Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE

Deliverable D7.1

Periodic Report on VA Activities 1

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	WP7 VA1 - Virtual Access
WORK PACKAGE LEADER	CNR
WORK PACKAGE PARTICIPANTS	CNR, EGI
DELIVERABLE NUMBER	D7.1
DELIVERABLE TITLE	Periodic Report on VA Activities 1
AUTHOR(S)	Valerio Grossi (CNR), Roberto Trasarti (CNR), Beatrice Rapisarda (CNR), Ilaria Barsanti (CNR)
CONTRIBUTOR(S)	Andrea Manzi (EGI)
EDITOR(S)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
REVIEWER(S)	Pasquale Pagano (CNR), Massimiliano Assante (CNR), Andrea Manzi (EGI)
CONTRACTUAL DELIVERY DATE	31/12/2020
ACTUAL DELIVERY DATE	15/01/2021
VERSION	1.2
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	42
KEYWORDS	Virtual Access, e-infrastructure, catalogue, key performance indicators

EXECUTIVE SUMMARY

The SoBigData RI portal allows the user to navigate and discover datasets and services using real applications and case studies as part of the exploratories developed in WP10. All the resources metadata will be accessible through the web interface for free and anonymously, but access to the existing resources of the e-infrastructure will require a free registration (using ad hoc or academic/social credentials supported by the EOSC portal). This registration will not require any moderation by the system administrators and will be used to track the resource usage and statistical purpose.

This deliverable reports the results of Virtual Access based on key performance indicators defined in Grant Agreement. The data on the accesses to SoBigData e-infrastructure allow scientists to assess resource usage in a very open and collaborative way. The data are taken from a dedicated dashboard page that enables to the administrators to access the detailed indicators.

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
EOSC	European Open Science Cloud
VA	Virtual Access
VRE	Virtual Research Environment
TNA	TransNational Access
KPI	Key Performance Indicators

TABLE OF CONTENTS

1 Relevance to SoBigData++

1.1 Purpose of this document

Virtual Access (VA) gives the user the possibility to navigate and discover datasets and services employing real applications and case studies. All metadata of the items are accessible through the web interface for free and anonymously. Access to the existing resources of the e-infrastructure requires a free registration (using ad hoc or academic/social credentials supported by the EOSC portal). This registration does not require any moderation by the system administrators and will be used to track the resource usage and statistical purpose. This document reports the current state of the e-infrastructure and the actions in progress to integrate new contents and attract new users.

1.2 Relevance to project objectives

The objective of the VA is to offer online services for big data and social mining research. SoBigData aims to focus on the reproducibility of results by making it easier to find, access, and replicate experiments under the FAIR and FACT principles. In the case of VA, the objective is to increase the number of datasets and methods integrated into the e-infrastructure. The available methods can be used in the cloud as a service (using the computational resources of the e-infrastructure) or the methods can be downloaded for local use. WP3 liaises with dissemination and outreach WPs to attract new potential users for local infrastructure sites. To this aim, with the new release of the SoBigData website (expected in spring 2021) a dedicated area in the project's website will offering an updated description of the e-infrastructure capacities (e.g., number and typologies of federated resources) and its current exploitation (i.e., the number of Virtual Research Environments created, number of registered users, outstanding results, latest posts).

1.3 Relation to other work packages

VA works in synergy with WP10 and WP9, which are responsible for the community building with the exploratories and the management, planning, and releasing of the e-infrastructure and VREs. Furthermore, WP8 delivers datasets, methods, and applications to the SoBigData++ platform for virtual and trans-national access. Finally, VA can also be used for dissemination of the SoBigData RI (WP3) and as a supporting tool for training events (WP4). The design and integration of resources are done in WP4 for the training modules and in WP8 for the datasets, libraries, and services. In the description of those WPs, there are examples of the resources which will be provided. The tools available to the users to discover, search, use and execute resources are developed (or upgraded from the existing one) by WP9, such as the catalogue and on-line coding and workflow design tool, and will be part of the web portal designed by the WP7.

1.4 Structure of the document

This document reports on operation activities and updates actions from January 2020 to December 2020. The deliverable contains the following main sections:

- Section 2 outlines the main components and services related to SoBigData RI.
- Section 3 introduces the engine where a user can execute methods and outlines the JupyterHub integrated into the platform.
- Section 4 reports the results of the review of the SoBigData e-infrastructure. The results will be used to design and guarantee a better interaction in the platform's next release.
- Section 5 outlines and analyses several key performance indicators in the first year of the project, SoBigData++. The result examines a period from 01 January 2020 to 31 December 2020.

2 The SoBigData Research Infrastructure

The SoBigData RI supports data science serving a cross-disciplinary community of researchers studying all the aspects of societal complexity from a data- and model-driven perspective. The SoBigData VA services have two main entry points: i) the main site: www.sobigdata.eu and ii) the gateway: sobigdata.d4science.org. The user starting from SoBigData.eu can access the catalogue or the gateway. The user can access the several Virtual Research Environments (VREs) related to Exploratories, Applications, and SoBigData Lab from the gateway. The gateway is designed to **emphasize** all the resources provided by the e-infrastructure. The following subsections give a short overview of the primary services that support the VA. The description of the overall architecture is available in SoBigData deliverable SoBigData deliverable "D7.3 - VA e-Infrastructure Service Provision and Operation Report 3"[1]. Several technical details introduced by the SoBigData++ project are available in deliverable "D6.4 SoBigData e-Infrastructure Common Facilities 1"[2].

2.1 WebSite

The website's main aim is to promote SoBigData RI events/activities and promote the usage of our services (www.sobigdata.eu). The SoBigData RI home page can adapt to the different methods of access, web, mobile, etc. The architecture used for the homepage is designed to highlight three macro-sections:

- *explore*: all the information about the resources provided by the SoBigData platform organized by the six exploratories: Sustainable Cities for Citizens, Societal Debates and Misinformation Analysis, Demography, Economy & Finance 2.0, Migration Studies, Sport Data Science, and Social Impacts of AI and Explainable Machine Learning;
- *discover*: everything about the events organized by SoBigData;
- *participate*: how to be involved deeply in SoBigData, e.g., transnational access and what it is essential to know about Ethical Social Mining.

2.2 e-Infrastructure

The e-infrastructure and the Catalogue of the SoBigData RI are based on D4Science[3] services, which provide researchers and practitioners with a working environment where open science practices are transparently promoted, and data science practices can be implemented by minimizing the technological integration cost highlighted above. Currently, the e-infrastructure of SoBigData has a complex architecture of 18 VREs, each of which implements a specific service. The gateway of the SoBigData RI publishes all the services related to e-infrastructure, and its style recalls the concepts and the style of the Website home page. As shown in Figure

[1] https://data.d4science.net/stH9
[2] https://data.d4science.net/sjUu
[3] https://www.d4science.org/

2.2.1, a user has a snapshot of the principal services provided by the e-infrastructure organized into six main areas:

- *catalogue*: enables the user to search an item given a set of keywords;
- *exploratories*: contains a short description of and a link for accessing VRE related to each exploratory;
- *applications*: contains a description of all the applications available and a link for accessing VRE implementing the application;
- *e-learning area*: this part enables the user to access all the training material related to the SoBigData community;
- *SoBigDataLab*: using this VRE, users can execute methods on the e-infrastructure with the support of an online file sharing workspace;
- *workspace*: this is an online environment to support secure and controlled data storage and sharing. Each VREs has associated a workspace where users can store, access, and share documents and results related to the activities inside a specific gateway and the VRE. Each user has a private space, where storing data and documents and creating folders and a public space, one for each subscribed VREs where sharing files.

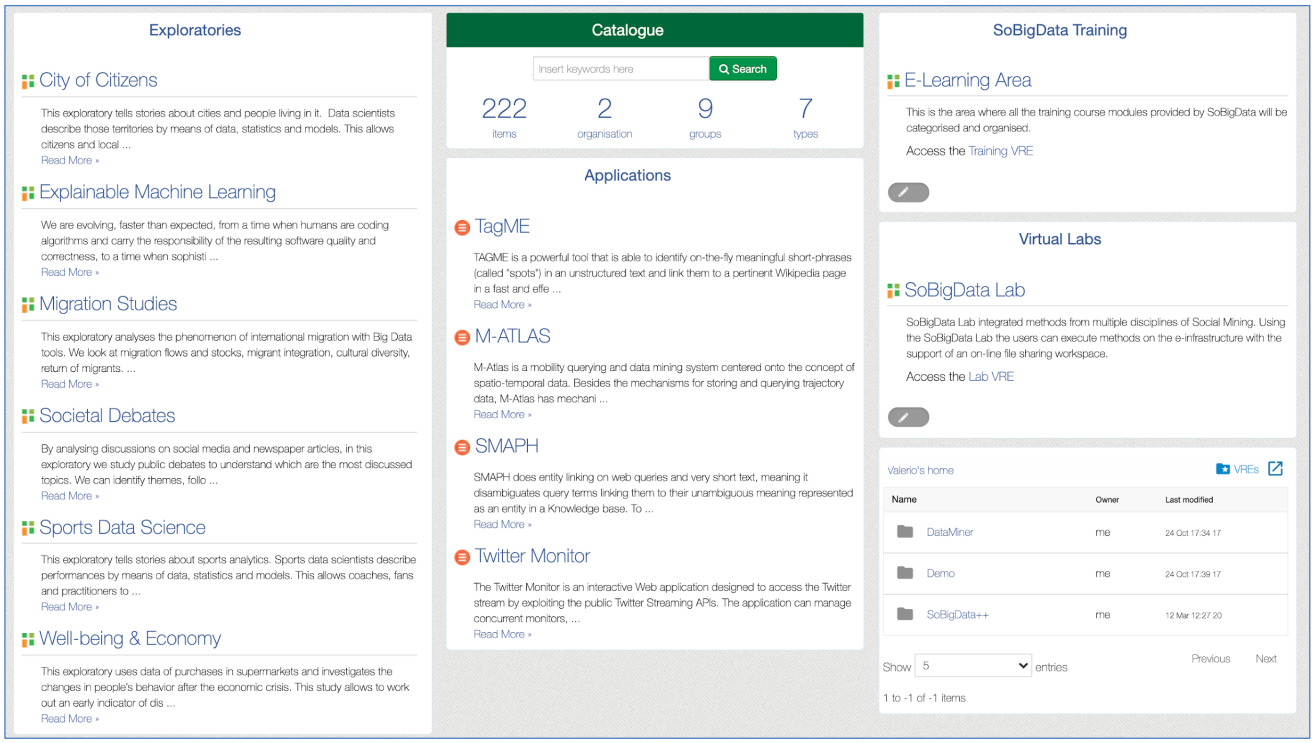

Figure 2.2.1. e-Infrastructure Gateway

2.3 Catalogue

The SoBigData Catalogue is a smart tool for finding and accessing all the datasets, services, publications provided by SoBigData RI. It contains all the resources, which may be accessed by VA (through web services) or by physically visiting the resource's publisher.

The catalogue is the primary tool for discovering and searching for an item inside the SoBigData RI. All the elements inside the SoBigData RI are discoverable through this service. The user can insert a set of keywords, and the list of the results will be visualized. The search result provides a list of items included in the catalogue and its classification (e.g., Method, Training Material, Dataset). The complete description is provided on the dedicated page, accessible by clicking on the item. These features can be added to the search filter, which will be recalculated in real-time. The search result can be sorted alphabetically concerning the insertion date or popularity.

As shown in Figure 2.3.1, the catalogue organizes products only by a set of predefined categories that the user can navigate by selecting the specific link. At the moment, we have defined the main organization of products. On the one hand, the datasets, services, methods, and applications, while on the other hand, we can find scientific papers and training material. Items can be selected based on the specific group they belong to, for example, if a user is interested only in a specific exploratory product. Currently, the catalogue includes two main classes of products (organizations): “SoBigData Services and Products” and “SoBigData Literacy”. Furthermore, the catalogue item is also organized by groups for searching quickly products related to an exploratory. Finally, we have also enabled the search by types, e.g., dataset, method, and much more.

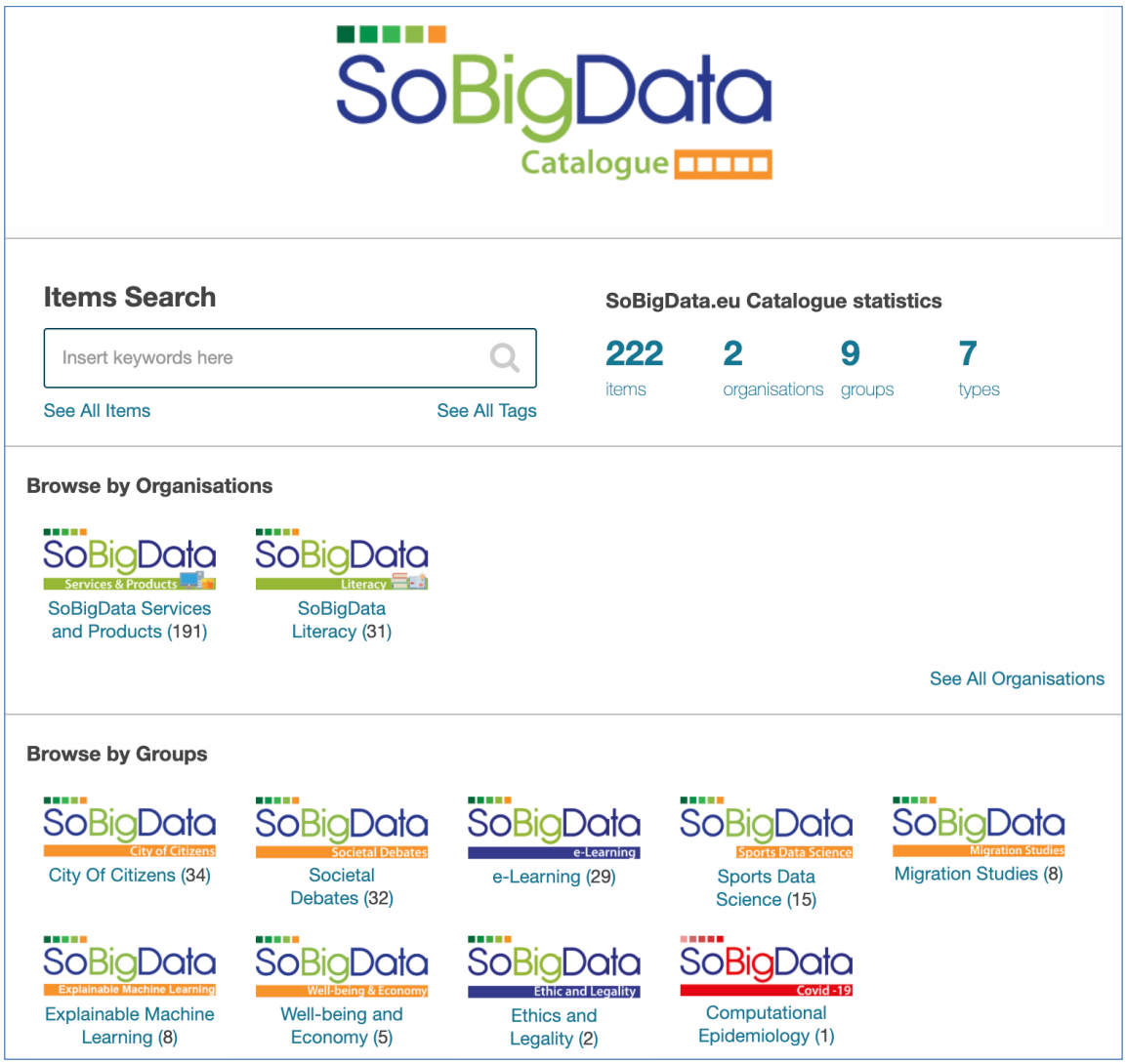

Figure 2.3.1. SoBigData Catalogue

3 The SoBigDataLab VRE & JupyterHub

The SoBigDataLab VRE will integrate under the same environment different methods that can be invoked through SoBigData e-infrastructure. A method is the implementation of an algorithm/procedure, or is an algorithm that requires an engine to be executed. Different kinds of integration are available based on the programming language in which a method is implemented. In any case, once a method is integrated into the platform, the final user has a homogeneous web-form for inserting parameters and for invoking it independently from the programming language employed.

The SoBigDataLab VRE is linked and accessible through the platform gateway as well as its items are linked through the catalogue. This environment enables a user to Execute an Experiment, Check the status of started computations and access the DataSpace for getting the results.

The SoBigDataLab VRE allows the scientific community to make its methods available. The integration of a new method into the e-infrastructure must be as simple as possible, otherwise the users may be discouraged from making their methods available. For this purpose, by clicking on the service method importer, the users have a guided procedure to integrate a new method.

This VRE provides many methods that can be selected and executed into the e-infrastructure. The methods are performed by loading the input data into the user workspace. It is possible to execute a method only if the required input file is already present in the workspace.

During this first year of SoBigData++ project, this VRE has been updated in several directions:

- the method importer now supports the following programming languages: R, Java, Python, Windows, or Linux compiled, or directly from a GitHub repository.
- a new environment for interactive computations called JupiterHub has been integrated into the lab (see Section 3.1 for further details)

3.1 JupyterHub

The project name, Jupyter, comes from the core supported programming languages: Julia, Python, and R. JupyterHub provides a convenient, cloud-hosted way to serve Jupyter Notebooks for multiple users. A Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

JupyterHub is easily accessible by clicking on the link on the top of SoBigData Lab VRE. After starting the server by selecting one of the default profiles available, the user can start to use Jupyter Notebook as the local version. Figure 3.1.1 shows the JupyterHub environment accessed by SoBigDataLab VRE.

The libraries currently usable in JupyterHub includes:

- NDlib: definition and simulation of a diffusive model on networks (epidemics, opinion dynamics) http://ndlib.readthedocs.io/;
- CDlib: community discovery on complex networks (~50 algorithms, ~30 evaluation functions, ~10 visual analytics facilities) - http://cdlib.readthedocs.io/;
- DyNetX: a library for modeling and analyzing dynamic networks - http://dynetx.readthedocs.io/;
- BiCM: bipartite configuration model - https://bicm-beta.readthedocs.io/.

Several new integrations are planned for the beginning of 2021, for example, scikit-mobility[4] and for March 2021 a library related to Explanailable Ai exploratory.

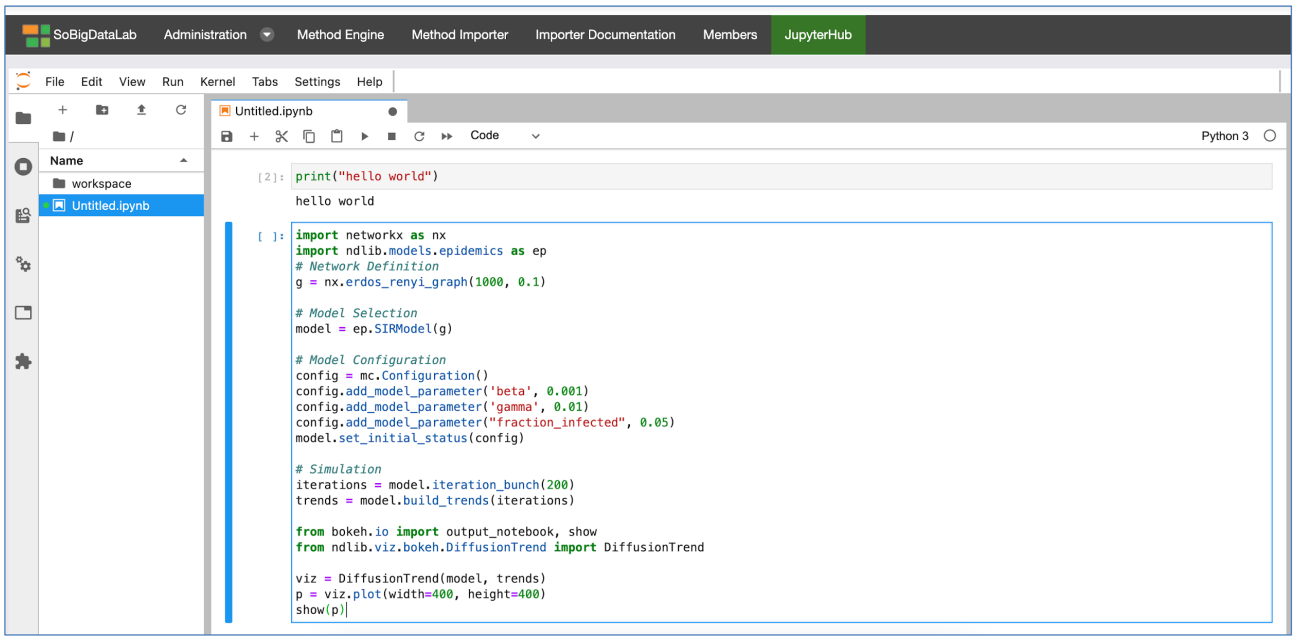

Figure 3.1.1. The JupyterHub notebook in SoBigData e-infrastructure

[4] https://scikit-mobility.github.io/

4 SoBigData++ Virtual Access Review

To improve the user experience, the WP7 periodically reviews all the e-infrastructure and website sections. The review is done by selecting internal and external reviewers, in order to give feedback useful to propose changes. The WP7 working group analyses the proposed changes and prioritizes the implementation of the modifications in collaboration with WP9 (responsible for the e-infrastructure services). The current review started in October 2020, and we are now collecting the results to start the implementation in January 2021. We plan to release the new version of the e-infrastructure and website in late February 2021.

4.1 e-infrastructure and website

A first review of the SoBigData Gateway was done in November 2017 (see SoBigData deliverable D7.2 VA e-Infrastructure Service Provision and Operation Report 2[5]), which triggered many improvements to the gateway. With the growth of the SoBigData community and the growth of users of the RI, a new review has been necessary to understand how to enhance the user experience to an exponentially growing user community.

In November 2020, we performed an in-depth internal review of the e-infrastructure to detect the critical aspects from the point of view of the Design, Content, Organization, Usability, and User-friendliness. Moreover, to speed up the growth of the RI in terms of available resources, in this review, particular attention was given to the integration process of resources (datasets, methods, applications, etc.) to understand how to improve and simplify it.

For this review, we selected a group of reviewers from the SoBigData++ consortium. In particular, we chose the reviewers from the new partners of the project's consortium as they have no in-depth knowledge of the gateway and they can be considered new users.

The reviewers are:

- Dmitry Gnatyshak (BSC)
- Carlos Castillo (UPF)
- Adriano Fazzone (UNIROMA1)
- Amleto di Salle (UAQ)
- Jesús A. Manjón (URV)

All reviewers selected are experts with coding skills (C#, Java, Python3, DBMSs: C++, C, Lisp, Prolog, R, SQL, PHP, BASIC, MySQL, MongoDB). The selection of reviewers with coding skills has been made to better evaluate the critical aspects of the integration process.

To each reviewer, 11 objectives (questions) have been set, where they had to describe the path followed, the weak and strong points, the usability, and an overall evaluation. Table 4.1.1 reports the score reported

[5] https://goo.gl/T6WAkQ

by the reviewers. The overall score is quite good, in fact we received only 3 Insufficient scores from the reviewers. In particular, the average score for the use of the Gateway for searching, or using resources is Very Good, while the importing process needs some improvement as it is tricky also for skilled people. As suggested in the comments, some ad hoc tutorial would help the process. The execution of a method, which is one of the most important features in the R. has been considered almost Excellent. The synthesis of the comments of the reviewers can be found in Appendix A.1.

Objective	Score	
1: Search for Datasets in the Catalogue	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Excellent Very Good Very Good Very Good Very Good
2: Search for Methods in the Catalogue	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Very Good Excellent Sufficient Very Good Very Good
3: Browse and Search for Exploratories datasets	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Very Good Very Good - Very Good Sufficient
4: Browse and Search for Exploratories Methods	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Very Good Excellent - Very Good Very Good
5: Use of Training Resources	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Sufficient Excellent - Sufficient Excellent
6: Using the Workspace (for sharing files and folders)	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Very Good Excellent Excellent Very Good Excellent
7: Use of Applications	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Very Good Excellent - Sufficient Excellent

8: Importing a Dataset in the Catalogue	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Insufficient Very Good Very Good Sufficient -
9: Importing a Method in the Catalogue	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Insufficient Very Good Very Good Sufficient -
10: Importing a Method in the SoBigData Lab	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Excellent Very Good - Very Good Insufficient
11: Execution of a Method in the SoBigData Lab	*Rev 1* *Rev 2* *Rev 3* *Rev 4* *Rev 5*	Excellent Excellent Very Good Very Good Excellent

Table 4.1.1 - The score obtained by the reviewers

4.2 Internal Audit on SoBigData++ Datasets

During the current e-infrastructure review, we also started an internal Audit in collaboration with WP2 to review the legal aspects of all the datasets in the catalogue. The audit aims also to improve the submission form and facilitate the users in specifying the relevant aspect when they add a new dataset and to highlight automatically critical points which need the intervention of the BOEL to clarify ethical and legal issues.

5 Key Performance Indicators (KPIs)

As reported in the SoBigData++ Grant Agreement, the KPIs' definition to be used in the SoBigData platform statistics is defined into two main objects:

- O1: Advancing the social mining platform,
- O2: Expanding the multidisciplinary community.

This section reports several indicators related to the Virtual Access service provided by the SoBigData RI during the period from 01 January 2020 to 31 December 2020. Since the SoBigData RI is available from 2016, we report (when possible) a comparison between the indicators available at the end of the previous project SoBigData GA. 654024 ended the 31 December 2019. For a complete description of the indicators related to the previous period of SoBigData RI, see SoBigData deliverable "D7.3 - VA e-Infrastructure Service Provision and Operation Report 3"[6].

5.1 Accounting dashboard VRE

The reported indicators are collected automatically by the platform and reported as dashboards in a dedicated portal for administrators[7]. The administrators can have statistics on accesses (daily updated) of all the VREs deployed in SoBigData RI by querying this service. The dashboard reports the statistics related to catalogue usage, the methods invocations, and social interactions. From the first January 2020, the dashboard also includes another service called "datastudio" where it is possible to have geospatial access statistics on SoBigData Gateway both on European and Worldwide level.

5.2 Users/Accesses statistics

As reported in Section 2.2, the SoBigData VA has supported 18 VREs. In this section, we report the total number of users registered at the gateway. It is important to highlight that the registration is not moderated, and it is required to keep track of the resource usage and for statistical purposes. All the services related to VA are free of charge and freely available. Figure 5.2.1 reports the number of users registered to the SoBigData e-infrastructure. Considering the period from January 2020 to December 2020, we can state that the SoBigData RI has continued to attract new users by inviting them to explore the Catalogue and the Exploratories. The trend of new users for this first year of the project reflects the same period of the previous SoBigData project. The increment of users is also due to the number of beneficiaries involved in the SoBigData++ project, as we passed from 12 in SoBigData to 31 in SoBigData++. This aspect has given a positive impact to the SoBigData RI dissemination in Europe, directly impacting the number of registered VRE users. This trend is influenced as well by the CoVid-19 crisis that is forcing online access; the number of users is almost twice with respect to the end of the first SoBigData project. Of course, this side effect partially covers

[6] https://data.d4science.net/stH9
[7] https://sobigdata.d4science.org/group/sobigdata/accounting-dashboard

the lack of face-2-face dissemination events, training courses, and transnational visits that typically provides an increment of registered VRE users into the platform. At the end of December 2019, the registered VRE users were 4814, while at the end of 2020 we registered 7284 users with an increment of 51%. The same trend is noticeable, also considering the users subscribed to different VREs. Figure 5.2.1 shows the positive trend of VRE users registered to the SoBigData Gateway and their distribution along the VREs implementing the services related to the e-infrastructure. Some peaks are presented in correspondence with specific events, for example, in October 2020 in relation to the SoCiNFO Conference or the summer school in November. The Catalogue and the SoBigDataLab & Services remain the VREs with more registered users.

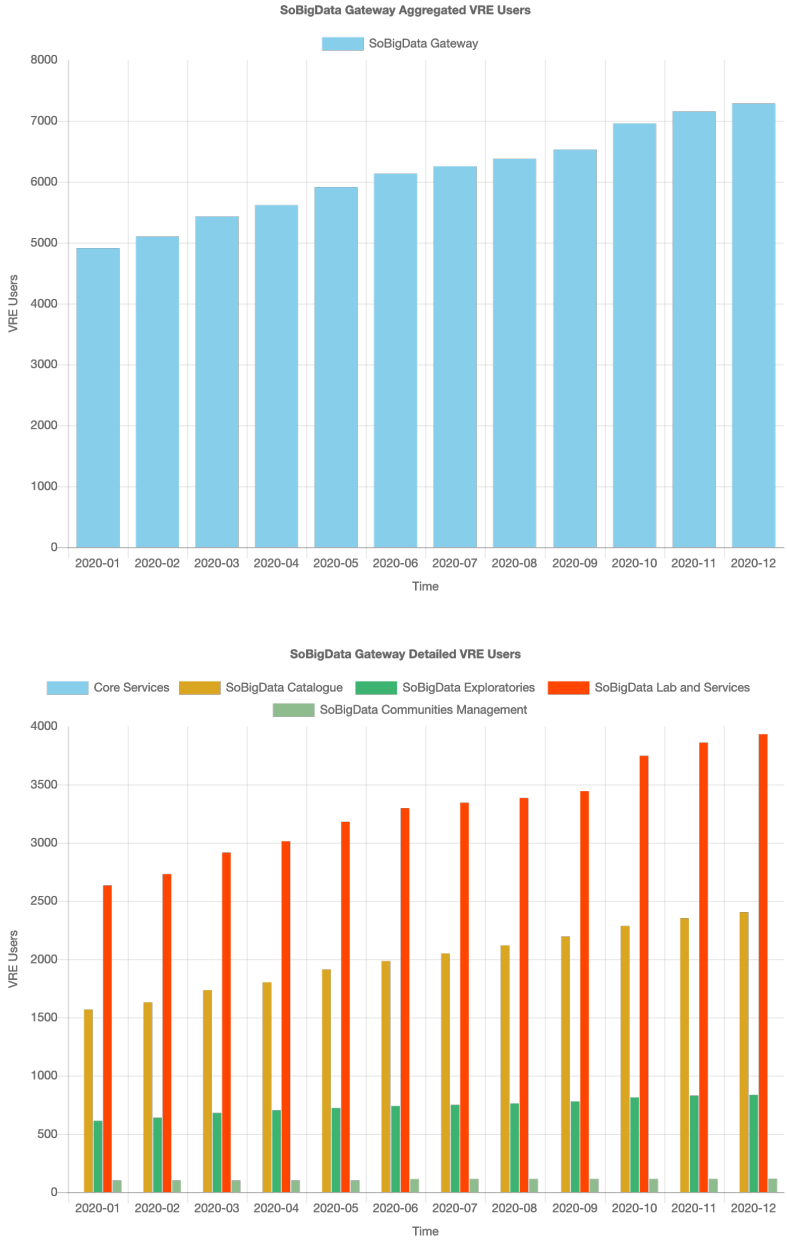

Figure 5.2.1. User registered in the gateway (all the VRE users) and Distribution of the users in the different parts of the platform

Figure 5.2.2 reports the number of accesses to the gateway and the relative distribution on the different VREs of the platform; also in this case, we can observe a positive trend in the accesses. In 2019, the SoBigData RI observed 1130 monthly average accesses while in 2020 we registered an increment of 6% with an average of 1195. With the introduction of new services such as JupiterHub, and the integration of new libraries we expect to have bigger increment of accesses in the next period.

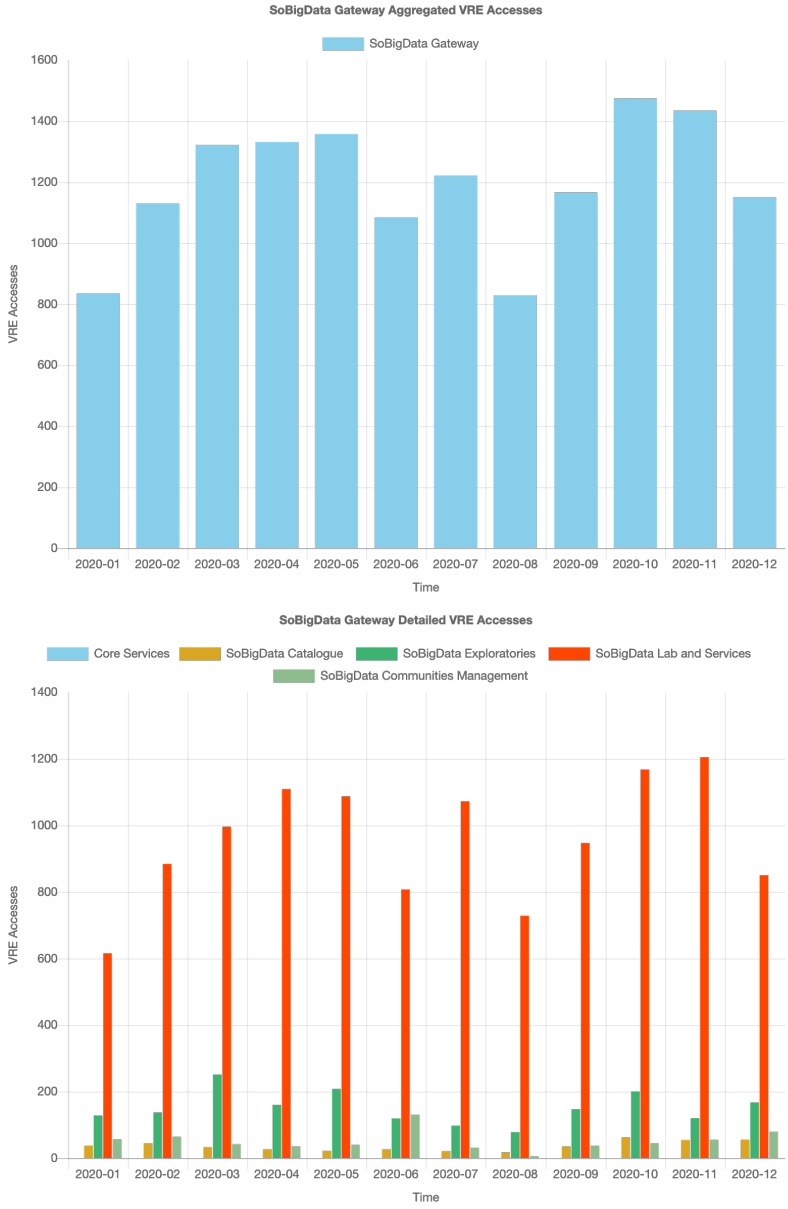

Figure 5.2.2. Number of accesses of the gateway (all the VRE users) and distribution of the accesses on the different parts of the platform

The current positive trend is related to the increased number of users. Positive trends are reported also considering Figure 5.2.3. In this case, it is worth noticing that non all the exploratories are accessed in the same way, and the number of products in the catalogue, and the services related to an exploratory influence the number of accesses (see Table 5.3.1 for details on the distribution of products along the exploratories). Sports analytics and Explainable Machine Learning (the last two introduced) have positive trends as emerging exploratories.

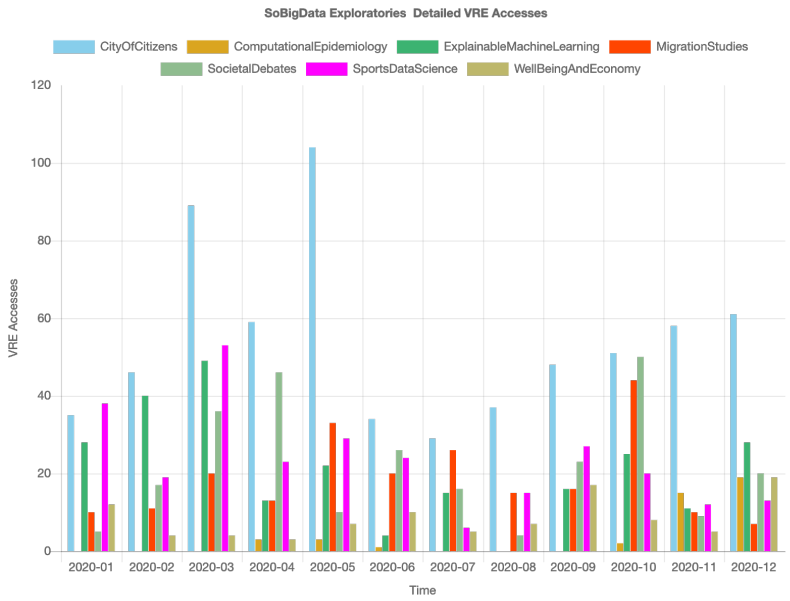

Figure 5.2.3. Number of accesses to the SoBigData Exploratories

Figure 5.2.4 reports the number of accesses to SoBigDataLab and Applications. As reported in Section 3, the SoBigDataLab is the VRE where a user can execute the social mining methods integrated into VRE. In 2019 we observed an average of 902 monthly accesses. During 2020, we registered an average of 955 monthly accesses with an increment of 6%. This increment, such as for the gateway accesses, is directly connected to the number of registered users and the increment of online services usage due to the pandemic situation.

The use of other services such as the workspace or the social area remains relatively low. We think that these two services are essential, on the one hand, for sharing and storing research results, on the other hand for providing support and exchange information inside a VRE or a specific community. We expect an increment of usage of workspace with the introduction of JupyterHub, and an improvement of social interaction with the creation of well-focused posts by User Community Activists related to exploratories in WP10.

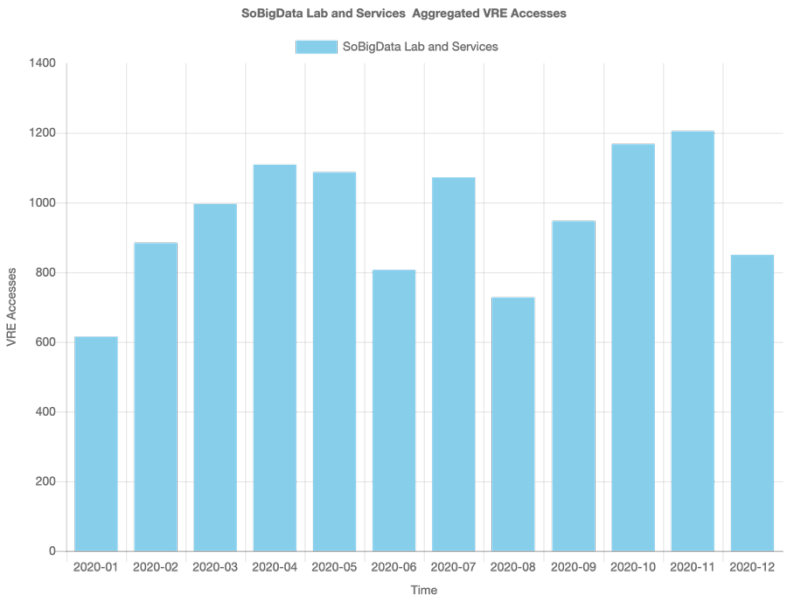

Figure 5.2.4. Number of accesses to SoBigDataLab & Applications

5.3 Catalogue Statistics

The catalogue is the main searching tool of the RI. As reported in Table 5.3.1, the SoBigData catalogue contains 222 resources: 92 datasets, 83 methods, 31 training materials and publications, 9 applications, 7 experiments. There are 107 online and 73 onsite resources (i.e., reachable only on-site visits). Considering the Exploratories, there are 34 items in Sustainable Cities for Citizens, 32 in Societal Debates and Misinformation Analysis, 8 in Migration Studies, 8 in Social Impacts of AI and Explainable Machine Learning, 5 in Demography, Economy & Finance 2.0, and 15 in Sports Data Science. The items are currently organized into two main organizations, 9 groups, and can also be browsed by 8 main types. All these statistics are available on the catalogue thanks to the filter system described in Section 2.3. In 2020, 20 new products have been added to the catalogue, and many others have been updated with new resources.

	Datasets	Methods	App. + Exp	Train	***Total***
Sustainable Cities for Citizens	17	15	1	1	***34***
Societal Debates and Misinformation Analysis	25	1	5	1	***32***
Demography, Economy & Finance 2.0	3	1	0	1	**5**
Migration Studies	1	6	0	1	**8**
Sport Data Science	5	5	4	1	**15**
Social Impacts of AI and Explainable Machine Learning	6	2	0	0	**8**
Generic	35	53	6	26	**120**
Total	***92***	***83***	***16***	***31***	***222***

Table 5.3.1 Number of integrated resources grouped by exploratories

Compared to 2019, when we registered a total of 533 accesses with 45 average monthly, in 2020, we observed an increment of more than 300% with a total of 2189 accesses (182 average monthly). Our catalogue responds to around 28k queries with an average of 2400 queries monthly.

As reported in Figure 5.3.1, we observed a stable trend in querying the catalogue that is free and without registration, while in our view it is still low in the number of downloaded elements with respect to the number of items accessed. This is due to two main factors, on the one hand to download an item the user has to be registered and this may discourage several users. On the other hand, several products do not have any registered resource to download; or are methods integrated into the platform and in this case is available a direct link to SoBigDataLab. For publications, training material and public dataset we expect that a downloadable resource is available. The internal audit performed by WP2 to review the legal aspects of all the datasets checked also this aspect.

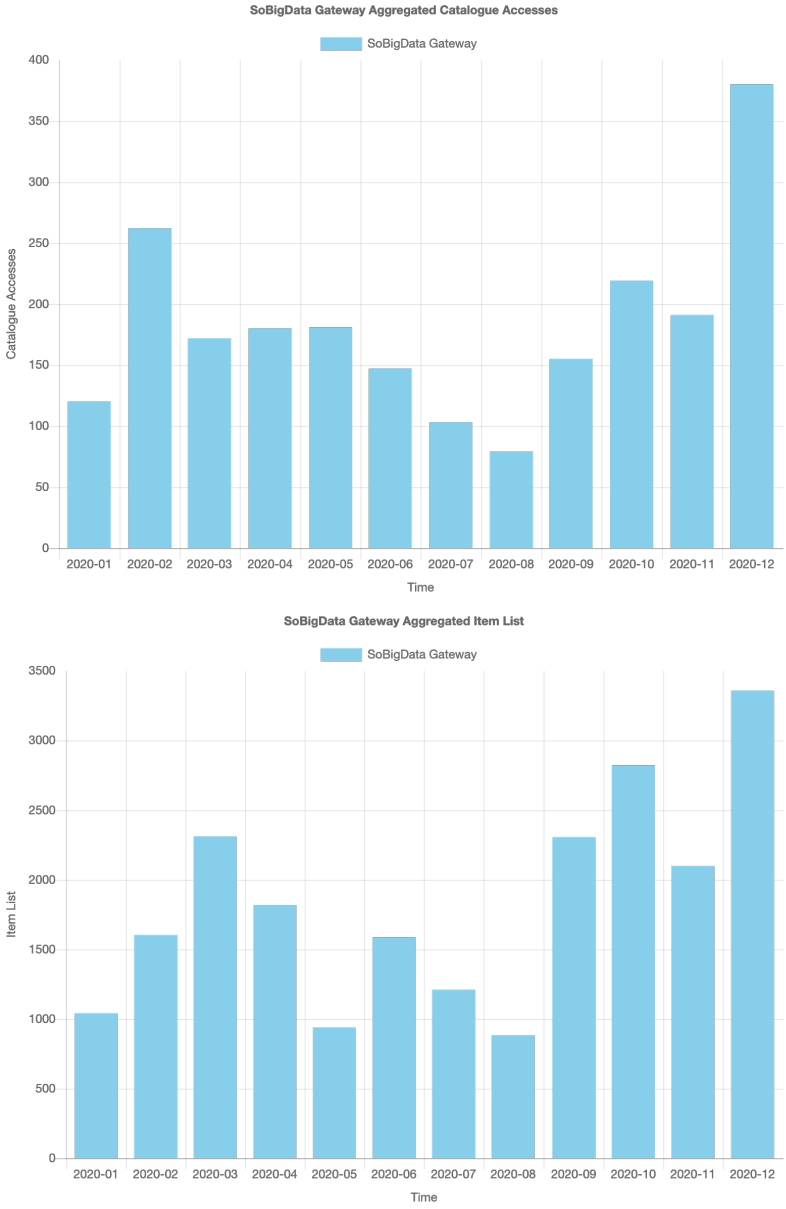

Figure 5.3.1. SoBigData RI Catalogue accesses and queried items list

5.4 Geo Localization of the accesses

This section reports the Geo Localization of the user accesses considering the period from January 2020, to December 2020.

Figure 5.4.1 reports the accesses from European Countries. It is noticeable that the users open sessions on the SoBigData RI from all countries in Europe and not only to the ones involved in the consortium. For

example, users from Norway, Denmark, and Poland access our services but they are not represented by any institution in SoBigData++ consortium. This shows how our platform can be relevant for all Europe in the context of Social Mining, big data analytics and artificial intelligence.

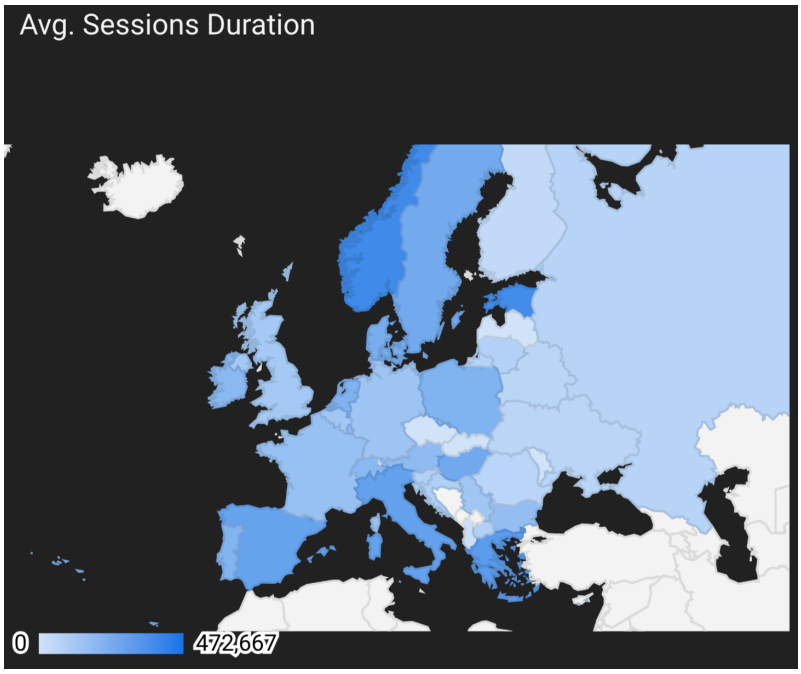

Figure 5.4.1. European Countries accesses Map Overlay

Figure 5.4.2 shows a World Map Overlay of the user accesses to the e-Infrastructure Gateway. The map shows that the e-infrastructure users come from all the continents, most of them from Europe and Asia.

To better understand the distribution of the accesses we added a pie chart in Figure 5.4.3 showing the top 10 World countries accesses distribution. From these two pictures we see that nearly 60% of the accesses comes from four countries, namely Italy, China, and USA, with accesses originated from Italy in the first place, followed by the United States. This high number of accesses from Italy is not surprising since the project consortium includes 4 institutions from Italy and also the number of TNA visits are mainly in Italy due to beneficiaries and research groups involved.

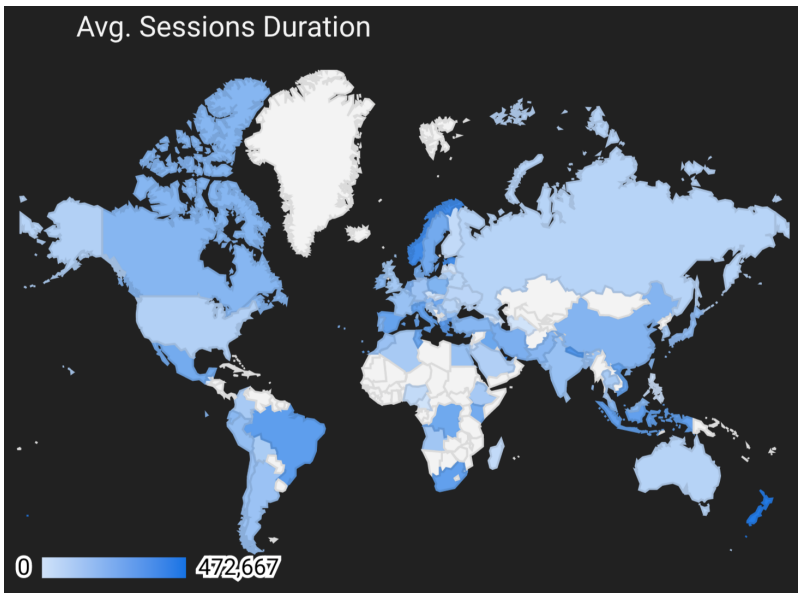

Figure 5.4.2. World Countries accesses Map Overlay

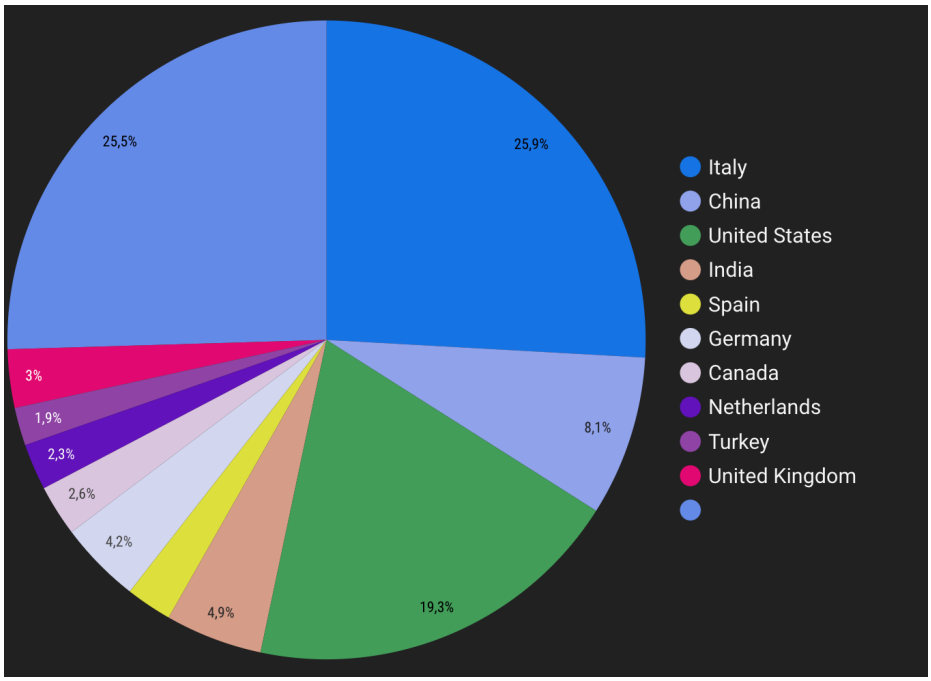

Figure 5.4.3. Top 10 World countries accesses distribution

6 Conclusions

This document reported the main components related to the e-infrastructure VA of SoBigData RI. It highlighted the main results in terms of registered users and accesses and highlighted some critical aspects that emerged in the first year of SoBigData++ project. All the data are compared with 2019, last year of SoBigData project.

As reported in Section 4, the catalogue products and the overall structure of SoBigData Gateway have been evaluated. This process will lead to the new version of the Gateway expected for mid-2021, introducing several improvements related to the findability of products and a better organization and explanation of the contents, and an improved user experience.

Appendix A. e-infrastructure review result & form

A.1 Overall comments from platform reviewers

Objective 1: Search for Datasets in the catalog

The reviewers found the page easy to use, intuitive and visually pleasant. The average evaluation is Very Good.

To improve:

1. navigation: too many click to download an item and not clear where to click in order to download;
2. the types of items that we can find in the catalogue is not immediately clear;
3. the search box should be on the top of the page;
4. shortcuts for datasets and methods would be useful;
5. a direct link for downloading the dataset should be available in the dataset main page;
6. the link to the gateway should be more visible in the SoBigData website;
7. suggest putting a clear flag that represents the fact that the data are available only upon request.

Other platforms:

- Kaggle datasets: search and filtering options are comparable (with a difference that Kaggle lets you filter by the dataset size and SBD further groups them by exploratories).
- UCI ML Repository: older and less intuitive UI, but more search options as they focus specifically on datasets (for instance, the number of instances and attributes, data type, etc.)

Objective 2: Search for Methods in the catalog

The reviewers found the page easy to use and intuitive. The average evaluation is Very Good.

To improve:

1. navigation: too many click to download an item and not clear where to click in order to download;
2. the types of items that we can find in the catalogue is not immediately clear;
3. shortcuts for datasets and methods would be useful;
4. make clear that this is not a repository of methods in the sense of downloading code, but methods you can run in the platform;

Objective 3: Browse and Search for Exploratories datasets

The reviewers found the page easy to use. The average evaluation is Very Good.

To improve:

1. On the VRE page the catalogue block shows the numbers for all the items. To do: show only the items related to the selected exploratory;
2. In the filtering menu exploratories are named “groups”. To do: consistency in the used terminology;
3. Page organization: the most part of the VRE is taken by (almost) empty news feed, while exploratory-related resources are cramped on the side. To do: give more visibility to resources and less to news-feed;
4. In Migration studies, there is “1 to -1 of -1 items” in the list on the right, but there are zero items;
5. The search from the Exploratory VRE seems not working. Whatever term is searched, the result is always the same;
6. The links under the search input field are not working either.

Objective 4: Browse and Search for Exploratories Methods

The reviewers found the page easy to use. The average evaluation is Very Good.

To improve:

1. On the VRE page the catalogue block shows the numbers for all the items. To do: show only the items related to the selected exploratory;
2. In the filtering menu exploratories are named “groups”. To do: consistency in the used terminology;
3. Page organization: the most part of the VRE is taken by (almost) empty news feed, while exploratory-related resources are cramped on the side. To do: give more visibility to resources and less to news-feed.

Objective 5: Use of Training Resources

The reviewers found the page easy to use. The average evaluation is a mix of Excellent and Sufficient.

To improve:

1. Reduce the number of clicks for selecting and downloading a training resource;
2. Make clear the content of the Catalogue, as it is not intuitive that training materials may be in the same catalogue as datasets and methods;
3. Not clear how to download a resource (click on the resource and then click on “URL” to download it, but this did not feel natural for a PDF file). To do: take directly to the resource.
4. As there are resources that are not in English (page 11 onwards was in Italian, as its name was in English in the “Data Mining and Machine Learning Module”) To do: put a flag to specify the language of the resources.
5. It would be better to add the training resources to the main menu, as the “E-learning_Area” link is a bit hard to find. It’s in the “Go to” menu but could be placed in a more relevant place;
6. To do: improve on-site learning. Overall, SBD provides training materials just as a catalogue of files, while MOOC platforms provide opportunities for on-site learning. It would be more convenient if training resources were accessible within the platform, integrated, not as downloadable files.

Other platforms:
Coursera offers similar functionality but it’s very restrictive and you can’t access many materials.

Objective 6: Using the Workspace (for sharing files and folders)

The reviewers found the page easy to use, they have full control over files and folders and all the necessary functionality. The average evaluation is Excellent.

To improve:

1. Navigation: hard to locate the Workspace button on the dashboard. Add the workspace page to the main menu;
2. “How-to” panel can be confused with the actual buttons for Searching, Sharing, etc. as they are located at the top center, while “How-to” name and controls for the panel are off to the sides;
3. Correct the permission to rename a shared file (only author and admin should be able to rename).

Other platforms:

- Google Drive, Dropbox. Compared to them SBD Workspace platform has way more intuitive controls, and navigation within Workspace seems smoother.
- Microsoft Teams.

Objective 7: Use of Applications
The reviewers found the Applications easy to find and easy to use. The overall evaluation is Excellent.

To improve:

1. Add a “see more” button, as the formatting of the Application panel on the main page dashboard gives the false impression that there are only 4 applications available;
2. Add link to the main menu;
3. Make it clear that the blue boxes below the item description are links to the resources (they seem simple tags);

Objective 8: Importing a Dataset in the Catalogue
The reviewers found the Importing not easy to use. The average evaluation is Sufficient.

To improve:

1. Add the Publish button directly in the catalogue. Very hard to locate the upload option through the workspace;
2. Consider change the mandatory fields (i.e., TimeCoverage field was mandatory but not all datasets have a time coverage)
3. The balloon help text for ProcessingDegree was unhelpful
4. Field/Scope of use and Basic rights interact with the License field and perhaps should be together. For instance, I think if I apply License: Academic Free License then I cannot say I want only non-commercial use.
5. Make it clear that the Creator field has structure. It would be better to ask for Creator Last Name, Creator First Name, Creator Email, Creator ORCID separately;
6. Make it clear what to write in the DataProtection field;
7. Make it clear that you have to upload the dataset to your workspace first;
8. Make it clear that it is not possible to upload a file but only add a URL. Would be better to be able to upload a file;
9. Add a shortcut for uploading, somewhere in the dashboard.

Objective 9: Importing a Method in the Catalogue
The reviewers found the Importing easy to use. The average evaluation is Sufficient.

To improve:

1. Add Publish button directly in the catalogue (Very hard to locate upload option);
2. Not very intuitive that the right-click menu is custom, not the system one;
3. Provide defaults (i.e., Creator and Owner could default to the current user; Creation Date could default to the current date/time);
4. Perhaps Basic rights could be set to all if the license chosen allows us;
5. Make it clear that you have to upload the file to your workspace first;
6. In the ‘add new resource” view add an option for selecting a file from your workspace. In this way, you already know that you should upload the file to your workspace previously. It would be better if I could choose a file from your computer and upload it;
7. The editing interface is different from the upload interface (the order of fields was different). Make it uniform;
8. Make it clear what is expected in each field and his purpose (i.e., what ‘UsageMode’ field and an ‘AccessibilityMode’ field are);
9. When I finish creating an item, I can see an alert box with a confirmation message and a ‘Go back’ button. Change this button to ‘Close’.
10. In the “Publish Item” dialog box the title is modified automatically with the name of the file removing its own title.

Objective 10: Importing a Method in the SoBigDataLab
The reviewers found Importing a complex process and the interface is quite simple considering that complexity. The use of the Tutorial makes things easier. The average evaluation is Sufficient.

To improve:

1. Make fields more intuitive;
2. The 'method importer' let you create a new folder, but it is not possible to upload a file on it;
3. Make it possible to select the required packages from those installed ones;
4. The help section doesn't help;
5. The main file of the program (selected with the 'set code' button') should be highlighted;
6. Not possible to edit a previously created project;
7. Some on-boarding procedure is needed for the first import, and the user would need some additional assistant for this.

Objective 11: Execution of a Method in the SoBigDataLab
The reviewers found the Execution page very easy to use. The use of the Tutorial makes things easier. The average evaluation is Excellent.

To improve:

1. Catalogue of methods needs filters;
2. Interface looks somewhat outdated;
3. In case of error, I would like to see the log, not download it;
4. Add a mandatory field containing the bibliographical information of the related paper;
5. After selecting the input file, the text in the splash window is wrapped in a not good way

A.2 Form provided to the reviewers

Auto-Profiling	
Name and Affiliation	
Describe your area of research and which kind of "user" you are. (anything you consider relevant for this review)	
Define your coding skills (if you have any) and which languages you use for your field of research.	

Objective 1: Search for Datasets in the catalogue	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigDat platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 2: Search for Methods in the catalogue	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 3: Browse and Search for Exploratories datasets	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 4: Browse and Search for Exploratories Methods	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 5: Use of Training Resources	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 6: Using the Workspace (for sharing files and folders)	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 7: Use of Applications	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 8: Importing a Dataset in the Catalogue	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 9: Importing a Method in the Catalogue	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 10: Importing a Method in the SoBigData Lab	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent

Objective 11: Execution of a Method in the SoBigData Lab	
Describe the path followed to reach your objective	
Strong Points	
Weak Points	
Usability and Opportunities given by the platform	
Do you know other platforms which provide the same kind of service (for this objective)? If yes please provide a comparison with SoBigData platform.	
Additional notes which are not well described in the previous fields.	
Overall Evaluation (Select one)	❑ Insufficient ❑ Sufficient ❑ Very Good ❑ Excellent