Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D9.2

# SoBigData e-Infrastructure Operation Report 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<br>Integrating Activities for Advanced Communities |
| GRANT AGREEMENT N. | 871042 |

| DELIVERABLE INFORMATION | |
|---|---|
| WORK PACKAGE | WP9 JRA2 - E-Infrastructure and Supercomputing Network |
| WORK PACKAGE LEADER | CNR |
| WORK PACKAGE PARTICIPANTS | BSC, EGI, Nubisware, OpenAIRE, USFD, UNIPI, FRH, UT, LUH, AALTO, ETH Zürich, TUDelft |
| DELIVERABLE NUMBER | D9.2 |
| DELIVERABLE TITLE | SoBigData e-Infrastructure Operation Report 2 |
| AUTHOR(S) | Massimiliano Assante (CNR), Leonardo Candela (CNR), Roberto Cirillo (CNR), Andrea Dell'Amico (CNR), Luca Frosini (CNR), Lucio Lelii (CNR), Francesco Mangiacrapa (CNR), Pasquale Pagano (CNR), Giancarlo Panichi (CNR), Tommaso Piccioli (CNR) |
| CONTRIBUTOR(S) | Marco Lettere (Nubisware), Mauro Mugnaini (Nubisware), Enol Fernandez (EGI), Andrea Manzi (EGI), Ignacio Lamata Martinez (EGI) |
| EDITOR(S) | Massimiliano Assante (CNR), Valerio Grossi (CNR) |
| REVIEWER(S) | Alessia Bardi (OpenAIRE), Valerio Grossi (CNR), Ilaria Barsanti (CNR) |
| CONTRACTUAL DELIVERY DATE | 31/12/2022 |
| ACTUAL DELIVERY DATE | 26/04/2023 |
| VERSION | V1.1 |
| TYPE | Report |
| DISSEMINATION LEVEL | Public |
| TOTAL N. PAGES | 28 |
| KEYWORDS | e-infrastructure, service, gateway, component |

# EXECUTIVE SUMMARY

This Deliverable D9.2 - “SoBigData e-Infrastructure Operation Report 2” is the revised version of the deliverable D9.1 - “SoBigData e-Infrastructure Operation Report 1” [3]. It reports on the activities carried out within Work Package 9 in the period from M19 (January 2021) to M36 (December 2022) for the SoBigData e-Infrastructure operation activity. It includes a detailed set of usage indicators (i.e., the number of users, access to resources, usage of resources from scientists, etc.). It also reports the deployment and procedures governing the operation of the Virtual Research Environments, the catalogue, and the services devoted to data analytics.

A total of 17 Virtual Research Environments (VREs) have been created and/or operated to serve the needs arising in the context of the project. The SoBigData gateway (https://sobigdata.d4science.org/) provide its users with: 6 Exploratories VREs paired with the use cases (Demography, Economy & Finance 2.0; Migration Studies; Societal Debates and Misinformation Analysis; Social Impacts of AI and Explainable Machine Learning; Sports Data Science; Sustainable Cities for Citizens); 4 Virtual Lab VREs - SoBigDataLab and the OpenScienceGraphLab to exploit and experiment tools and solutions, the SoBigData-PlusPlus at DSAA 2021 Lab and the XAISS VLab, conceived to be the working environment for Hands-on Tutorials showing the services provided by SoBigData for the new generation of Responsible data scientists; 3 Applications VREs - TagME, SMAPH, M-Atlas; 2 Project Internal VREs - SoBigData.eu VRE for the communications and collaboration among project and initiative members and SBD-InfraCore VRE for supporting SoBigData++ WP9; 2 Literacy And Training VREs - the SoBigDataLiteracy, supporting Critical Data Literacy of task T.2.4, creating a curated collection of literature of interest for the SoBigData Community, and the e-Learning_Area VRE to host training materials developed within the SoBigData project.

As of mid-December 2022, the e-infrastructure served more than 10,000 users by a total of more than 47,000 working sessions, with an average of 1350 working sessions per month with stable trend. This required to deal with approximately 130 issue tracker tickets (65 requests for support, 4 requests for incidents and bugs, 22 requests for new features, and 39 requests for Tasks, Virtual Machine or Container creations).

# 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

| | |
|---|---|
| DM | DataMiner |
| EC | European Commission |
| EU | European Union |
| H2020 | Horizon 2020 EU Framework Programme for Research and Innovation |
| SMAE | Social Mining Analytics Engine |
| VRE | Virtual Research Environment |
| WP | Work Package |

# TABLE OF CONTENTS

# 1 Relevance to SoBigData++

## 1.1 Purpose of this document

This Deliverable D9.2 - "SoBigData e-Infrastructure Operation Report 2" is the revised version of the deliverable D9.1 - "SoBigData e-Infrastructure Operation Report 1" [3]. It reports on the activities carried out within Work Package 9 in the period from M19 (January 2021) to M36 (December 2022) for the SoBigData e-Infrastructure operation activity. E-infrastructure operation activity intended since its deployment, and including a detailed set of usage indicators (i.e. the number of users, access to resources, usage of resources from scientists, etc.). It also reports the e-Infrastructure deployment and procedures governing the operation of the Virtual Research Environments, the Catalogue, and the services devoted to data analytics.

## 1.2 Relevance to project objectives

One of the main goals of the SoBigData++ project is to support cross-disciplinary research and innovation on the multiple aspects of social complexity from combined data-driven and model-driven perspectives, possibly implementing Open Science practices by means of exploiting an integrated platform where executions can be repeated, compared, discussed and logged. The deployment and procedures governing the operation of the VREs, the Catalogue as well as the procedures for the services devoted to data analytics listed in this deliverable facilitate and support the above-mentioned activities.

## 1.3 Relation to other work packages

The e-infrastructure operation activity is a fundamental and necessary activity for the maintenance of a platform where interdisciplinary tools, methods, and services can be contributed by Work Package 8 and Work Package 10, towards a view where these tools, methods, and services can be shared according to tailored policies, and easily combined. Additionally, the e-infrastructure is the tool by which VA1 - Virtual Access of Work Package 7 is realised.

## 1.4 Structure of the document

The remainder of the document is as follows: Section 2 introduces the e-infrastructure, the plans and procedure governing its parts and resources. Section 3, Section 4, and Section 5 report respectively on the e-infrastructure VREs, Catalogue, and Analytics services deployment and operation activity, including a set of usage indicators for each of the three. Finally, Section 6 concludes the report.

# 2 SoBigData e-infrastructure Planning and Procedures

The SoBigData e-Infrastructure is built on the D4Science infrastructure [2] and the gCube open-source technology [1]. From the end-user point of view, it manifests in the SoBigData gateway (accessible at https://sobigdata.d4science.org), the access point to the Virtual Research Environments, services and methods, available to the SoBigData++ project.

The development of the SoBigData e-Infrastructure counts on 2 main driving points:

1. **The enabling services:** the availability of new versions of the enabling services technology that are made available by the releases of new software, that is available in https://code-repo.d4science.org/gCubeCI/gCubeReleases. These versions are produced by taking into account the requirements (with the relative priority) formulated by the SoBigData community via the specification of the Exploratories and Virtual Lab VREs that might correspond to new facilities to be developed or requests for enhancements of existing facilities as well as requests for resolving malfunctions;

2. **The methods, tools and service integration:** interdisciplinary tools, methods, and services can be contributed by WP 8 and WP 10. By integrating these tools, methods, and services in the e-infrastructure these can be shared according to tailored policies, and easily combined.

The technology supporting the development of the SoBigData e-Infrastructure was included in the following 14 gCube open-source software releases that have been deployed into the D4Science production infrastructure powering the VRE: 5.0 (Feb. 2021), 5.1 (Mar. 2021), 5.2 (May. 2021), 5.3 (June. 2021), 5.4 (Aug. 2021), 5.5 (Oct. 2021), 5.6 (Nov 2021), 5.7 (Jan. 2022), 5.8 (Mar. 2022), 5.9 (Mar. 2022), 5.10 (Apr. 2022), 5.11 (May. 2022), 5.13 (Jul. 2022), 5.13.1 (Sep. 2022) and 5.14 (Dec. 2022).

All the requests are modelled and managed by an activity tracker operated by D4Science and available at https://support.d4science.org. For the needs of the SoBigData community, two specific activity tracker projects have been created within the first 18 months of the project, and one additional issue tracker has been created within the second period:

1. the SoBigData.eu activity tracker project: this activity tracker preexisted SoBigData++ project https://support.d4science.org/projects/sobigdata-eu, it was used during the previous SoBigData project and SoBigData++ kept using it to track activities not directly involved in the e-Infrastructure domain;

2. the SoBigData Infrastructure Core activity tracker project: created during the first 18 months, this tracker available at https://support.d4science.org/projects/sbd-infracore (Figure 1), was specifically conceived to support the WP 9 Joint Research Activities on e-Infrastructure and Supercomputing Network core facilities.

3. the SoBigData Support tracker project: this is a new activity tracker available at https://support.d4science.org/projects/sobigdata-support (Figure 1), open not only to the project members but also to the SoBigData Research Infrastructure at large, i.e. Summer School participants, students and any practitioner interested in the SoBigData technology.

Both the above-mentioned activity trackers are configured to make it possible to create tickets for tasks, requests for support, incidents, VRE creation, and request for specific services provisioning. Moreover the SoBigData.eu activity tracker project is the parent of the SoBigData Infrastructure Core activity tracker, this makes it possible to visualise the child activities in the parent project tracker.

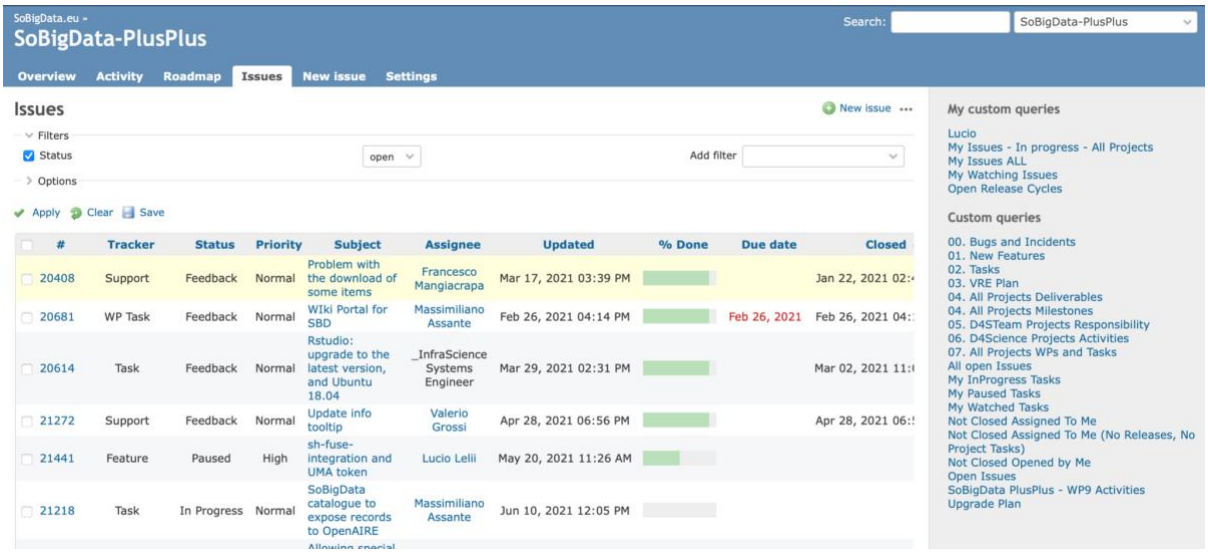


Figure 1a. A screenshot of the SoBigData++ infrastructure core activity tracker

The operation of VREs requires the management of requests for support, of issues and malfunctions, but also the creation of new Virtual Machines and Containers (e.g., Docker). Figure 1 shows screenshots of the issue trackers reporting the tickets for these typologies of tickets. During the reporting period, a total of 130 of such tickets have been resolved (65 requests for support, 4 requests for incidents and bugs, 22 requests for new features, and 39 requests for Tasks, Virtual Machine or Container creations).

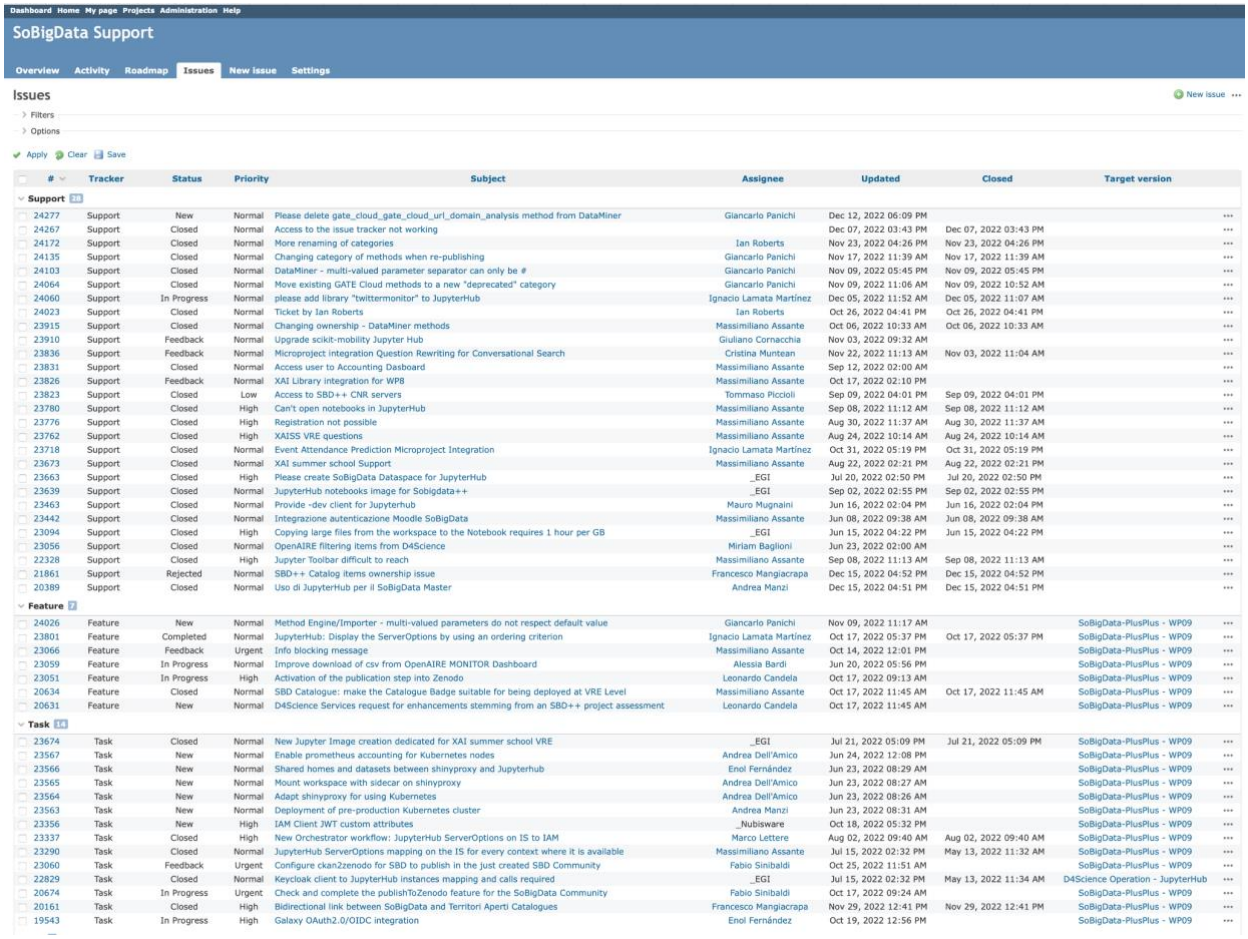


**Figure 1b. A screenshot of the SoBigData Support activity tracker**

## 2.1 Procedures

Deployment and operation of VREs is a collaborative effort involving the WP9 Task 9.1 team called to deploy and configure the technology to create VREs providing access to the interdisciplinary tools, methods, and services expected by the work packages working to develop them, i.e., WP8 and WP10.

The procedure leading to VRE deployment is a consolidated one, i.e., it is the procedure inherited from the D4Science infrastructure[1] and described in the D4Science Wiki:

https://wiki.d4science.org/index.php?title=Virtual_Research_Environments_Deployment_and_Operation

For the needs of SoBigData++, it was decided to support this activity with the project activity tracker. A specific VRE tracker has been created with the goal of capturing the entire process from specification to

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

operation. The specification of the VRE is produced by the VRE designer/requester. This specification must contain:

- VRE name and abstract;
- Membership policy, i.e. whether the VRE is open or restricted, who is allowed to invite members; VRE expected datasets;
- VRE expected functionalities;
- VRE due date;

The following statuses are supported:

***Planned***: the WP9 team is fine with the specification, i.e. the specification contains enough details to proceed with the creation, and acknowledges that the creation of the VRE is feasible by the due date initially requested (or liaise with the designer/requester to find a mutually suitable date);

***Available***: the VRE is up and running and ready to be validated by the VRE designer/requester;

***Released***: the VRE has been validated and the target community can start using it;

***Removed***: the VRE has been disposed as for the request of its manager;

***Rejected***: the requested VRE cannot be created as the requirements outlined for it cannot be satisfied.

# 3 SoBigData VREs Deployment, and Operation

This section briefly describes the facilities used by VRE creators for the actual deployment of VREs, reports the complete list of deployed and operated VREs during the first 36 months of the project, and offers a characterisation of each available VRE. In addition, since SoBigData++ project builds up on the previous SoBigData Project, the VREs deployed and operated during SoBigData have been maintained and enhanced and are part of the list of operated VREs.

The procedure leading to VRE deployment is a consolidated one, i.e., it is the procedure inherited from the D4Science infrastructure and described in the D4Science Wiki:

https://wiki.d4science.org/index.php?title=Virtual_Research_Environments_Deployment_and_Operation

The act of definition and deployment of a new VRE is supported by a wizard (cf. Figure 2) that enables authorised users to transform the opened requests according to the procedure described in Sec. 2 into an actual specification and then, automatically, into a working VRE made available by the SoBigData e-Infrastructure gateway. Through the wizard, the user is requested to specify: (i) the descriptive information characterising the expected VRE (i.e., name, description, duration), and (ii) the functionalities and datasets to be made available in the specific VRE by selecting among the available ones. The resulting list of functionalities is derived from the feasible functionalities created thanks to the software version and services hosted by the underlying infrastructure.

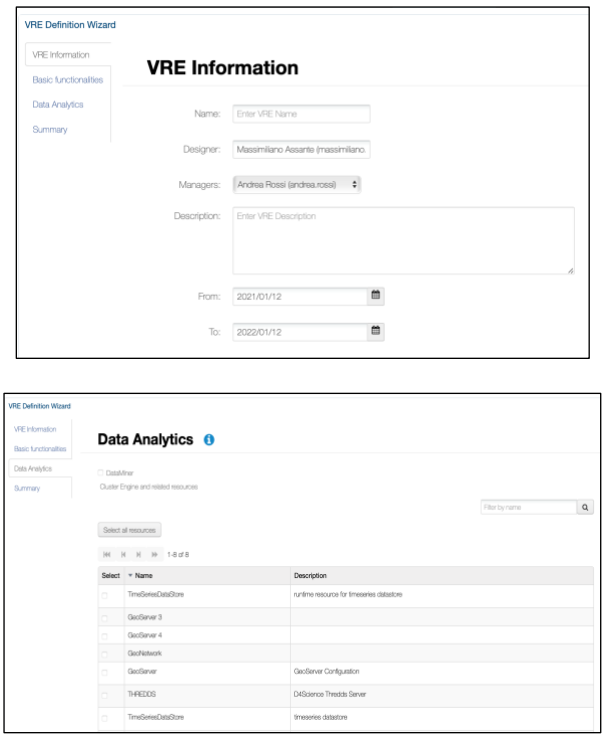


**Figure 2. VRE Creation Wizard Screenshots**

A total of 17 Virtual Research Environments (VREs) have been created and/or operated to serve the needs arising in the context of the project (the complete list is in Table 1). Specifically, a total of 15 VREs during the first period until M18, and 2 VREs during the second period until M36.

These VREs have been classified following the offering type, namely Exploratories, Applications, Lab, Training and Project Internal:

- 6 Exploratories VREs: the list of Exploratories is inherited by the WP10 definition and tasks and are:
  - Demography, Economy & Finance 2.0;
  - Migration Studies;
  - Societal Debates and Misinformation Analysis;
  - Social Impacts of AI and Explainable Machine Learning;
  - Sports Data Science;
  - Sustainable Cities for Citizens.

  To simplify the access and the organization of the gateway, It is important to notice that these 6 VREs are not visible by newcomers but in fact are still accessible and part of the infrastructure and kept in operation. All the resources related to the above VREs are accessible through the Catalogue, and the methods accessible by data miner engine.

- 2 Virtual Lab VREs: the **SoBigDataLab VRE** where the user can develop algorithms in an interactive Python notebook, integrating an algorithm (written in any programming language) or executing experiments on the SoBigData cloud computing centre; and the **OpenScienceGraphLab VRE** conceived for the analysis of Open Science Graphs with Big Data tools from complex networks to descriptive statics, machine learning and Natural Language Processing.

- 3 Applications VREs: the list of Applications which are present in the Catalogue have been organised in 3 main VREs according to the type of services provided and are: **TagME**, **SMAPH**, **M-Atlas**;

- 2 Project Internal VREs: **SoBigData.eu VRE** conceived to provide the SoBigData project members with a VRE-based working environment useful for the communications and collaboration among project and initiative members; **SBD-InfraCore VRE** for supporting the operation of the SoBigData++ WP9 including the editing of HPC Portal Available Resources.

- 2 Literacy And Training VREs: the **SoBigDataLiteracy VRE**, conceived to be the working environment supporting the activities of the Critical Data Literacy task T.2.4 of SoBigData++ Project WP2, aiming at creating a curated collection of literature of interest for the SoBigData Community . The **e-Learning_Area VRE** specifically dedicated to the online training hosts training materials developed within the SoBigData project.

Additionally, in order to support the development of algorithms in an interactive Python notebook, integrating them and/ or executing experiments on the SoBigData cloud computing centre, for specific courses or workshops, the following VRE Labs have been created during the second reporting period until M36:

- the **SoBigData-PlusPlus at DSAA 2021 VRE**, conceived to be the working environment for an Hands-on Tutorial showing the services provided by SoBigData for the new generation of Responsible data science, in the context of the 8th IEEE International Conference on Data Science and Advanced Analytics[2].

- The **XAISS VRE**, conceived to be the working environment for the eXplainable AI Summer School 2022[3] held by TUDelft, that involved both lectures and hands-on activities. A Screenshot of the XAISS VRE Home page is depicted in Figure 3.

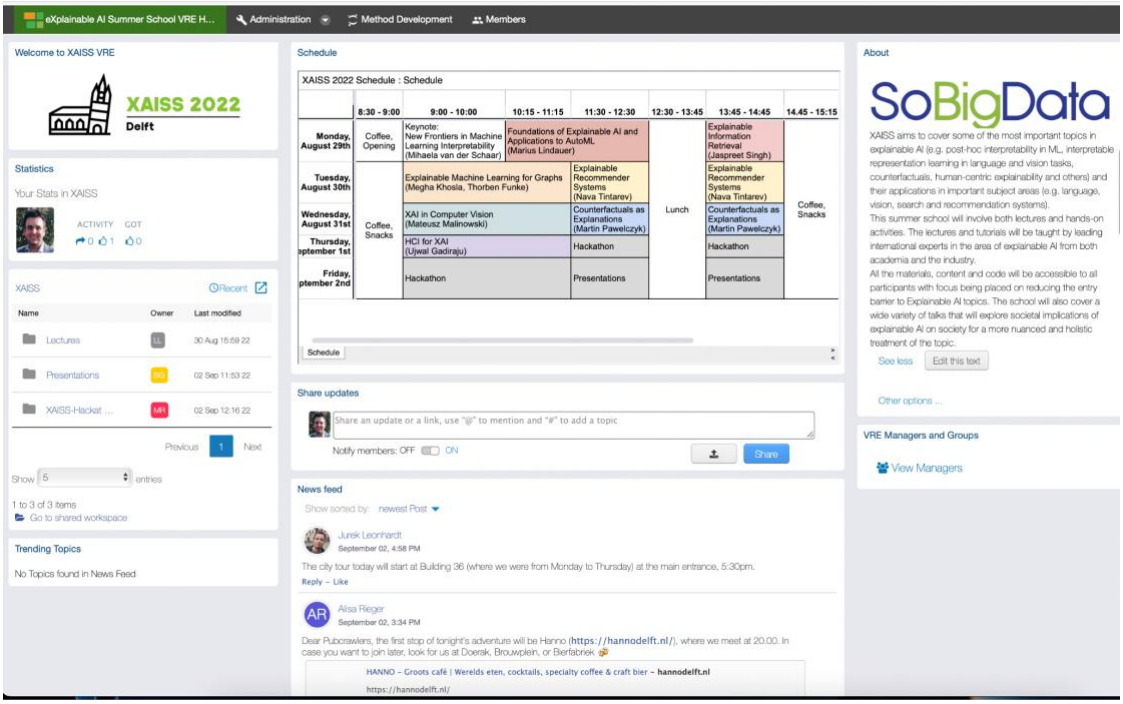


Figure 3. the eXplainable AI Summer School 2022 VRE Lab Home

## 3.1 Operation Activity Indicators

Figure 4 reports the number of VREs operated per month. During the first months of the project, available VREs include those inherited by the previous project (namely, SoBigData) and those created for supporting project activities. From June '20, new VREs began being deployed to serve the needs of SoBigData++ Work Packages.

[2] https://dsaa2021.dcc.fc.up.pt/
[3] https://xaiss.eu/

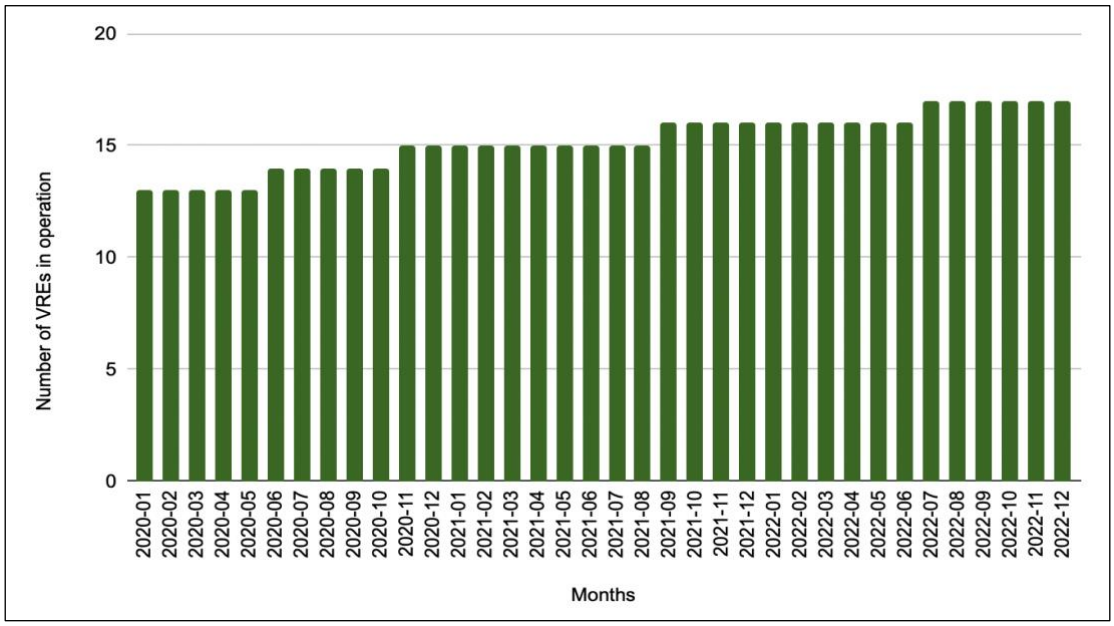


Figure 4. Number of VREs operated per month (January 2020 - December 2022)

In Figure 5, the overall number of users benefitting from the facilities offered by the existing SoBigData VREs is reported, i.e., as of mid-December 2022, the 17 existing VREs are serving more than 10,000 users.

By analysing the email addresses of the users (which is what they are using to log in), it can be observed that: 56% of the users are exploiting an email address that can be attributed to national domains (e.g., .it, .fr, .be) while the remaining 44% of the users are exploiting email addresses provided by commercial providers (e.g., google.com). The users exploiting an email address that can be assimilated to national domains are spread across 20 countries. Between December 2021 and December 2022, the top 3 countries are the United States of America (18%), Italy (16%), and the United Kingdom (9%).

Figure 6 reports the overall number of working sessions initiated per month via the SoBigData VREs. Up to November 2022, a total of more than 47,000 working sessions have been executed by the users, with an average of 1350 working sessions per month with stable trend.

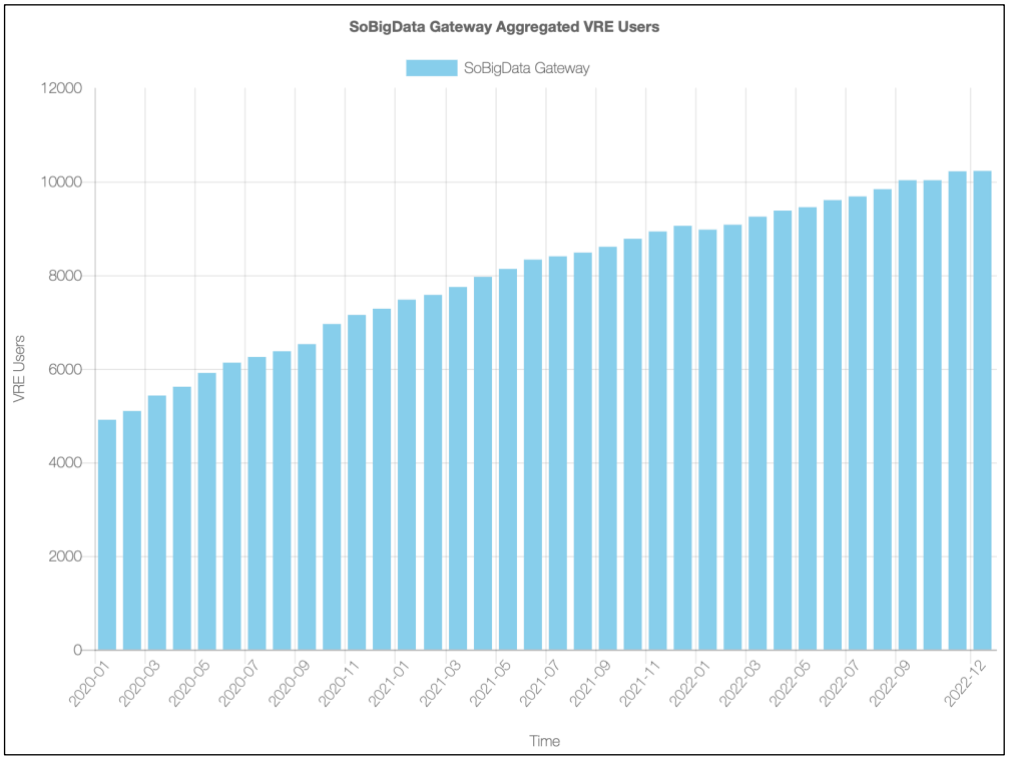


Figure 5. Number of users served by SoBigData VREs (January 2020 - mid December 2022)

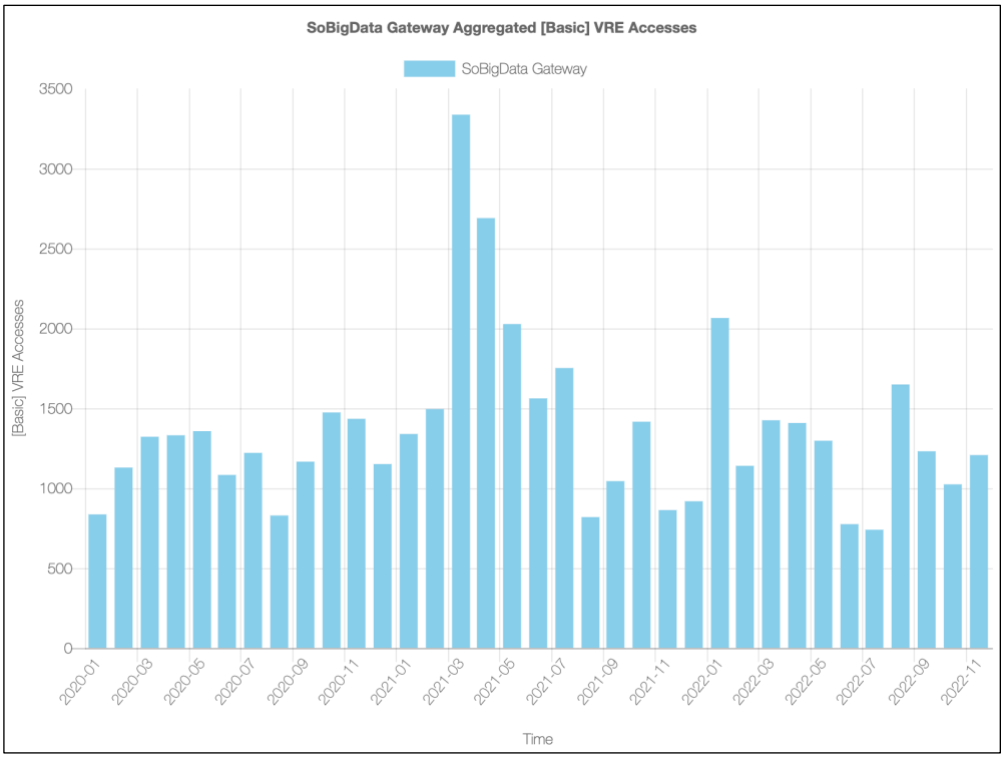


Figure 6. Number of VRE Accesses per month (January. '20 - November '22)

# 4 SoBigData Catalogue Deployment and Operation

The SoBigData Catalogue (https://sobigdata.d4science.org/catalogue-sobigdata) represents a primary place for users to be informed on what is available. It is a core service of the SoBigData e-infrastructure where all the resources contributing to form this e-Infrastructure can be registered, thus making it possible for clients to discover them and be informed on their characteristics for, e.g. properly using them. This catalogue serves both (a) human users willing to know the offering of the e-Infrastructure in terms of datasets, services and methods and (b) other services willing to dynamically discover resources to consume or interact with.

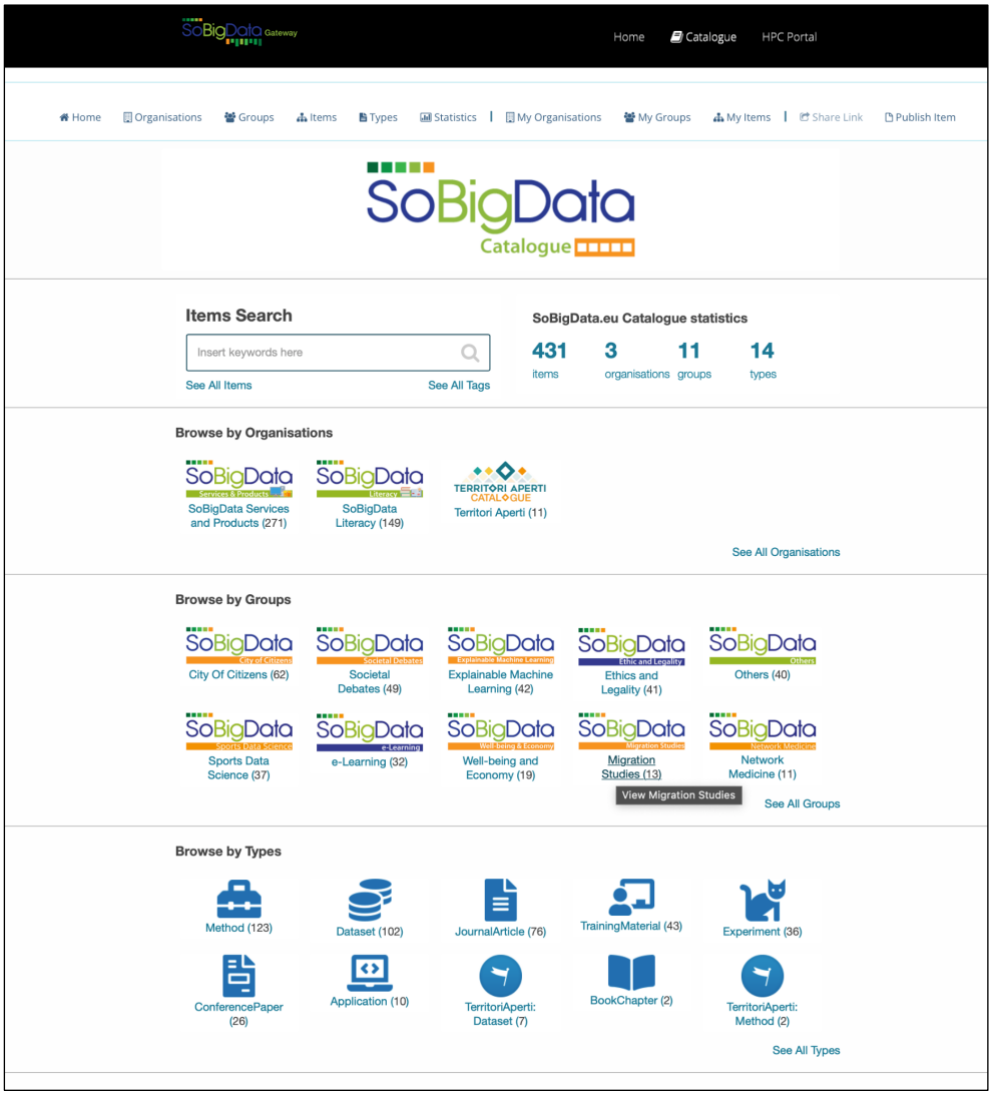


Figure 7. The SoBigData Catalogue welcome page at mid December '22

The central part of the Catalogue solution of the e-infrastructure is based on the open-source technology for data catalogues (CKAN ckan.org). It has been extended to (a) be integrated with SoBigData e-infrastructure services and (b) support a rich, community-defined, and extensible set of catalogue item typologies.

Catalogue item types (aka profiles) are specifications of additional metadata fields to be added to the common metadata characterising every catalogue item. Each profile consists of a list of fields each having a

name, a mandatory directive (whether the field is mandatory or optional), a type (e.g., string, number, spatial extent), a max occur directive to specify whether the field can be instantiated one time only or many times), a default value, a descriptive note helping to understand the intended meaning of the field, a controlled vocabulary (if any) of allowed values to use to compile the field, and a validator (if any) to check the inserted value adherence to specific validation rules.

At the time of writing this deliverable (December 2022) there exist 10 different item types (see Table 1) defined during the operation of the SoBigData Catalogue. Each of these types is characterised by a specific set of attributes, controlled vocabularies and formats carefully describing the specific class of items.

Table 1 also shows the Catalogue evolution, since the beginning of this project: it is important to highlight that not only the item types number has doubled since the beginning of the project, in fact there were 5 item types in January 2020, resulting from the previous SoBigData.eu project, namely: Dataset, Method, Training Material, Application, and Experiment, but also the total number of published items has more than doubled since the beginning of the project.

| Item Type name | Jan 2020 (M1) Number of occurrences | Period 1 (M18) Number of occurrences | Period 2 (M36) Number of occurrences |
|---|---|---|---|
| Method | 76 | 96 | 123 |
| Dataset | 81 | 91 | 102 |
| Journal Article | N.A. | 75 | 76 |
| Training Material | 22 | 30 | 43 |
| Conference Paper | N.A. | 26 | 26 |
| Experiment | 3 | 13 | 36 |
| Application | 8 | 9 | 10 |
| Book Chapter | N.A. | 2 | 2 |
| Research Article | N.A. | 1 | 1 |
| Deliverable | N.A. | N.A. | 1 |
| **Total** | **190** | **343** | **420** |

Table 1. The SoBigData Item Types available and their occurrences in the catalogue

## 4.1 Operation Activity Indicators

In order to quantify the operation activity related to the SoBigData Catalogue, at the time of writing this deliverable (December 2022) the indicators in Table 2 have been collected.

| Indicator Type | Period 1 (M1-M18) Value | Period 2 (M19-M36) Value | Description |
|---|---|---|---|
| Catalogue Accesses | 3.992 | 10.500 | This is the total number of accesses to the SoBigData catalogue in the period January 2020 – mid December 2022. A chart reporting the per month figures is in Figure 8. |
| Catalogue Item Metadata Views | 9.317 | 18.246 | This is the total number of views to catalogue item metadata to the SoBigData catalogue in the period January 2020 – mid December 2022. A chart reporting the per month figures is in Figure 9. |
| Catalogue Item Resource Views | 1.105 | 2.265 | This is the total number of views to catalogue item resources (e.g. linked resources, payloads etc.) to the SoBigData catalogue in the period January 2020 – mid December 2022. A chart reporting the per month figures is in Figure 10. |
| Catalogue search / browse tasks | 37.891 | 81.249 | This is the total number of search and browse operations performed to the SoBigData catalogue in the period January 2020 – mid December 2022. A chart reporting the per month figures is in Figure 11. |

Table 2. The SoBigData Catalogue Operation Activity Indicators up to mid December 2022

Figure 8, 9, 10 and 11 report column charts related to the monthly distribution of the operation activity indicators described in Table 2.

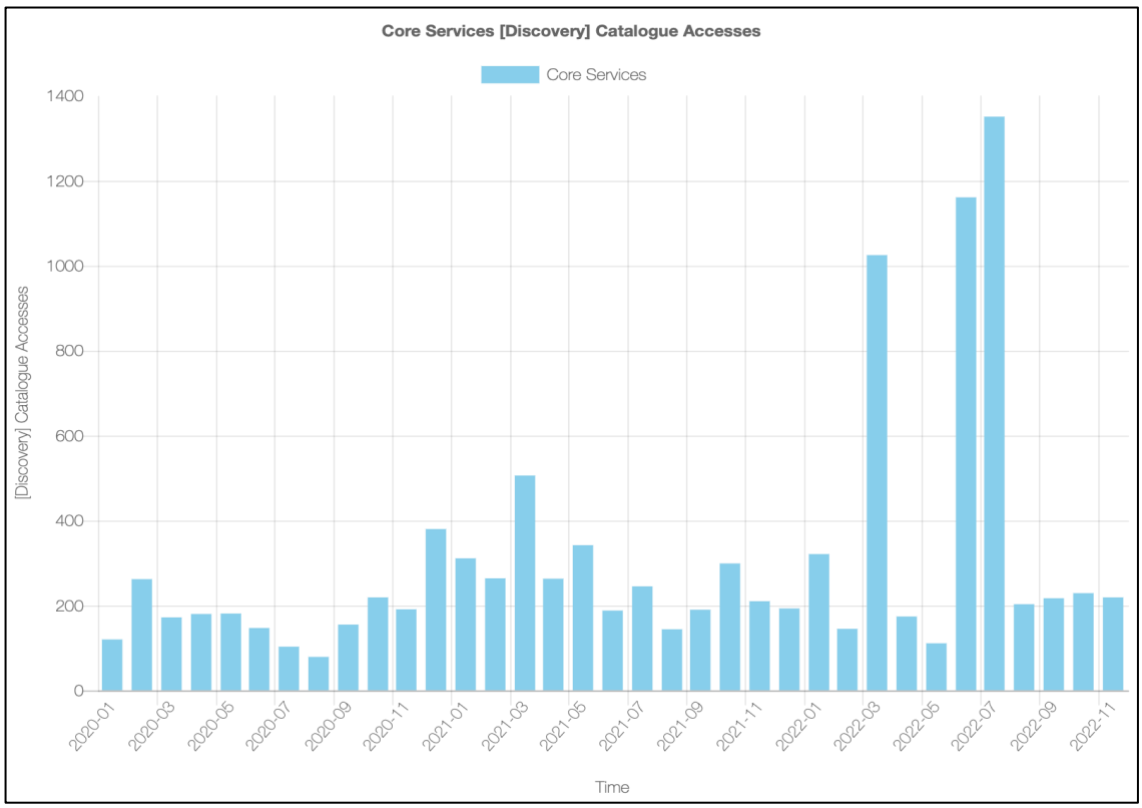


Figure 8. Catalogue Accesses monthly distribution during the period M1 - M35 (Jan 2020 to November 2022)

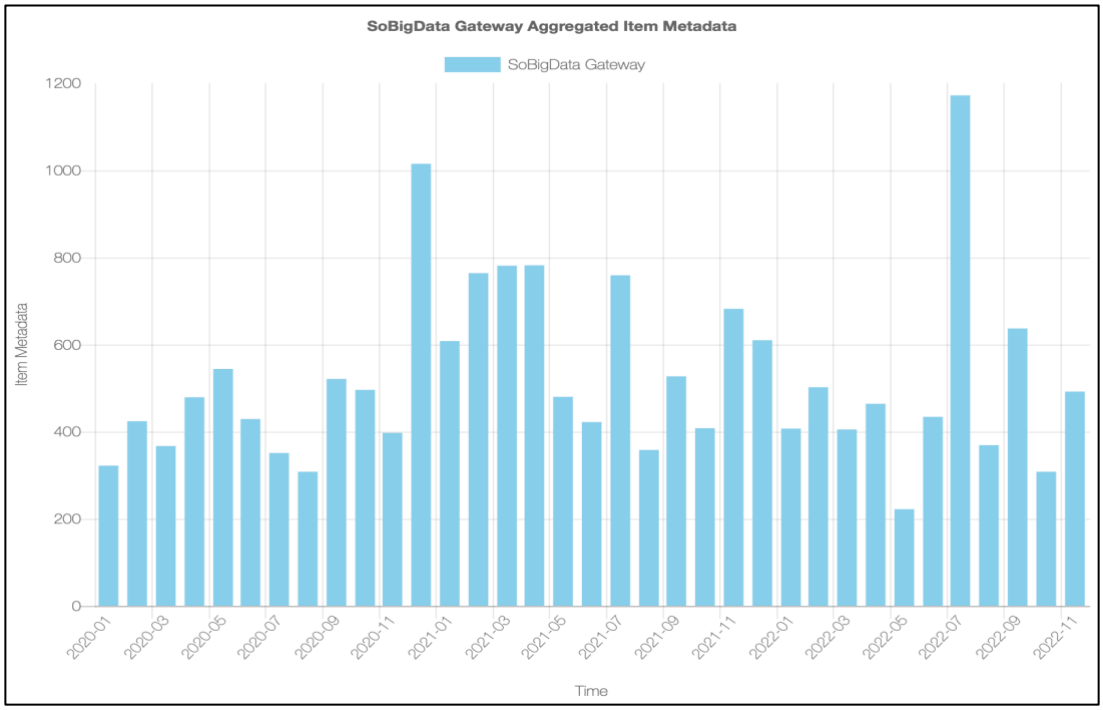


Figure 9. Catalogue Metadata views monthly distribution during the period M1 - M35 (Jan 2020 to November 2022)

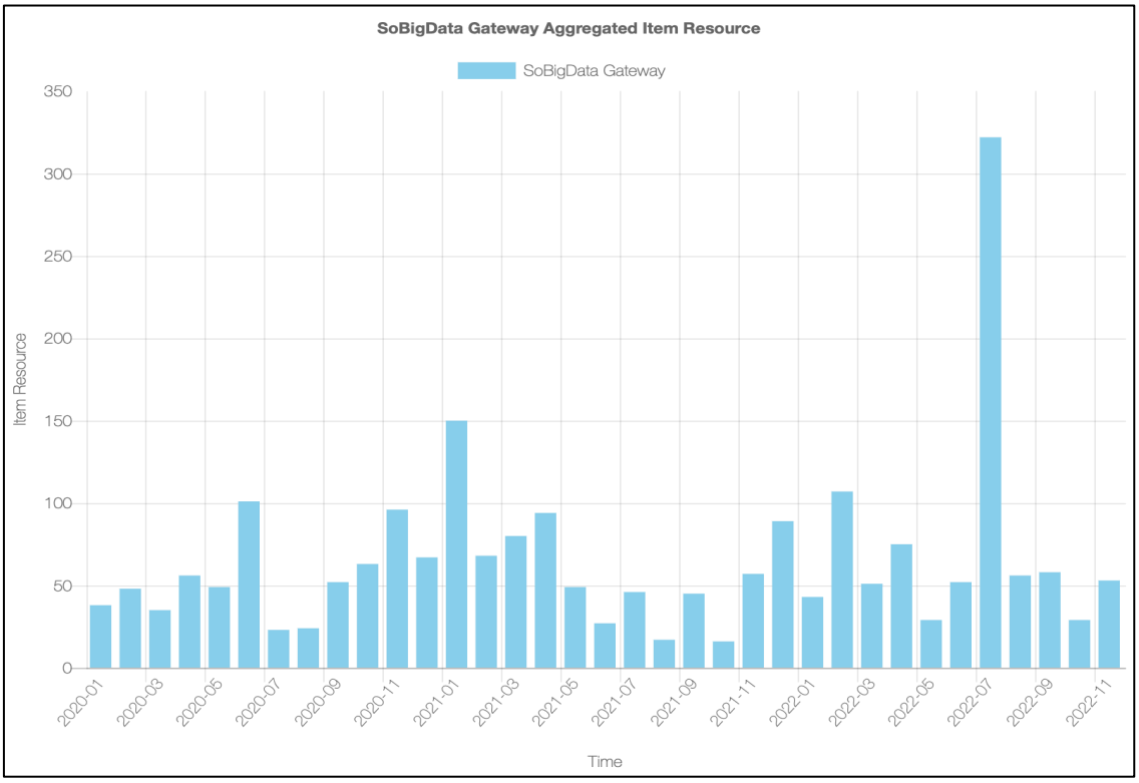


Figure 10. Catalogue Item Resource views monthly distribution during the period M1 - M35 (Jan 2020 to November 2022)

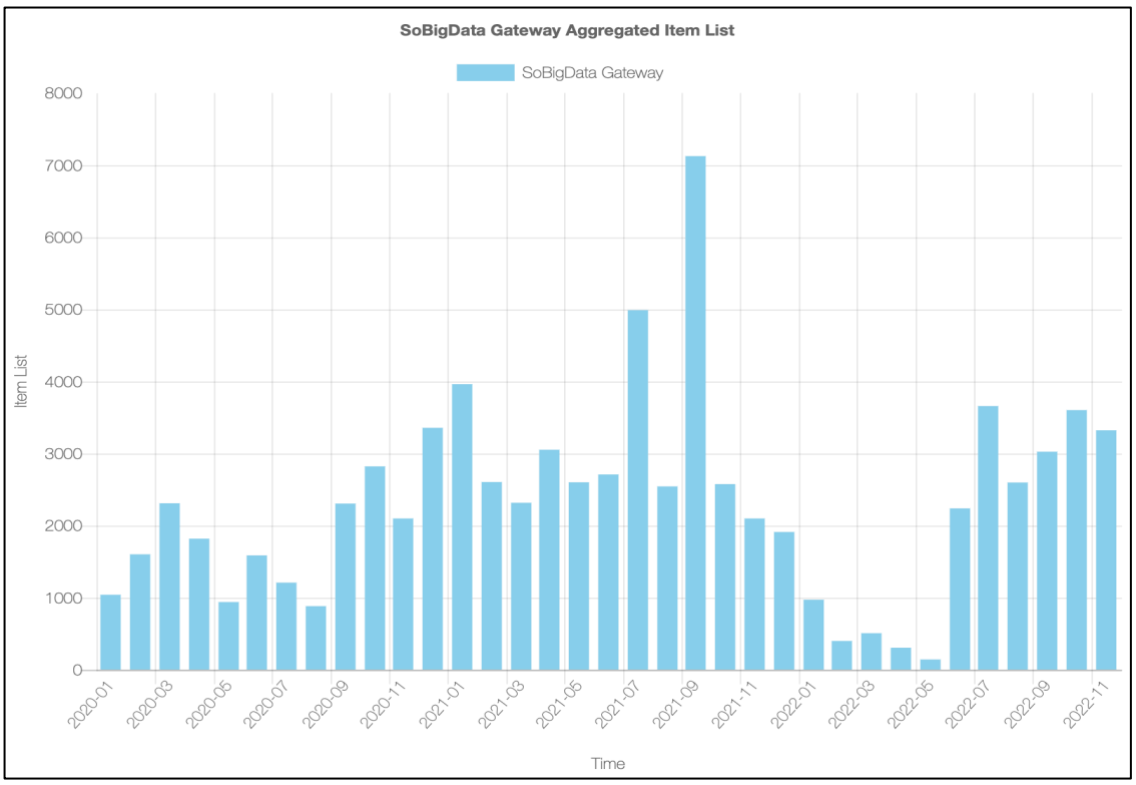


Figure 11. Catalogue Search & Browse tasks monthly distribution during the period M1 - M35 (Jan 2020 to November 2022)

# 5 SoBigData Analytics Services Deployment and Operation activity indicators

## 5.1 Social Mining Analytics Engine

As reported in Deliverable "D9.4 e- Infrastructure Common Facilities" [4], the Social Mining Analytics Engine (SMAE), or Method Engine, includes a set of services and components for performing data processing and mining on information sets. The Method Engine deployment is made up of two sets of clusters: the master and the worker clusters. Each of the two clusters is composed by 16 servers managed by a load balancer - HA-Proxy - that distributes the requests uniformly to these servers. The worker cluster serves cloud computations

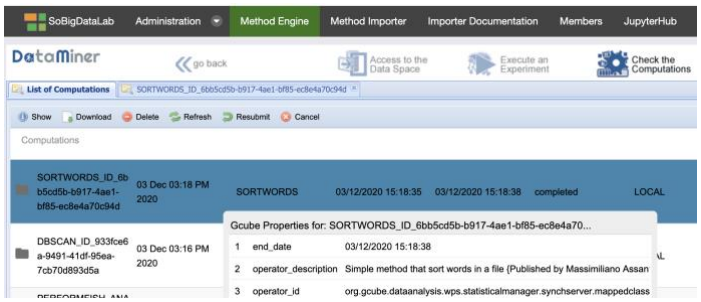


Figure 12. Method Engine instance available in SoBigData Lab VRE

As shown in Figure 12, the Method Engine is a service offered on the SoBigData Lab VRE. However, it is available also in those Exploratory VREs (cf. Sec. 3) that have imported related methods into the e-infrastructure.

| Methods Category | Period 1 (M19-M36) Methods number | Period 2 (M19-M36) Methods number |
|---|---|---|
| Text Learning/Processing/Classification/Analytics | 6 | 21 |
| Archaeological Text Processing | 0 | 6 |
| Misinformation Detection | 0 | 5 |
| Web Analytics | 2 | 2 |
| Representation Learning | 1 | 1 |
| Image Analysis And OCR | 0 | 1 |
| Networks and Metrics | 1 | 1 |
| Examples | 1 | 2 |

Table 3. The SoBigData imported methods available for executions up to mid-December 2022

Table 3 reports indicators on the SoBigData imported methods in the Method Engine available for executions up to Mid-December 2022. The number of methods has grown (almost doubled) during the reporting period.

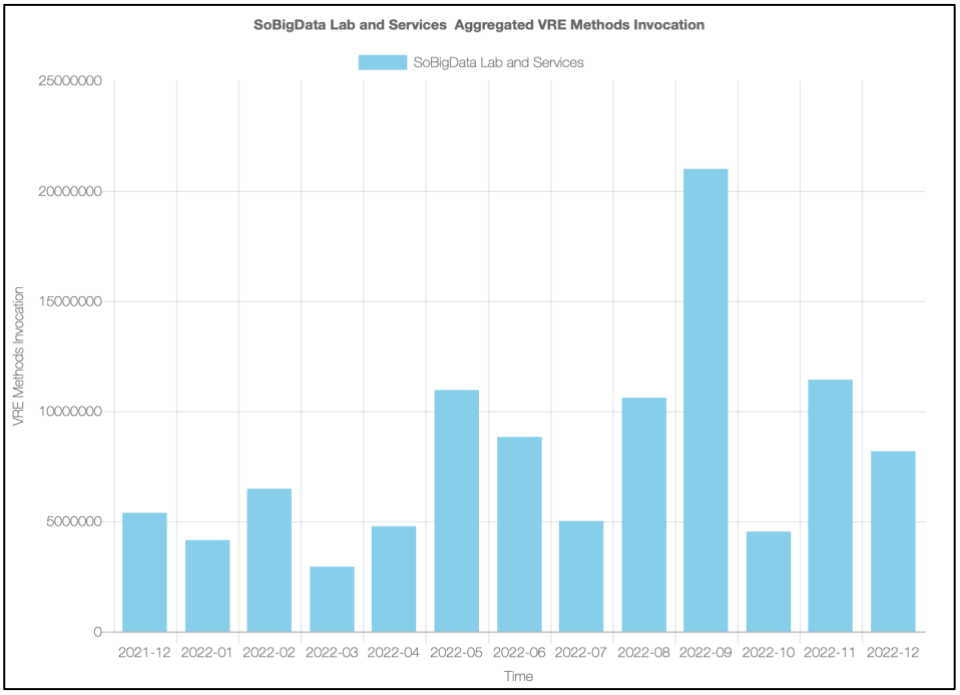


**Figure 13. Number of Methods Executions monthly distribution during the last 12 months**

Figure 13 reports on the number of method executions during the last 12 months. The monthly average number of executions is about 8M invocations, with peaks up until more than 20M in Sept. 2022.

## 5.2 Jupyterhub

The online coding and workflow system enables users to create live documents with code, text and visualisations that capture the whole research process: developing, documenting, and executing code, as well as communicating the results. As reported in Deliverable "D9.4 e-Infrastructure Common Facilities" [4], among the SoBigData++ coding common facilities available and integrated, JupyterHub allows executing Jupyter notebooks providing users with access to computational environments and resources of the e-infrastructure. A Kubernetes[4] deployment (an open-source container-orchestration system for automating deployment, scaling, and management) has been selected to offer a way to automatically provision notebooks servers as containers, with the ability to select the image flavours to run and the resource limits (in terms of CPU and RAM). JupyterHub allows users to run different server instances (which might differ in hardware specifications or pre-installed libraries) according to the user's need. SoBigData++ offers two

[4] https://kubernetes.io

different server options, an *Official* instance and a *Staging* instance (see Figure 14). These instances are similar in capacity and have ready-to-use libraries that are useful to conduct the activities of SoBigData++. The main difference is that the *Staging* instance is used to test the installation of new libraries and software components before publishing them for general use. Images are automatically built and published to a container registry (DockerHub) to make them available for the Kubernetes cluster. The system is highly flexible, so new servers can be easily allocated as needed in SoBigData++, in order to meet the demands and needs of the users.

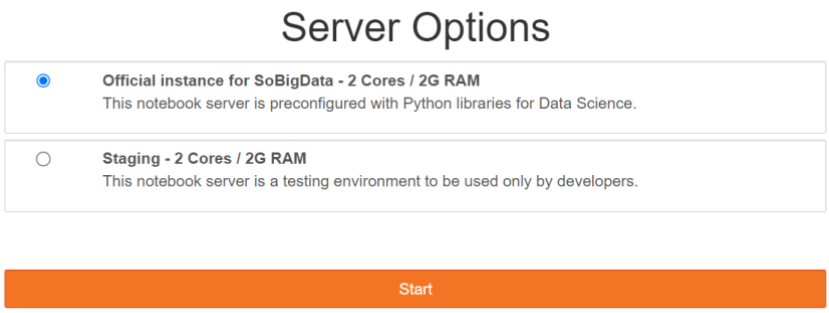


Figure 14. JupyterHub instance available in SoBigData Lab VRE with its server options available

Apart from the regular Notebook servers, new servers are deployed to meet user needs whenever necessary. Figure 15 shows an example of the dedicated server created for the eXplainable AI Summer School[5] during August 2022.

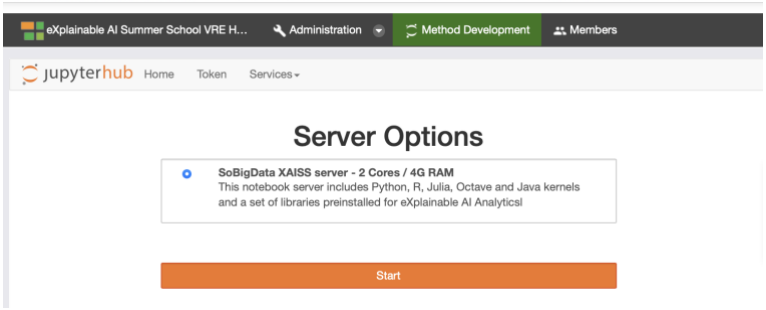


Figure 15. Dedicated server options for eXplainable AI Summer School

JupyterHub is a service offered in SoBigData, since its release in December 2020. The uptake of the service has been monitored to check whether the resources available on the cluster needed to be extended. As

[5] XAISS: eXplainable AI Summer School 2022 https://xaiss.eu

shown in Figure 16, the JupyterHub cluster resources initially allocated did not require to be extended during the period.

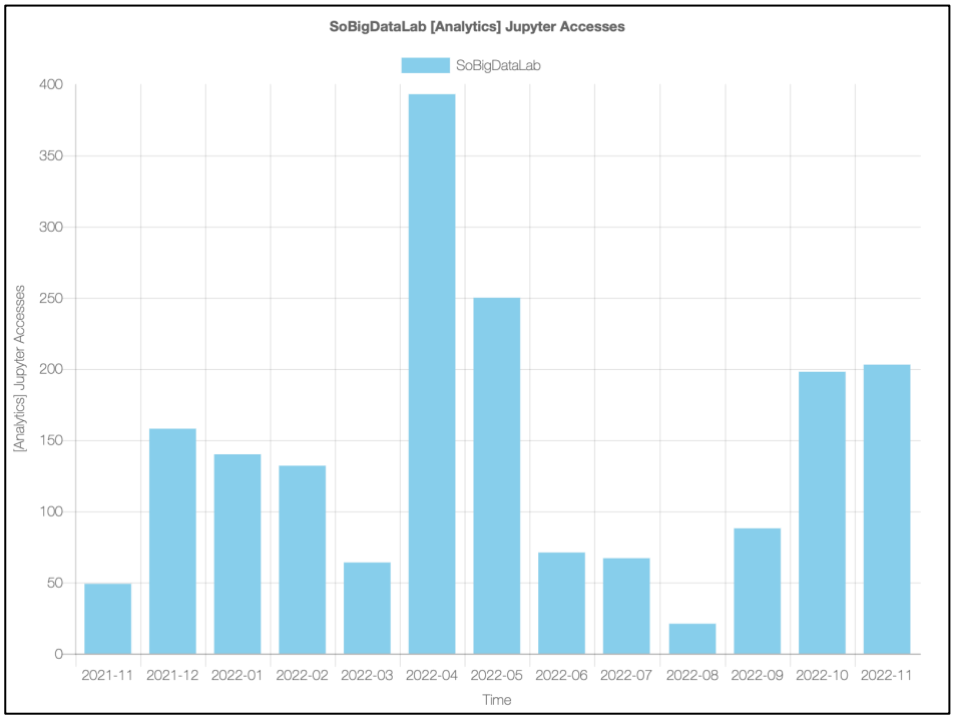


Figure 16. JupyterHub accesses monthly distribution in the last 12 months

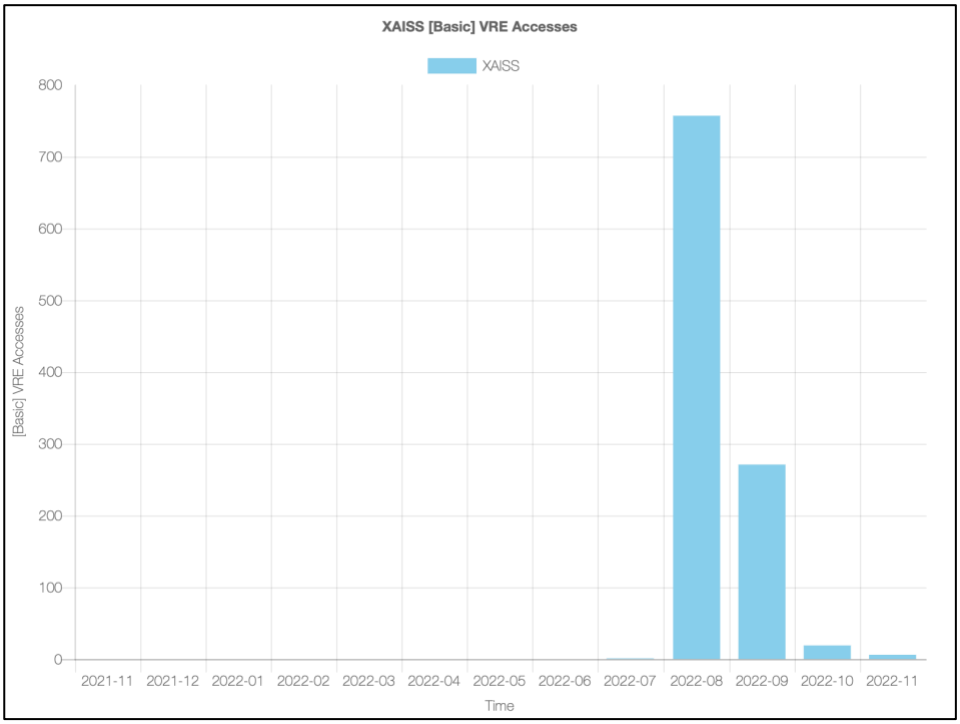


Figure 17. JupyterHub dedicated server accesses created for the eXplainable AI Summer School during August-September 2022.

Figure 17 shows a typical scenario when allocating a dedicated server to specific courses, to meet user needs whenever necessary. The JupyterHub resources initially allocated were not needed until the eXplainable AI Summer School started at the end of August 2022 (https://xaiss.eu), these resources were largely used in August and beginning of September.

# 6 Conclusions

The SoBigData e-infrastructure is a key product to be delivered by the SoBigData++ project to meet the needs of its target community and application scenarios. This deliverable has detailed the e-Infrastructure operation activity in the second reporting period, from M19 (January 2021) to M36 (December 2022), namely the Virtual Research Environments, the Catalogue, and the Analytics services.

As of mid-December 2022, the e-infrastructure served more than 10,000 users by a total of more than 47,000 working sessions, with an average of 1350 working sessions per month with stable trend. This required to deal with approximately 130 issue tracker tickets (65 requests for support, 4 requests for incidents and bugs, 22 requests for new features, and 39 requests for Tasks, Virtual Machine or Container creations).

# References

[1] Assante M., Candela L., Castelli D. , Cirillo R., Coro G., Frosini L. , Lelii L. , Mangiacrapa F., Marioli V. , Pagano P. , Panichi G., Perciante C., Sinibaldi F. (2019) ***The gCube system: Delivering Virtual Research Environments as-a-Service***. Future Gener. Comput. Syst. 95: 445-453 https://doi.org/10.1016/j.future.2018.10.035

[2] Assante M., Candela L., D. Castelli D. , R. Cirillo R., G. Coro G., L. Frosini L. , L. Lelii L. , F. Mangiacrapa F., Pagano P. , Panichi G., Sinibaldi F. (2019) ***Enacting open science by D4Science***. Future Gener. Comput. Syst. 101: 555-563 https://doi.org/10.1016/j.future.2019.05.063

[3] Assante M. and Candela L. and Cirilli R. and Dell'Amico A. and Frosini L. and Lelii L. and Mangiacrapa F. and Pagano P. and Panichi G. and Sinibaldi F. (2021) **SoBigData-PlusPlus - D9.1: SoBigData e-Infrastructure Operation Report 1** https://openportal.isti.cnr.it/doc?id=people ::9da112ec3396c6ea85c705b40c322889

[4] Assante M. and Bardi A. and Fernandez E. and Manzi A. and Pagano P. (2020) **SoBigData-PlusPlus - D9.4: SoBigData e- Infrastructure Common Facilities 1** https://openportal.isti.cnr.it/doc?id=people ::b82acc0fd6422e22b160a5b2f8295196