Deliverable D9.3

SoBigData e-Infrastructure Operation Report 3

SoBigData Plus Plus receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 871042.

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData Plus Plus
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	WP09 - E-Infrastructure and Supercomputing Network
WORK PACKAGE LEADER	CNR
WORK PACKAGE PARTICIPANTS	BSC, EGI, Nubisware, OpenAIRE, USFD, UNIPI, FRH, UT, LUH, AALTO, ETHZ, TUDelft
DELIVERABLE NUMBER and TITLE	D9.3 SoBigData e-infrastructure Operation Report 3
AUTHOR(S)	Massimiliano Assante (CNR), Leonardo Candela (CNR), Andrea Dell'Amico (CNR), Luca Frosini (CNR), Francesco Mangiacrapa (CNR), Elisa Molinaro (CNR), Alfredo Oliviero (CNR), Pasquale Pagano (CNR), Giancarlo Panichi (CNR), Tommaso Piccioli (CNR)
CONTRIBUTOR(S)	Marco Lettere (Nubisware), Mauro Mugnaini (Nubisware), Enol Fernandez (EGI), Ignacio Lamata Martinez (EGI)
EDITOR(S)	Massimiliano Assante, Valerio Grossi (CNR)
REVIEWER(S)	Marco Braghieri (KCL), Valerio Grossi (CNR), Ilaria Barsanti (CNR)
CONTRACTUAL DELIVERY DATE	31/12/2024
ACTUAL DELIVERY DATE	23/12/2024
VERSION	V1.1
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	27
KEYWORDS	Facilitates, computing platform, e-infrastructure

EXECUTIVE SUMMARY

This Deliverable builds upon and updates the previous reports, D9.2 - "SoBigData e-Infrastructure Operation Report 2" [5] and D9.1 - "SoBigData e-Infrastructure Operation Report 1" [3].

The SoBigData e-Infrastructure has been pivotal in enabling the core services and research support required for the SoBigData++ project, including Virtual Research Environments (VREs), the Catalogue, and Analytics Services. It is accessible through the SoBigData gateway (https://sobigdata.d4science.org), which provides end-users with seamless access to tools, datasets, and services. The SoBigData e-Infrastructure is built upon the D4Science infrastructure, offering a comprehensive platform that facilitates collaborative, transparent, and interdisciplinary research. The deployment and operation of VREs followed a well-defined procedure, leveraging the consolidated process inherited from D4Science.

Throughout the 60 months of the project, a total of 27 VREs were created and operated to meet project and community needs. These VREs were classified into five categories: Exploratories, Applications, Virtual Labs, Training, and Management. Notable examples include, *(i)* SoBigDataLab and SoBigDataLab-PlusPlus for method development and experiments, *(ii)* Training VREs created for events like Summer Schools and specialised workshops, and *(iii)* Research spaces (formerly known as Exploratories) supporting targeted domains, such as Migration Studies, Sports Data Science, and Social Impacts of AI.

The SoBigData Catalogue (https://sobigdata.d4science.org/catalogue-sobigdata) emerged as a critical resource for both human users and integrated services, enabling access to datasets, services, and analytical methods. The catalogue supports customisable item profiles enriched with metadata fields, controlled vocabularies, and validation rules. By end of term, the Catalogue recorded significant growth, particularly in key item types such as Methods (192 items) and Datasets (250 items). This expansion underscores the Catalogue's role in promoting resource discoverability and supporting research workflows. Its usage indicators demonstrate its active adoption, with 31,909 total accesses, 29,595 metadata views, and 4,171 resource views recorded. Monthly trends reveal consistent engagement, highlighting its importance in the research ecosystem.

The Social Mining Analytics Engine (SMAE) transitioned through the development of a new service, namely Cloud Computing Platform (CCP), offering enhanced scalability and automation through container orchestrations. Methods hosted on the SMAE span multiple categories, such as Text Processing, Web Analytics, and Image Analysis. Over the last year, the platform executed an average of 6.4 million method invocations per month, peaking at 16 million executions in July 2024.

As of mid-December '24, the e-infrastructure serves more than 13,000 users, with an overall trend in the use of the SoBigData VREs from January 2020 to December 2024, highlighting their importance for the research community. The steady engagement through 2023 and 2024, with peaks like July 2024 (2,592 sessions), underscores the VREs continued relevance and utility.

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

CCP	Cloud Computing Platform
DM	DataMiner
EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
SMAE	Social Mining Analytics Engine
VLab	Virtual Laboratory
VRE	Virtual Research Environment
WP	Work Package

TABLE OF CONTENTS

1. Relevance to SoBigData++

This Deliverable D9.3 - "SoBigData e-Infrastructure Operation Report 3" builds upon and updates the previous reports, D9.2 - "SoBigData e-Infrastructure Operation Report 2" [5] and D9.1 - "SoBigData e-Infrastructure Operation Report 1" [3]. It provides an updated account of the activities carried out throughout the 60 months of the project from M1 (January 2020) to M60 (December 2024) for the SoBigData e-Infrastructure operation activity. The report details the operation and deployment of the e-infrastructure, along with a comprehensive set of usage indicators (e.g., the number of users, resource access, and resource utilization by scientists). Additionally, it outlines the procedures governing the operation of the Virtual Research Environments, the Catalogue, and the services supporting data analytics.

1.1 Relevance to project objectives

A key objective of the SoBigData++ project is to enable cross-disciplinary research and innovation on various aspects of social complexity, integrating data-driven and model-driven approaches while promoting Open Science practices. This is achieved through an integrated platform that allows for the repetition, comparison, discussion, and logging of executions. The deployment and operational procedures for the VREs, the Catalogue, and the data analytics services outlined in this deliverable play a crucial role in facilitating and supporting these activities.

1.2 Relation to other work packages

The e-infrastructure operational activity is essential for maintaining a platform that supports the contribution of interdisciplinary tools, methods, and services developed under Work Packages 8 and 10. This platform ensures that these resources can be shared in accordance with tailored policies and seamlessly combined to foster collaborative research. Furthermore, the e-infrastructure serves as the mechanism through which VA1 - Virtual Access, under the scope of Work Package 7, is effectively implemented.

1.3 Structure of the document

The structure of this document is as follows:

- Section 2 provides an overview of the e-infrastructure, detailing the plans and procedures governing its components and resources.
- Sections 3 and 4 focus on the deployment and operation activities of the e-infrastructure's VREs, Catalogue, and Analytics services, respectively, while also presenting a set of usage indicators for each. Lastly, Section 5 concludes the report, summarising key findings and insights.

2. SoBigData e-infrastructure Overview, Planning and Procedures

The SoBigData e-Infrastructure is built upon the D4Science infrastructure [2, 6] and the open-source gCube technology [1]. From an end-user perspective, it is accessible through the SoBigData gateway (available at https://sobigdata.d4science.org), which serves as the central access point to Virtual Research Environments (VREs), services, and methods provided by the SoBigData++ project.

The development of the SoBigData e-Infrastructure is guided by two main principles:

1. **The enabling services:** the e-Infrastructure leverages new versions of enabling service technologies made available through periodic software releases. These releases, accessible at https://code-repo.d4science.org/gCubeCI/gCubeReleases, incorporate updates based on the priorities and requirements gathered from the SoBigData community. These requirements stem from the Research spaces (formerly known as Exploratories) and Virtual Lab VREs, encompassing new functionalities, enhancements of existing features, and resolutions to reported malfunctions.
2. **The methods, tools and service integration:** Work Packages 8 and 10 contribute interdisciplinary tools, methods, and services to the e-Infrastructure. Through integration with the SoBigData platform, these resources become accessible under tailored policies, enabling seamless sharing and the ability to combine them efficiently. This integration supports collaborative research and innovation across diverse disciplines, enhancing the platform's overall utility and impact.

The technology supporting SoBigData's e-Infrastructure development was included in the following gCube open-source software releases that have been deployed into the D4Science production infrastructure powering the VRE. The updates span across 2023 and 2024, ensuring continuous improvements to meet the evolving needs of the SoBigData community. In 2024, the following versions were deployed: 5.17.3 (October 2024), 5.17.2 (July 2024), 5.17.1 (June 2024), 5.17.0 (May 2024), and 5.16.1 (March 2024). The releases in 2023 include 5.16.0 (September 2023), 5.15.5 (August 2023), 5.15.4 (June 2023), 5.15.3 (May 2023), 5.15.2 (April 2023), 5.15.1 (April 2023), 5.15.0 (March 2023), 5.14.4 (February 2023), 5.14.3 (February 2023), 5.14.2 (January 2023), and 5.14.1 (January 2023). These versions have brought new features, enhancements, and bug fixes to ensure the e-Infrastructure remains robust, flexible, and aligned with the requirements of interdisciplinary research activities facilitated by the SoBigData project. Further details about each release can be found at the following link: https://code-repo.d4science.org/gCubeCI/gCubeReleases.

All 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, three specific activity tracker projects have been created in the previous reporting periods and have been maintained in operation to ensure seamless tracking and management of activities:

1. **The SoBigData.eu activity tracker project**: This activity tracker pre-existed the SoBigData++ project (https://support.d4science.org/projects/sobigdata-eu). Initially used during the previous SoBigData project, it has been actively maintained and utilized during SoBigData++ 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 of the SoBigData++ project, this tracker (https://support.d4science.org/projects/sbd-infracore) was specifically designed to support the Work Package 9 Joint Research Activities on e-Infrastructure and Supercomputing Network core facilities. It has remained in operation to ensure continuous support for these activities.

3. **The SoBigData Support tracker project**: Introduced during the second period of the current SoBigData++ project, this new activity tracker (https://support.d4science.org/projects/sobigdata-support) is open not only to project members but also to the broader SoBigData Research Infrastructure community, including Summer School participants, students, and practitioners interested in SoBigData technology. Its continued operation ensures effective communication and support for this extended audience.

These trackers have been instrumental in maintaining structured workflows and facilitating collaboration across the SoBigData community.

The activity trackers mentioned above are configured to allow the creation of tickets for various purposes, including tasks, support requests, incident reporting, VRE creation, and requests for specific service provisioning. Additionally, the SoBigData.eu activity tracker project serves as the parent for the SoBigData Infrastructure Core activity tracker, enabling the visualization of child activities directly within the parent project tracker. Figure 2.1 displays screenshots of the issue trackers summarizing these ticket types. During the extended reporting period, a total of 180 such tickets have been resolved, including 90 requests for support, 8 incident and bug reports, 32 requests for new features, and 48 requests for tasks related to Virtual Machine or Container creations.

2.1 Procedures

The deployment and operation of VREs is a collaborative effort led by the WP9 Task T9.1 team, responsible for deploying and configuring the necessary technology to create VREs. These VREs provide access to the interdisciplinary tools, methods, and services developed by other work packages, such as WP8 and WP10.

The process for deploying VREs follows a well-established and standardized procedure, inherited from the D4Science infrastructure and detailed 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 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:

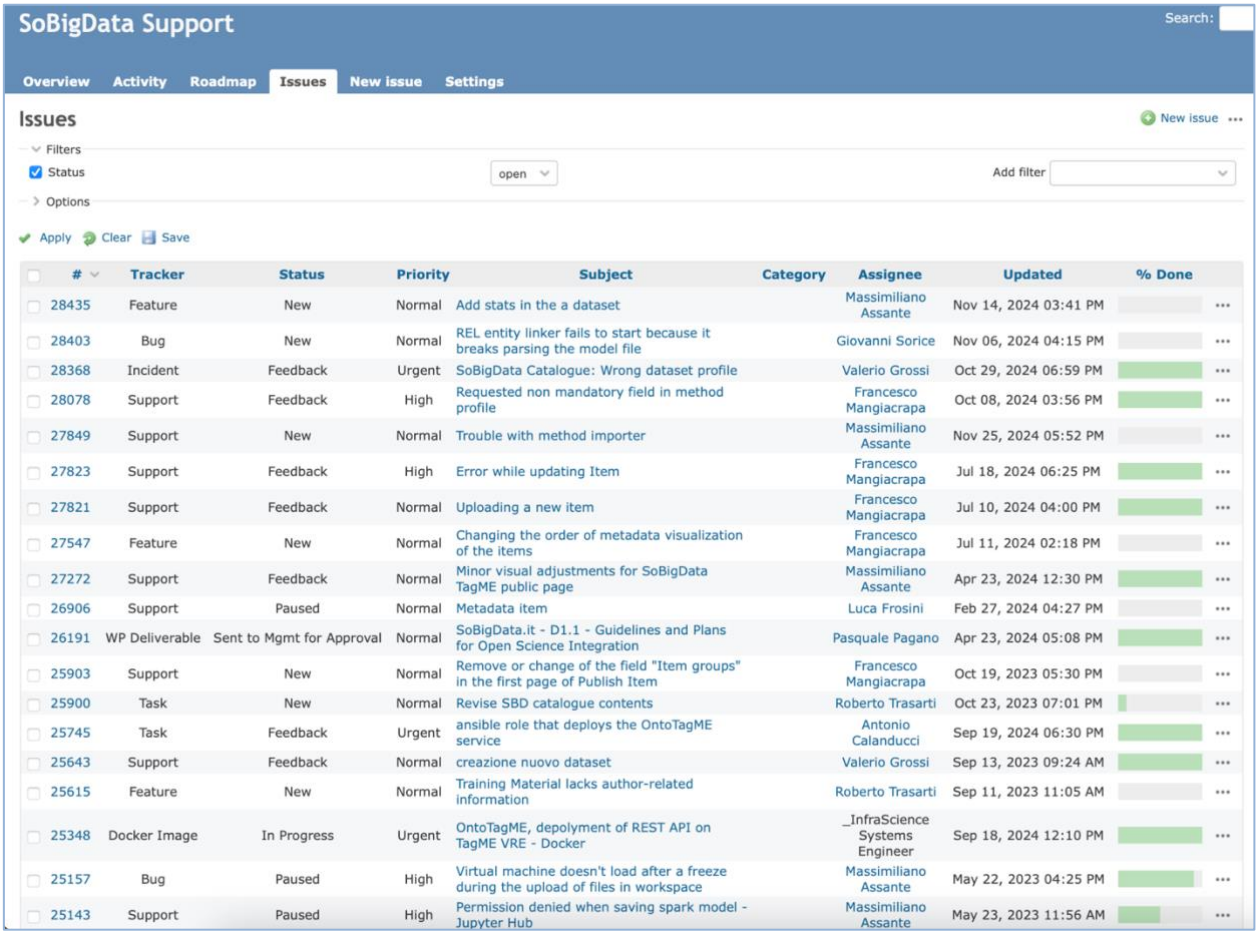
Planned: the WP9 team has checked the specification, i.e. the specification contains enough details to proceed with the creation, and acknowledges that the creation of the VRE is feasible by the due date initially requested (or liaise with the designer/requester to find a mutually suitable date);

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.



#	Tracker	Status	Priority	Subject	Category	Assignee	Updated	% Done
28435	Feature	New	Normal	Add stats in the a dataset		Massimiliano Assante	Nov 14, 2024 03:41 PM	
28403	Bug	New	Normal	REL entity linker fails to start because it breaks parsing the model file		Giovanni Sorice	Nov 06, 2024 04:15 PM	
28368	Incident	Feedback	Urgent	SoBigData Catalogue: Wrong dataset profile		Valerio Grossi	Oct 29, 2024 06:59 PM	
28078	Support	Feedback	High	Requested non mandatory field in method profile		Francesco Mangiacrapa	Oct 08, 2024 03:56 PM	
27849	Support	New	Normal	Trouble with method importer		Massimiliano Assante	Nov 25, 2024 05:52 PM	
27823	Support	Feedback	High	Error while updating Item		Francesco Mangiacrapa	Jul 18, 2024 06:25 PM	
27821	Support	Feedback	Normal	Uploading a new item		Francesco Mangiacrapa	Jul 10, 2024 04:00 PM	
27547	Feature	New	Normal	Changing the order of metadata visualization of the items		Francesco Mangiacrapa	Jul 11, 2024 02:18 PM	
27272	Support	Feedback	Normal	Minor visual adjustments for SoBigData TagME public page		Massimiliano Assante	Apr 23, 2024 12:30 PM	
26906	Support	Paused	Normal	Metadata item		Luca Frosini	Feb 27, 2024 04:27 PM	
26191	WP Deliverable	Sent to Mgmt for Approval	Normal	SoBigData.it - D1.1 - Guidelines and Plans for Open Science Integration		Pasquale Pagano	Apr 23, 2024 05:08 PM	
25903	Support	New	Normal	Remove or change of the field "Item groups" in the first page of Publish Item		Francesco Mangiacrapa	Oct 19, 2023 05:30 PM	
25900	Task	New	Normal	Revise SBD catalogue contents		Roberto Trasarti	Oct 23, 2023 07:01 PM	
25745	Task	Feedback	Urgent	ansible role that deploys the OntoTagME service		Antonio Calanducci	Sep 19, 2024 06:30 PM	
25643	Support	Feedback	Normal	creazione nuovo dataset		Valerio Grossi	Sep 13, 2023 09:24 AM	
25615	Feature	New	Normal	Training Material lacks author-related information		Roberto Trasarti	Sep 11, 2023 11:05 AM	
25348	Docker Image	In Progress	Urgent	OntoTagME, depolyment of REST API on TagME VRE - Docker		_InfraScience Systems Engineer	Sep 18, 2024 12:10 PM	
25157	Bug	Paused	High	Virtual machine doesn't load after a freeze during the upload of files in workspace		Massimiliano Assante	May 22, 2023 04:25 PM	
25143	Support	Paused	High	Permission denied when saving spark model - Jupyter Hub		Massimiliano Assante	May 23, 2023 11:56 AM	

Figure 2.1. A screenshot of the SoBigData Support activity tracker

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 60 months of the project, and offers a characterisation of each available VRE. In addition, since the 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 defining and deploying a new VRE is facilitated by a wizard (cf. Figure 3.1) that allows authorised users to transform open requests, following the procedure outlined in Section 2, into a detailed specification. This specification is then automatically converted into a fully functional VRE, accessible through the SoBigData e-Infrastructure gateway. Using the wizard, users are guided to specify: (i) the descriptive details of the VRE, such as its name, description, and duration, and (ii) the functionalities and datasets to be included, selected from the options available. The resulting list of functionalities is determined based on the capabilities provided by the underlying infrastructure's software versions and hosted services.

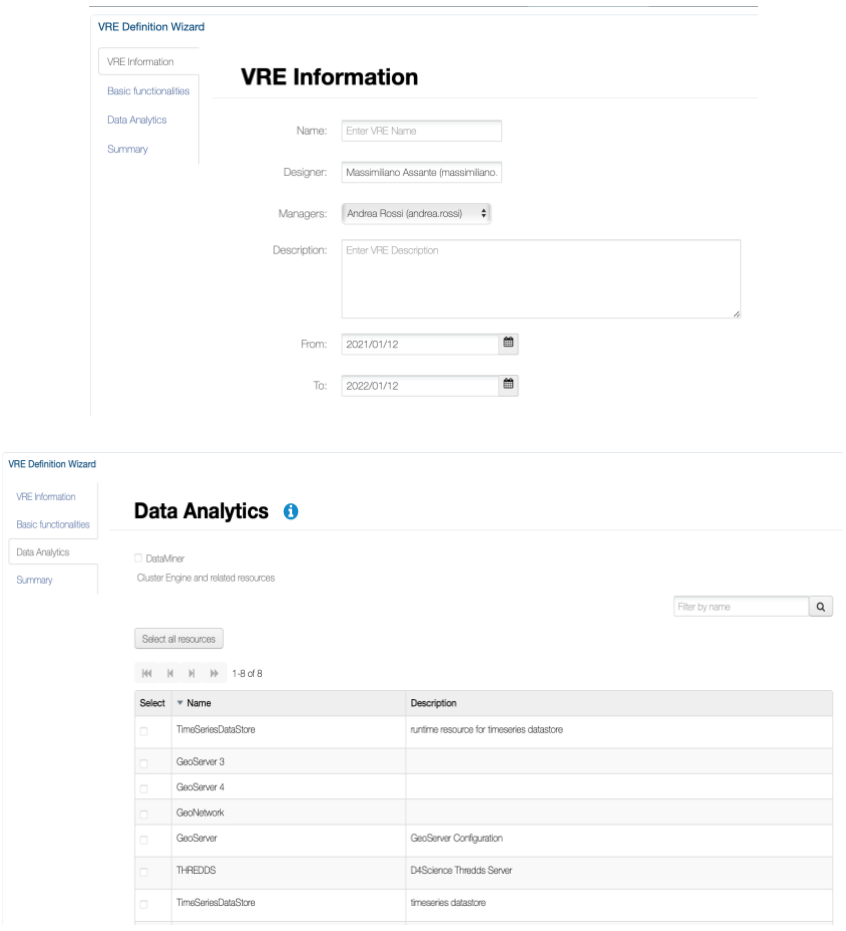


Figure 3.1. VRE Creation Wizard Screenshots

A total of 27 Virtual Research Environments (VREs) have been created and/or operated to serve the needs arising in the context of the project. Specifically, a total of 15 VREs during the first period until M18, 2 VREs during the second period until M36, 10 VREs during the third and final period up to M60.

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

- 6 Exploratories VREs: the list of original Exploratories is inherited by the WP10 definition and tasks and were supported by the following VRES:
    - 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 VRE access and the gateway's organisation, 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 Catalogue, and the methods accessible by data miner engine.

- 3 Virtual Lab VREs: the **SoBigDataLab** VRE, which enables users to develop algorithms using interactive python notebooks, integrate algorithms written in any programming language, or run experiments on the SoBigData cloud computing centre; the **SoBigDataLab** VRE, offering the same features as SoBigDataLab but with access to more computing resources, including options for dedicated jobs; and the **OpenScienceGraphLab** VRE, designed for analysing Open Science Graphs with Big Data tools, covering areas such as complex networks, descriptive statistics, machine learning, and Natural Language Processing.
- 4 Applications VREs: grouping the applications available in the Catalogue into four main VREs based on the type of services they provide: **TagME**, **SMAPH**, **M-Atlas**, and **NetME**. Notably, NetME was integrated during the last reporting period and focuses on on-the-fly knowledge network construction from biomedical literature, enabling efficient extraction and synthesis of relationships among biological elements.
- 3 SoBigData RI Management VREs: the **SoBigData.eu VRE** was 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 WP9 including the editing of HPC Portal Available Resources, and the **SoBigData.it VRE** conceived to provide the SoBigData RI Italian Node project members with a VRE-based working environment useful for the communications and collaboration among project and initiative members.
- 2 Literacy and Training VREs: the **SoBigDataLiteracy VRE** was conceived to be the working environment supporting the activities of the Critical Data Literacy task T.2.4 of 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.

To further support the development of algorithms in an interactive Ipython 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 a hands-on Tutorial showcasing 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[1].
- The **XAISS VRE**, conceived to be the working environment for the eXplainable AI Summer School 2022[2] held by TUDelft, that involved both lectures and hands-on activities.

Additionally, the following VRE Labs were created during the third reporting period, spanning from M36 to M60:

- The **SoBigDataAtDIPSCO VRE**, established to be an introductory course on Python programming (focused on Data Mining and Network applications) at the Psychology and Cognitive Science (DIPSCO) dept. of the University of Trento, that involved both lectures and hands-on activities.
- The **SoBigData Lipari Summer School 2023 VRE,** supporting the summer school for "Responsible Data Science for Society: Models, Algorithms, Trustworthy AI" introduces participants to selected topics to better understand the complexity of our world from the data scientist's perspective

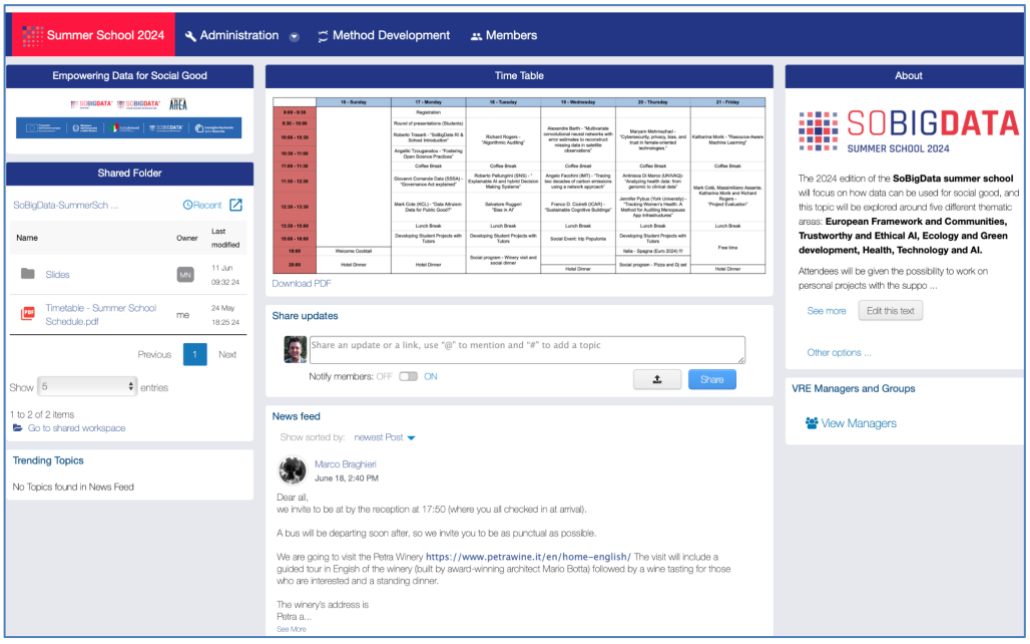


Figure 3.2. the SoBigData Summer School 2024 VRE Lab Home

- The **Laboratory on Graphs Counterfactual Explainability at AAAI 2024 VRE**, dedicated to the hands-on Laboratory at the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24), for developing and evaluating novel graph counterfactual explanation (GCE) methods using the simple and modular framework, GRETEL (https://github.com/aiim-research/GRETEL).
- The **SoBigData Summer School 2024 VRE**, supported the 2024 summer school focussing on how data can be used for social good, around five different thematic areas: European Framework and

[1] https://dsaa2021.dcc.fc.up.pt/

[2] https://xaiss.eu/

Communities, Trustworthy and Ethical AI, Ecology and Green development, Health, Technology and AI, a screenshot of this VRE is given in Figure 3.2.

- The **Privacy Risk Assessment in Mobility Applications VRE**, supported a seminar to provide an overview of the main privacy-preserving techniques, and a hands-on session where participants exploit the practical tools of the VRE.
- The **SoBigData AI & Society 2024 Summer School VRE**, supported the school AI & Society 2024 Summer School activities, including a hands-on session where participants take advantage of a practical tool of the VRE.
- The **SoBigData "Il pensiero computazionale 2024" VRE**, supporting the course for secondary school teachers, focused on developing skills in problem analysis and resolution through the design of algorithms and their experimentation on computers. Details and programme available on the website: http://ilpensierocomputazionale.di.unipi.it/2024/

3.1 Operation Activity Indicators

Figure 3.1.1 illustrates the number of VREs operated each month. During the initial months of the project, the available VREs consisted of those inherited from the previous project (SoBigData) and others created to support project activities. Starting from June 2020, additional VREs were progressively deployed to address the requirements of SoBigData++ Work Packages and events such as Summer Schools. By the end of the reporting period, the total number of VREs has reached 24.

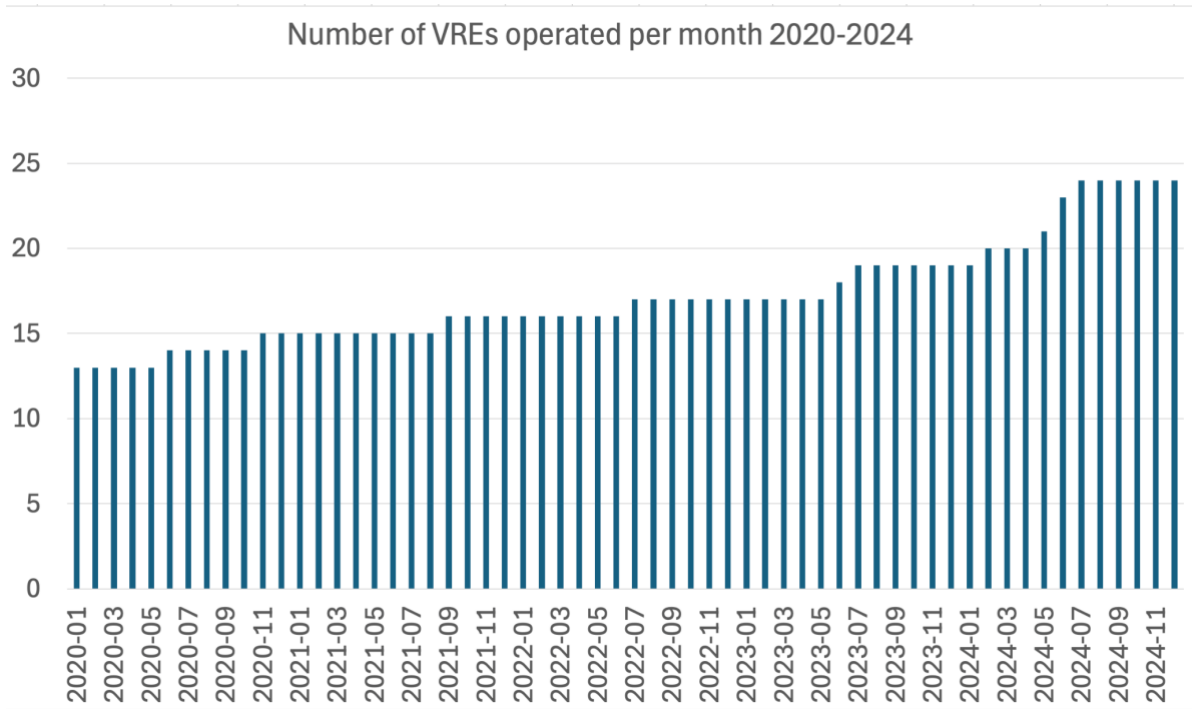


Figure 3.1.1. Number of VREs operated per month (January 2020 - December 2024)

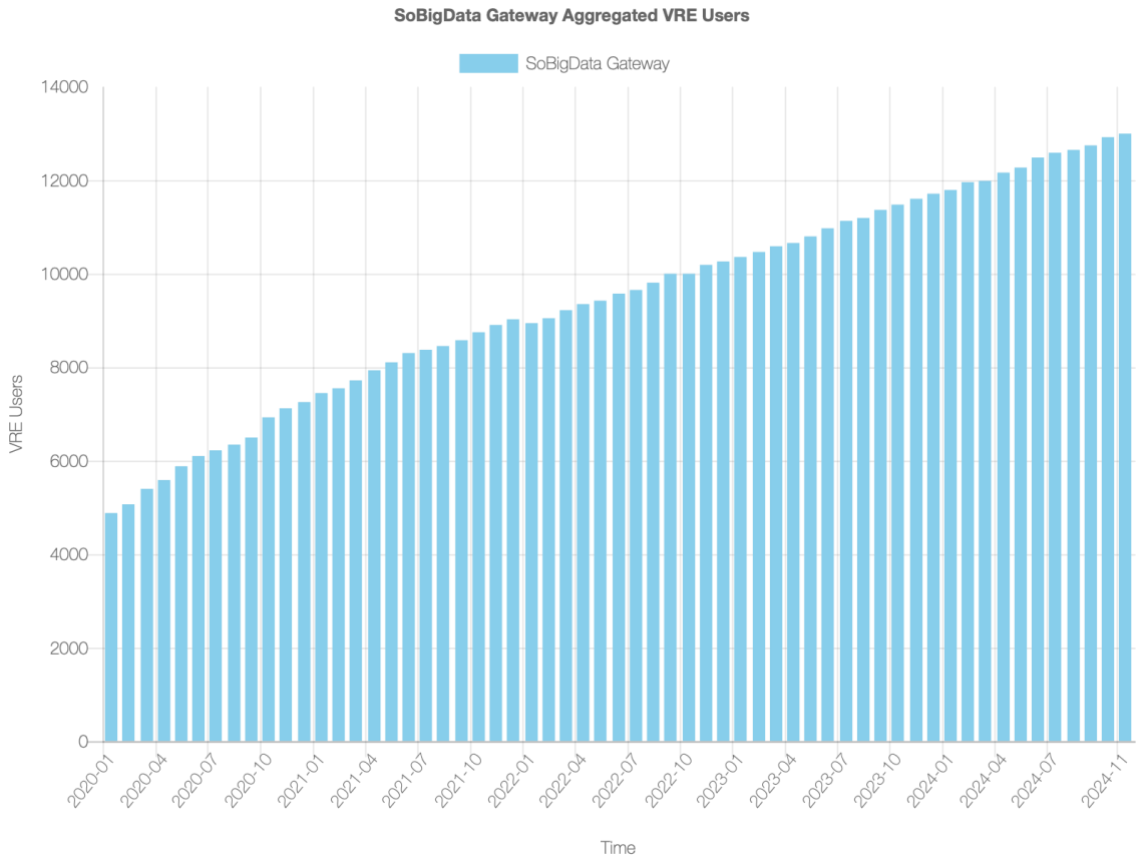


Figure 3.1.2. Number of users served by SoBigData VREs (Jan. 20 - Dec. 24)

In Figure 3.1.2, the overall number of users benefitting from the facilities offered by the existing SoBigData VREs is reported, i.e., as of mid-December '24, the 24 existing VREs are serving more than 13,000 users.

By analysing the email addresses used by users to log in, it can be observed that 58% of users utilise email addresses associated with national domains (e.g., .it, .fr, .de), while the remaining 42% use email addresses provided by commercial providers (e.g., gmail.com). Users with email addresses linked to national domains are spread across 22 countries. Between December 2021 and December 2024, the top three countries are the United States of America (20%), Italy (15%), and the United Kingdom (10%).

Figure 3.1.3 reports the overall number of working sessions initiated per month via the SoBigData VREs. A positive overall trend in the use of the SoBigData VREs from January 2020 to November 2024, highlighting their importance for the research community. The initial growth in working sessions, particularly between 2020 and early 2021, demonstrates increasing adoption and engagement. Notably, significant spikes in March 2021 and July 2024 reflect successful outreach efforts or events driving user activity. The steady engagement through 2023 and 2024, with peaks such as the one in July 2024 (2,592 sessions), underscores the VREs continued relevance and utility.

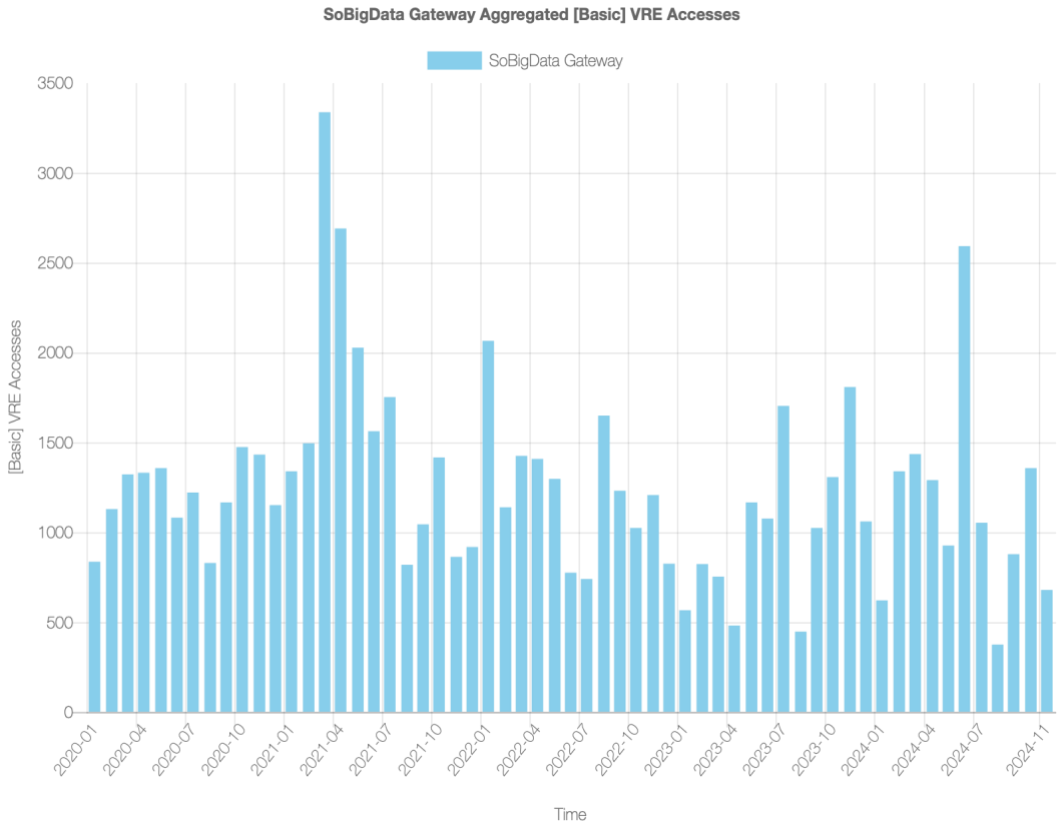


Figure 3.1.3. Number of VRE Accesses per month (Jan. '20 - Nov. '24)

Moreover, VREs operation require continuous management of support requests, issues, and malfunctions, as well as the creation of new Virtual Machines and Containers (e.g., Docker).

4. SoBigData Catalogue Deployment and Operation

The SoBigData Catalogue (https://sobigdata.d4science.org/catalogue-sobigdata) serves as a central hub for users to explore and access available resources. As a core service of the SoBigData e-infrastructure, it enables the registration of all resources that contribute to this infrastructure, allowing users to discover these resources and understand their characteristics for effective utilisation.

The catalogue caters to two primary audiences: (a) researchers who want to gain insights into the offerings of the e-infrastructure, such as datasets, services, and analytical methods, and (b) other services that require a dynamic means of discovering resources for consumption or interaction. This dual functionality ensures that the catalogue not only supports individual researchers in their workflows but also enables seamless integration with other digital infrastructures.

The SoBigData Catalogue is a key component of the e-infrastructure, built on the open-source CKAN technology (https://ckan.org) and extended to integrate with the SoBigData e-infrastructure services. It supports a rich and extensible set of item types (or profiles), which define additional metadata fields tailored to specific classes of catalogue items. 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 – Figure 4.1.

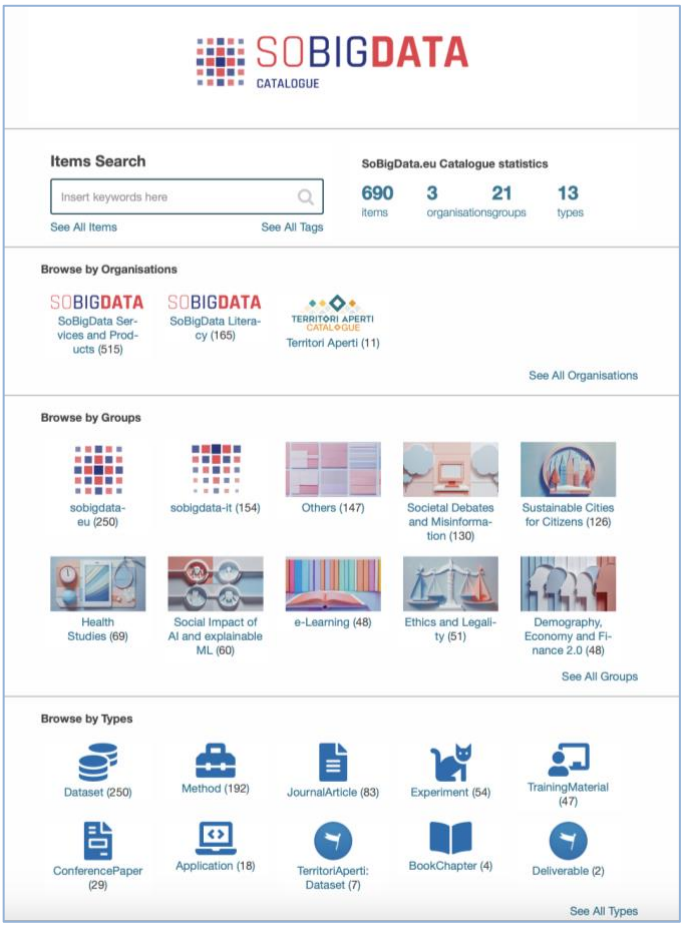


Figure 4.1. The SoBigData Catalogue welcome page mid-December 2024

As the project concludes, the number of item types and their occurrences within the SoBigData Catalogue have grown significantly, demonstrating expanded content and increasing adoption. Table 4.1 highlights this evolution, comparing the state of the catalogue at M36 (mid-term) to the current and final reporting period at M60:

- The Method and Dataset item types have seen the most substantial growth, increasing from 123 to 192 and 102 to 250 occurrences, respectively. This reflects a concentrated effort to document methods and datasets, which are essential for enabling reproducible and transparent research;
- other item types, such as Experiment and Application, have also shown considerable growth, nearly doubling their occurrences (36 to 54 and 10 to 18, respectively);
- incremental increases can be seen for Journal Articles, Training Materials, and Conference Papers, showcasing the diversity of contributions to the catalogue.

Item Type name	Number of occurrences at M36	Number of occurrences at M60
Method	123	192
Dataset	102	250
Journal Article	76	83
Training Material	43	47
Conference Paper	26	29
Experiment	36	54
Application	10	18
Book Chapter	2	4
Deliverable	1	2

Table 4.1. The SoBigData Item Types available and their occurrences at medium term (M36) and final term (M60)

4.1 Operation Activity Indicators

To quantify the operation activity related to the SoBigData Catalogue, at the time of writing this deliverable (December 2024) the indicators in Table 4.1.2 have been collected.

Indicator Type	Value	Description
Catalogue Accesses	31.909	This is the total number of accesses to the SoBigData catalogue in the period January 2020 – mid December 2024. A chart reporting the per month figures is in Figure 4.1.1.

Catalogue Item Metadata Views	29.595	This is the total number of views to catalogue item metadata to the SoBigData catalogue in the period January 2020 – mid December 2024. A chart reporting the per month figures is in Figure 4.1.2.
Catalogue Item Resource Views	4.171	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 2024. A chart reporting the per month figures is in Figure 4.1.3.
Catalogue search / browse tasks	132.491	This is the total number of search and browse operations performed to the SoBigData catalogue in the period January 2020 – mid December 2024. A chart reporting the per month figures is in Figure 4.1.4.

Table 4.1.2. The SoBigData Catalogue Operation Activity Indicators up to mid-December 2024

Figure 4.1.1, Figure 4.1.2, Figure 4.1.3 and 4.1.4 report column charts related to the monthly distribution of the operation activity indicators described in Table 4.1.2.

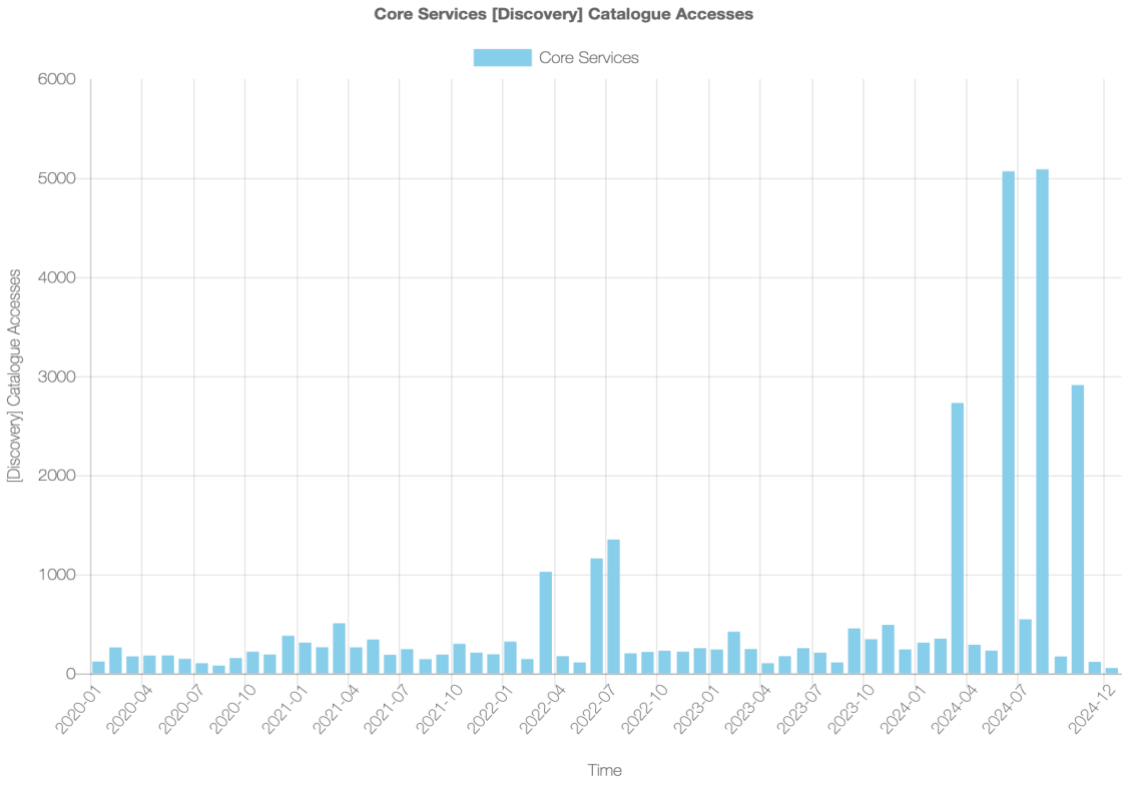


Figure 4.1.1. Catalogue Accesses monthly distribution during the period (Jan '20 to Dec. '24)

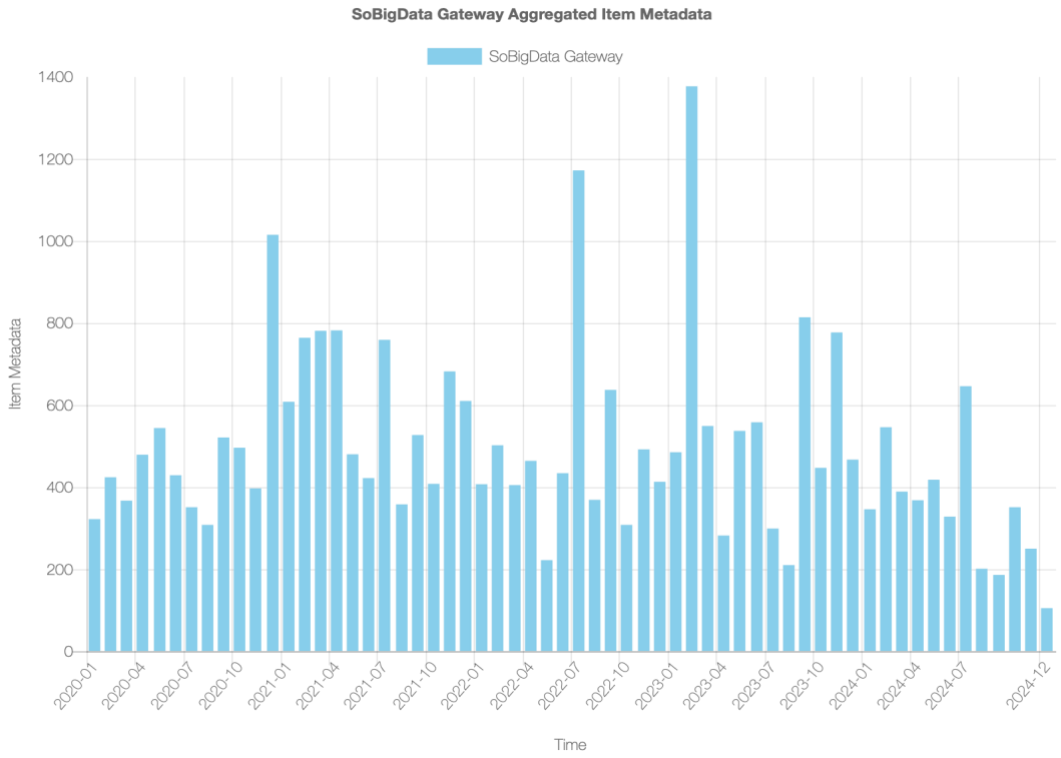


Figure 4.1.2. Catalogue Metadata views monthly distribution in the period (Jan. '20 to Dec. '24)

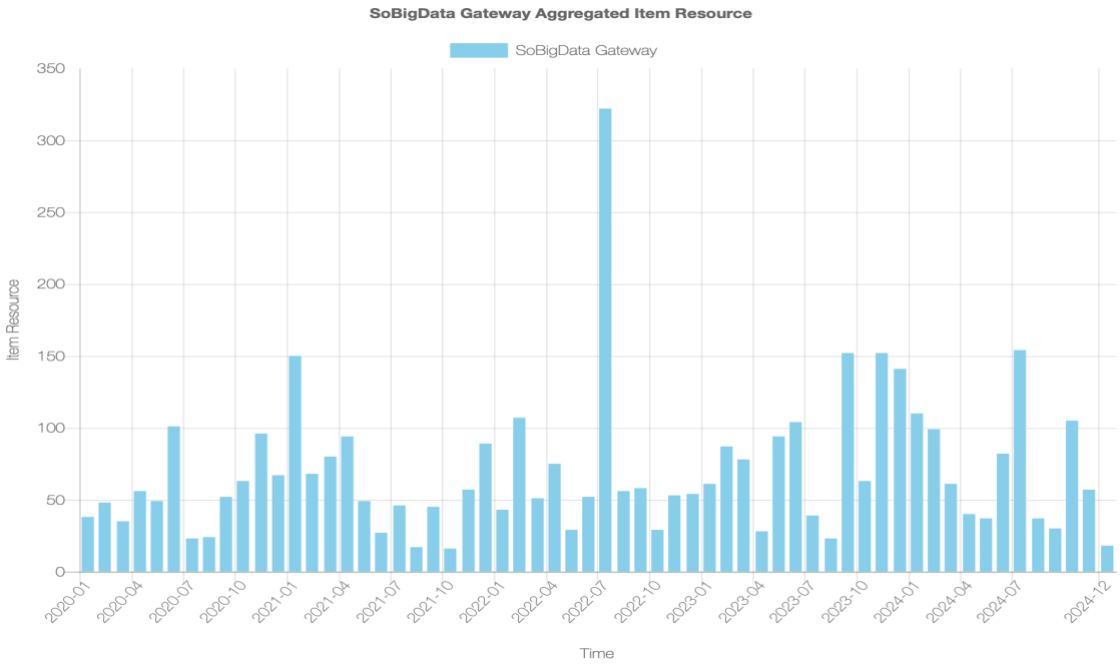


Figure 4.1.3. Catalogue Item Resource views monthly distribution during the period (Jan. '20 to Dec. '24)

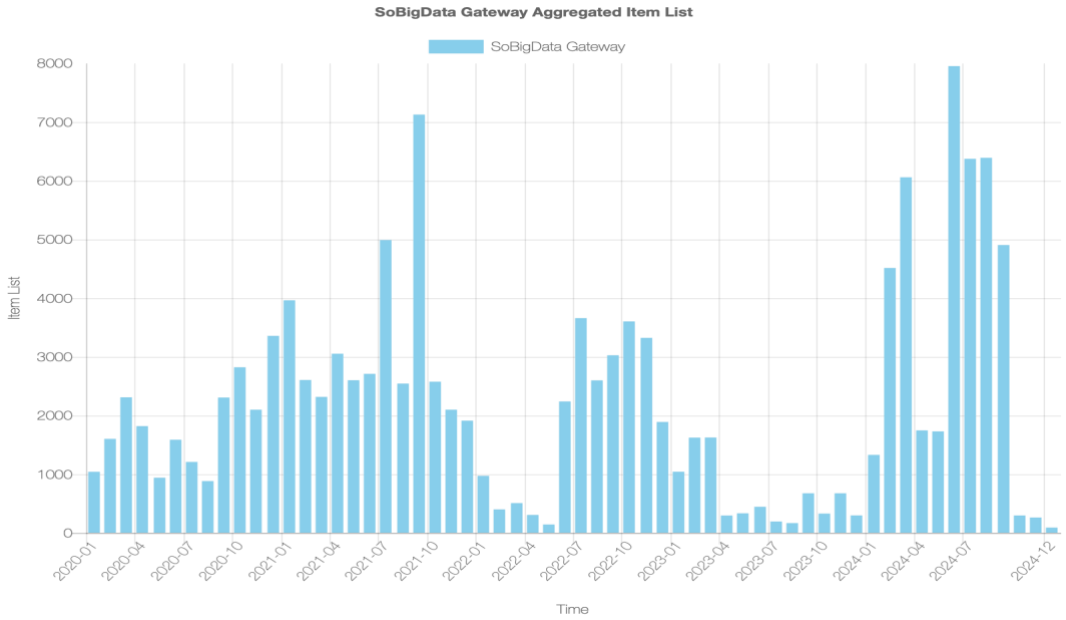


Figure 4.1.4. Catalogue Search & Browse tasks monthly distribution during the period (Jan. '20 to Dec. '24)

5. SoBigData Analytics Services Deployment and Operation activity indicators

5.1 Social Mining Analytics Engine

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. As reported in "D9.5 - M36 e- Infrastructure Common Facilities 2", the Cloud Computing Platform (CCP) represents an evolution of the current Social Mining Analytics Engine – Figure 5.1.1. The deployment of the CCP is distributed, containerised infrastructure utilising **Docker Swarm**. The deployment process begins by encapsulating methods within Docker containers, which are then orchestrated by the platform's **Workflow Orchestrator**. The Orchestrator manages the deployment and execution of methods across the clusters. The platform automatically provisions the necessary resources, installs dependencies, and ensures that each computation is isolated and optimised for performance.

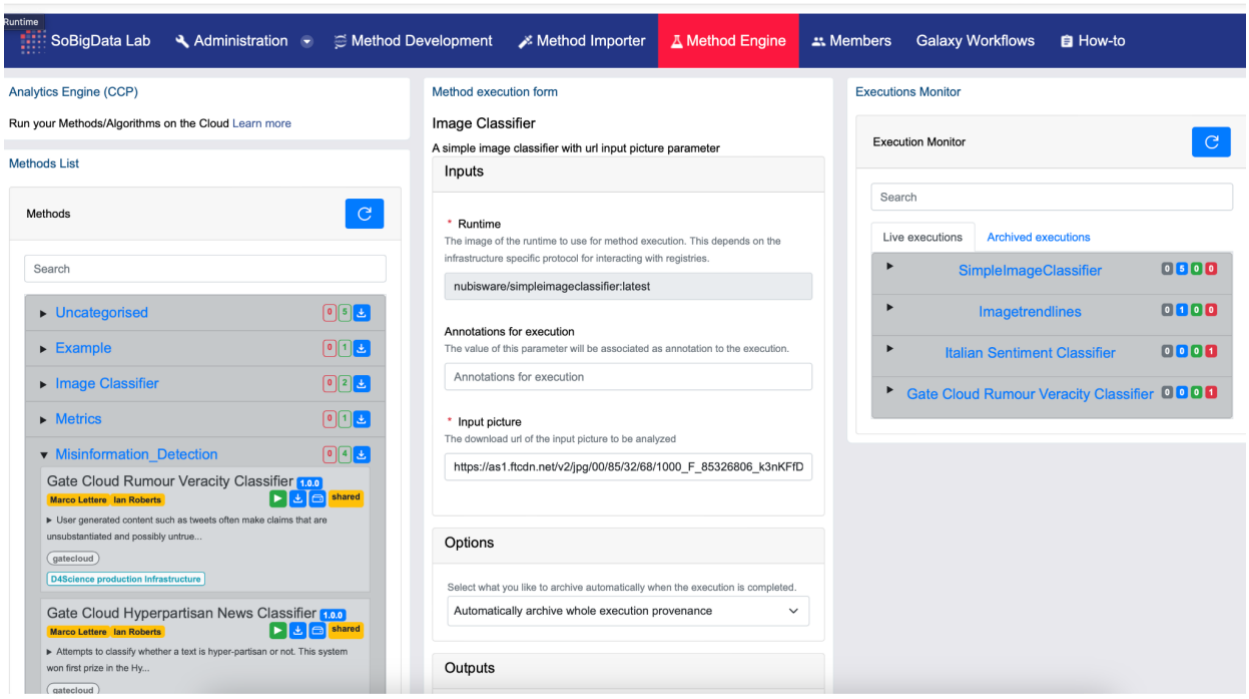


Figure 5.1.1. Method Engine (CCP) instance available in SoBigData Lab VRE

Methods Category	Methods number
Text Learning / Processing / Classification / Analytics	25
Archaeological Text Processing	6
Misinformation Detection	4
Web Analytics	2
Chemical Text Processing	1

Image Analysis And OCR	1
Networks and Metrics	1
Examples	2
Visual Computing	1
Uncategorized	5

Table 5.1.1. The SoBigData imported methods available for executions up to mid-December 2024

Table 5.1.1 reports indicators on the SoBigData imported methods in the Method Engine available for executions up to Mid-December 2024. The number of methods has increased during the reporting period, previous methods available in the SMAE have been ported to CCP.

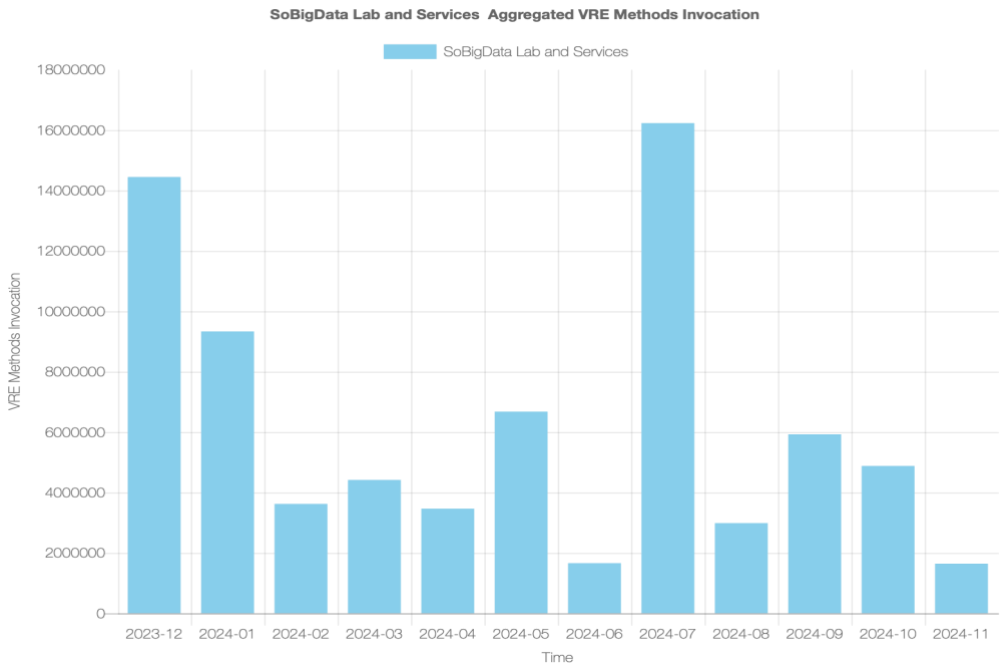


Figure 5.1.2. Number of Methods Executions monthly distribution during the last 12 months

Figure 5.1.2 reports on the number of method executions during the last 12 months. The monthly average number of executions is about 6.4M invocations, with peaks up tomore than 16M in July 2024.

5.2 Jupyterhub

The online coding and workflow system empowers users to create live documents combining code, text, and visualisations, effectively capturing the entire research process: from developing, documenting, and executing code to communicating the results. As outlined in Deliverable *"D9.5 e-Infrastructure Common Facilities 2"* [4], JupyterHub is one of the key coding facilities integrated into the SoBigData infrastructure. It provides users with access to computational environments and resources through Jupyter notebooks.

To automate the provisioning, scaling, and management of notebook servers, a Kubernetes deployment (an open-source container orchestration system) has been adopted. This deployment is supported by two cloud platforms: the D4Science Platform and the Google Cloud Platform. This dual support ensures high reliability and scalability, offering users the flexibility to select containerised notebook server images, specify resource limits such as CPU and RAM, and choose pre-configured image flavours to suit their requirements. JupyterHub further enables users to run multiple server instances tailored to different hardware configurations or pre-installed libraries, depending on their needs – Figure 5.2.1.

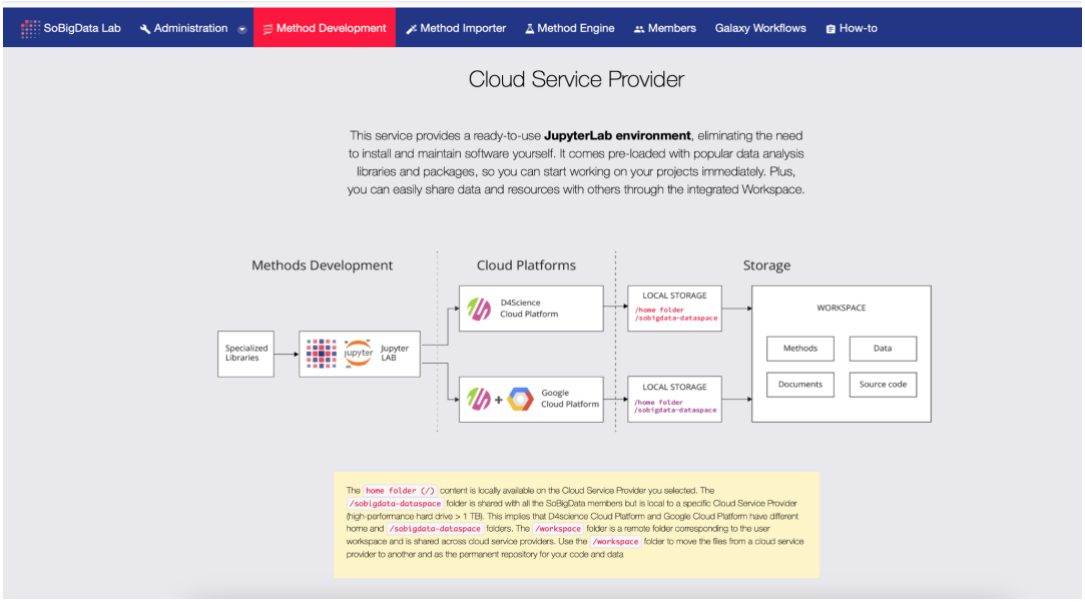


Figure 5.2.1. JupyterHub available in SoBigData Lab VRE on D4Science and Google Cloud Platform

SoBigData++ provides two server options: an Official instance and a Staging instance (see Figure 5.2.2) on both previously mentioned Cloud Platforms. Both offer similar computational capabilities and come preloaded with libraries essential for SoBigData++ activities. The key difference lies in their purpose: the Staging instance is dedicated to testing new libraries and software components before making them publicly available. Once validated, images are automatically built and published to a container registry (DockerHub), ensuring their availability for the Kubernetes cluster.

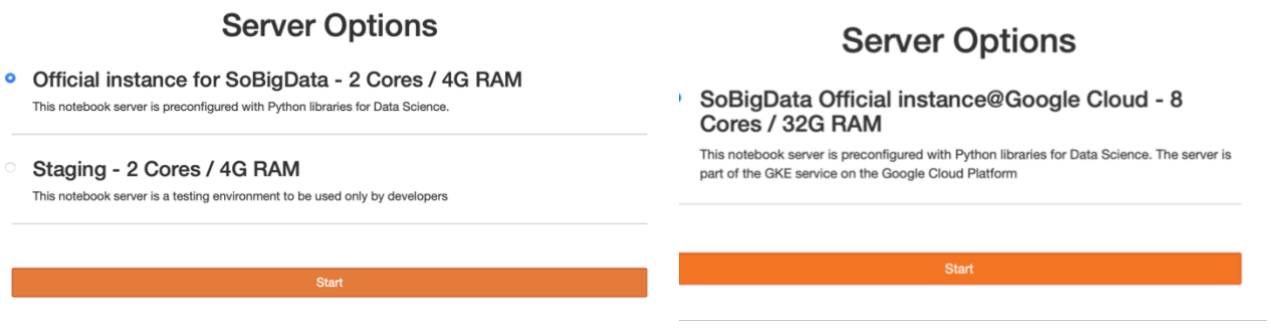


Figure 5.2.2. JupyterHub instance available in SoBigData Lab VRE with its server options available on D4Science (sx) and on Google Cloud Platform (dx)

The system's high flexibility, empowered by the two infrastructures, allows for the seamless allocation of additional servers as needed, ensuring that the computational demands and evolving needs of SoBigData++ users are consistently met.

JupyterHub has been a key service offered within the SoBigData infrastructure since its release in December 2020. The usage and uptake of the service have been closely monitored to ensure that the resources allocated on the cluster are sufficient to meet user demands. As illustrated in Figure 5.2.3, the initially allocated JupyterHub cluster resources have proven adequate, and no extensions were required during the monitoring period. This indicates that the current infrastructure setup effectively supports the computational needs of the SoBigData user community.

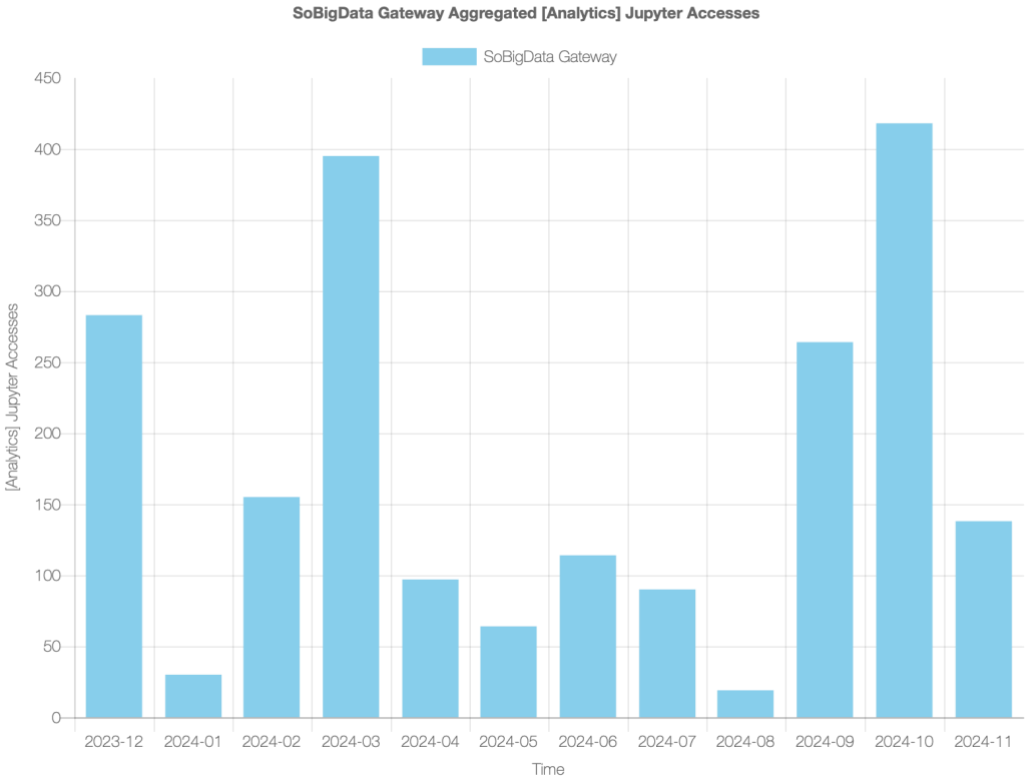


Figure 5.2.3. JupyterHub accesses monthly distribution in the last 12 months

6 Conclusions

The SoBigData e-infrastructure is a key product delivered by the SoBigData++ project to meet the needs of its target community and application scenarios. This deliverable provides a comprehensive overview of the e-infrastructure operation activities carried out throughout the entire reporting period, spanning from M1 (January 2020) to M60 (December 2024). Specifically, it details the operation and evolution of key components, including the Virtual Research Environments (VREs), the Catalogue, and the Analytics services.

The development of the e-Infrastructure has followed two key principles: (1) enabling services, where periodic software releases have introduced new features, improvements, and bug fixes aligned with user requirements; and (2) integration of methods, tools, and service , where tools and resources contributed by Work Packages WP8 and WP10 have been effectively integrated to promote cross-disciplinary research.

As of mid-December, 2024, the e-infrastructure served more than 13,000 users, with an overall positive trend in the use of the SoBigData VREs from January 2020 to November 2024, highlighting their importance for the research community. Initial growth in working sessions, particularly between 2020 and early 2021, demonstrates increasing adoption and engagement. Notably, significant spikes in March 2021 and July 2024 reflect successful outreach efforts or events driving user activity. The steady engagement through 2023 and 2024, with peaks like July 2024 (2,592 sessions), underscores the VREs continued relevance and utility. This growing activity also implied that the WP had to deal with approximately 180 tickets that have been resolved, including 90 requests for support, 8 incident and bug reports, 32 requests for new features, and 48 requests for tasks related to 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., Bardi A. Pagano P. (2023) **SoBigData-PlusPlus - D9.5: SoBigData e- Infrastructure Common Facilities 2 https://iris.cnr.it/handle/20.500.14243/453867**

[5] 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 - D9.2: SoBigData e-Infrastructure Operation Report 2** https://iris.cnr.it/handle/20.500.14243/453865

[6] M. Assante et al. (2023) Virtual research environments co-creation: The D4Science experience. Concurrency Computat Pract Exper. 2023; 35(18):e6925. https://doi.org/10.1002/cpe.6925