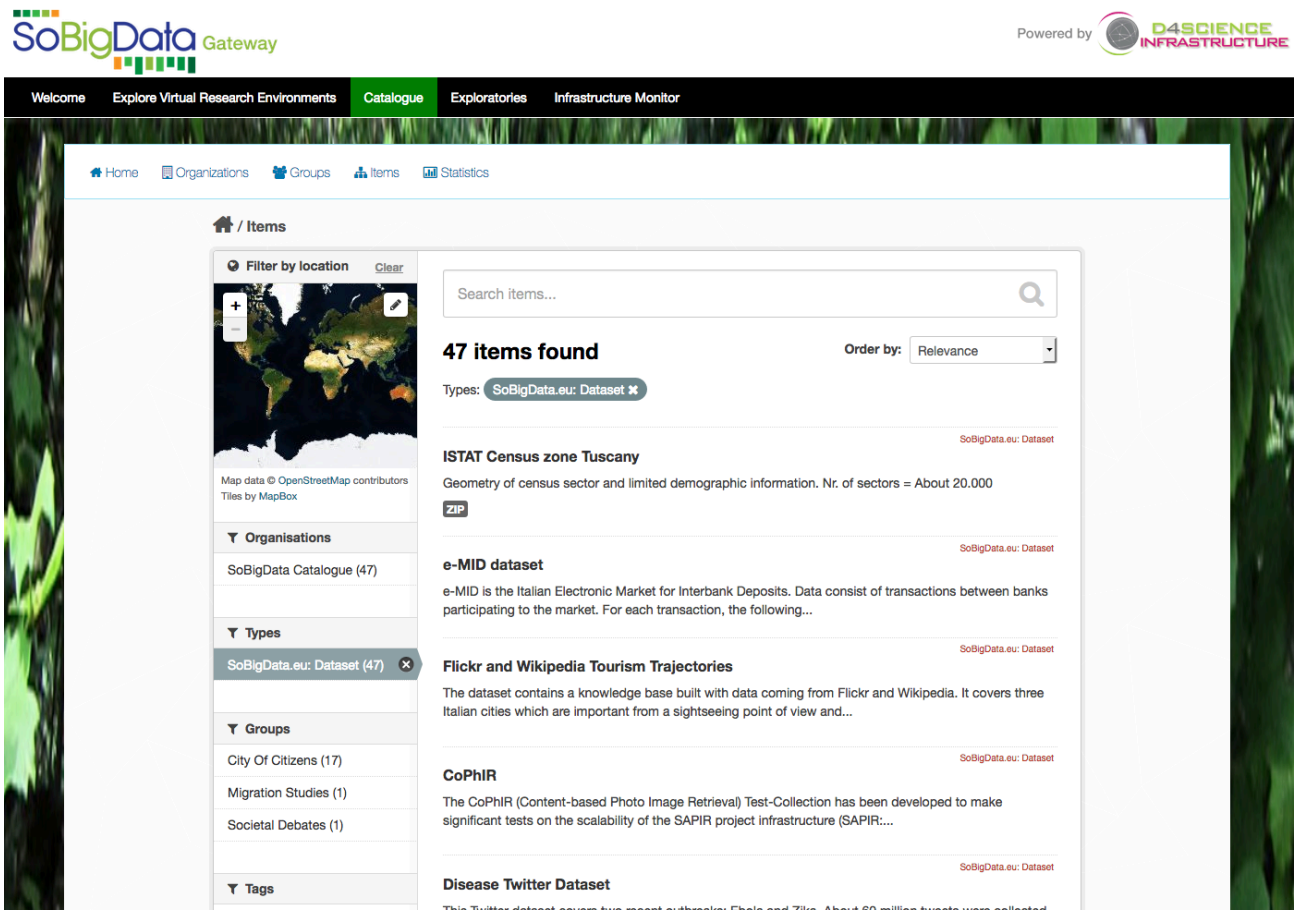

Figure 15. SoBigData Catalogue: Datasets

4 SOBIGDATA E-INFRASTRUCTURE SECOND RELEASE

Subsequent versions will be continuously released thanks to the integration of datasets and methods already identified as well as new ones that will be identified on due course. Development teams at local infrastructures will integrate their resources by following the guidelines and best practices developed by the project [3] with the active support of the CNR technical team. The next sections list the resources that are to be integrated in the next phase of the project (second release of the e-infrastructure) and extensions of the e-infrastructure planned for the future.

4.1 RESOURCES TO BE INTEGRATED

4.1.1 METHODS

4.1.1.1 ELIANTO

Elianto[10] is a platform hosting an open-source, user friendly and re-active Web interface to support the crowdsourced creation of gold standard datasets for entity linking and salient entities recognition. It supports human labelling of semi-structured documents through a guided two-step process. Collections of unstructured or structured documents can be uploaded on the platform and the guided annotation task easily monitored.

4.1.1.2 M-ATLAS

M-Atlas[11] is a mobility querying and data mining system centered onto the concept of trajectory. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechanisms for mining trajectory patterns and models that, in turn, can be stored and queried. M-Atlas is equipped with a querying and mining language making the analytical process possible and providing the mechanisms to master the complexity of transforming raw GPS tracks into mobility knowledge.

Other important facets of M-Atlas include (i) the privacy-preserving data publishing and mining techniques, designed to transform trajectory datasets into anonymous forms in such a way that strong privacy-protection guarantees can coexist with high data utility (ii) the analysis of different forms of mobility data, such as mobile phone call records, characterised by complementary weaknesses and strengths with respect to GPS trajectories.

4.1.1.3 SNA TOOLKIT

The SNA Toolkit is a collection of datasets and methods offered by the various project partners, with particular attention to scalable methods for Community Discovery, Link Analysis, Evolutionary Analysis and Multidimensional Network modeling. The methods in the SNA toolkit are currently used by many students and scientists through the various interfaces provided by the individual partners. For instance, the DEMON[12] software developed by CNR has become a reference in the area of community discovery in complex networks, it is widely used.

[10] http://elianto.isti.cnr.it/
[11] http://www.m-atlas.eu/
[12] http://kdd.isti.cnr.it/~giulio/demon/

4.1.2 DATASETS

The datasets mentioned in the DoA are being incrementally added to the registry and progress is suggesting that all datasets will be registered to the e-infrastructure by the end of the project. On top of this, scientists have suggested to start an activity of registration of datasets that may be relevant to the community. These datasets may be public and already available on line, but not from a thematic registry like the one of SoBigData. The idea is to collectively grow a domain-specific registry of all resources relevant to the community.

4.2 FUNCTIONALITIES TO BE ADDED OR REFINED

The technology (gCube) and infrastructure (D4Science) enabling the building and operation of the SoBigData infrastructure (cf. Sec. 3) is in continuous evolution. This evolution results from requirements and suggestions originating from the various contexts and domains the underlying infrastructure is called to serve. Some of the requirements emerged in the context of the SoBigData and the planned activities are discussed below:

- The integration between the catalogue and the rest of facilities should be reinforced. This might lead to various changes including the following: (a) the links between the catalogue entry and the resource instance the entry is representing should be explicitly there and in both the places. The catalogue certainly represents a primary place for users to be informed on what is available, yet once a user manages to identify a resource of interest she/he would like to have a seamless access to it. As a consequence of this the catalogue entries will be reinforced with links and information aiming at helping users in accessing the discovered resource; (b) the various manifestations a resource can exist, e.g. replicas of a method, should be properly captured and represented. In case the instances are exact replicas, the catalogue entry should contain a link to every existing instance operated by the infrastructure. Whenever possible, the publication of replica instances should be automatic (no human intervention is expected). In case of instances of the same technology yet characterised by diverse policies or other features, dedicated catalogue entries are expected to be published into the catalogue and these entries should link each other. It is likely to have sort of overall abstract entities representing families of homogeneous resources; (c) catalogue entries sharing mechanisms should be reinforced. Catalogue entries are already equipped with unique and actionable identifiers (URLs) yet mechanisms facilitating the sharing of such identifiers should be added. Such mechanisms include the automatic generation of QR codes out of the catalogue entry URLs and user friendly mechanisms for posting the catalogue URI of the item by the social networks (e.g. Twitter) including the social networking area of the SoBigData Virtual Research Environments and Exploratories; (d) making the publication mechanism as much automatic as possible. When a new resource is integrated in the SoBigData infrastructure (or an existing one is enhanced) the characterisation of such a resource should contextually appear into the catalogue. In the currently existing workflows this is not systematically performed, i.e. in some cases there is a need for a human intervention for having the resource published. While avoiding this human intervention is not always possible (e.g., in all the case the resource has no single link with infrastructure services), in some cases the mechanism can be enhanced. This is particularly relevant for methods integrated and operated by DataMiner. A mechanism will be in place to make every instance of a method operated by DataMiner automatically published into the catalogue too.
- The first user experience for newcomers should be simplified. This might lead to various changes including the following: (a) the various Virtual Research Environments and

Exploratories will be systematically equipped with a "public" page aiming at clarifying (to both expert and non-expert users) what is the goal of the specific environment, what they get when accessing such dedicated working environments (e.g. methods and datasets), what are the policies governing such working places (e.g. if access to it is open to everybody or restricted to certain users only), links to any helps, demos and tutorials aiming at supporting users willing to use the environment; (b) the development of per-resource help and tutorials to go in tandem with any catalogue entry thus to provide the users with material aiming at simplifying the understanding and exploitation of every published resource; (c) the enrichment of the typologies of resources published into the catalogue. Besides the currently published applications, deliverables, datasets, methods other typologies of resources worth documenting and sharing are going to be added to the catalogue. Among the new resource typologies to be published there are Exploratories and Virtual Research Environments (these are sort of complex applications resulting from a combination of applications, datasets and methods); (d) simplification of the mechanisms and workflows governing registration / log in / use, both that preceding these phases and that following these phases. The SoBigData Gateway[13] (playing the role of human-oriented access point to the entire SoBigData infrastructure and its facilities) will be equipped with a set of "welcome messages" aiming at clarifying what is the role of the whole gateway, what is expected to be the role of the catalogue and the rest of offerings made available by the gateway. In addition to that, it will be equipped with a set of "what if" pages aiming at providing the users with a typical steps / actions to be performed in order to benefit from the offered facilities depending on who is the user and what he/she is willing to do.

These enhancements will be further detailed, developed and recorded by relying on the ticketing system[14]. In addition to that, requirements are expected to stem from both the day by day exploitation of the offered facilities and an internal assessment exercise that has been promoted by WP7 (user experience evaluation). All the requirements and activities leading to further developments of the overall infrastructure will be capture by dedicated tickets.

[13] https://sobigdata.d4science.org/
[14] https://sobigdata.d4science.org/group/sobigdata.eu/activity-tracker

5 CONCLUSION

This is the second of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable describes the status of the SoBigData e-infrastructure after one year of project life time and sets its future developments.

REFERENCES

[1] Candela, L., Castelli, D., Manzi, A., Pagano, P. (2014) Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. International Symposium on Grids and Clouds (ISGC) 2014, Proceedings of Science PoS(ISGC2014)022

[2] Candela, L., Castelli, D., Pagano, P. (2013) Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, Vol. 12, pp. GRDI75-GRDI81, DOI 10.2481/dsj.GRDI-013

[3] Candela, L., Manghi, P., Pagano, P. (2016) Best practices and guidelines towards interoperability. SoBigData Project Deliverable D10.1, March 2016

[4] Capannini, G., Dato, D., Lucchese, C., Mori, M., Nardini, F. M., Orlando, S., Perego, R., Tonellotto, N. (2015) QuickRank: a C++ Suite of Learning to Rank Algorithms. Proceedings of the 6th Italian Information Retrieval Workshop (IIR 2015). Cagliari (Italy)

[5] Cornolti, M., Ferragina, P., Ciaramita, M. (2013) A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web* (WWW '13). ACM, New York, NY, USA, 249-260. DOI 10.1145/2488388.2488411

[6] Ferragina, P., Scaiella, U. (2010) TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, 1625-1628. DOI 10.1145/1871437.1871689

[7] Grossi, V., Romano, V., Trasarti, R. (2015) Data Management report. SoBigData Project Deliverable D8.1, December 2015

[8] Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., Pagano, P. (2014) The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. Program: electronic library and information systems, Vol. 48 Iss: 4, pp.322 – 354, DOI 10.1108/PROG-08-2013-0045