H2020-EINFRA-2015-1

VI-SEEM

VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean

Deliverable D4.4

Final report on data, services, availability and usage

Author(s): Silviu Panica (editor)
Status –Version: Final – a
Date: Aug 31st, 2018
Distribution - Type: Public

Abstract: Deliverable D4.4 provides the description of the final data sets and services provided to the VRE as well as a usage and performance assessment report for those sets and services.

© Copyright by the VI-SEEM Consortium

The VI-SEEM Consortium consists of:

GRNET	Coordinating Contractor	Greece
CYI	Contractor	Cyprus
IICT-BAS	Contractor	Bulgaria
IPB	Contractor	Serbia
NIIF	Contractor	Hungary
UVT	Contractor	Romania
UPT	Contractor	Albania
UNI BL	Contractor	Bosnia-Herzegovina
UKIM	Contractor	FYR of Macedonia
UOM	Contractor	Montenegro
RENAM	Contractor	Moldova (Republic of)
IIAP-NAS-RA	Contractor	Armenia
GRENA	Contractor	Georgia
BA	Contractor	Egypt
IUCC	Contractor	Israel
SESAME	Contractor	Jordan

The VI-SEEM project is funded by the European Commission under the Horizon 2020 e-Infrastructures grant agreement no. 675121.
This document contains material, which is the copyright of certain VI-SEEM beneficiaries and the European Commission and may not be reproduced or copied without permission. The information herein does not express the opinion of the European Commission. The European Commission is not responsible for any use that might be made of data appearing herein. The VI-SEEM beneficiaries do not warrant that the information contained herein is capable of use, or that use of the information is free from risk and accept no liability for loss or damage suffered by any person using this information.

Document Revision History

Date	Issue	Author / Editor / Contributor	Summary of main changes
17/05/2018	a	Silviu Panica	ToC
18/06/2018	a	Silviu Panica, Ognjen Prnjat	Added content
20/08/2018	a	Silviu Panica, Dusan Vudragovic, Kyriakos Gkinis, Eleni Katragkou, Vladimir Dimitrov, Tamás Kazinczy	Added and integrated contributions from the partners
28/08/2018	a	Georgios Artopoulos, Tamas Maray	Added partner contribution
28/08/2018	a	Silviu Panica, Ognjen Prnjat	Final review, quality control, final updates

© VI-SEEM consortium

Table of contents

 © VI-SEEM consortium

References

[1] ownCloud – Distributed Cloud Collaboration Suite (https://owncloud.org/)

[2] Globus Toolkit GridFTP - High-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks (http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/)

[3] Apache Hadoop - open-source software for reliable, scalable, distributed massive computing (http://hadoop.apache.org/)

[4] Gatling – Load and performance testing tool (https://gatling.io/)

 © VI-SEEM consortium

List of Tables

 © VI-SEEM consortium

List of Figures

 © VI-SEEM consortium

Glossary

AAI	Authentication and Authorization Infrastructure
API	Application Programming Interface
CA	Certification Authority
CDI	Collaborative Data Infrastructure
CDMI	Cloud Data Management Interface
DSI	Data Storage Interface
DSS	Data Staging Script
EM	Eastern Mediterranean
EPIC	European Persistent Identifier Consortium
EUDAT	European Data Infrastructure
GB	Gigabyte
GPFS	General Parallel File System
GRIB	GRIdded Binary
GridFTP	File Tranfer Protocol for Grid computing
HDFS	Hadoop Distributed File System
HPC	High Performance Computing
HTTP	Hypertext Transfer Protocol
iRODS	integrated Rule-Oriented Data System
iCAT	iRODS metadata catalogue
MB	Megabyte
MPI	Message Passing Interface
netCDF	network Common Data Form
NFS	Network File System
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OPeNDAP	Open-source Project for a Network Data Access Protocol
pbdR	Programming with Big Data in R
PID	Persistent Identifiers
PEP	Policy Enforcement Point
PID	Persistent Identifier
REST	Representational State Transfer
SC	Scientific Community
SEE	South East European
SEEM	South East Europe and Eastern Mediterranean
SLA	Service Level Agreement
SPMD	Simple program, Multiple data
SQL	Structured Query Language
TB	Terabyte
UI	User Interface
UUID	Universally Unique Identifier
VAS	VI-SEEM Archival Service
VDAS	VI-SEEM Data Analysis Service

 © VI-SEEM consortium

VDDS	VI-SEEM Data Discovery Service
VI-SEEM	VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean
VLS	VI-SEEM Work Storage Space / Local Storage and Data Staging
VM	Virtual Machine
VRS	VI-SEEM Repository Service
VRE	Virtual Research Environment
VSS	VI-SEEM Simple Storage Service
WP	Work Package

 © VI-SEEM consortium

Executive summary

What is the focus of this Deliverable?
The D4.4 deliverable presents an updated version on the final data sets and services provided by the VREs of VI-SEEM and it also tackles their usage and performance aspects.

What is next in the process to deliver the VI-SEEM results?
The generic data services defined in D4.1 and detailed in D4.3 were successfully deployed and maintained during the project life cycle. Each service is hosted by one of the VI-SEEM consortium partners, as outlined in D4.3. The D4.3 information is updated in the current deliverable. The final VI-SEEM platform is operational and was functionally tested in production by the internal VI-SEEM consortium research teams and externally through the projects' open call users. The platform is currently operated, monitored and regularly updated continuously.

What are the deliverable contents?
This deliverable gives a final high-level description of each of the services of the final VI-SEEM data platform by highlighting the latest and final changes carried out to the data services: all of these with respect to the last reporting period of D4.3. Moreover, this deliverable will also provide usage data regarding the data sets and services' performance.
The deliverable is split in three parts: (a) brief data services description together with the usage reports, (b) datasets reports that include the final descriptive information and usage statistics and (c) user-perspective performance analysis of the VI-SEEM data platform services.

Conclusions and recommendations
The final VI-SEEM data services address a complex data management problem were different types of data models have to be stored, indexed and easily searched by the end-users (either final users or third-party integrating services). The VI-SEEM data platform was intensively used in production (by external users through the open calls) and proved to successfully manage the data management problem at hand. Moreover, the performance assessment report shows that the current data services are capable of handling large number of workloads, for the current VI-SEEM supported use cases and for the potential future scenarios. Moreover, important scientific results, that cover different research fields, are available to download and use to sustain further research activities.

 © VI-SEEM consortium

1. Introduction

This deliverable provides the updates on the description of the final VI-SEEM data platform, with respect to the information already provided in D4.3. Moreover, it also covers information about the usage and performance assessment.

The VI-SEEM data platform now consists of six general data services as well as the persistent identifier service. The services are also integrated with the VI-SEEM authentication and authorization infrastructure.

The deliverable presents, in Chapter 2, an updated description of the VI-SEEM data platform services, with respect to D4.3, together with service usage reports. Chapter 3 presents the final description of the available datasets officially registered in the VI-SEEM repositories, and usage statistics based on the logging reports provided by the repository services. In Chapter 4 we tackle the performance assessment problem by providing the results of the tests conducted to verify how the data services handle different workload scenarios. These scenarios simulate users or third-party services that have to consume the available datasets. Finally, in Chapter 5 we draw the final conclusions in what regards the final VI-SEEM data platform.

 © VI-SEEM consortium

2. VI-SEEM data platform – description and usage reports

The final VI-SEEM data platform has been envisioned to bring the generic data services of VI-SEEM together, provision a wide range of datasets made available by selected applications, and provide adequate support that helps scientific communities in utilizing them.
The data services of the final data platform are the following:

- VI-SEEM simple storage service (VSS)
- VI-SEEM repository service (VRS)
- VI-SEEM archival service (VAS)
- VI-SEEM work storage space / local storage and data staging (VLS)
- VI-SEEM data discovery service (VDDS)
- VI-SEEM data analysis service (VDAS)

The usage report will contain metrics dedicated for each data service provider apart, based on the particularity of the specific data provider. Therefore, the metrics are not generalized to cover common usage and performance aspects but mapped to the specific functionalities that a data service provider offers.

2.1. VI-SEEM Simple Storage Service (VSS)

The VI-SEEM Simple Storage service (https://simplestorage.vi-seem.eu/) allows VRC (Virtual Research Community) members to keep and sync their research data on various devices, as well as to share them, thus making it a useful tool in a collaborative environment. The access is enabled via web browsers (Figure), desktop and mobile clients. The service is operational and in use since September 2016.
The service is based on ownCloud platform [1], and therefore it inherits all its features:

- user-friendly web interface that allows connection from any web browser;
- desktop clients for popular operating systems;
- selective synchronization and version control;
- access and management of deleted and encrypted files;
- a video, PDF, ODF viewer.

VI-SEEM Simple Storage service (VSS) is a secure data service based on ownCloud technology that helps the VI-SEEM community in storing and sharing short-lived research data. By its nature, this service supports versioning and synching across different computers/devices.

Service complete description:

https://wiki.vi-seem.eu/index.php/Simple_Storage_Service

Service endpoint:

https://simplestorage.vi-seem.eu/

Service usage report:

Metric	Value
Total used storage size (GB)	12500
Total number of files	80000
Total number of registered users	163

Table 1: VSS service usage report

Figure 1: Number of users (blue), occupied storage space in GB (green), and number of registered files (orange) at the VI-SEEM Simple Storage service over time

2.2. VI-SEEM Repository Service (VRS)

VI-SEEM Repository Service (VRS) is the main storage service of the VI-SEEM community that holds "Regional Community Datasets". The VRS is also the platform to host all kinds of additional data such as publications (and their associated data), software (or references to software), workflow descriptions (e.g. how to generate research data) or even materials targeting the general public. (e.g. images, videos etc.) VRS is integrated with the VI-SEEM persistent identifier service as an assigned PID is required for each digital object (item, collection, community).

Service complete description:

https://wiki.vi-seem.eu/index.php/Repository_Service

Service endpoint:

https://repo.vi-seem.eu

Service usage report:

Metric	Value
Total used storage size (GB)	19000
Total number of items (a group/dataset of multiple files)	260
Total number of registered users	46
Total number of files uploads	4915
Total number of files downloads	3990

Table 2: VRS service usage report

2.3. VI-SEEM Archival Service (VAS)

VI-SEEM Archival Service (VAS) targets data that is selected for long term retention and future reference. This service is provided at multiple sites forming a federation making geographical redundancy of archived data possible. This service is also coupled with VI-SEEM work storage space / local storage and data staging to help VI-SEEM users making safe data replication part of their workflows (e.g. in the case where a result data set of a computation is selected for long term preservation). VAS is deployed at six sites (BA, GRNET, IICT-BAS, IPB, IUCC and NIIF) where each has its local policies as restrictions may apply regarding data sets.

Service complete description:

https://wiki.vi-seem.eu/index.php/Archival_Service

Service endpoints:

1. baicat.bibalex.org:1247
2. irods.aris.grnet.gr:1247
3. icat.avitohol.acad.bg:1247
4. irods.ipb.ac.rs:1247
5. icat.vi-seem.iucc.ac.il:1247
6. niificat.niif.hu:1247

Service usage report:

Metric	Value
Total used storage size (GB)	3000
Total number of ingested digital objects	1000

Table 3: VAS service usage report

This service usage reports are not final but intermediate values. The final results will be archived after the end of the project as a procedure for long terms data preservation and further data usage in different other non-VI-SEEM related scenarios.

 © VI-SEEM consortium

2.4. VI-SEEM Work Storage Space / Local Storage and Data Staging (VLS)

VI-SEEM work storage space / local storage and data staging service (VLS) offers storage for short-term workloads near grid and/or HPC facilities on one hand and data staging capability on the other so that users could readily have their input for computation and may stage out computation results as part of their scientific workflow. Most of the local storage services are using GridFTP technology [2] for exposing the data flow operations.

Service complete description:

https://wiki.vi-seem.eu/index.php/Accessing_VI-SEEM-Storage_resources

Service endpoints:

1. aa112642.archive.bibalex.org:2811
2. login2.cytera.cyi.ac.cy:2812
3. se.sg.grena.ge:2811
4. gftp.aris.grnet.gr:2811
5. gridgtp.grid.am:2811
6. gftp.avitohol.acad.bg:2811
7. paradox.ipb.ac.rs:2811
8. login.debrecen2.hpc.niif.hu:2811
9. gridftp.renam.md:2811
10. se.hpgcc.finki.ukim.mk:2811
11. gridftp.viseem.hpc.uvt.ro:2811

Service usage report:

Service usage reports for VLS endpoints are not relevant in context of VI-SEEM data services architecture. GridFTP endpoints are used to store temporary data, downloaded from the VRS. This temporary data is used by the researchers, on the computational clusters, to run custom simulations and the output data is again registered in the central VI-SEEM repository if is intended for further usage.
In this context, the usage report is not relevant from the VI-SEEM core data services and repositories usage analysis.

2.5. VI-SEEM Data Discovery (VDDS)

VI-SEEM data discovery service (VDDS) is a service provided to VI-SEEM users for flexible searching for data discovery. It is based on harvesting various research and other repositories (including VRS) for metadata. As a result, it is possible for the users of VI-SEEM to search for keywords, partial phrases, creator, organization, publisher, time of publishing, versions, tags, research areas and communities etc. and see results in a user-friendly way. It is also possible to refine a search based on a previous result.

 © VI-SEEM consortium

Service complete description:

https://wiki.vi-seem.eu/index.php/Data_Discovery_Service

Service endpoint:

https://search.vi-seem.eu/

Service usage report:

Metric	Value
Total number of queries	529
Total number of indexed metadata’s	253
Total number of the endpoint hits:	54206

Table 4 - VDDS service usage report

2.6. VI-SEEM Data Analysis Service (VDAS)

VI-SEEM data analysis service (VDAS) provides the capability to carefully and efficiently investigate and analyse even very large, unstructured datasets. VDAS is based on Apache Hadoop [3]. Users gain access to the login node of the analysis cluster where they could manage data upload and interact with the distributed file system through command line utilities.

Service complete description:

https://wiki.vi-seem.eu/index.php/Data_Analysis_Service

Service endpoint:

hadoop.ipb.ac.rs (via command line or API)

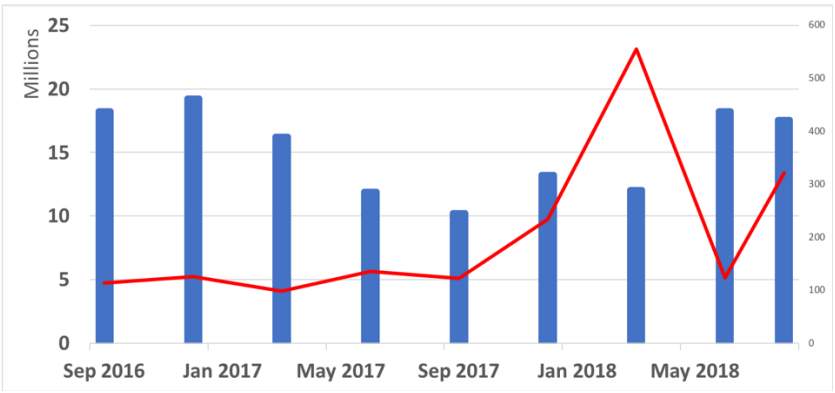

Figure 2: Usage of the VI-SEEM data analysis service: total execution time (blue bars) and total number of jobs (red line) per month.

Service usage report:

Metric	Value
Total HDFS space used for the completed jobs (GB):	2800
Number of jobs	1826
Execution time for the completed jobs	Average 37 hours

Table 5: VDAS service usage report

2.7. VI-SEEM Persistent Identifier Service (VPID)

VI-SEEM persistent identifier service provides globally unique identifiers for digital objects and other internet resources. This service helps the VI-SEEM community by making data findable as the PID could be resolved through the resolution service which redirects the user to the registered location of the resource.

Service endpoints:

- https://epic.grnet.gr/api/v2/handles/11239
- https://epic.grnet.gr/api/v2/handles/11500

Service usage report:

Metric	Value
Total number of hits	2048

Table 6: VPID service usage report

 © VI-SEEM consortium

3. VI-SEEM Datasets - description and usage reports

In this chapter we present the datasets used by researchers, from different research communities, in their dedicated work supported by the VI-SEEM platform. The dedicated communities deal with complex scenarios which involve both computational power and large storage requirements. VI-SEEM supports the latter by offering different storage types that will satisfy both community specific data domain and the type of application execution model that might involve a different types of storage architecture to be used. In this way the VI-SEEM data platform offers flexibility and supports different data models to be used. In the following subsections we will present each research community datasets with their usage characteristics based on the logging reports from the data service providers.

The set of metrics used in this chapter usage reports are:

- **Total storage (GB)** – the total amount of storage space used by the dataset;
- **Total number of hits/downloads** – the total number of dataset endpoint reach, either by other users or third-party services that consume those datasets;
- **Total number of datasets/files** – the total number of objects (dataset, files or similar) that are part of the current dataset.

3.1. VI-SEEM Climate Sciences

The climate modelling and weather forecasting community has traditionally very strong computational needs. In particular, the integration of various computational resources such as HPC and Grid jointly with data infrastructure that is addressed in VI-SEEM greatly supports research and operational activity of regional relevance. VI-SEEM will have strong impact on the Climate Modelling and weather forecasting communities. First, there is significant potential to share best practice and data for local and regional Climate Modelling, Weather forecasting and air quality simulations. The community will benefit from the combination of HPC and Grid computing jointly with the storage facilities as it heavily relies on data from very scattered locations.

ACIQLife

Dataset description: Atmospheric Composition Impact on Quality of Life and Human Health (WRF/CMAQ)
Dataset scope: Development of a methodology and performing reliable, comprehensive and detailed studies of the impact of lower atmosphere parameters and characteristics on the quality of life (QL) and health risks (HR) for the population in our country.

Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-263

 © VI-SEEM consortium

Dataset usage:

Metric	Value
Total storage (GB)	0.003
Total number of hits	169
Total number of datasets	6

Table 7: ACIQLife - Dataset usage report

ClimStudyArmenia

Dataset description: Accurate Prediction and Investigation of Weather and Climate in Armenia and South Caucasus (WRF)
Dataset scope: Methods and methodologies for accurate weather prediction and climate change based on series of experiments, as mountainous terrain of the country, the apparent ruggedness of the terrain, the big difference between relative altitudes, as well as atmospheric general circulation features make it challenge.

Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-222

Dataset usage:

Metric	Value
Total storage (GB)	0.807
Total number of hits	440
Total number of datasets	33

Table 8: ClimStudyArmenia - Dataset usage report

RCM Mena-CORDEX

Dataset description: Regional Climate Modelling (Middle East, North Africa) – WRF
Dataset scope: Very high horizontal resolution climate projections for the Middle East, North Africa and the eastern Mediterranean, improved climate change projections that will drive important vulnerability, impact and adaptation studies for the region, and resolution of smaller scale meteorological features critical for the realistic simulation of regional climate
Dataset endpoint:

- https://repo.vi-seem.eu/handle/21.15102/VISEEM-141
- https://repo.vi-seem.eu/handle/21.15102/VISEEM-183
- https://repo.vi-seem.eu/handle/21.15102/VISEEM-117
- https://repo.vi-seem.eu/handle/21.15102/VISEEM-174

Dataset usage:

Metric	Value
Total storage (GB)	3473.69
Total number of hits	4797
Total number of datasets	2410

Table 9: RCM Mena-Corder - Dataset usage report

 © VI-SEEM consortium

DREAMCLIMATE

Dataset description: Dust Regional Atmospheric Model Climatology
Dataset scope: Plan to downscale the atmospheric-dust DREAM model to fine horizontal resolution of 5-10 km, covering wide North Africa, Southern Europe and Middle East regions. Such model setup should be used to perform a decadal model execution and to produce corresponding dust concentration climatology.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-85

Dataset usage:

Metric	Value
Total storage (GB)	2.52
Total number of hits	145
Total number of datasets	14

Table 10: DREAMCLIMATE - Dataset usage report

ECMWF-ERAINT

Dataset description: AUTH_WRF371M_EUR0.44
Dataset scope: The dataset includes climatic variables of regional climate model simulations over Europe for the CORDEX initiative (www.cordex.org). The climatic simulations have been performed with the regional climate model WRF-AUTH for the hindcast simulation from 1990 to 2008 and are driven by the ERA interim reanalysis. The spatial resolution of the climate data is 50 Km and the temporal resolution 3 hours.

Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-28

Dataset usage:

Metric	Value
Total storage (GB)	55.91
Total number of hits	1448
Total number of datasets	47

Table 11: ECMWF-ERAINT - Dataset usage report

STARDEX and ETCCDI Climate Indices Datasets

Dataset description: Suite of STARDEX and ETCCDI Climate Indices Datasets based on E-OBS and CARPATCLIM Gridded Data
Dataset scope: The oncoming climate changes are the biggest challenge the mankind is faced with. The impacts of climate change are manifold and vary regionally, even locally, in their severity. For decades, most analyses of long-term global climate change using observational temperature and precipitation data have focused on changes in mean values. However, immediate damages to humans and their properties as well as the ecosystems, are not obviously caused by gradual changes in these variables but mainly by so-called extreme climate events. The

 © VI-SEEM consortium

relative rare occurrence of extremes makes it necessary to investigate long data records to determine significant changes in the frequency and intensity of extreme events. There are various methods to investigate extreme events, but the computation and analysis of climate indices (Cis) derived from daily data is probably the most widely used non-parametric approach.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-343

Dataset usage:

Metric	Value
Total storage (GB)	1.52
Total number of hits	169
Total number of datasets	10

Table 12: STARDEX and ETCCDI Climate Indices - Dataset usage report

TVRegCM

Dataset description: Tuning and Validation of the RegCM
Dataset scope: Adaptation and tuning of the RegCM model for the Balkan Peninsula and Bulgaria and thus development of a methodology able to predict possible changes of the regional climate for different global climate change scenarios and their impact on spatial/temporal distribution of precipitation, hence the global water budgets, to changes of the characteristics and spatial/temporal distribution of extreme, unfavourable and catastrophic events (drought, storms, hail, floods, fires, sea waves, soil erosion, etc.). All these changes will have influence on the ecosystems and on practically all sectors of the economy and human activity and consequently on the quality of life.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-23

Dataset usage:

Metric	Value
Total storage (GB)	0.639
Total number of hits	118
Total number of datasets	27

Table 13: TVRegCM - Dataset usage report

VIAM/NEA Regional Chemistry-Climate Model

Dataset description: VIAM/NEA Regional Chemistry-Climate Model
Dataset scope: Improving research in process-level understanding considering to the coupling and feedbacks above the territory of Georgia: dust emission (influence of climate, land surface state, etc.); dust ageing (cloud processing, physical and chemical interactions with other aerosols and gases); dust deposition (land use, lifecycles).
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-77

 © VI-SEEM consortium

Dataset usage:

Metric	Value
Total storage (GB)	1.46
Total number of hits	63
Total number of datasets	5

Table 14: VIAM/NEA Regional Chemistry-Climate Model - Dataset usage report

3.2. VI-SEEM Digital Cultural Heritage

The Cultural Heritage researchers pursue activities on a number of common themes and topics that will be impacted by the shared e-Infrastructure. Common data repositories and software, such as content management system MEDICI, algorithms for remote sensing image classification, idPromo for automatic object recognition etc., will advance the research capacity of the various groups to optimally utilize them. Beyond the data needs, VI-SEEM will also facilitate the slow transition of the Cultural Heritage community towards computational more intensive activities, such as high detail rendering of 3D modelling, and simulations of environmental influence on historical buildings. Shared datasets, easy remote access and visualization enabled by the VI-SEEM platform will offer a novel approach to Cultural Heritage research that can foster innovation in methodologies and applications used.

3DInv

Dataset description: Electrical Resistivity Tomography
Dataset scope: Electrical Resistivity Tomography (ERT) comprises one of the most important modern techniques of near surface applied geophysics. The method has met an increasing interest within the geophysical community due to its robustness and applicability in solving diverse problems related to: a) the mapping of geological formations and stratigraphy; b) the identification of underground zones related to water and mineral resources and geothermal activity; c) the detection of pollution zones and pollutants that flow in the earth’s subsurface; d) the extraction of quantitative information for buried archaeological relics; and, e) the spatial and temporal monitoring of the subsurface resistivity.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-335

Dataset usage:

Metric	Value
Total storage (GB)	0.0072
Total number of hits	54
Total number of files	26

Table 15: 3DInv - Dataset usage report

 © VI-SEEM consortium

AutoGr

Dataset description: Images matching with GRID system used by AutoGR application.
Dataset scope: The specific application is particularly suited for large image datasets, such as the aerial photographs collected with UAVs or during systematic aerial surveys. The GRID system is going to speed up the georeferencing process. The purpose of AutoGR is to be used as online service for image georeferencing.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-323

Dataset usage:

Metric	Value
Total storage (GB)	0.001
Total number of hits	32
Total number of files	4

Table 16: AutoGr - Dataset usage report

Banatica Collections

Dataset description: Banat Virtual Library in various picture format.
Dataset scope: The motivation of the CUL is that Banat region is characterized by the most diverse cultural, linguistic, national and religious in Romania. Thus, there are now in Banat region the following national communities: Romanian, Hungarian, German, Serbian, Italian, Slovak, Hebrew, Bulgarian, Russian, Ukrainian, Gypsy, Arabic, Persian, Indian, Chinese, Greeks, Africans, Asians, etc. All those communities left a footprint in the cultural common heritage. For a better understanding between different communities is mandatory to identify the common heritage, to know the history and the events one community can identify with, and a digital archive of the evolution of language and dialects in the region would enable mapping identities and stories of the different communities that populate the area.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-337

Dataset usage:

Metric	Value
Total storage (GB)	14.6
Total number of hits	183
Total number of files	1374

Table 17: Banatica Collections - Dataset usage report

CNN Features for Remote Sensing Image Classification

Dataset description: CNN Features for Remote Sensing Image Classification
Dataset scope: Collection SAT CNN Models contains code and pretrained convnet models for classification of satellite images. The convnets are trained on publicly available SAT-4 and SAT-6 datasets of satellite images (http://csc.lsu.edu/~saikat/deepsat/).

 © VI-SEEM consortium

Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-329

Dataset usage:

Metric	Value
Total storage (GB)	1.25
Total number of hits	29
Total number of files	10

Table 18: CNN Features for Remote Sensing Image Classification - Dataset usage report

CSAD

Dataset description: Centre for the Study of Ancient Documents, Oxford University
Dataset scope: Provide a focus for the study of ancient documents, while holding one of the largest epigraphical archives in the world. A wealth of archive materials that were previously unavailable online will be accessed by a worldwide audience.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-320

Dataset usage:

Metric	Value
Total storage (GB)	16.34
Total number of hits	64
Total number of files	802

Table 19: CSAD - Dataset usage report

DataCrowds

Dataset description: Data crowds tracking and information retrieval
Dataset scope: Training sets for data crowd automatic information retrieval and extraction.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-314

Dataset usage:

Metric	Value
Total storage (GB)	0.632
Total number of hits	162
Total number of files	9

Table 20: DataCrowds - Dataset usage report

ELKA

Dataset description: Electronic Corpus of Karamanlidika Texts
Dataset scope: The final product of the Electronic Corpus of Karamanlidika (ELKA) will offer access to graphic, phonetic and morphemic varieties search, as well as

ceramics and pottery study and restoration data management, visualization and presentation to wider communities.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-333

Dataset usage:

Metric	Value
Total storage (GB)	0.0054
Total number of hits	40
Total number of files	43

Table 21: ELKA - Dataset usage report

PETRA

Dataset description: Characterization and Conservation of Paintings on walls and sculptures from Nabataean Petra
Dataset scope: Characterization of newly excavated painted marble sculptures from Petra and of gilded wall paintings from Petra. The application will develop a method for confocal μXRF and μXANES, and experimental conservation material for gold on painted surfaces. Moreover, the application will be surveying, documenting and condition assessment of wall and sculpture paintings in Petra.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-325

Dataset usage:

Metric	Value
Total storage (GB)	5.02
Total number of hits	64
Total number of files	1339

Table 22: PETRA - Dataset usage report

RIVEEL3D

Dataset description: A mobile viewing platform for viewing artifacts and media in appropriate locations using GPS coordinates.
Dataset scope: Images for showing artifacts and media based on GPS location.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-331

Dataset usage:

Metric	Value
Total storage (GB)	5.89
Total number of hits	29
Total number of files	19

Table 23: RIVEEL3D - Dataset usage report

 © VI-SEEM consortium

3.3. VI-SEEM Life Sciences

Advances in computational infrastructure during the last decade have facilitated the development of biological data analysis for big data and computational biology as key research methodologies in both academia and industry. The use of computers in biology has enabled our better understanding of mechanistic aspects in health and disease and has accelerated the development of novel therapeutics. In this project, the Life Sciences Research Community is chosen because of its central role in achieving a higher quality of life in the SEEM region. The aim of the VRE is to create and provide the necessary services over a capable infrastructure to facilitate research for understanding of disease mechanisms in the SEE and EM populations.

Arachidonic Acid

Dataset description: This dataset contains Molecular Dynamics simulations trajectories of the effect of a mono trans isomer of arachidonic acid (C20:4-5trans,8cis,11cis,14cis). The input and output files are in GROMACS format.
Dataset scope: To investigate the effect of a mono trans isomer of arachidonic acid (C20:4-5trans,8cis,11cis,14cis) produced by free radicals in physiological concentration on a model erythrocyte membrane using a combined experimental and theoretical approach. Molecular Dynamics (MD) simulations of two model lipid bilayers containing arachidonic acid and its 5-trans isomer in 3% mol. were carried out for this purpose.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-312

Dataset usage:

Metric	Value
Total storage (GB)	375.81
Total number of hits	52
Total number of datasets	1

Table 24: Arachidonic Acid - Dataset usage report

D3R

Dataset description: Predicting the structure of the FXR-ligand inhibitor complexes and affinities using computer-aided drug design within the D3R challenge
Dataset scope: Enhancing the predictability of ligand-protein binding pose prediction using computer-aided drug design. Delivering guidelines for lead optimization using free energy perturbation calculations.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-272

Dataset usage:

Metric	Value
Total storage (GB)	32.43

 © VI-SEEM consortium

Total number of hits	76
Total number of datasets	4

Table 25: D3R - Dataset usage report

MD Simulations

Dataset description: MD Simulations of Biomolecules
Dataset scope: Modelling and Molecular Dynamics (MD) study of identified drug targets and computer-aided drug design. Simulations of biomolecules important in health and disease.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-115

Dataset usage:

Metric	Value
Total storage (GB)	50.7
Total number of hits	71
Total number of datasets	1

Table 26: MD Simulations - Dataset usage report

PI3Ka

Dataset description: PI3Ka datasets used to assess the mechanism of protein overactivation by performing extensive MD simulations, to examine conformational changes differing among wild type (WT) and mutant proteins as they occur in the microsecond timescale.
Dataset scope: The PIK3CA gene is one of the most frequently mutated oncogenes in human cancers. It encodes p110a the catalytic subunit of phosphatidylinositol 3-kinase alpha (PI3Ka which activates signalling cascades leading to cell proliferation, survival, and cell growth. The most frequent mutation in PIK3CA is H1047R, which results in enzymatic overactivation. Understanding how the H1047R mutation causes the enhanced activity of the protein in atomic detail is central to developing mutant-specific therapeutics for cancer.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-171

Dataset usage:

Metric	Value
Total storage (GB)	12520
Total number of hits	200
Total number of datasets	16

Table 27: PI3Ka - Dataset usage report

PSOMI

Dataset description: Protein-small-organic-molecules-interaction
Dataset scope: Connect pure theoretical and practical organic chemistry research with practical application and usage of newly synthesized organic molecules. So far

 © VI-SEEM consortium

newly synthesized molecules or group of molecules have never been tested for biological activity. The results of research will be of great importance for the understanding of ligand-receptor in simulated “live” system.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-76

Dataset usage:

Metric	Value
Total storage (GB)	0.173
Total number of hits	45
Total number of datasets	3

Table 28: PSOMI - Dataset usage report

RXRa

Dataset description: This dataset contains Molecular Dynamics simulations trajectories of the dimer RAR-RXRa in the RXRa normal (wild-type) and mutated RXRa S427F form. The input and output files are in GROMACS format.
Dataset scope: Retinoic acid receptors (RARs) and Retinoid X nuclear receptors (RXRs) are ligand-dependent transcriptional modulators that execute their biological action through the generation of functional heterodimers. RXR acts as an obligate dimer partner in many signalling pathways, gene regulation by rexinoids depending on the ligated state of the specific heterodimeric partner. One of these dimers is formed with the retinoic acid receptor (RAR) or with, which is a type of nuclear receptor, which acts as a transcription factor. The dimer formation is protective in cancer. A single point mutation on RXRa, S427F, which is found in 5% of patients with bladder cancer, is located exactly at the dimerization interface; however, its mechanism of action is unknown. To address the question of the effect of mutation on the dimerization of RXRa and RAR, we performed MD simulations to understand how the change of serine to phenyalanine at position 427 is suppressing the dimer function and is thus implicated in cancer.
Dataset endpoint: https://repo.vi-seem.eu/handle/21.15102/VISEEM-167

Dataset usage:

Metric	Value
Total storage (GB)	438.19
Total number of downloads/hits	58
Total number of datasets	2

Table 29: RXRa - Dataset usage report

 © VI-SEEM consortium

4. Data services performance reports

The following chapter will present the report over the performance tests conducted to measure the ability of the data services to react in case of a traffic burst. The increasing trend of data services traffic can be triggered either by final users that access the services directly or by the computational services that programmatically use the data services API to discover and download scientific datasets for processing.

The performance tests are aiming to measure the two aspects: (a) the maximum number of requests per time frame (in seconds) that the current data services deployments can handle and (b) the maximum number of successful responses the services are able to return before going in the blocking state because of the high traffic loads. The performance tests are applied on the data services that are publicly exposed, were the users can directly interact and has impact over the performance. The data services to be tested are the following:

- VI-SEEM Simple Storage Service (VSS)
- VI-SEEM Repository Service (VRS)
- VI-SEEM Data Discovery Service (VDDS)

For the other data services these tests are irrelevant because of the architectural aspects of the services. Direct users cannot directly perform high traffic loads that has impact over the performance. Moreover, in some cases, like VI-SEEM Data Analysis Service, it uses a distributed architecture that has mechanisms to overcome high traffic load or usage by adjusting the hosting resources dynamically. In this way the service is able to increase or decrease its’ performance capabilities based in the external load.

For the performance tests we are using Gatling tool [4]. This tool is written in Java and uses Scala language to describe the tests to be performed. Beside the tests, Gatling is able to also draw performances charts by addressing all the metrics defined by the testing scenario. Gatling’s architecture is asynchronous, so it can deliver high performance execution times by simulating hundreds of thousand users on a single machine. The machine from which the tests were conducted has a 16 cores Intel® Core-i7® CPU at 3.2Ghz, with 16GB of RAM, 512GB SSD storage and an internet connection of 10Gbps
In the tests, Gatling was configured to feed the data services with different requests, with a linear ramp over 120 seconds. The number of potential users was increased during the instances starting with 600 users and going up to 4800 users (at each step the number of users was doubled).

VI-SEEM Simple Storage Service (VSS)

This service runs on a machine with 24 Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz cores, 64 GB of RAM, and 16 TB of storage space.

Number of users	Response time in ms (mean)	Throughput (requests/s)	Successful requests	Errors
300	72	174.99	21146	28
600	70	351.22	43200	0
1200	74	685.71	86400	0
2400	476	640.33	91055	4355
4800	482	716.92	106303	8405

Table 30: VSS Data Service - performance report table

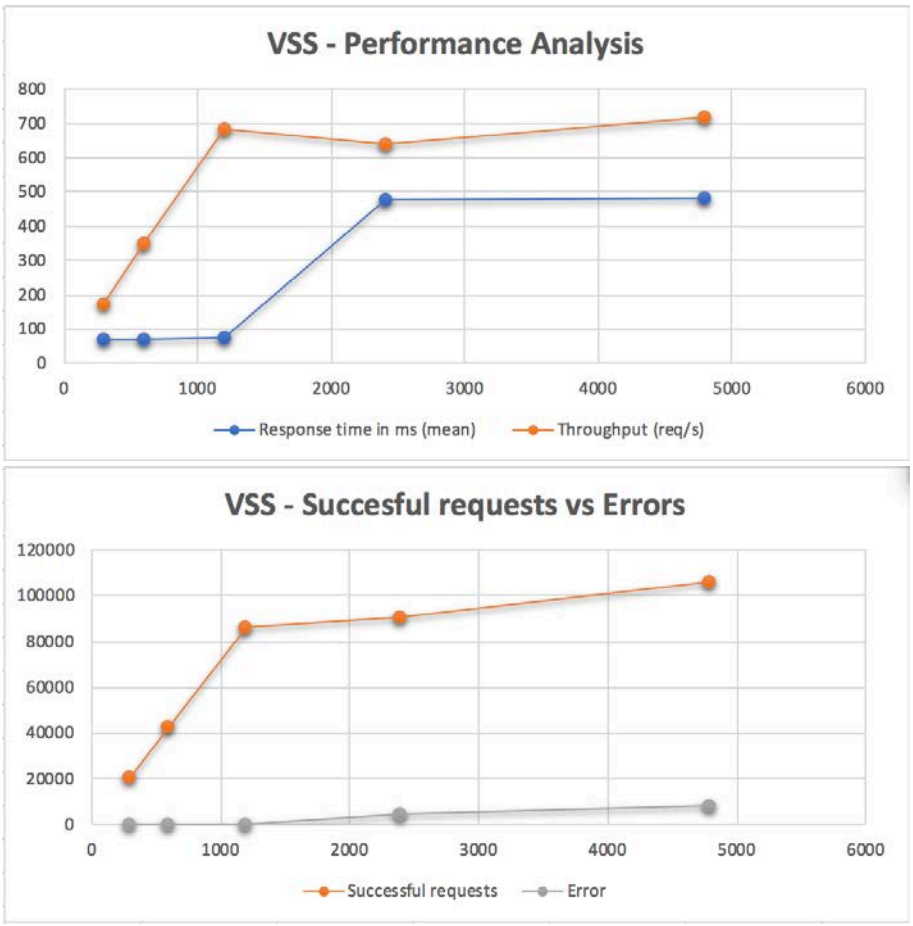

Figure 3: VSS performance analysis chart

VI-SEEM Repository Service (VSS)

VI-SEEM Repository is hosted on a group of VMs provided by the infrastructure of GRNET’s HPC[1] service. VI-SEEM Repository is connected to the GEANT network and therefore the research communities via 2x10Gbit/s connections. The bit stream store is connected to ARIS parallel file system (GPFS) that has a 1 PB of disk capacity.

[1] GRNET HPC - http://hpc.grnet.gr/

Number of users	Response time in ms (mean)	Throughput (requests/s)	Successful requests	Errors
300	384	22.13	2700	0
600	474	43.9	5398	2
1200	833	81.23	11022	273
2400	1302	122.92	16033	439
4800	7131	62.72	5426	4257

Table 31: VRS Data Service - performance report table

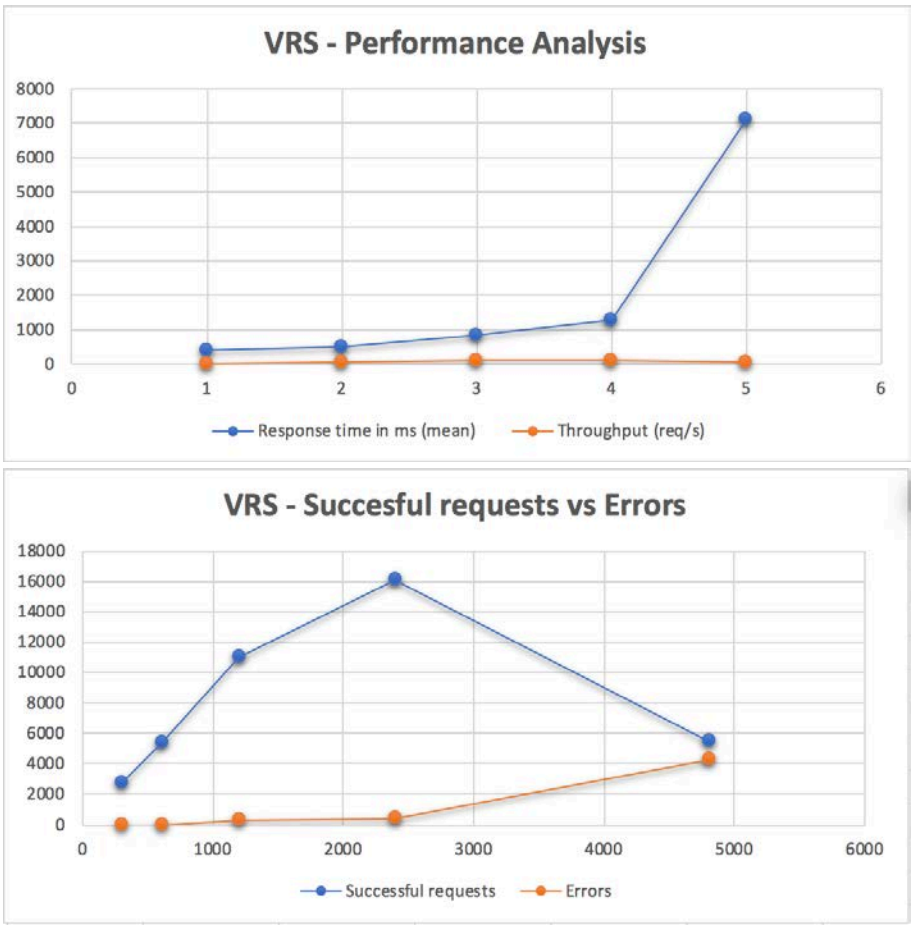

Figure 4: VRS performance analysis chart

VI-SEEM Data Discovery Service (VDDS)

This service runs on a machine with two 2 Intel(R) Xeon(R) CPU E5-5430 with 4 cores each, 16 GB of RAM, and 250GB of local storage and 2TB external storage.

Number of users	Response time in ms (mean)	Throughput (requests/s)	Successful requests	Errors
300	29,05	151	3300	300
600	58.065	153	6600	600
1200	2619	82.67	11595	1716

2400	2834	111.75	11094	5111
4800	2494	106	8996	11039

Table 32: VDDS Data Service - performance report table

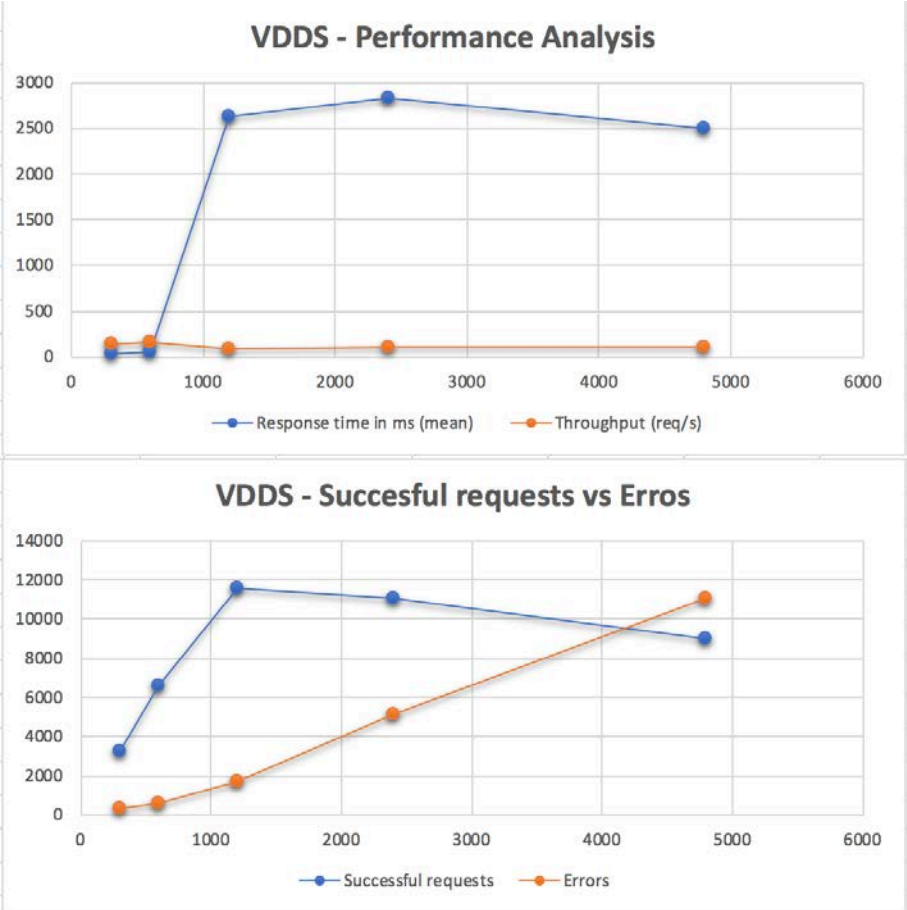

Figure 5: VDDS Performance analysis chart

Conclusions

The performance tests reveal that the data services are able to handle a high traffic load up to 2400 users that happen to access the resources in a very short time frame of 120 seconds. From that point above the service performance starts to degrade. This performance equals to a medium 120 requests per seconds, with bigger values for VSS that can handle up to 680 requests per second without the performance being affected. VSS performed much better than the other services because of the better hardware resources used for hosting the service.

These numbers are more than sufficient to support a normal targeted load for VI-SEEM data services. Nevertheless, if the situation imposes these performance limits can be pushed higher by changing the underlying hardware hosting infrastructure with resources that can handle extra traffic loads.

 © VI-SEEM consortium

5. Conclusions

The VI-SEEM data platform consists of seven high-level integrated services that cover all the different data storage and management needs of the user communities. These services are:

- the VI-SEEM simple storage service (VSS)
- the VI-SEEM repository service (VRS)
- the VI-SEEM archival service (VAS)
- the VI-SEEM work storage space/local storage and data staging (VLS) service
- the VI-SEEM data discovery service (VDDS)
- the VI-SEEM data analysis service (VDAS)
- the VI-SEEM persistent identifier service (VPID)

These services facilitated the data infrastructure for the research communities. The research communities were able to store and use the scientific data in different research activities. VI-SEEM managed to set up a multi-purpose datastore platform were different types of information can be stored, indexed and searched using web user interface or advanced APIs for programmatic integration with other third-party services. These facilities are offered by VSS and VRS services.
For scientific activities VI-SEEM integrated the data platform with the computational cluster through VLS service. This service allows easy data transfer from the main repository or archive to the computational cluster where the datasets are processed. Also, VI-SEEM offers a data-as-service service through VDAS, where the user only uploads its’ analysis program, defines some parameters and specifies the required input datasets references. The platform will handle all the data related operations (data search, copy and stage) and perform the analysis based on the user scenario.

In terms of datasets, VI-SEEM currently hosts different data types for various research fields. This data types are grouped on three research communities:

- Climate Sciences
- Digital and Cultural Heritage
- Life Sciences

The report presented in Chapter 3 demonstrates that the communities made use of VI-SEEM data platform by storing different scientific data on the dedicated repositories.

In terms of performance, in Chapter 4 we analysed the main data services that are publicly exposed and are in direct contact with the users or other consuming services. The tests outlined that the data services are correctly sized to support sufficient traffic load for the use case scenarios that VI-SEEM supports.

 © VI-SEEM consortium