H2020-EINFRA-2015-1

VI-SEEM

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

Deliverable D4.3

Description of the final data platform available to VRE users

Author(s): Tamas Maray (editor), Tamas Kazinczy

Status –Version: Final – d

Date: Sep 9, 2017

Distribution - Type: Internal

Abstract: Deliverable D4.3 – A detailed description of the final set of deployed data services, listing also the usage scenarios and their documentation for the newly available or modified services.

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.

Document Revision History

Date	**Issue**	**Author/Editor/Contributor**	**Summary of main changes**
04/08/2017	a	Tamas Maray	ToC
23/08/2017	b	Tamas Maray, Tamas Kazinczy, Ioannis Liabotis, Dusan Vudragovic, Todor Gurov	Compilation of the different partners’ contributions
28/08/2017	c	Tamas Maray	Draft version
15/09/2017	d	Ioannis Liabotis, Tamas Maray, Tamas Kazinczy	Final additions and corrections

Table of contents

References

[1] Project VI-SEEM-675121 - Annex I - Description of the action

[2] VI-SEEM deliverable D4.2

Description of the initial deployed data services

[3] VI-SEEM deliverable D4.1

Data sources and services deployment plan

[4] VI-SEEM deliverable D3.1

Infrastructure and services deployment plan

[5] VI-SEEM deliverable D5.1

Detailed technical implementation plan for VRE services and tools

[6] VI-SEEM deliverable D5.2

Data management plans

[7] EUDAT

http://www.eudat.eu/

[8] CKAN

http://ckan.org

[9] OwnCloud

http://owncloud.org/

[10] GRNET HPC

http://hpc.grnet.gr/

[11] DSspace

http:/dsspace.org/

[12] iRODS

http://www.irods.org/

[13] handle.net

http://www.hdl.net/

[14] GRNET’s Handle service

http://epic.grnet.gr/guides/overview/

List of Figures

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
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

Executive summary

What is the focus of this Deliverable?

The D4.3 deliverable is focusing on the design and implementation of the final data platform for the VRE users of VI-SEEM, while it also introduces the corresponding usage scenarios.

What is next in the process to deliver the VI-SEEM results?

All of the VI-SEEM generic data services that were defined in D4.1 have been successfully developed and deployed at one or more of VI-SEEM consortium partners' sites. Built on the experience gathered during the “initial service deployment” stage and pilot user projects, the data services have been further refined and tailored to the needs of the VRE users, and they have been integrated with the AAI infrastructure. Thus the final VI-SEEM data platform is now available, which will be maintained, operated, monitored and regularly updated during the rest of the project.

What are the deliverable contents?

This deliverable gives a high level description of each of the services of the final VI-SEEM data platform, and highlights the latest developments of the services carried out since the initial service deployment stage. Moreover the deliverable also provides the usage scenarios and their documentation of the services.

In Chapter 2 a short overview of our data services is presented. Next, in Chapter 3 the deployment of the final versions of the data services is discussed, including the technical parameters and also the use cases. Chapter 4 gives information about the latest developments on the different services.

Conclusions and recommendations

In the first period of the project the main goal of WP4 was to gather and to analyze the user requirements regarding the potential data services. The results of this work proved that the following generic data services are required by most of the user projects:

- a simple storage service that allows to store short lived data and to share this data by the users enabling efficient collaboration between them
- a repository service which makes it possible to publish datasets, data collections, software, documentation, publications, etc. for general availability and wider use
- an archival service to safely store valuable data, especially results and findings of research projects for longer term and future reference
- work storage space to temporarily store the data during the processing
- a data discovery service that provides an efficient tool to search the data
- a data analysis service which allows to perform sophisticated processing on even very large and unstructured data sets

Building upon the user requirements, in the next phase of the project the different data services have been selected, specified, designed and implemented resulting the initial version of the VI-SEEM data platform.

The usage of the storage services – especially the pilot projects – demonstrates the need for operating and maintaining these services. The experience resulted from this initial usage also helped to refine and improve the services after the initial deployment.

Although the development of the VI-SEEM data platform reached its production stage and can fully serve the VRE users, the experience of the pilot user projects and the observations gained from the daily operations of the services, during the period left from the VI-SEEM project, will lead to further fine tunings and minor modifications in order to improve the quality and usability of the services for the benefit of the VRE users.

1. Introduction

This deliverable provides a description of the final VI-SEEM data platform. It also describes the use cases of the services, and highlights the development carried out on each of the services since the initial deployment phase (described in D4.2).

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

The deliverable is organized as follows: In Chapter 2 a short overview of our data services is presented. Next, in Chapter 3 the deployment of the final versions of the data services is discussed, including the technical parameters and also the use cases. Chapter 4 gives information about the latest developments on the different services.

2. The VI-SEEM data platform

In deliverable D4.2 ("Description of the initial deployed data services") [1] a final 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. A brief description for each of those services of the final data platform is presented in this chapter.

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)

Besides several updates and configuration refinements, the main progress made since the initial deployment:

- full integration with VI-SEEM AAI for VSS and VRS
- PID service integration for VRS
- deployment of data services at new sites / broadening the VI-SEEM iRODS federation
- DSI gridFTP servers deployed at new VAS provider sites
- VI-SEEM AAI integration efforts for VDDS

2.1. *VI-SEEM simple storage service*

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.

VSS is hosted at the Institute of Physics Belgrade and is available at:

https://simplestorage.vi-seem.eu

The infrastructure:

CPU:	2x Intel E5-2620, 6 cores
Memory:	64 GB
Storage:	16 TB (RAID-6) + 2 TB (RAID-1)

Documentation and training material:

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

2.2. VI-SEEM repository service

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).

VRS is hosted at GRNET and is available for users at:

https://repo.vi-seem.eu

The infrastructure:

CPU:	vCPU
Network:	2x 10Gbit
Storage:	50 TB (GPFS)

Documentation and training material:

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

https://repo.vi-seem.eu/bitstream/handle/21.15102/VISEEM-21/VI-SEEM-Repository-Training-Material-VI-SEEM-Template.pdf

2.3. VI-SEEM archival service

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). As part of the final data platform, VAS is deployed at six sites (BA, GRNET, IICT-BAS, IPB, IUCC and NIIF) where each has its local policies (to provide controlled access to resources) as restrictions may apply regarding data sets (e.g. obligatory choice of sites for replication when geographical redundancy is required).

The infrastructure:

NIIF	
CPU:	24x vCPU* *: Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz
Memory:	48 GB
Network:	10 Gbit for each VM
Storage:	480 GB local + 50 TB iSCSI + 272 TB tape (HSM)

IICT-BAS

CPU :	2x Intel Xeon E5430, 4 cores, 2.6GHz, 12 MB cache
Memory:	16 GB
Network:	1 Gbit
Storage:	160 GB local

GRNET

CPU:	2x Intel(R) Xeon(R) CPU E5-2640 v2 @ 2.00GHz
Memory:	128 GB
Network :	10 Gbit
Storage :	1.6 TB local

IPB

CPU:	quad-core Intel Xeon E3-1220 v3 @ at 3.1 GHz
Memory:	4 GB
Storage:	500 GB local (RAID-1 mirror array) + 30 TB external

Documentation and training material:

https://wiki.vi-seem.eu/index.php/Archival_Service
http://wiki.vi-seem.eu/index.php/VI-SEEM_iRODS_installation_and_operation#Working_with_files_using_iCommands
http://wiki.vi-seem.eu/index.php/VI-SEEM_iRODS_installation_and_operation#Working_with_files_using_GridFTP

2.4. VI-SEEM work storage space / local storage and data staging

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.

Details of deployment had been already discussed in chapter 4.4.1 - as part of 4.4 ("VI-SEEM work storage space / local storage and data staging") - of the deliverable D4.2 ("Description of the initial deployed data services").

The VLS service is deployed at the following 12 sites:

BA, CYI, GRENA, GRNET, IIAP-NAS, IICT-BAS, IUCC, IPB, NIIF, RENAM, UKIM, UVT.

2.5. VI-SEEM data discovery service

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.

This service is deployed at IICT-BAS and is available at:
https://discovery.vi-seem.eu

The infrastructure:

CPU:	2x Intel Xeon E5430, 4 cores
Memory:	16 GB
Network:	2x1 Gbit
Storage:	250 GB local + 2 TB external

Documentation and training material:
https://wiki.vi-seem.eu/index.php/Data_Discovery_Service

2.6. VI-SEEM data analysis service

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.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. Analysis (in terms of defining map-reduce operations) could be done via a Java API. A template helps users in creating such projects.

VDAS is deployed at IPB and is available on the machine:
hadoop.ipb.ac.rs

The infrastructure:

CPU:	Xeon E3-1220-v3, 4 core 3x Xeon E5 2620 v3, 24 core
Memory:	64GB/node
Storage:	7 TB (HDFS)

Documentation and training material:
https://wiki.vi-seem.eu/index.php/Data_Analysis_Service
https://training.vi-seem.eu/index.php/data-and-visualizations/data

2.7. VI-SEEM persistent identifier service

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.

VI-SEEM persistent identifier service is provided for VI-SEEM by GRNET's Handle Service deployed on two different IaaS infrastructures for high availability.

PID management is done through EPIC [2] API endpoints:

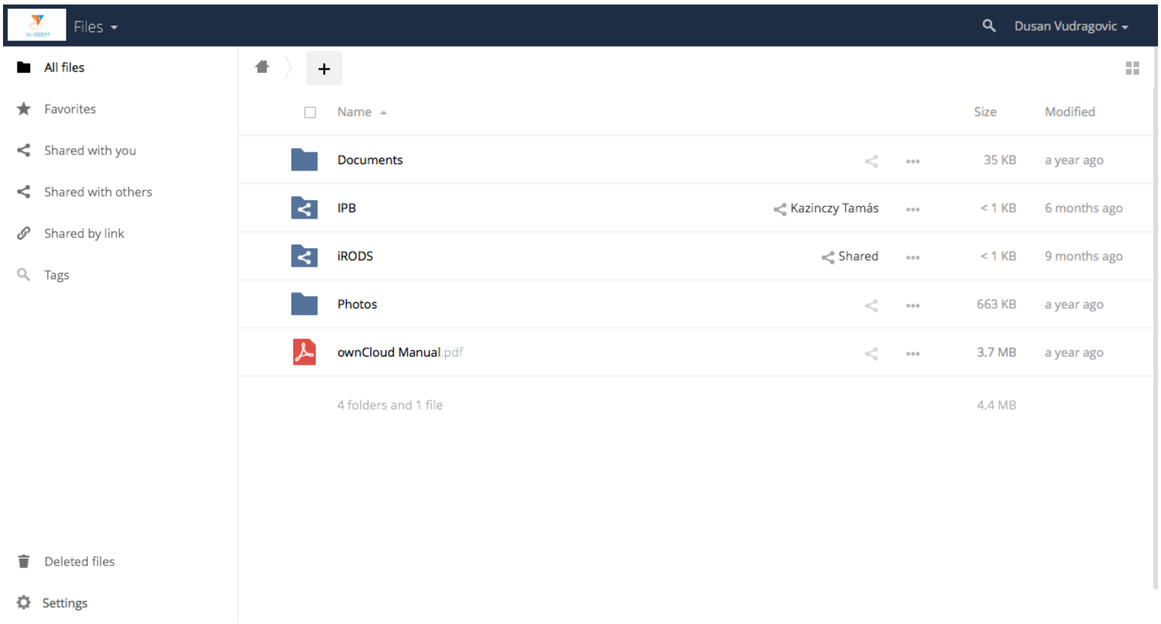
https://epic.grnet.gr/api/v2/handles/11239

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

3. Description and usage scenarios of the specific services of the final VI-SEEM data platform

3.1. 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 research data on various devices, as well as to share their data, thus making it a useful tool in collaborative environment. Access is enabled via web browsers (Figure), desktop and mobile clients.



Name		Size	Modified
Documents		35 KB	a year ago
IPB	Kazinczy Tamás	< 1 KB	6 months ago
iRODS	Shared	< 1 KB	9 months ago
Photos		663 KB	a year ago
ownCloud Manual.pdf		3.7 MB	a year ago
4 folders and 1 file		4.4 MB	

Figure 1 - VI-SEEM Simple Storage service web interface

The service is based on ownCloud platform (https://owncloud.org/), 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;
- video, PDF, ODF viewer.

Also, it supports different versions of Linux operating system (Red Hat/Centos, Debian, SUSE, Ubuntu). VI-SEEM instance is deployed on Debian 8 operating system. It runs under the Apache 2.4 web server, while the MySQL database stores metadata information.

The VI-SEEM Simple Storage service is deployed at the Institute of Physics Belgrade, and is based on version 9.0 of ownCloud. It is installed from the ownCloud repository (https://download.owncloud.org/download/repositories/). Starting from version 9.0, ownCloud deployment packages are divided into owncloud-files and owncloud-deps.

Package owncloud-files installs only ownCloud, without Apache, database, or PHP dependencies, while owncloud-deps package installs all dependencies (Apache, PHP, and MySQL). This package is not meant to be installed by itself but pulled in by the additional metapackage named owncloud.

After deployment of packages, the security of the VI-SEEM Simple Storage service is improved via a useful script that can be found on ownCloud documentation pages https://doc.owncloud.org/server/9.0/admin_manual/installation/installation_wizard.html. The script sets correct permissions on the deployed ownCloud files.

Installation is finalized by the Installation Wizard: setting username and password of ownCloud super-user, defining location of ownCloud data directory (where users files will actually be stored), choosing database that will be used (MySQL/MariaDB is recommended).

VI-SEEM Simple Storage service runs on HTTPS protocol, and has a browser-friendly SSL certificate that is issued by a trusted certificate authority. By default, ownCloud server is accessible under the /owncloud web path, and this is changed to / path via the Apache virtual host configuration. In addition, in order to improve performance of the service, we have enabled and tuned memory caching, where frequently-requested objects are stored in memory for faster retrieval.

VI-SEEM Simple Storage 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. Currently the service has 37 registered users that occupy approximately 1.5 TB of available storage space. The service is used during the project’s integration phases, as well as by the applications accepted in the 1st call for production use of resource and services.

3.2. VI-SEEM repository service (VRS)

The main storage service that will allow the users of the VI-SEEM VRE to deposit and share data is the VI-SEEM Repository Service (VRS). Such a repository in VI-SEEM is the main repository for hosting the "Regional Community Datasets" and therefore provide a component to host one of the main services of the VRE as specified in D5.1. It can also be used to host publications and their associated data as well as software or references to software and workflows, used to generate such data and publications.

The VRS is also the service for storing simplified data formats such as images, videos or others suitable also for the general public. The VRS is therefore the platform to host all of the types of data specified in the VI-SEEM data management plan, D5.2 [5], when users consider it suitable i.e. for sharing.

GRNET has deployed the service as a VM provided by the infrastructure of GRNET’s HPC [5] 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 [5] parallel file system (GPFS) that has a 1 PB of disk capacity. This storage capacity is shared with other services provided by GRNET to the project and at the national level. The available VI-SEEM repository storage capacity will depend on demand and the usage of the capacity of other storage services offered by GRNET not exceeding GRNET’s storage commitments as specified in VI-SEEM D3.1 [4], i.e. 50TB of disk space.

The VI-SEEM Repository has been deployed by GRNET and is available for all users at https://repo.vi-seem.eu/

The VI-SEEM repository is implemented using DSpace [6]. DSpace open source software is a turnkey repository application used by more than 1000+ organizations and institutions worldwide to provide durable access to digital resources. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets.

There are five main entities which are the hub of information structure and aggregation in VRS (DSPace)

- Communities: an administrative/logic groping of one or more collections (and sub-communities); In VI-SEEM we have defined 4 communities. The first three represent the 3 different scientific communities relevant to VI-SEEM, namely Climate Research, Life sciences and Digital Cultural heritage, while the last one is a general VI-SEEM community that holds generic data sets and documents. Communities and sub-communities are made up by:
    - A set of collections and/or sub-communities
    - A customizable home page
    - Dedicated Feed RSS, Set OAI-PMH, search and browsing
    - A group of users with administrative and managerial role
- Collections : a grouping of items which correspond to similar datasets, or to datasets that belong to a specific VI-SEEM application
- Item: a box which contains both a document metadata and one or more bitstream bundles
- Bundle: a groping of bistreams used to separate the original documents, those obtained from automatic process, archival and Creative Commons licence
- Bitstream i.e. digital content, or a fulltext

3.2.1. Submission process

The VRS submission follows three main steps:

1. description of the metadata (3 pages),
2. upload (file by file),
3. verification and the license agreement.

At the top of the submit pages users find 7 buttons representing each step in the submission process.

As users move through the process these buttons will change color. Once the user has started they can also use these buttons to move back and forth within the submission process by clicking on them.

After the submission has finished the quality assurance manager will get a notification and will have to make all the necessary checks based on the management plan until they approve the dataset and it ıs published in VRS.

3.3. VI-SEEM archival service (VAS)

Data archiving is the practice of moving data that is no longer being used or are being used on a less frequent fashion into a separate storage device. It is a single set or a collection of historical records specifically selected for long term retention and future reference. Additionally, data archives contain data that are important for future reference or it is important to preserve them for regulatory and audit purposes. In science archived data are important for future reference and reproducibility of scientific simulations. Data archives are indexed and have search capabilities so that files and parts of files can be easily located and retrieved.

The VI-SEEM archival service provides the capability of safe data replication for the users of the VI-SEEM VRE. The VI-SEEM Archival Service is implemented using iRODS.

3.3.1. iRODS zones of the final data platform

The final deployment of VAS consists of six distinct iRODS zones at BA, GRNET, IICT-BAS, IPB, IUCC and NIIF. These zones are connected together in a federation. This allows controlled access to resources at remote partners, i.e. local policies apply for remote zone users. Furthermore, it also allows better availability and performance for the users of the VI-SEEM VRE as each zone has its own iCAT server so it is less likely that one of them becomes a bottleneck.

The final deployment of VAS consists of the following six distinct iRODS zones:

Site	Zone name	Disk storage (TB)	Tape storage (TB)	Remarks
BA	BA	100	0	shared with VLS service
GRNET	GRNET_ARIS	50	210	shared with VRS and VLS
IICT-BAS	IICT_Zone	5	0	shared with VDDS
IPB	IPB	10	0	shared with VSS
IUCC	iuccZone	5	0	shared with VLS
NIIF	NIIF	50	272	

Federation is configured on respective catalog providers of each zone by setting:

```
* 'catalog_provider_hosts' (FQDN of the iCAT server)
* 'zone_name' (as seen in the table: "iRODS zones of final data
platform")
* 'negotiation_key' and 'zone_key'
    (these serve as the basis of zone-to-zone trust establishment)
```

After configuration zones of the VI-SEEM iRODS federation are visible:

```
$ ils /
/:
   C- /BA
   C- /GRNET_ARIS
   C- /IICT_Zone
   C- /IPB
   C- /iuccZone
   C- /NIIF
$
```

This does not mean though that access is granted as well. Controlled access to resources at remote partners means that local policies of the partner site apply. By default there is no policy that allows access so an error is expected when trying to list the root of another zone:

```
$ ils /BA
remote addresses: 193.224.20.45 ERROR: rcObjStat of /BA failed status =
-92111 CROSS_ZONE_SOCK_CONNECT_ERR, Connection refused
```

3.3.2. Data replication

There are several forms of safe data replication possible depending on user needs and restrictions:

- all data of a project or a specific data set (for which the equivalent iRODS term is collection) need to be replicated to one or more other zone(s)

- some data may be required to be replicated only inside a specific zone

Besides safe data replication, VAS is integrated with the VI-SEEM work storage space / local storage and data staging service so that users of the VI-SEEM VRE could move replicated data to local storage as part of a pre-compute step, or the reverse, safely replicate computation results to different zones. The distributed nature of VAS has the potential of a "connect to nearest zone" approach for users, e.g. this way, staging or transfer of results could use the nearest iRODS zone for best performance.

Integration with the VI-SEEM work storage space / local storage and data staging service is done installing the iRODS DSI plugin for gridFTP server.

The iRODS DSI plugin makes it possible for gridFTP users to access iRODS through gridFTP by tools such as globus-url-copy. As iRODS DSI enabled gridFTP servers operate on the iRODS namespace only, they are deployed as separate gridFTP server instances (in most cases beside non-DSI server instances running on the same physical resource).

3.3.3. Enforcing resource allocation policy

VI-SEEM contributors are granted resources through the open calls for access. Regarding VAS this means that specific amounts of archival storage are available for each of the projects and therefore quotas need to be used and local policies have to be implemented to enforce such allocation policies.

Examples of quota management and local policies are shown below.

- Quotas

  The zone administrator could check/set quotas for a user/group.
  Quotas could be set for a particular resource or globally.

- Listing quotas (as an unprivileged user)

  ```
  $ iquota –a

  Resource quotas for users:
  Resource: testResc
  User:  testuser
  Zone:  testZone
  Quota: 104,857,600 (104 million) bytes
  Over:  -58,760,814 (-58 million) bytes (Nearing quota)
  ```

  ```
  Global (total) quotas for users:
   None

  Group quotas on resources:
   None

  Group global (total) quotas:
   None

  Information was set at 2016-08-26.11:38:59
  ```

- The zone administrator could also list quotas with the admin command:

  ```
  `iadmin lq`
  ```

- Setting quota (requires administrator privileges)

  ```
  iadmin {suq <user> <resourceName or 'total'> <value> | sgq <group>
  <resourceName or 'total'>}
  ```

  where <user> is the name of iRODS user,
  <group> is the name of iRODS group,
  `<resourceName or 'total'>` is either the name of the resource, or 'total'
  if a global quota is being set, and <value> is the quota value in bytes.

- Enabling quota enforcement
  The `acRescQuotaPolicy` could be used to turn on Resource Quota enforcement.
  By default it is turned off.

- Calculating usage

The iRODS admin could calculate usage by issuing `iadmin cu`.
A simple rule could also be set to regularly calculate usage:

```
myTestRule {
#Administrator command to cause update to iCAT quota tables
   delay("<PLUSET>30s</PLUSET><EF>24h</EF>") {
     msiQuota;
     writeLine("serverLog","Updated quota check");
   }
}
INPUT null
OUTPUT ruleExecOut
```

If this rule is executed then after 30 seconds and once a day (*EF*: _Execution Frequency_) a task is executed that calculates usage and put a notice about this in the server log.

3.4. VI-SEEM work storage space/local storage and data staging (VLS)

As already discussed in D4.1, efficient computing at VI-SEEM partners offering grid and/or HPC facility requires quasi-local storage for short-term workloads on one hand, and a data staging capability on the other.

While the former is provided as is - ie. existing solutions that were already implemented at the partners – the latter was expected to be provided by all sites by a separate gridFTP server for each of them.

D4.2 provides all the details about gridFTP server implementation at VI-SEEM sites.

In the final data platform VI-SEEM has six sites with VAS, so those sites have deployed an iRODS DSI enabled gridFTP server each.

Access points and dedicated storage at each site:

BA	aa112642.archive.bibalex.org:2811 aa112643.archive.bibalex.org:2811	100 TB
CYI	login2.cytera.cyi.ac.cy:2812	20 TB
GRENA	se.sg.grena.ge:2811	2 TB
GRNET	gftp.aris.grnet.gr:2811	50 TB
IIAP-NAS	gridgtp.grid.am:2811	3 TB
IICT-BAS	gftp.avitohol.acad.bg:2811	5 TB
IPB	paradox.ipb.ac.rs:2811	10 TB

NIIF	login.debrecen2.hpc.niif.hu:2811	6 TB
RENAM	gridftp.renam.md:2811	1 TB
UKIM	se.hpgcc.finki.ukim.mk:2811	2 TB
UVT	gridftp.viseem.hpc.uvt.ro:2811	5 TB

3.4.1. Examples for data staging / transfer of computation results

- Stage data from local computer to HPC facility

```
globus-url-copy /path/to/input/file gsiftp://my.hpc.site:<gridFTP
port>/path/to/destination/dir/
```

where `<gridFTP port>` is the port in use by gridFTP (e.g. 2811)

- Stage data from iRODS to HPC facility

```
globus-url-copy gsiftp://my.irods.site:<DSI gridFTP
port>/myZone/path/to/input/file \
gsiftp://my.hpc.site:<gridFTP port>/path/to/destination/dir/
```

where `<gridFTP port>` is the port in use by gridFTP (e.g. 2811) and `<DSI gridFTP port>` is the port in use by iRODS DSI enabled gridFTP (e.g. 2812)

- Transfer computation results to local computer

```
globus-url-copy gsiftp://my.hpc.site:<gridFTP
port>/path/to/my/result /path/to/destination/dir/
```

- Transfer computation results to iRODS

```
globus-url-copy gsiftp://my.hpc.site:<gridFTP
port>/path/to/my/result \
       gsiftp://my.irods.site:<DSI gridFTP
port>/myZone/path/to/destination/dir/
```

- Staging/transferring data sets (collections)

To copy files in subdirectories, the '`-r`' option of `globus-url-copy` shall be used.

- Secure (encrypted) transfer

For encrypted transfers, the '`-dcpriv`' option of `globus-url-copy` shall be used.

3.5. VI-SEEM data discovery service (VDDS)

3.5.1. Service endpoint

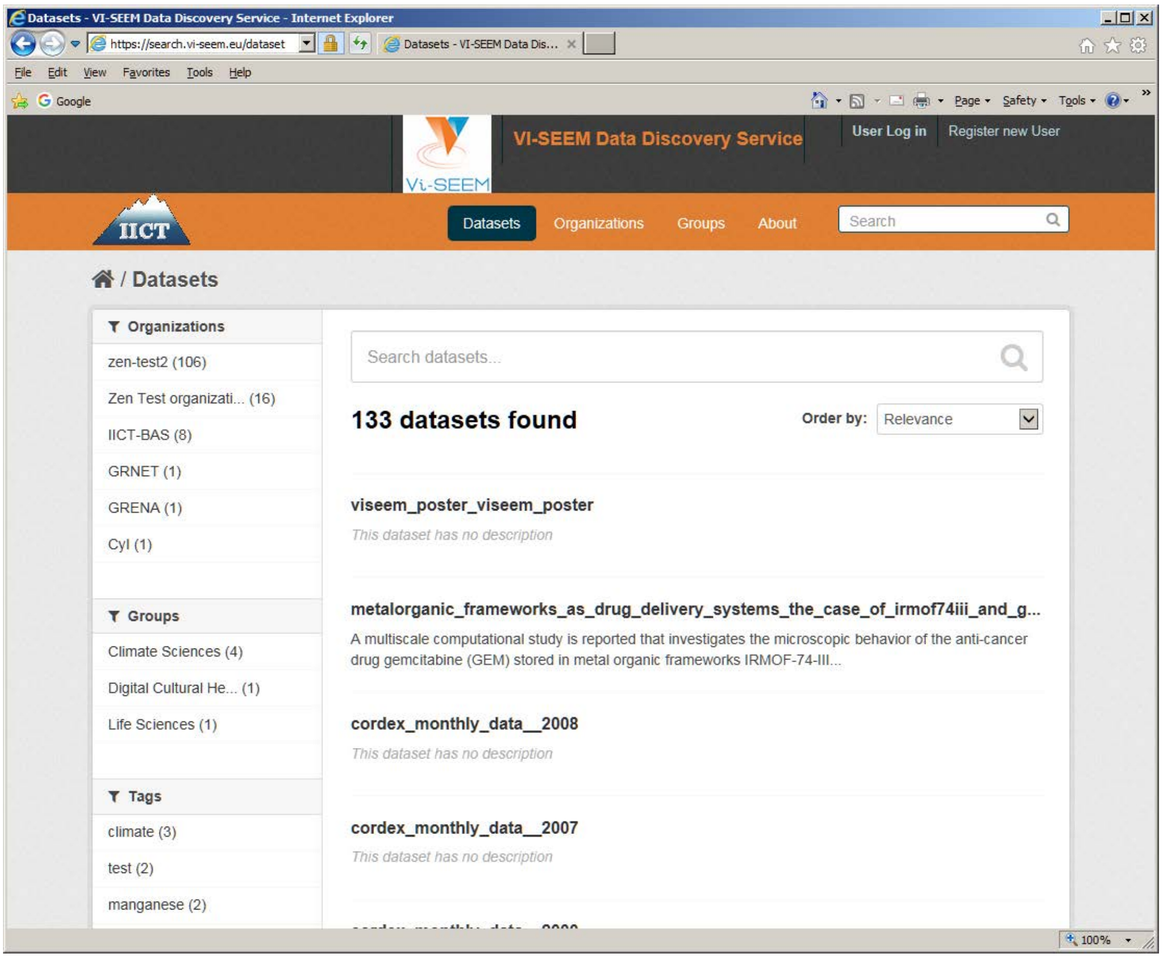
The service endpoint is changed from what was declared in the previous VI-SEEM deliverable D4.2. Currently it is https://search.vi-seem.eu.

The HTTPS communication is secured by a globally accepted certificate signed by TERENA SSL Certification Authority 3. An alternate service endpoint http://dds.avitohol.acad.bg is also available.

3.5.2. Detailed description

Follows a description of the usage of the service that demonstrates its main features.

The initial screen is presented below:


Datasets - VI-SEEM Data Discovery Service - Internet Explorer
https://search.vi-seem.eu/dataset
Datasets - VI-SEEM Data Dis...
File Edit View Favorites Tools Help
Google
Page Safety Tools
Vi-SEEM
VI-SEEM Data Discovery Service
User Log in
Register new User
IICT
Datasets
Organizations
Groups
About
Search
100%

/ Datasets

Organizations
- zen-test2 (106)
- Zen Test organizati... (16)
- IICT-BAS (8)
- GRNET (1)
- GRENA (1)
- CyI (1)

Groups
- Climate Sciences (4)
- Digital Cultural He... (1)
- Life Sciences (1)

Tags
- climate (3)
- test (2)
- manganese (2)

Search datasets...

133 datasets found

Order by: Relevance

viseem_poster_viseem_poster

This dataset has no description

metalorganic_frameworks_as_drug_delivery_systems_the_case_of_irmof74iii_and_g...

A multiscale computational study is reported that investigates the microscopic behavior of the anti-cancer drug gemcitabine (GEM) stored in metal organic frameworks IRMOF-74-III...

cordex_monthly_data__2008

This dataset has no description

cordex_monthly_data__2007

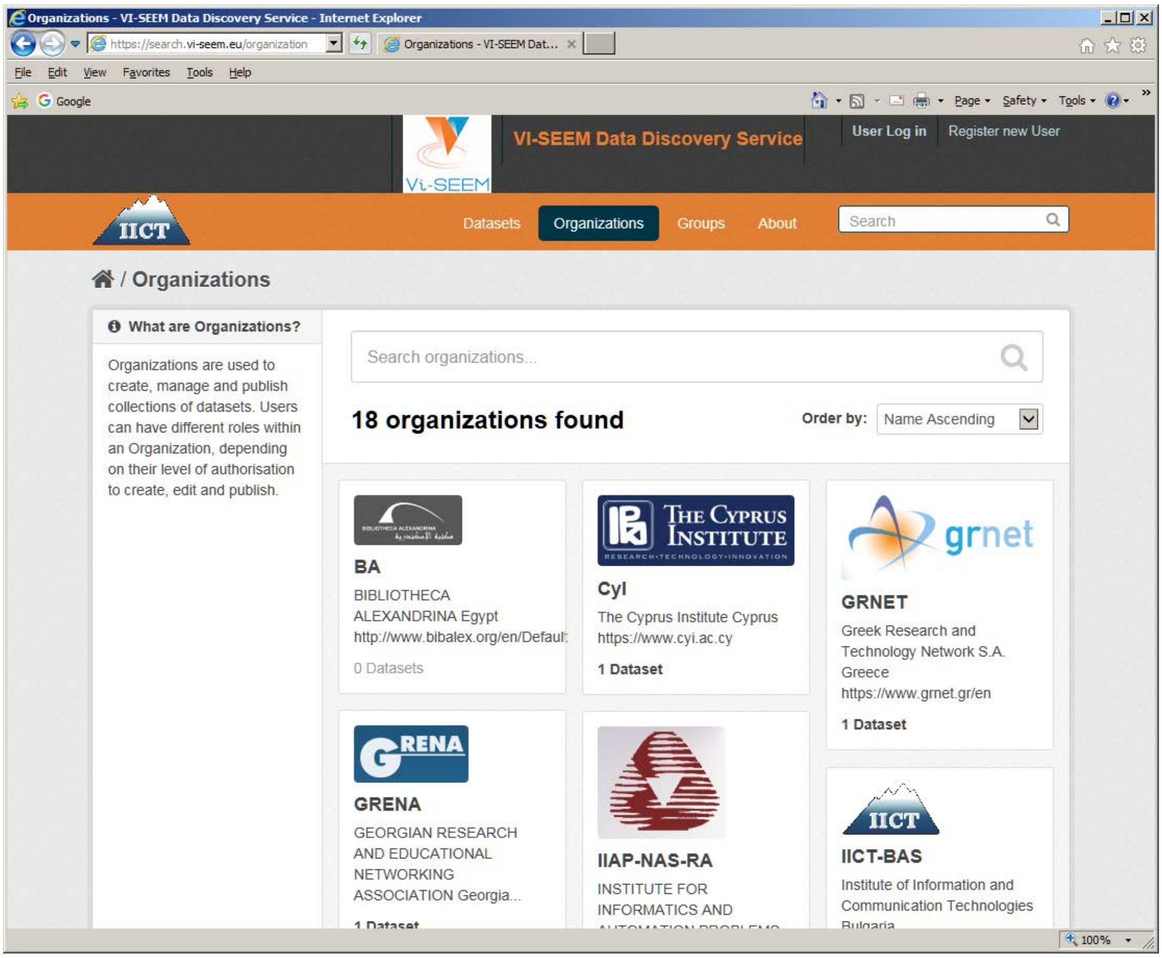
This dataset has no description

Figure 2: VI-SEEM data discovery service home page (Datasets section)

This page is seen from a non-registered user who can access all of the public datasets which are indexed in the site. There are 3 main sections on the orange ribbon: Datasets, Organizations and Groups.

The Datasets section shows a list of all datasets on site which could be sorted using the dropdown list control "Order by:" on the right. Sorting can be by name, relevance and last modified time. On the left there is a panel which provides some statistics: number of datasets per organization, per group, per tags, per file format etc.

The Organization section shows a list for registered organizations (research institutes) which are most of the partners in the VI-SEEM project. Organizations are used to create, manage and publish collections of datasets. Users can have different roles within an Organization, depending on their level of authorization to create, edit and publish. An example screenshot is presented below.


Organizations - VI-SEEM Data Discovery Service - Internet Explorer
https://search.vi-seem.eu/organization
Organizations - VI-SEEM Dat...
File Edit View Favorites Tools Help
Google
Page Safety Tools
VI-SEEM Data Discovery Service
Vi-SEEM
User Log in
Register new User
IICT
Datasets
Organizations
Groups
About
Search
/ Organizations
100%

What are Organizations?

Organizations are used to create, manage and publish collections of datasets. Users can have different roles within an Organization, depending on their level of authorisation to create, edit and publish.

Search organizations...

18 organizations found

Order by: Name Ascending

BA
BIBLIOTHECA ALEXANDRINA Egypt
http://www.bibalex.org/en/Default
0 Datasets

CyI
The Cyprus Institute Cyprus
https://www.cyi.ac.cy
1 Dataset

GRNET
Greek Research and Technology Network S.A. Greece
https://www.grnet.gr/en
1 Dataset

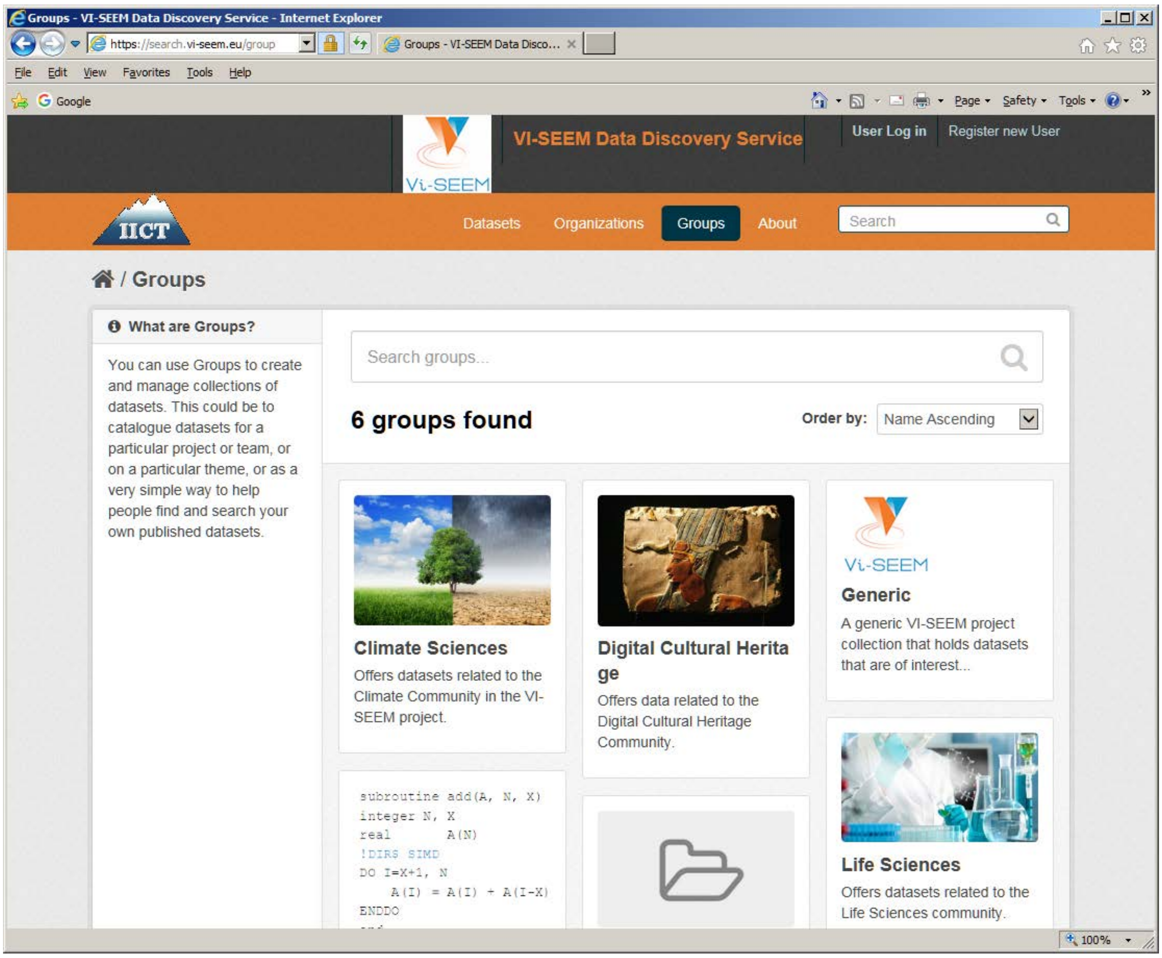
GRENA
GEORGIAN RESEARCH AND EDUCATIONAL NETWORKING ASSOCIATION Georgia...
1 Dataset

IIAP-NAS-RA
INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS

IICT-BAS
Institute of Information and Communication Technologies Bulgaria

Figure 3: VI-SEEM Organizations section

The Groups section contains groups which correspond to the main research directions as it is defined in the VI-SEEM project description of work. An example screenshot of the Groups section is presented below. There is also a group called Software projects which will contain information for software applications and tools developed during VI-SEEM project and a group for testing purposes only.


Groups - VI-SEEM Data Discovery Service - Internet Explorer
https://search.vi-seem.eu/group
Groups - VI-SEEM Data Disco...
File Edit View Favorites Tools Help
Google
Page Safety Tools
VI-SEEM Data Discovery Service
Vi-SEEM
User Log in
Register new User
IICT
Datasets
Organizations
Groups
About
Search
/ Groups
What are Groups?
Search groups...
Order by:
Name Ascending
100%

You can use Groups to create and manage collections of datasets. This could be to catalogue datasets for a particular project or team, or on a particular theme, or as a very simple way to help people find and search your own published datasets.

6 groups found

Climate Sciences
Offers datasets related to the Climate Community in the VI-SEEM project.

Digital Cultural Heritage
Offers data related to the Digital Cultural Heritage Community.

Generic
A generic VI-SEEM project collection that holds datasets that are of interest...

Life Sciences
Offers datasets related to the Life Sciences community.

```
subroutine add(A, N, X)
integer N, X
real        A(N)
!DIRS SIMD
DO I=X+1, N
    A(I) = A(I) + A(I-X)
ENDDO
```

Figure 4: VI-SEEM Groups section

The main purpose of Groups section is to catalogue datasets for a particular research direction or team, or on a particular theme, or as a very simple way to help people find and search the published datasets.

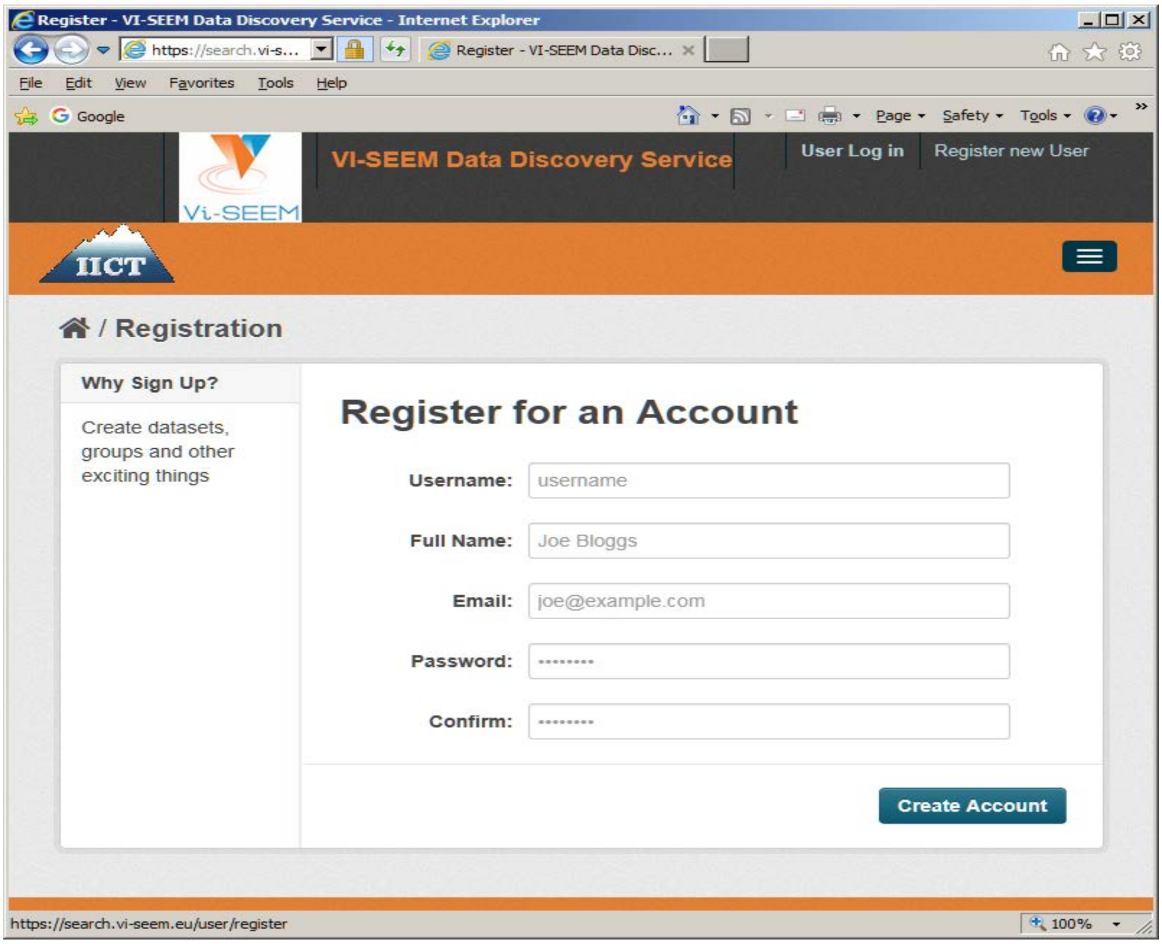
3.5.3. Searching for a dataset

The searching for a particular dataset can be performed in section Dataset with selecting the field "Search dataset". It is possible to search for keywords, tags, partial phrases, organization, creator, publisher, timestamps, versions, tags, research areas and communities etc. Similar searches can be executed in Oragnization and Group sections.

Another way to perform automated or semi-automated queries is to execute programming functional calls through the well documented CKAN API. However it is outside of the scope of this document beacause most of the end users are not professional software developers.

3.5.4. User registration and authorization

Unregistred user can not add or edit datasets. The registration process will be invoked when selecting "Register new User" which is located upper right and a registration form is opened.


Register - VI-SEEM Data Discovery Service - Internet Explorer
https://search.vi-s...
Register - VI-SEEM Data Disc...
File
Edit
View
Favorites
Tools
Help
Google
Page
Safety
Tools
Vi-SEEM
VI-SEEM Data Discovery Service
User Log in
Register new User
IICT
https://search.vi-seem.eu/user/register
100%

/ Registration

Why Sign Up?

Create datasets, groups and other exciting things

Register for an Account

Username:	username
Full Name:	Joe Bloggs
Email:	joe@example.com
Password:	·········
Confirm:	·········

Figure 5: New user registration form

The user must provide username which must cover the requirements from Unix, full name, Email and a password and press the button "Create account". When this process completes successfully, the user will be able to add their organization and new datasets to it. If the user needs to add or edit a dataset which belongs to an existing organization, he or she must contact the system administrator of VDDS for the right to do so.

3.6. VI-SEEM data analysis service (VDAS)

In the analysis of very large datasets, the movement of data can present a far more severe bottleneck than the actual computation. Therefore, one of the design goals of the VI-SEEM data analysis service is to overlap computation and data storage operations, i.e., to enable performing of computation on the same machine(s) that store the corresponding data.

Apache Hadoop was chosen as a system that fits those requirements nicely. It is based on Google’s MapReduce computation paradigm in which the data is broken into chunks and stored across multiple machines, and the analysis is broken into two parts: a mapper and a reducer. The splitting and distribution of data is handled by the Hadoop Distributed File System (HDFS), which also handles data replication for fault tolerance and faster access speed.

The mapper is a function executed over every data point in parallel, creating an intermediate key-value pairs. Those pairs are then sorted by keys and given to the reducer to calculate the final result on them. This kind of computation minimizes the need for data to be accessed over network from different machines, as much of the computation is done locally.

While this model of computation could seem overly simple, it supports a reasonably large number of use cases in the analysis of large data sets. It also has no special hardware requirements, and can make use of heterogeneous hardware, so scaling the cluster up can be done with commodity machines.

Besides the basic MapReduce system, Vi-SEEM data analysis service supports entire ecosystem of different data analysis tools and approaches inherited from the Hadoop technology. These tools support higher level abstractions, such as Query Languages, which automatically create low-level map-reduce analysis steps in a more concise manner.

Some of the more popular tools in this ecosystem, listed together with the service they provide, are:

- Pig, a scripting engine that supports execution of parallel data flows,
- Hive, which provides an SQL-like database,
- Hbase, a non-relational distributed database inspired by Google’s BigTable,
- Mahout, a scalable machine learning and data mining algorithms library,
- Spark, a general computation engine for in-memory analysis of data,
- Solr, a search engine with indexing and search.

VI-SEEM data analysis service consists of a “name node” and a number of “data nodes”. The name node hosts a resource manager (Figure) that schedules the analysis jobs, and handles the allocation of data nodes to a particular task. It also holds the structure of the distributed file system, it’s filenames, directories and their metadata. The data nodes hold the actual chunks of data stored in the HDFS and execute computation steps on them.

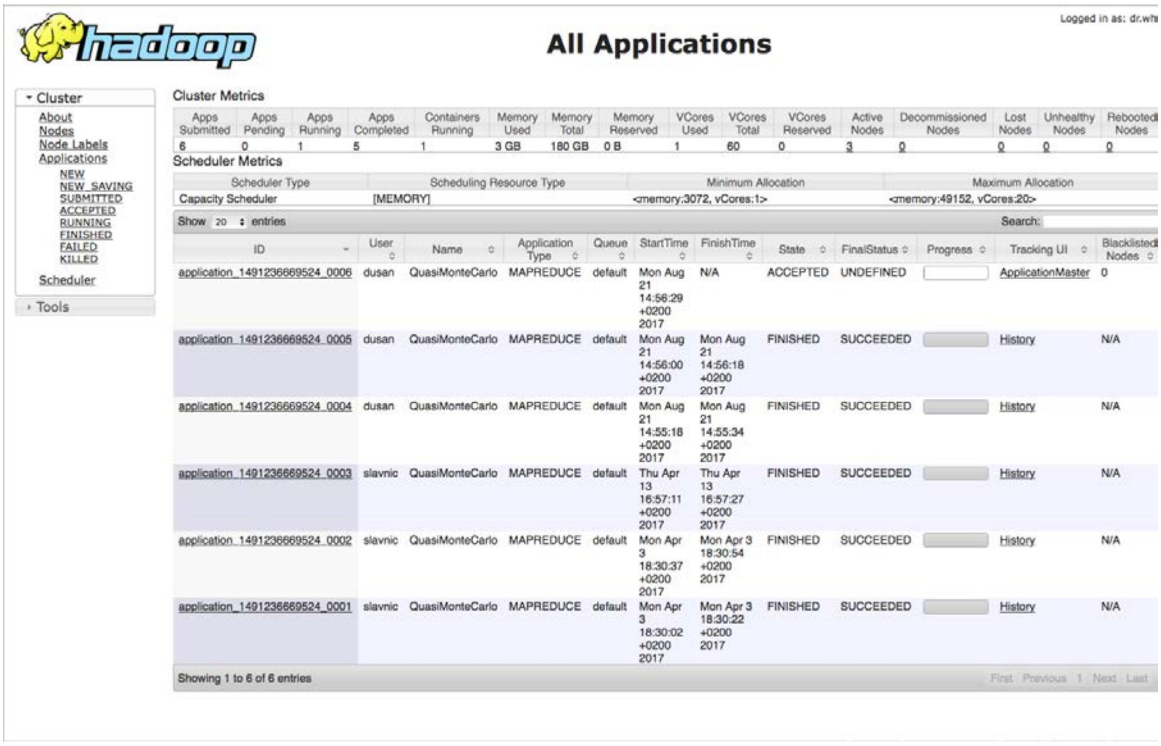
VI-SEEM data analysis service at the Institute of Physics Belgrade consists of a single name node which runs YARN resource manager, and three additional data nodes. The name node is hosted on a machine with 4-core Intel Xeon E3-1220v3 CPU running at 3.1 GHz, with 4 GB of RAM, and 500 GB of local hard disk storage. Each of the data nodes, which perform the computation and storage, are on machines with 24-core Intel Xeon E5-2620 CPUs at 2.4 GHz, with 64 GB of RAM and 2 TB of storage.

In total, the Vi-SEEM data analysis service provides access to 60 CPU cores, 180 GB of RAM and 5.3 TB of storage in HDFS.

A training webinar, which covered introduction to Hadoop and few hands-on examples of data analysis with MapReduce, has been organized by IPB on 24th June 2016. The corresponding training material is available online at https://events.hpc.grnet.gr/event/19/.

Code for the examples used in the training is also available at https://code.vi-seem.eu/petarj/hadoop-training.git.

Additionally, there is a Maven archetype for creation of Java projects that implement data analysis on Hadoop, available at the VI-SEEM Code repository: https://code.vi-seem.eu/petarj/hadoop-archetype.



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
6	0	1	5	1	3 GB	180 GB	0 B	1	60	0	3	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:3072, vCores:1>	<memory:49152, vCores:20>

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1491236669524_0006	dusan	QuasiMonteCarlo	MAPREDUCE	default	Mon Aug 21 14:56:29 +0200 2017	N/A	ACCEPTED	UNDEFINED		ApplicationMaster	0
application_1491236669524_0005	dusan	QuasiMonteCarlo	MAPREDUCE	default	Mon Aug 21 14:56:00 +0200 2017	Mon Aug 21 14:56:18 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1491236669524_0004	dusan	QuasiMonteCarlo	MAPREDUCE	default	Mon Aug 21 14:55:18 +0200 2017	Mon Aug 21 14:55:34 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1491236669524_0003	slavnic	QuasiMonteCarlo	MAPREDUCE	default	Thu Apr 13 16:57:11 +0200 2017	Thu Apr 13 16:57:27 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1491236669524_0002	slavnic	QuasiMonteCarlo	MAPREDUCE	default	Mon Apr 3 18:30:37 +0200 2017	Mon Apr 3 18:30:54 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1491236669524_0001	slavnic	QuasiMonteCarlo	MAPREDUCE	default	Mon Apr 3 18:30:02 +0200 2017	Mon Apr 3 18:30:22 +0200 2017	FINISHED	SUCCEEDED		History	N/A

Showing 1 to 6 of 6 entries

Figure 6: Resource manager of the VI-SEEM data analysis service

3.7. *VI-SEEM persistent identifier service (VPID)*

GRNET’s Handle Service is provided to VI-SEEM for the purposes of persistently identifying digital objects across their lifecycle. Such digital objects are maintained in the VI-SEEM repository as well as other VI-SEEM generic or application specific data services. The handle service supports the management and resolution of:

- Resources: Digital objects and other internet resources.
- Part Identifiers: computes an unlimited number of handles on the fly.
- Multiple locations in a single handle: structured alternatives, e.g. multiple locations, in a single handle value.

A Persistent Identifier, also known as PID, is an identifier that is effectively permanently assigned to the object. It provides a means of connecting and distinguishing between an identifier for an object (a permanent identity) and the object's location (although it may change over time). PIDs introduce a level of indirection and complexity, since apart from managing PID a separate service needs to be used so as to resolve it.

The European Persistent Identifier Consortium (EPIC) provides persistent identifier (PID) services for European scientific and cultural heritage communities, using the

Handle System (http://www.handle.net). The Handle System consists of a Global Handle Service (GHS) and Local Handle Services (LHS). It provides a resolution system consisting of a distributed infrastructure of global, local, and caching servers. GRNET provides a high availability of the PID service as a LHS. GRNET service supports the EPIC REST web service for issuing and managing PIDs.

VI-SEEM, has requested access accounts in order to be able to use the GRNET PID Service. VI-SEEM is responsible to create, maintain and update its PID collection by using the REST web service when it is necessary.

Handles are persistent identifiers for Internet resources. In the handle system the syntax of a PID handle consists of a Prefix and a Suffix.

Prefix: is used to access the service information that describes the "home" service (each organization may have one or more prefixes under its ownership).

Suffix: is a unique "local name" under the prefix. The uniqueness of a prefix and a local name under that prefix ensures that any identifier is globally unique within the context of the Handle System.

4. Summary of the new or updated functions of the final data services, compared to the initial deployment

4.1. VI-SEEM simple storage service

The VI-SEEM Simple Storage service has been recently fully integrated with the VI-SEEM AAI (Authentication and Authorization Infrastructure). The service is configured to act as a SAML Service Provider (SP) via the Shibboleth software, so that user_shibboleth ownCloud app, which enables Shibboleth authentication for ownCloud users, can be used.

Since the VI-SEEM Simple Storage service is deployed under the Debian OS, Shibboleth Apache module (libapache2-mod-shib2) is installed directly from the repository. Important configuration files are stored in /etc/shibboleth folder, namely /etc/shibboleth/shibboleth2.xml and /etc/shibboleth/attribute-map.xml. First file is the main configuration file that contains data about the SP and also about Shibboleth Identity Provider (IdP, VI-SEEM Login IdP Proxy in VI-SEEM environment), while second file tells the SP how to map SAML attributes received from IdP to environment variables that can be then used in web applications.

After the setup, SP metadata is sent to the AAI team so that their SP can connect to the VI-SEEM Login IdP Proxy, as it is stated on the VI-SEEM Wiki page dedicated to integration of Service Providers into VI-SEEM AAI infrastructure

http://wiki.vi-seem.eu/index.php/VI-SEEM_Login_integration_guide_for_Service_Providers.

This metadata includes entityID and Metadata URL, and in the case of the VI-SEEM Simple Storage Service instance these values are

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

(for entityID) and

https://simplestorage.vi-seem.eu/Shibboleth.sso/Metadata

(for Metadata URL).

VI-SEEM Login IdP Proxy metadata should be present in /etc/shibboleth/shibboleth2.xml file.

The OwnCloud application user_shibboleth (https://github.com/EUDAT-B2DROP/user_shibboleth) was added on top of the service, so that it can use Shibboleth service for user authentication. During the integration, we found several inconsistencies related to the mapping of the attributes coming from the VI-SEEM Login to the Simple Storage Service. In order to fix this, we have created a fork of the project at the VI-SEEM source code repository (https://code.vi-seem.eu/). The VI-SEEM user_shibboleth module is now available at:

https://code.vi-seem.eu/petarj/user_shibboleth.

4.2.VI-SEEM repository service

The VI-SEEM repository service data sets are organized in an hierarchical way focusing on the main user categories of the service. Besides the main categories specified, the repository is configured to require metadata in the Dublin core format for each data set added to it. Documenting and describing research data is time-consuming and hence expensive work. Thus, increasing the complexity of the documentation process would impact the usability of the repository service, and consequently potential users might lose their interest in using it. One of the main objectives was to make the tool as simple as possible for the data depositors as well as data consumers, while at the same time gathering as much meaningful information about the data as possible. In order to keep the metadata simple, we use the Dublin Core whose schema is simply flat and has no complex hierarchical structure. Dublin core elements provide understandable information about complex objects and help data consumers to become acquainted with the research data. In cases where this is applicable we are extending the Dublin Core schema. Extension of Dublin core data has been implemented for the case of Climate Research data as will be described late in this section. To be compatible with the data management plan that is developed in the VI-SEEM project the VI-SEEM repository requires from data contributors the following meta data items to be associated with each data sets that is being deposited in the service: The contributor (author), the accessioned date (that the data sets is being added to the repository), the date of availability(the date the data is available to be accessed), the issued date (the date fo publication), the identifier (an automatically generated PID based on GRNET PID service), a description of the data set, the format of the dataset, the publisher (organization) of the dataset, the subject of the dataset, the title, the type and the accompanying licence of the dataset. Several other Dublin Core metadata fields can be used on a voluntarily basis.

As mentioned before specifically for the Climate Research community the metadata set has been extended to better support the NetCDF metadata format. Some of the required NetCDF metadata has been mapped to Dublin Core metadata while some others have been defined in an additional metadata schema as follows:

- Project (could be CORDEX or any other project): We use the DC property “mediator": An entity that mediates access to the resource and for whom the resource is intended or useful.
- Domain (that could be EUR, MED etc). We use the DC property “coverage" (spatial) Spatial characteristics of content.
- Institute (Institute that contributes the simulation eg AUTH-MC). We use the SC property “publisher”: An entity responsible for making the resource available
- Experiment (could be evaluation - what we all uploaded so far, historical or projection). We use a new attribute in the new metadata schema.
- Driving model (e.g. WRF381). We use a new attribute in the new metadata schema.
- Time frequency (e.g. 3h, mon, seas etc). We use a new attribute in the new metadata schema.
- Temporal coverage. i.e. period of data applicability. We use the DC attribute coverage (temporal): temporal characteristics of content.

- Variable long name (e.g. Temperature at 2 m, precipitation etc). We use a new attribute in the new metadata schema.

In general the data follow the ESGF specifications used in all climate modeling studies.

https://www.esrl.noaa.gov/psd/repository/entry/show?entryid=0621e6b0-de3a4ecc-9e9a-42c3305c671b

All the metadata are searchable and therefore the dataset consumers can generate queries using such meta data in order to retrieve the data sets that are interesting to them.

Further to that the VI-SEEM repository has been fully integrated with GRNETs PID service therefore each dataset deposited in the VI-SEEM repository is automatically being assigned a persistent identifier. The integration involved the development from GRNET, of a special DSPACE plugin that uses the PID API. The prefix that has been associated with the VI-SEEM repository service is “21.15102/VISEEM-“.

Since its production operation in Q1 2016, the VI-SEEM repository service hosts around 16 TB of data that represent 45 collection of 205 different items. The datasets have 17,930 views while users have performed 177,011 searches.

4.3. VI-SEEM archival service

While the initial deployment had four zones (GRNET, IICT-BAS, IPB and NIIF), the final data platform consists of six zones (those mentioned before plus BA and IUCC). For now, federation is configured in a star topology between NIIF and all other partners. It is expected though that other connections will be established as needed by the VI-SEEM community.

Further tasks will be handled locally (local policies) or in co-operation (in case of geographically redundant replication scenarios). There are local initiatives as well: a standalone iRODS server has been deployed in Armenia (irods.asnet.am) to provide service for the life sciences community. They will use it to store results and associated metadata of their molecular dynamics simulations. It is also planned to integrate their local iRODS with the VI-SEEM VAS in some way.

4.4. VI-SEEM work storage space / local storage and data staging

The VI-SEEM work storage space / local storage and data staging service in the final data platform has iRODS DSI enabled gridFTP server instances at all sites where the VI-SEEM Archival Service is deployed.

4.5. VI-SEEM data discovery service

A set of Python scripts which collect metadata from defined sources, e.g. harvesting and register it in VDDS is developed at a testing phase. Some of the datasets which are visible on the site are collected in that way.

Many cosmetic enhancements have been made in this CKAN based complex site to make it more user-friendly.

The VRE AAI service is not yet fully integrated in VDDS, it is work in progress.

4.6. VI-SEEM data analysis service

Initially, the VI-SEEM data analysis service was provided as a standalone Hadoop cluster, consisting of four 8-core machines (32 cores in total). Recently, service’s name node was separated from the cluster, and deployed separately on a machine with 4-core Intel Xeon E3-1220v3 CPU at 3.1 GHz, with 4 GB of RAM, and 500 GB of local hard disk storage. Also, each of the data nodes (which perform computation and storage) are upgraded to machines with 24-core Intel Xeon E5-2620 CPUs at 2.4 GHz, with 64 GB of RAM and 2 TB of storage. Therefore, starting from January 2017, the VI-SEEM data analysis service provides access to 60 CPU cores, 180 GB of RAM and 5.3 TB of storage in HDFS.

4.7. VI-SEEM persistent identifier service

At GRNET the service provided to VI-SEEM among others, is hosted in virtual machines hosted in two different IaaS infrastructures offered by GRNET. The VMs are running version 8.1.0 of the handle.net software [ref: http://www.hdl.net/download_hnr.html] and they run several handle service instances at a primary / mirror handle service configuration. At the same VMs the relevant instances of the epic api are running i.e.

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

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

Replication between primary and mirror handle server is being implemented in an automatic way by the handle service. During the initial installation of a mirror server the administrator needs to manually bootstrap the mirror by using the hdl-dumpfromprimary tool. This tool downloads all the handles from the primary server. Then the mirror handle server automatically pulls the changes from the primary server at pre-defined, frequent intervals.

5. Conclusions

The VI-SEEM data platform consists of 7 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 have been developed and deployed at different partners of the VI-SEEM consortium, resulting a geographically distributed – thus geo-redundant – data infrastructure that is available for all the VI-SEEM user communities and able to satisfy the different needs.

Based upon the experiences and the lessons learnt from the initial deployment, the services have been further improved and tailored in order to reach a mature, final stage.

Some of the VI-SEEM data services are already used for a long time by the scientific projects, while some others – especially those that are designed to store and manage the results and the scientific products of the user projects (eg. VAS, VDAS) – will be intensively used in the future. For this reason it is expected that further improvements and/or minor changes will become necessary in the final period of the VI-SEEM project.