Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE

Deliverable D7.2

Periodic and Assessment Report on VA Activities 2

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 (60 months)
ENDING DATE	31/12/2024
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	D7.2
DELIVERABLE TITLE	Periodic and Assessment Report on VA Activities 2
AUTHOR(S)	Valerio Grossi (CNR), Michela Natilli (CNR), Roberto Trasarti (CNR), Beatrice Rapisarda (CNR), Ilaria Barsanti (CNR)
CONTRIBUTOR(S)	
EDITOR(S)	Valerio Grossi (CNR), Michela Natilli (CNR)
REVIEWER(S)	Massimiliano Assante (CNR), Ignacio Lamata Martinez (EGI)
CONTRACTUAL DELIVERY DATE	31/12/2023
ACTUAL DELIVERY DATE	16/01/2023
VERSION	V1.3
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	31
KEYWORDS	Virtual Access, e-infrastructure, catalogue, key performance indicators

EXECUTIVE SUMMARY

The VA services related to SoBigData RI are accessible by a portal that allows the user to navigate and discover datasets, methods, and services using real applications and case studies as part of the exploratories developed in WP10.

All resources metadata are accessible through the web interface for free and anonymously, but access to the existing resources of the e-infrastructure requires a free registration (using either ad hoc or academic/social accounts supported by the EOSC portal).

This deliverable D7.2 “Periodic Report Periodic and Assessment Report on VA Activities 2” updates and integrates deliverable D7.1 “Periodic Report on VA Activities 1”. It describes the newly implemented features, the technological upgrade, and a report and discussion on user access to the resources in the period from 1st January 2021 to 31st December 2022.

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

EC	European Commission
EOSC	European Open Science Cloud
EU	European Union
ESFRI	European Strategy Forum on Research Infrastructures
FACT	Fair, Accurate, Confidential, and Trustworthy
FAIR	Findable, Accessible, Interoperable, and Reusable
GA	Grant Agreement
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
KPI	Key Performance Indicators
RI	Research Infrastructure
TNA	TransNational Access
VA	Virtual Access
VRE	Virtual Research Environment

TABLE OF CONTENTS

1 Relevance to SoBigData++

1.1 Purpose of this document

Virtual Access (VA) offers users the possibility to navigate and discover datasets and services employing real applications and case studies. All metadata of the items are accessible through a Web interface for free and anonymously[1]. Access to the existing resources of the e-infrastructure requires 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 for resource usage tracking and statistical purposes. This document reports and updates the current state of the e-infrastructure and the actions in progress to integrate new content and attract new users. In this context, the document includes a detailed discussion of the access of the resources from 1st January 2021 to 31st December 2022, together with a comparison of the performance reported in Deliverable D7.1 "Periodic Report on VA Activities 1"[2].

1.2 Relevance to project objectives

The objective of VA is to offer online services for big data and social mining research. SoBigData RI 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 locally (as downloaded methods). WP3 liaises with dissemination and outreach WPs to attract new potential users for local infrastructure sites. To this aim, the new release of the SoBigData RI website[3] will include a dedicated area to display an updated description of the e-infrastructure capacities (e.g., number and typologies of federated resources, such as methods and datasets) and its current exploitation (i.e., number of registered users, outstanding results, latest blog posts, and so on).

1.3 Relation to other work packages

VA works in strict synergy with WP8, WP9, and WP10, which are responsible for 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 transnational access. Finally, VA can also be used to disseminate 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 WP8 for the datasets, libraries, and services. In the description of those WPs, there are examples of the

[1] https://sobigdata.d4science.org/

[2] Available at: https://data.d4science.net/9uND

[3] Originally expected in spring 2021, it was rescheduled (and re-engineered) for early 2023 due to the entrance of SoBigData RI in the ESFRI roadmap.

resources which will be provided. The tools are available to the users to discover, to search, to use and execute resources being 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 WP7.

1.4 Structure of the document

This document reports on operation activities and updates actions from January 2021 to December 2022. The deliverable contains the following main sections:

- Section 2 describes the service provided through VA by the SoBigData RI in detail.
- Section 3 outlines the KPI and analyses the performances in terms of registered users and access to the different resources. It also compares the performance obtained in the previous and current periods.
- Section 4 presents some conclusions and briefly explains the plan to expand current services.
- Appendix A includes the assessment report of the project advisory board.

2 Online open science VA services

The SoBigData RI supports data science serving a cross-disciplinary community of researchers studying all aspects of societal complexity from a data & model-driven perspective. One of the goals of SoBigData RI is to build a common environment where scientists can create, validate, assess, compare and share their digital scientific results, such as research data and methods, by using a common "digital laboratory" composed of agreed services and tools. SoBigData VA services work towards this direction with a single main entry point[4], in which users can freely interrogate the catalogue and access the RI gateway. The SoBigData Gateway is based on D4Science[5] 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 minimising the technological integration cost highlighted above.

Concerning the version outlined in Deliverable "D7.1 - Periodic Report on VA Activities 1", the gateway has been simplified and includes the link and a short description of all RI-related services. Figure 2.1 shows the new design of the gateway, divided into 4 main areas. In the upper section, the user can find the catalogue menu, containing a search box and shortcuts for navigating the different kinds of items available. The main area is divided into three columns: on the left, the user can access the personal workspace (including also all public folders). In the middle, the access to SoBigData Lab, the training material, and the services related to high-performance computing features can be found. Finally, on the right, the stand-alone applications hosted and served by SoBigData RI are accessible.

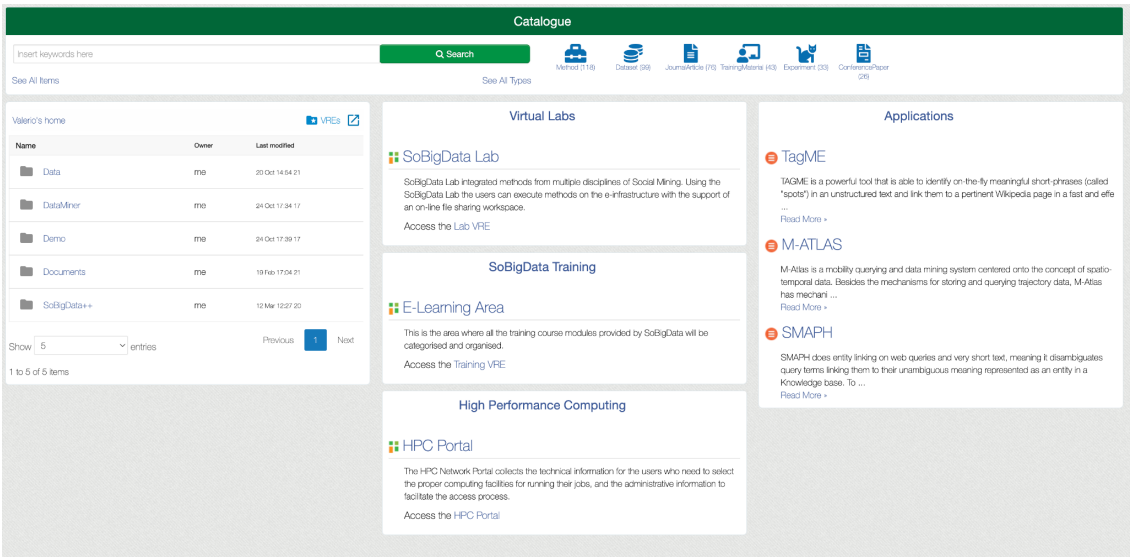


Figure 2.1 - The SoBigData RI Gateway

The following subsections give a short overview of the primary services that support the VA. Extended information can be found in Deliverable "D7.3 - VA e-Infrastructure Service Provision and Operation Report

[4] Accessible at: https://www.sobigdata.eu
[5] https://www.d4science.org/ [1]

3"[6] (and previous [6]), where the overall SoBigData architecture is described, and Deliverable "10.4 SoBigData e-Infrastructure Common Facilities 1"[7], in which several technical details of the SoBigData++ project are introduced.

With the launch of the SoBigData PPP and SoBigData.it projects, we are rethinking all our online services to increase their usability and provide applications in the perspective of open science diffusion. The following subsection proposes an updated overview of the principal services provided by the e-infrastructure and freely available for registered users.

2.1 Catalogue

The catalogue is a tool for finding and exploring datasets, methods, applications, experiments, and publications through different navigable views. All the elements inside SoBigData RI are discoverable through this service. The complete description of each item is provided on the dedicated page. These features can be added to the search filter, which will be recalculated in real-time, and search results can be sorted alphabetically concerning the insertion date or popularity.

The catalogue enables users to search for an item given a set of keywords, and filters can be used to find products more efficiently and access methods, such as those related to each exploratory. Currently, the catalogue is divided into two main areas:

- ***SoBigData Services and Products***, which permits access to datasets, methods, experiments, and applications.
- ***SoBigData Literacy***, which enables users to find and freely download papers related to the SoBigData community[8].

Given an item, the catalogue shows different information and facilities (*e.g.,* accessibility features, some basic rights on the usage, and the creation date, to mention a few) and a set of refereed resources such as PDF files or references to the method engine. Figure 2.1.1 presents the main page of the catalogue, where a user can follow the different links available for searching items using free text. The new version can harvest catalogues of different RIs powered by D4Science, as in the case of *Territori Aperti*[9].

[6] https://data.d4science.net/stH9
[7] https://data.d4science.net/sjUu
[8] Unregistered users cannot access the resources but only see the metadata related to an item.
[9] https://territoriaperti.d4science.org

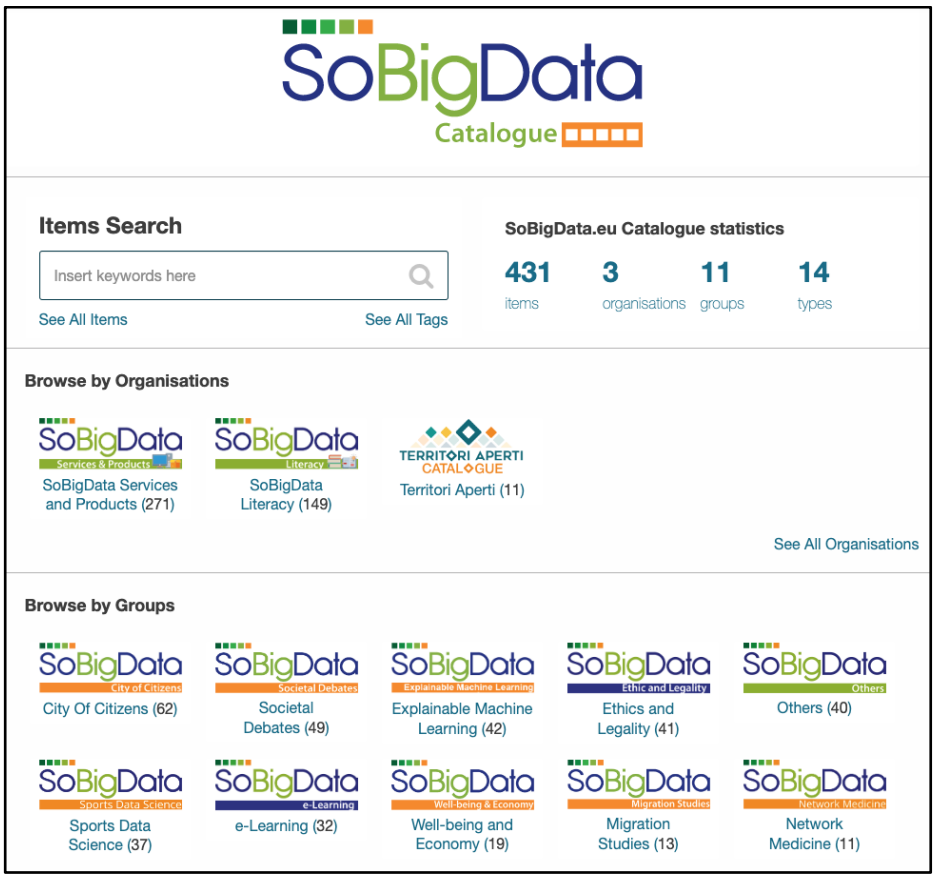


Figure 2.1.1 SoBigData Catalogue main page

The new release of the catalogue (expected in early 2023) will introduce a moderation workflow to enable all registered users to publish items into the catalogue. The introduction of this new feature requires an update to the defined roles for catalogue management. Any catalogue user has one or more of the following roles:

- ***Catalogue-Member***: a user with such a role is mainly capable of listing and reading items. Additionally, the new catalogue version allows users to start the moderated-publication flow for adding a new item to the catalogue. This role is assigned automatically to all registered users.
- ***Catalogue-Editor***: a user with such a role can manage the items that he/she creates and using other APIs. This role is reserved to the management team of the catalogue and to some selected people inside the consortium of SoBigData++.
- ***Catalogue-Admin***: a user with such a role can configure and managing the catalogue. This role is reserved for management users only.
- ***Catalogue-Manager***: a user with such a role can use every APIs exposed by the service except item moderation APIs (*e.g.,* approve, reject, …). This role is reserved for the management people and developers inside the consortium.
- ***Catalogue-Moderator***: a user with such a role can invoke the item moderation procedure. Only the VRE Manager can assign roles to VRE users, and only some selected users from the consortium will be involved as moderators.

The roles Catalogue-Editor, Catalogue-Admin, and Catalogue-Manager are configured in a hierarchy, meaning that the Manager role encompasses the Admin role, and the Admin role encompasses the Editor role.

In the moderation process, the *Catalogue-Moderator* must approve any submitted items to make them available to the other users of the catalogue. An item can be in the following states:

- ***Pending***: the item has been submitted by an author (a Catalogue-Editor or above) but is still not available to other users of the catalogue. A Catalogue-Moderator has to approve or reject it.
- ***Approved***: the publication of the item has been approved by a Catalogue-Moderator, and can be seen by other users of the catalogue.
- ***Rejected***: the publication of the item has been rejected by a Catalogue-Moderator.

2.2 Virtual Research Environments

Online services of SoBigData RI are served by "special containers" called Virtual Research Environments (VREs). VREs are web-based working environments built to support communities' needs by assembling and exposing the resources required by their scientists and users. D4Science VREs provide users with domain-specific resources and are equipped with core services supporting data analysis and collaboration. Users subscribe to specific VREs, which offer different services such as sending private messages, interacting directly with the community of the exploratory, and having private and public storage for sharing documents, data sets, and results. A VRE can be seen as a Virtual Laboratory, a dedicated area where users may share information and experiences focused on a specific topic.

2.3 Workspace

The workspace is a cloud storage and online environment to support secure and controlled data storage and sharing. Each VRE has a dedicated 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 access to a private space to store data and documents and public spaces (one for each subscribed VRE), to share files. Access to these workspaces is granted to the user upon registration to SoBigData RI.

2.4 SoBigData Interactive Programming environment - SoBigData Lab

SoBigData users develop their methods on Jupyter Notebook[10] using cutting-edge libraries developed by experts. Using SoBigData Lab, users can also execute methods on the e-infrastructure with the support of an online file-sharing workspace. This VRE will integrate different methods that can be invoked under the same environment through SoBigData e-infrastructure. A method is the implementation of an

[10] https://jupyter.org/

algorithm/procedure or an algorithm that requires an engine to be executed. Different kinds of integration are available based on the programming language in which the method is implemented. Once a method is integrated into the platform, the final user has a homogeneous web form for inserting parameters and invoking methods independently from the programming language employed.

SoBigDataLab is linked and accessible through the platform gateway, and its items are linked through the catalogue. This environment enables a user to Execute an Experiment, Check the status of started computation, and access the DataSpace to get the results. The SoBigDataLab allows the scientific community to make its methods available. Integrating a new method into the e-infrastructure must be as simple as possible; otherwise, the users may be discouraged from making their methods available. By clicking on the service method importer, the user has 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 project's second period, this VRE has been updated in several directions:

- The method importer has been enriched to support 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 of D7.1 for further details).

Several new integrations have been done from the previous period, and now the method engine includes 103 methods fully integrated and executable through the RI computational node, while the libraries available in JuperHub include:

- *NDlib*, a Network diffusion library that contains a set of models to simulate disease spreading and opinion dynamics.
- *CDlib*, a library that allows users to extract, compare and evaluate communities from complex networks.
- *DyNetX,* a library that extends networks with dynamic network models and algorithms.
- *BiCM*, a Python module for calculating an entropy-based null model for the analysis of bipartite networks. It also includes the unbiased projection of the information contained in the bipartite network on one of the two layers.
- *Scikit-Mobility*, a library for human mobility analysis in Python.
- NEMtropy, a library that collects different entropy-based methods for analyzing complex networks, provides an unbiased and flexible benchmark for network analysis.

2.5 Training: SoBigData Academy

The current version of the e-learning area available in the gateway enables users to access all training material, courses, lectures, and tutorials produced by the SoBigData community. It also makes freely available the material related to the Master in Big Data Analytics and Social Mining[11] held at the University of Pisa, Italy.

It is planned for 2023 to complete the training offer by providing more advanced products based on the Moodle[12] platform, such as a set of online modules for the training of a responsible data scientist. Moodle will enable the SoBigData RI community to publish more interactive and structured lessons and courses. This work is done in collaboration with WP4 "NA3 - Training" and WP9 "JRA2 - E-Infrastructure and Supercomputing Network". Currently, the authentication is being implemented to use SoBigData user accounts, and the functionalities related to the role of Moodle Teacher is being tested.

2.6 Applications

The SoBigData gateway provides a set of applications powered by a stand-alone VRE. One of the most noteworthy examples is the *TagME application, which* includes a complete set of tools for tagging text. From the TagMe - The Entity Linking tools by Acube lab[13], the user accesses the SoBigData RI powerful entity-linking functionalities[14], and can access the API documentation pages and share the experience in the news feed. TagMe was designed and published in 2010 (ACM CIKM, IEEE Software), and it is probably the most famous and used public entity-linker worldwide. To date, it has served over 800 million queries with peaks of tens of million queries per month. TagMe is currently used in several other software tools; one notable example is *Ruby Star*, which has been shortlisted among the finalists of the Alexa Prize 2017[15]. *WAT* is a sophisticated evolution of TagMe based on the Wikipedia Knowledge Graph and some algorithms based on Word2Vect and Enty2Vect techniques. *WAT* significantly improves TagMe's efficacy and efficiency (ERD@SIGIR 2014) [2] over well-formed texts. *SWAT* is an add-on of TagMe and WAT that adds saliency evaluation to entity linking so that one can not only annotate an input document with entities drawn from Wikipedia (à la TagMe or WAT), but also can assign a saliency score to every annotated entity that could be subsequently used for specific post-processing tasks, such as document clustering and classification, entity filtering, *etc.* (NLDB 2017) [3]. SMAPH is an entity-annotator for open-domain web search queries. It is based

[11] https://masterbigdata.it/

[12] Moodle is a Learning Platform or course management system, a free Open-Source software package. https://moodle.org/

[13] http://acube.di.unipi.it/

[14] https://sobigdata.d4science.org/web/tagme/service-overview

[15] https://www.amazon.science/alexa-prize

on a second-order approach that, by piggybacking on a web search engine (either Bing or Google), alleviates the noise and irregularities that characterise the language of queries and places queries in a larger context in which it is easier to make sense of them (WWW 2016) [4]. SMAPH was the winner of the short track in the ERD@SIGIR2014 benchmark (ERD@SIGIR 2014) [5]. Figure 2.6.1 shows the overall picture of all the tools available through the VRE.

Figure 2.6.1. The Entity Linking tools[16]

[16] Current and former members of this lab who contributed to developing and deploying these services include Paolo Ferragina, Marco Cornolti, Francesco Piccinno, Marco Ponza, Ugo Scaiella and Daniele Vitale.

3 Evaluation of the VA Performance

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*, and *O2: Expanding the multidisciplinary community*. This section reports the evaluation of VA performance by analysing several indicators related to the Virtual Access service provided by the SoBigData RI **during the period from 1st January 2021 to 31st December 2022**. SoBigData RI has been available since 2016 and, we report (when possible) a comparison between the indicators available at the end of the previous project SoBigData GA 654024 ended the 31st of December 2019[17], and the VA performances reported in Deliverable D7.1 "Periodic Report on VA Activities 1" for the period from 1st January 2020 to 31st December 2020.

The reported indicators are collected automatically by the platform and reported through dashboards in a dedicated portal for administrators[18]. 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 1st of January 2020, the dashboard also includes another service called "datastudio", where geospatial access statistics are available in the SoBigData Gateway both on European and global level.

3.1 Users/Accesses statistics

In this section, the total number of users registered at the gateway is reported. 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 publicly available. Figures 3.1.2 and 3.1.3 report the number of users registered to SoBigData e-infrastructure. Considering the period from January 2021 to December 2022, it can be stated that 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 period of the project reflects the previous ones observed for SoBigData and SoBigData++ projects. The increased number of beneficiaries from SoBigData to SoBigData++ has provided a positive impact of SoBigData RI dissemination in Europe, directly impacting the number of registered VRE users. This trend is influenced by the CoVid-19 that enforces online products; the number of users is almost twice with respect to the end of the first SoBigData projects. Of course, this side effect partially covers the lack of face-to-face dissemination events, training courses, and transnational visits that typically provide an increment of registered VRE users into the platform. At the end of December 2019, the registered VRE users were 4,814, increasing to 7,284 users by the end of 2020 and 10,292 by the end of 2022, an increment of 41% compared to the previous period (more than double if compared to December 2019). The positive trend is extremely evident when looking at Figure 3.1.1 in which the number of users from January 2018 is shown, when the gateway was considered stable and functioning. The same trend is noticeable, also considering the users subscribed to different VREs. Figures 3.1.2 and 3.1.3 show the positive

[17] For a complete description of the indicators related to the previous period of SoBigData RI, see SoBigData Deliverable 7.3 "D7.3 - VA e-Infrastructure Service Provision and Operation Report 3", https://data.d4science.net/stH9.
[18] https://sobigdata.d4science.org/group/sobigdata/accounting-dashboard

trend of VRE users registered on the SoBigData Gateway and their distribution across the VREs implementing the e-infrastructure related services. The data confirm that the Catalogue and the SoBigDataLab & Services remain the VREs with the most registered users.

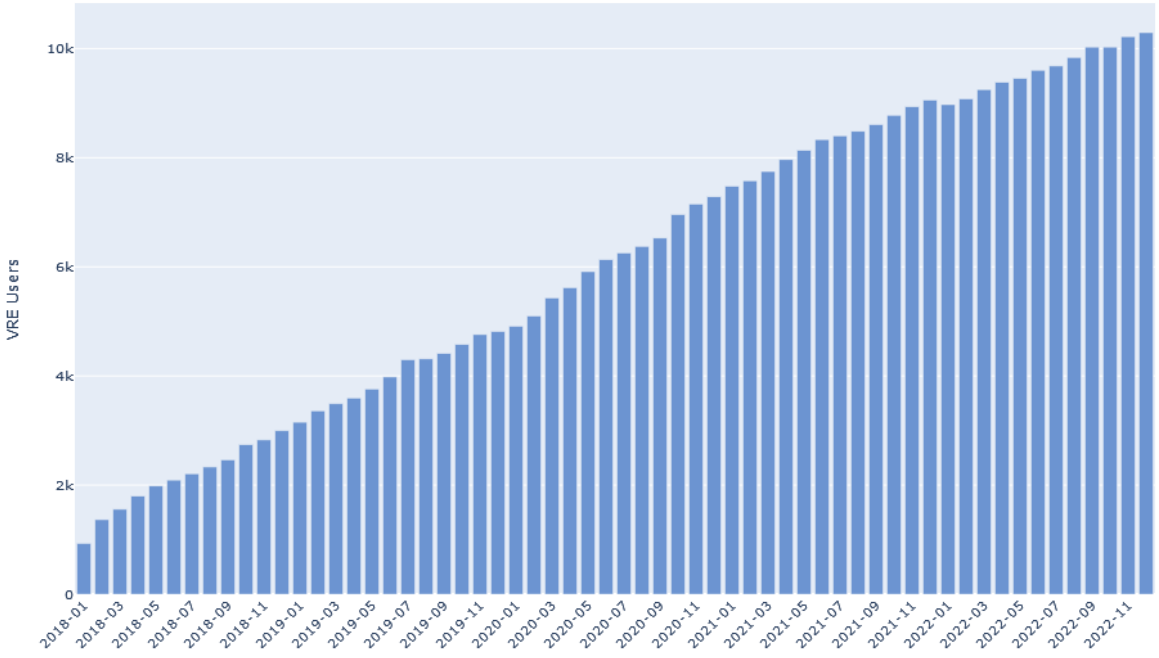


Figure 3.1.1 User registered in the gateway from January 2018

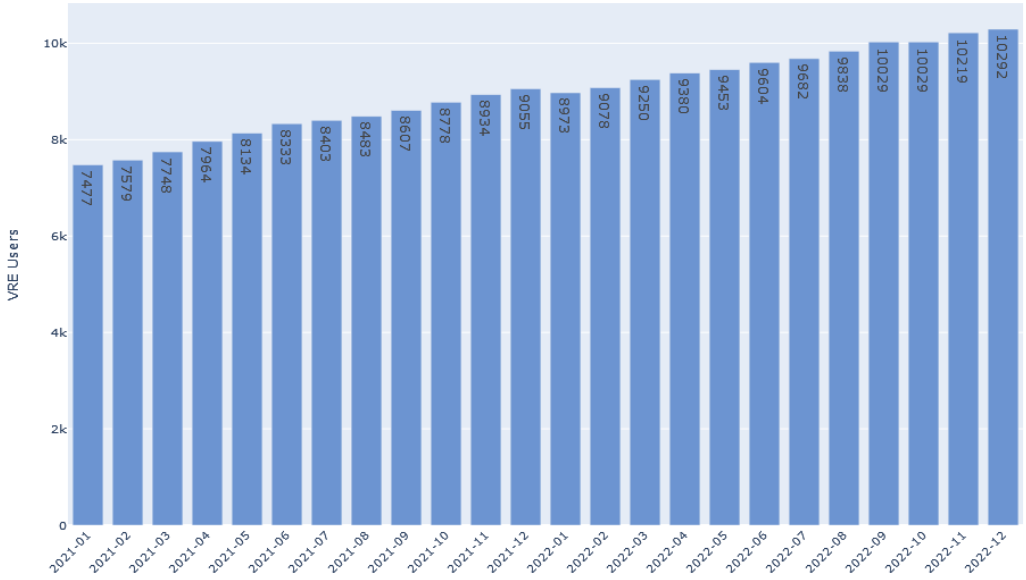


Figure 3.1.2 User registered in the gateway in the reporting period 2 (2021 - 2022)

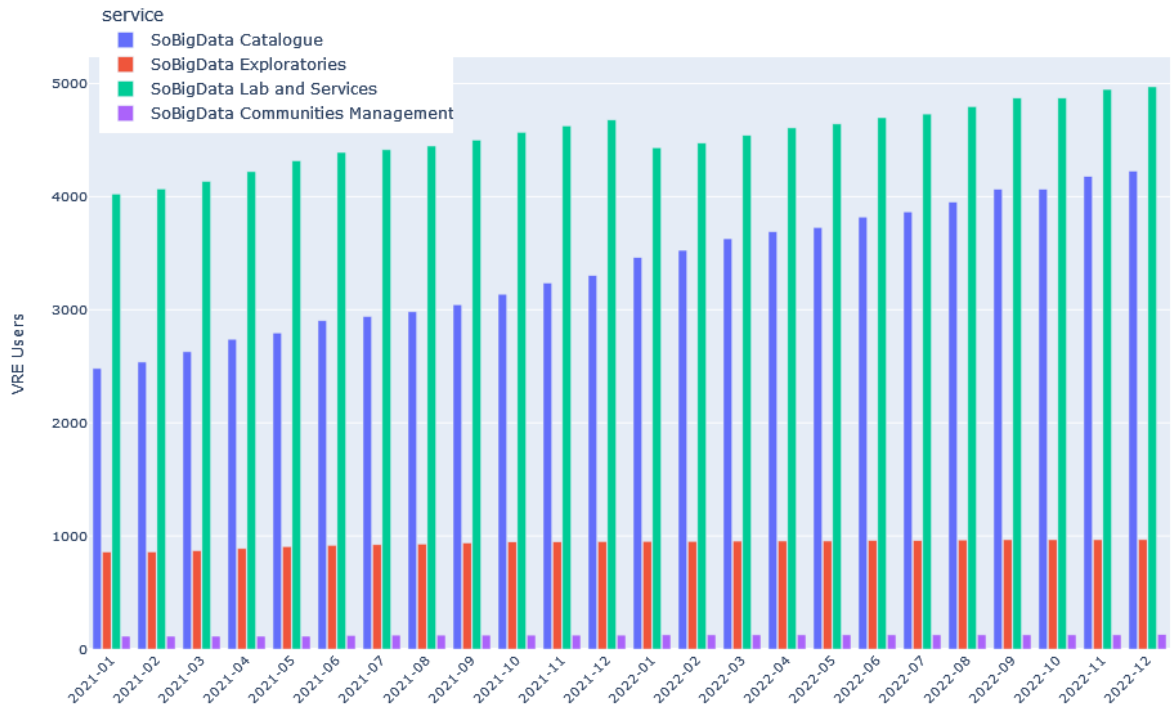


Figure 3.1.3 User registered in the gateway (all the VRE users) and Distribution of the users in the different parts of the platform

Figure 3.1.4 (first graph) reports the number of accesses to the gateway and the relative distribution on the different VREs of the platform. Again, it can be observed a quite stable number of accesses to the platform. In 2021, the SoBigData RI observed 1,605 monthly average accesses, while the accesses on average for 2022 is 1,232. Both numbers are greater than the average for 2020 (1,195 accesses) registering an increment of 34% from 2020 to 2021 and a small decrease from 2021 to 2022 (23%). From these numbers, it can be seen the impact of the introduction of new services such as JupiterHub, and the integration of new libraries from 2020 to 2021. Also, Figure 3.1.4 (second graph) confirms this trend, showing that the introduction of SoBigData Lab and Services heavily influenced the total number of accesses.

3.2 Catalogue Statistics

The catalogue is the searching tool of the RI. As reported in Table 3.2.1, the SoBigData catalogue contains 313 resources: 102 datasets, 123 methods, 42 training materials and publications, 10 applications and 36 experiments. There are 203 online and 81 onsite resources (i.e., reachable only by an on-site visits), reporting the fact that the online resources are almost twice with respect to the ones available in the previous period. The table also reports the breakdown of the resources based on the exploratory. It is important to highlight that the catalogue provides links for navigating all the products considering a specific exploratory. The items are currently organised into three main organisations with the harvesting of the *Territori Aperti* catalogue, 11 groups, and can also be browsed by 13 main types.

Table 3.1 Number of integrated resources grouped by exploratories. On brackets the values from the previous period.

	Datasets	Methods	App. + Exp	Train	***Total***
Sustainable Cities for Citizens	19 (17)	25 (15)	5 (1)	4 (1)	***53** (34)*
Societal Debates and Misinformation Analysis	27 (25)	1 (1)	13 (5)	1 (1)	***42** (32)*
Demography, Economy & Finance 2.0	6 (3)	5 (1)	4 (0)	1 (1)	***16** (5)*
Migration Studies	2 (1)	7 (6)	2 (0)	1 (1)	**12** (8)
Sport Data Science	6 (5)	9 (5)	14 (4)	2 (1)	***31** (15)*
Social Impacts of AI and Explainable Machine Learning	6 (6)	7 (2)	0 (0)	0 (0)	***13** (8)*
Network Medicine	4 (-)	4 (-)	3 (-)	1 (-)	***12** (-)*
Generic	32 (35)	65 (53)	5 (6)	32 (26)	***134** (120)*
Total	*102 (92)*	*123 (83)*	*46 (16)*	*42 (31)*	*313 (222)*

A total of 533 accesses have been registered, with a monthly average of 45 for 2019. A slight decrease in 2020 has been observed, with 434 accesses (36 on average monthly), while 506 and 533 accesses were registered respectively for 2021 and 2022.

Our catalogue responded to more than 61k queries (see Figure 3.2.1) for the biennium 2021/2022 with an average of 2,549 queries per month (an increase of 6% compared to the previous reporting period).

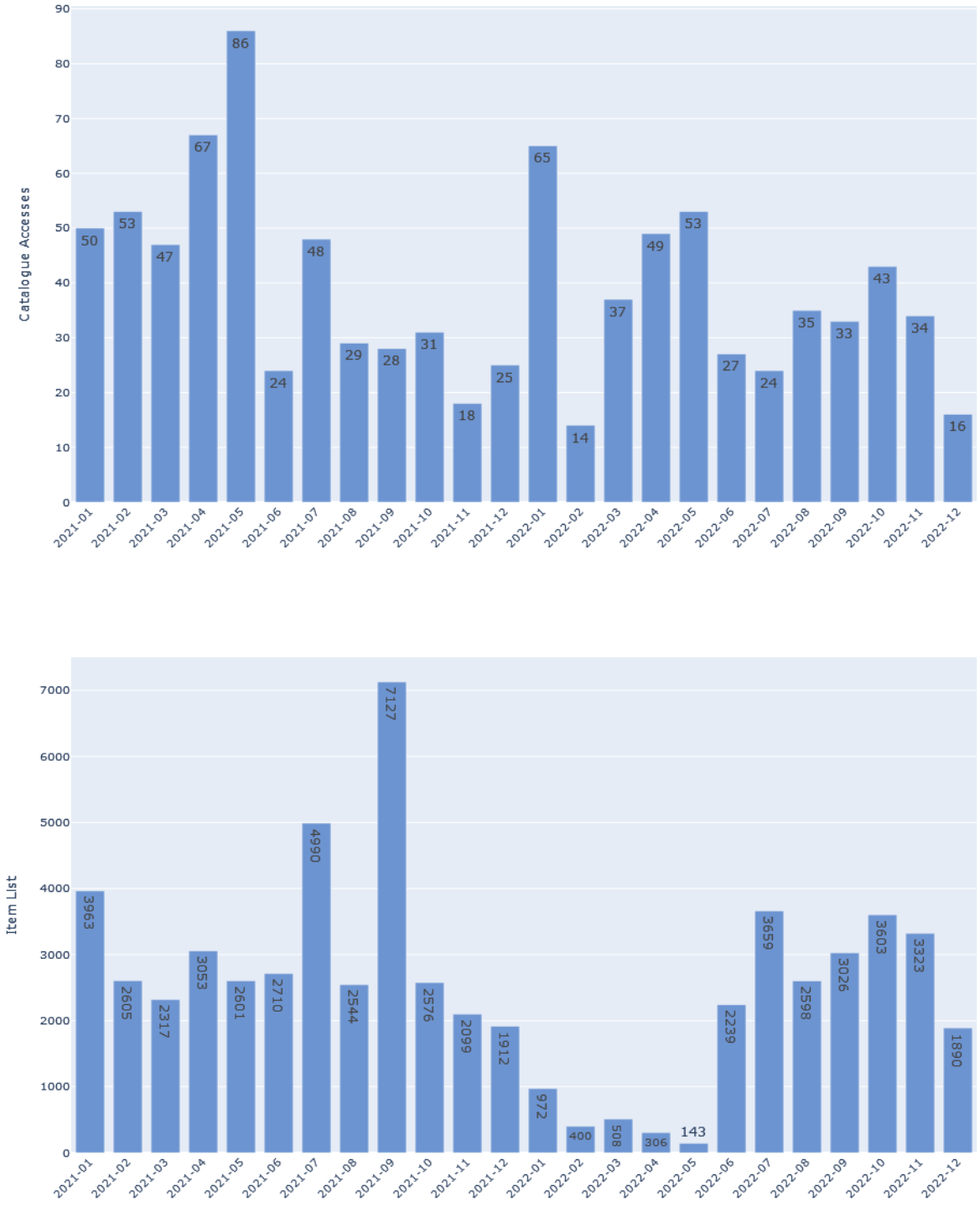


Figure 3.2.1. SoBigData RI Catalogue accesses and queried items list

3.3 SoBigData Lab and Methods Invocation Statistics

JupyterHub is a service offered in SoBigData Lab since December 2020, and it allows the execution of Jupyter notebooks, providing users with access to computational environments and resources of the e-infrastructure.

Figure 3.3.1 presents the number of accesses in the last biennium, having 1,664 and 1,796 accesses respectively for 2021 and 2022 (on average 139 for 2021 and 150 for 2022), with an increase of around 8% from the two years.

Looking at the method invocations, for the biennium 2021/2022 there were almost 500 millions of requests with a monthly average of 19,417,509 invocations.

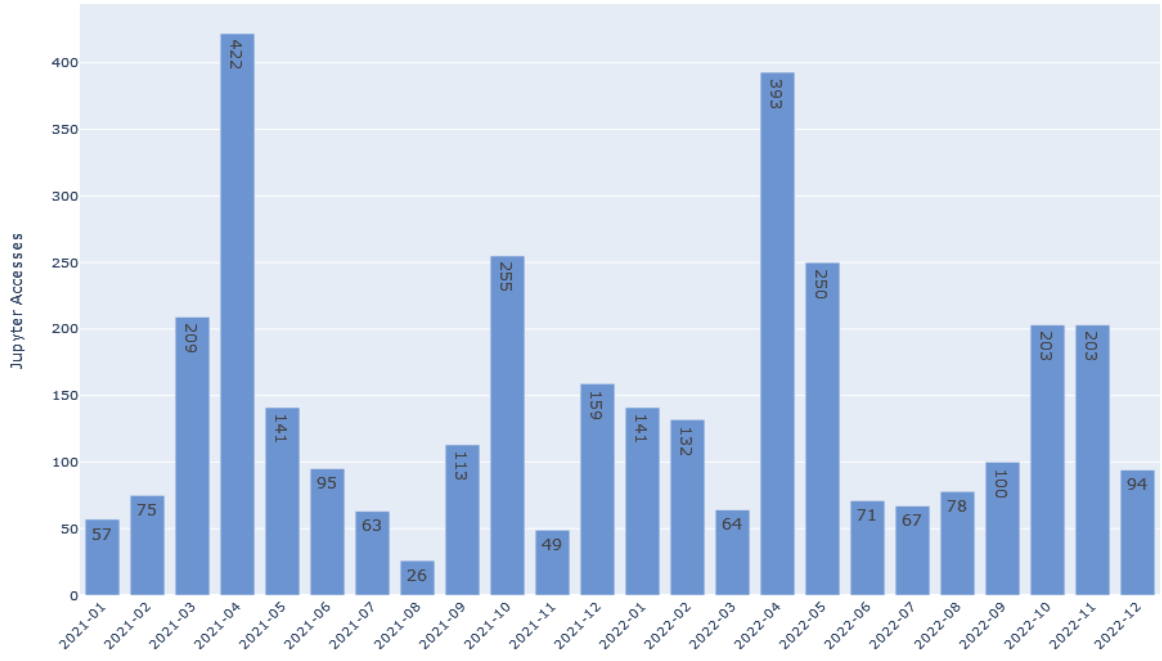


Figure 3.3.1 JupyterHub Accesses for 2021 and 2022

3.4 Applications Statistics

As explained in Section 2, the catalogue comprises different areas (VREs) containing some applications (namely TagME, SMAPH, M-Atlas), the SoBigDataLab and an e-Learning Area. Access metrics to these catalogue areas for 2021 and 2022 are depicted in Figure 3.4.1.

Among the applications it can be seen how TagMe plays a predominant role with a number of accesses exceeding 7,000 for both 2021 and 2022, with an average number of monthly accesses of 633. Also significant is the impact of SoBigDataLab with a total number of accesses of 9,607 for the two-year period, corresponding to an average number of monthly accesses of 400.

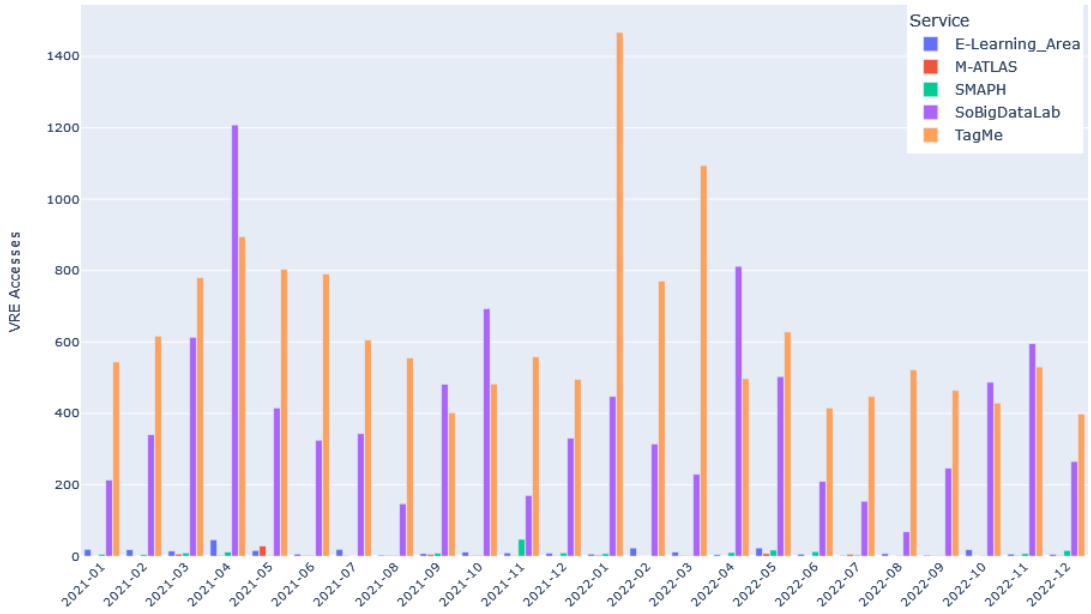


Figure 3.4.1 VRE Accesses by type of application for 2021 and 2022

3.5 Geographical location of accesses

This section reports the geolocation of the user accesses considering the period from January 2021 to December 2022.

Figure 3.5.1 reports the access from European Countries. It is noticeable that users access the SoBigData RI from all countries in Europe and not only from those participating in the consortium. For example, users from Norway, Denmark, Portugal, and Poland access the services but are not represented by any institution in the SoBigData++ consortium. This shows how the platform can be relevant for the whole of Europe in the context of Social Mining, Big Data analytics and Artificial Intelligence.

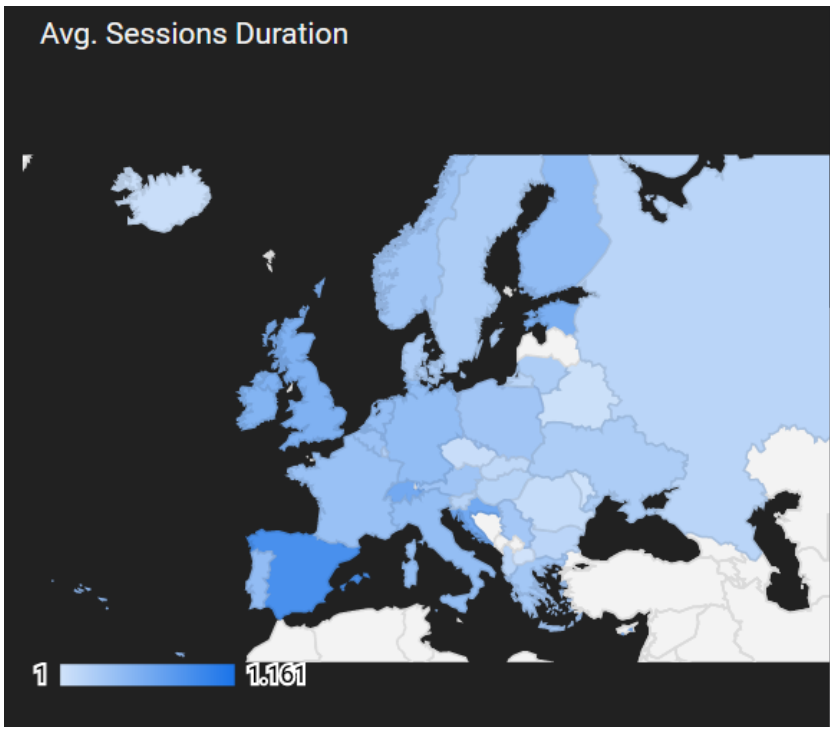


Figure 3.5.1. European Countries accesses Map Overlay

Figure 3.5.2 shows a World Map Overlay of user access to the e-Infrastructure Gateway. The map shows that the e-infrastructure users come from all the continents, especially from Europe and Asia, followed by North America.

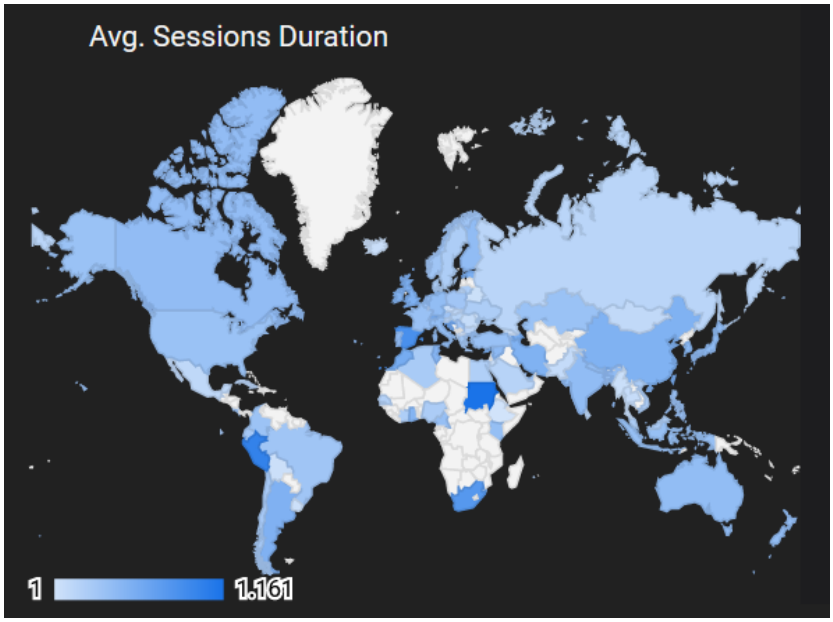


Figure 3.5.2. World Countries accesses Map Overlay

To better understand the distribution of the accesses, a pie chart has been created in Figure 3.5.3, which shows the distribution of the top 10 accesses by country. More than 60% of the accesses come from three countries: Italy, the Netherlands and China. This high number of accesses from Italy is not surprising, as the project consortium includes four institutions from Italy, and the number of TNA visits is mainly in Italy due to the beneficiaries and research groups involved.

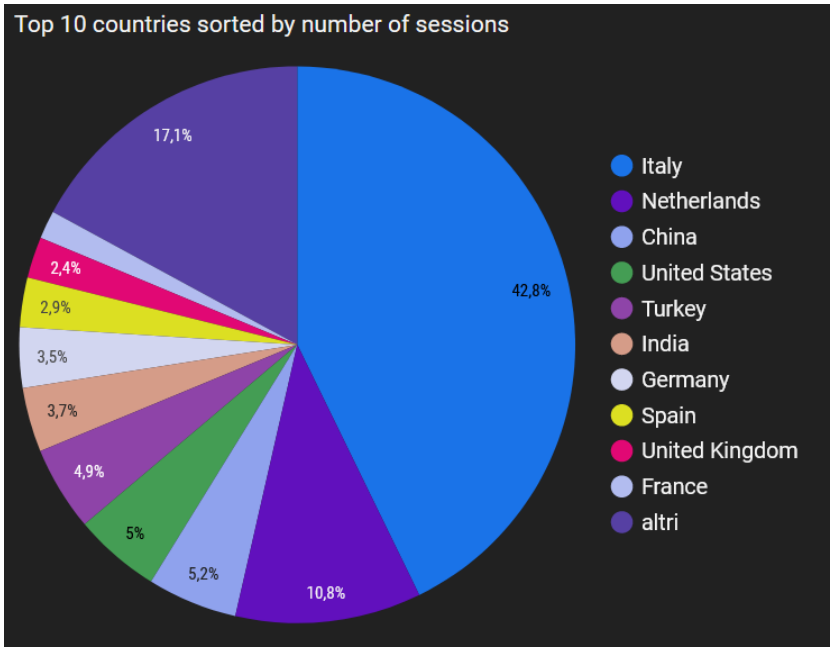


Figure 3.5.3. Top 10 World countries access distribution

4 Conclusions

The main components related to the e-infrastructure VA of SoBigData RI have been described in Section 2, which include the product catalogue, the VREs, the workspaces, the interactive programming environment, the training facilities, and the applications. A new version of the gateway has been deployed in mid 2021, which introduced improvements in terms of findability of products, content organisation and description and user experience.

Usage metrics and indicators of the e-infrastructure have been presented in Section 3, highlighting the main results in terms of registered users and user access. These data have been compared to those of the previous period and are actively used to suggest areas of focus and implement future improvements to the infrastructure.

The evaluation statistics of VA is a crucial part of understanding the usage of the services related to the e-infrastructure, as they reveal which tools and resources are used the most, how many users access them and from which countries; and they are an important indicator of the success of the RI itself.

References

1. M. Assante et al. *Enacting open science by D4Science*. Future Gener. Comput. Syst. 101: 555-563 10.1016/j.future.2019.05.063, 2019
2. F. Piccinno, P. Ferragina: *From TagME to WAT: a new entity annotator*. ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia. ACM 2014, ISBN 978-1-4503-3023-7, 55-62, 2014
3. M. Ponza, P. Ferragina, F. Piccinno. *Document Aboutness via Sophisticated Syntactic and Semantic Features*. Natural Language Processing and Information Systems - 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings. Lecture Notes in Computer Science 10260, Springer 2017, ISBN 978-3-319-59568-9, 441-453, 2017
4. M. Cornolti, P. Ferragina, M. Ciaramita, S. Rüd, H. Schütze. *A Piggyback System for Joint Entity Mention Detection and Linking in Web Queries*. Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016. ACM 2016, ISBN 978-1-4503-4143-1, 567-578, 2016
5. M. Cornolti, P. Ferragina, M. Ciaramita, S. Rüd, H. Schütze.. *The SMAPH system for query entity recognition and disambiguation.* Natural Language Processing and Information Systems - 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings. Lecture Notes in Computer Science 10260, Springer 2017, ISBN 978-3-319-59568-9, 25-30, 2017
6. R. Trasarti, P. Pagano, C. Falchi, V. Grossi, B. Rapisarda. SoBigData - VA e-Infrastructure service provision and operation report 1, https://openportal.isti.cnr.it/doc?id=people_____::57a8c45ad867f2bd7ed8d57a274ac81f, 2017

Appendix A. Project Advisory Board report

SOBIGDATA
RESEARCH INFRASTRUCTURE

Amsterdam, 29 June 2022

SoBigData++ Project Advisory Board assessment & recommendations report

In this report, the Project Advisory Board (PAB) reviews the progress of the SoBigData++ project from January 2020 to June 2022. The document highlights the evaluation of some main results obtained by the project during the reference period and the recommendations related to project execution and subsequent actions.

The main objective of the H2020 SoBigData++ project is to consolidate the SoBigData research infrastructure (RI) for the social mining and big data Ecosystem. The aim is to deliver an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining which refer to various dimensions of social life. The Advisory board assesses the high quality of the organized events and research production of the project. SoBigData++ is a timely, excellent and very stimulating project with great merits in research, teaching, community building.

In this context, particular attention has been devoted to all the events done by the project and the proposed planning for future ones. For the past ones, we can highlight the following training and dissemination events:

- The **12th International Conference on Social Informatics (SocInfo 2020)** took place in Pisa, Italy, from 6 to 9 October 2020. SocInfo is an interdisciplinary venue for researchers from Computer Science, Informatics, Social Sciences, and Management Sciences to share ideas and opinions and present original research on studying the interplay between socially-centric platforms and social phenomena.
- The **Real-Time Epidemic Datathon** took place virtually from 6 April to 31 May 2020, aimed at joining forces to develop real-time and large-scale epidemic forecasting models.
- The **SoBigData Summer School on Machine Learning of Dynamic Processes and Time Series Analysis** took place in Pisa at Scuola Normale Superiore from 26 to 27 November 2020. The School aimed to present recent developments in Machine Learning, focusing on data-driven approaches to statistical learning and dynamical systems.

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

1 / 6

- The **International Forum on Digital and Democracy**, which was supposed to take place in Venice, ended up as an online event due to covid19 pandemic. The two-day conference aimed to be a meeting place between politics and academics, promoting international collaboration through exchanging information, ideas, and best practices.
- The scheduled **SoBigData++ Conference at IFDaD 2022** - second edition, November 17th - 18th, 2022. A virtual and on-site event in Rome, Italy.

The following event toward industry:

- The **1° Challenge Us** initiative took place in 2021. The program allows companies to further benefit from the data they share for analysis purposes. The program will run for the next three years and will help companies by providing free-of-charge service to accomplish the POC for their proposals.

and 5 Awareness Panels about data protection and legal and ethical issues.

Furthermore, considering the VA and TNA services, the number of published papers and the items (datasets, methods, training material) published in the catalog, the project is reaching good scientific results in the reference community. The VA services has been assessed with a special attention on the introduction of JupyterHub inside SoBigData Lab VRE.

The project demonstrated to be active also on sustainability issues and now has a plan for releasing a reliable and lasting RI with the inclusion of the SoBigData RI in the ESFRI RoadMap 2021. This aspect shows that sustainability actions are essential to providing a clear vision of the future of the SoBigData RI and the future development of the SoBigData++ project results. The EU launched a specific call to support the preparatory phase of new ESFRI research infrastructure projects identified in the ESFRI RoadMap 2021. A new project called SoBigData PPP will start in October 2022. The main aim of SoBigData PPP (Preparatory Phase Project) is to guarantee that the RI becomes sustainable, remaining an open science research infrastructure funded by countries (members of the future ERIC) and partially with services toward the EU market/projects. The ERIC allows the establishment and operation of new or existing Research Infrastructures on a non-economic basis. Synergies between SoBigData++ and SoBigData-PPP are expected and necessary.

Recommendations

Even though the results of the project are excellent and in line with the scientific expectation, this board believes that the consortium should give some special attention to the following aspects:

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

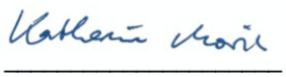
- Considering the growth of the SoBigData++ community and the next steps for the sustainability of the infrastructure, reaching relevant stakeholders, including policymakers, MEPs, civil society organisations, NGOs, government officials, and representatives of other organisations, is considered an act of high importance for the future of the RI. Thus, there is a need to improve some communication actions to valorize the RI and its resources. For example, multiple access links could be introduced to find material from different places.
- The user experience of a website is crucial to involve people and make available the RI features easily. For this reason, it is crucial to improve the user experience on the SoBigData.eu website as the access points of the e-infra. At the moment, it is not easy to find all the various services provided by the RI and reach the resources available. We recommend making it more explicit where to find information by enhancing the structure of the home page and the main menu to give more accessible access to the resources. For example, at the moment, it is impossible to find (on the homepage nor in the main menu) any info regarding the Challenge Us 2021 Program (or a report or video on the event). We recommend giving easy access to the Calls program by having a page with a list of past and future calls.
- The micro-projects are a very interesting way to narrate the research done within the project, and they should be also used by making them visible on the website promoting their outcome and results. We recommend bringing them out on the homepage to give them the visibility they deserve.
- Considering the next steps that SoBigData RI will face, we also recommend having a section more service oriented and providing data privacy services: i.e., if some company or organization donates data, SoBigData RI can support them by providing anonymization.
- Regarding networking, SoBigData RI should give more visibility to the connections and collaborations you have with other EU-funded projects and RIs by displaying them on the website. SoBigData++ project should also consider advertising the TNA program at the leading conferences in the field, like the KDD conference or similar.
- Given the good number and quality datasets and methods the RI got, their visibility should be improved by doing some communication actions like enhancing the use of Twitter and other social media to show off the resources of the e-infra. An overall consideration is that the importance of the RI and the resources could not be very intuitive for stakeholders other than researchers in these fields. For this reason, we recommend creating use cases to show how to use the RI resources.
- Finally, introducing the SoBigData Award for Diversity and Inclusion is a great idea, and we suggest promoting it now, starting from the ECML PKDD Conference.

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

SOBIGDATA
RESEARCH INFRASTRUCTURE

The Advisory Board Members:

Prof. ***Katharina Morik***
University of Dortmund



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

4 / 6

SOBIGDATA
RESEARCH INFRASTRUCTURE

Prof. ***Yücel Saygın***
Sabanci University

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