H2020-EINFRA-2015-1

VI-SEEM

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

Deliverable D5.2

Data management plans

Author(s): Andreas Athenodorou (editor), VI-SEEM consortium

Status –Version: Final – j

Date: October 30, 2017

Distribution - Type: Public

Abstract: Deliverable D5.2 – "Data Management Plans" presents the data management plan for the VI-SEEM project. The data to be collected is described in detail, and policies regarding the collection, quality control and assurance, access and sharing, storage and preservation, privacy and security and ethics and legal compliance of the data are outlined. This is an updated version of the deliverable, following the first project review.

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.

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
19/02/2016	a	Constantinos Lazarou	First draft ToC
26/02/2016	b	Constantinos Melachrinos	First version of the Data Management Plans
13/03/2016	c	Constantinos Melachrinos, Ioannis Liabotis, Constantinos Lazarou	Second version of the Data Management Plans
23/03/2016	d	Constantinos Melachrinos, Ioannis Liabotis, Zoe Cournia, Theodoros Christoudias, George Artopoulos, Constantinos Lazarou	Third version of the Data Management Plans
03/08/2017	e	Andreas Athenodorou	Create the first draft, TOC
08/08/2017	f	Andreas Athenodorou, Tamás Kazinczy, Ioannis Liabotis	Updated TOC
10/09/2017	g	Andreas Athenodorou	Added Parts
26/09/2017	h	Andreas Athenodorou, Theodoros Christoudias	contributions
17/10/2017	i	Andreas Athenodorou, Theodoros Christoudias, George Artopoulos, Tamás Kazinczy.	contributions
30/10/2017	j	Andreas Athenodorou, George Artopoulos, Alexandr Golubev, Zoe Cournia, Valentina Vassallo, Mihajlo Savic, Daniel Pop, Ognjen Prnjat.	comments

Table of contents

References

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

[2] Creative Commons Licenses - https://creativecommons.org/licenses/

[3] Project VI-SEEM Deliverable 3.1 “Infrastructure and services deployment plan”

[4] “Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics”, Association for Clinical Genetic Science (2013) - http://www.acgs.uk.com/

[5] “Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation”, Association for Clinical Genetic Science (2015) - http://www.acgs.uk.com/

[6] “Guidelines for diagnostic next-generation sequencing”, European Journal of Human Genetics (2016) 24, 2–5

[7] “PDBx/mmCIF – Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework - See more at: http://www.dcc.ac.uk/resources/metadata-standards/pdbxmmcif-%E2%80%93-protein-data-bank-exchange-dictionary-and-macromolecular-cr#sthash.xYI0Wbrh.dpuf” or http://www.dcc.ac.uk/resources/metadata-standards/pdbxmmcif-–-protein-data-bank-exchange-dictionary-and-macromolecular-cr

List of Tables

List of Figures

Glossary

AMBER	Assisted Model Building with Energy Refinement molecular simulation programs
ASCII	American Standard Code for Information Interchange
CADD	Computer-aided drug design
CBIR	Content-Based Image Retrieval
COSMO	Consortium for Small-scale Modeling
CPU	Central Processing Unit
CRM	Conceptual Reference Model
CRYSTAL	Computational tool for solid state chemistry and physics
DICOM	Digital Imaging and Communications in Medicine
DREAM	The Dust Regional Atmosphere Model
ECHAM	Global Climate Model developed by the Max Planck Institute for Meteorology
ECMWF	European Centre for Medium-Range Weather Forecasts
EMAC	ECHAM/MESSy Atmospheric Chemistry
ERT	Electrical Resistivity Tomography
FERRET	Interactive computer visualization and analysis environment
FFTW	Fastest Fourier Transform in the West, library for computing the discrete Fourier transform
FIREFLY	Ab initio and density functional theory chemistry program.
GAMESS	General Atomic and Molecular Electronic Structure System is a general ab initio quantum chemistry package
GATK	Genome Analysis Toolkit
GIS	Geographic Information System
GPS	Global Positioning System
GPU	Graphics Processing Unit
GrADS	Grid Analysis and Display System
GROMACS	Molecular Dynamics Software Toolkit
GUI	Graphical User Interface
HPC	High Performance Computing
IDL	Interactive Data Language, a programming language used for data analysis
ISBD	International Standard Bibliographic Description
LAS	Live Access Server
MEDICI	A multimedia content management system
MESSY	Modular Earth Sub-model System
MM5	The PSU/NCAR mesoscale model
NAMD	Scalable Molecular Dynamics Toolkit

NCL	NCAR Command Language
NetCDF	Network Common Data Form
NWCHEM	High Performance Computational Chemistry Software
OCR	Optical character recognition
OpenCV	Open Source Computer Vision library
OPENFOAM	Open source Field Operation And Manipulation toolbox for continuum mechanics
PIDs	Persistent Identifiers
RDF	Resource Description Framework
RegCM	The Regional Climate Model system
RTI	Reflectance Transformation Imaging
SCP	Secure Copy
SEEM	South East Europe and Eastern Mediterranean region
SFTP	Secure File Transfer Protocol
SOL	Soft Ontology Layer
SQL	Structured Query Language
SSH	Secure Shell
UI	User Interface
UNIMARC	Universal MARC format
VI-SEEM	VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean
VRE	Virtual Research Environment
WEST	Wind Energy Simulation Toolkit
WMO	World Meteorological Organization
WRF	Weather Research and Forecasting Model
XML	Extensible Markup Language

Executive summary

What is the focus of this Deliverable?

This document constitutes a high level Data Management Plan for the VI-SEEM project that defines the governance of all data sets to be provided in the scope of the project. The data to be collected is described in detail, and policies regarding the collection, quality control and assurance, access and sharing, storage and preservation, privacy and security and ethics and legal issues of the data are outlined. Furthermore, the question of dealing with cross-disciplinary datasets is also discussed along this manuscript, providing a plan of developing the required ontologies which will provide interoperability to metadata. Furthermore, the question of adopting strategies for dealing with data with restricted access such as clinical phenotypic datasets is also addressed. In addition, the document discusses the overall QA/QC strategy.

This manuscript provides an update of the deliverable D5.2 incorporating the comments given by the reviewers as a result of the 1st periodic review which took place on the 11th of May 2017 at Brussels.

The Data Management Plan is a "live document" meaning that it continuously evolves according to new additions of datasets as well as to new developments in addressing the interoperability of metadata for cross-disciplinary applications. Hence, this document is expected to be occasionally updated according to the progress of the project.

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

This deliverable provides the Data Management Plan for the datasets in the VRE platform. Based on this, WP5 will work in collaboration with WP4 to collect, pre-process, curate and make available these datasets through the VRE platform. The results and conclusions from this deliverable will be used by the following activities:

- WP4.1-Data services design
- WP4.2-Data access, preservation and re-use
- WP4.3-Data collection and provisioning
- WP4.4-Data analysis
- WP5.3-Development of the VRE platform

The complete description of the activities and dependencies can be found in the VI-SEEM DoA[1].

What are the deliverable contents?

The Data Management Plan describes the data to be collected by VI-SEEM in detail. The general categories of datasets are discussed, demonstrating different types of datasets expected to be uploaded in the VI-SEEM VRE. The data is classified in three categories, depending on the Scientific Community (SC) it serves; namely Climate, Digital Cultural Heritage and Life Science data sets. Upon data collection, the contributors will fill out a form with information about the data, such as file format and metadata. The policies for data access and sharing are outlined, with each specific dataset designated as Open, Restricted, or Closed, depending on the conditions of access to the data. VI-SEEM has an interdisciplinary character and, thus, cross-disciplinary data is discussed. In addition a plan for evolving a metadata structure for cross-disciplinary data is presented. Moreover, the storage and preservation of the data is described, dividing the datasets in Short term, Medium term, or Long term, depending on the length of the expected data preservation. Furthermore, the security and privacy of datasets, the use of Persistent Identifiers, as well as legal and ethics issues are discussed. Finally, the responsibilities for the implementation and following the policies outlined in the Data Management Plan are clearly defined.

Conclusions and recommendations

This Data Management Plan serves to guide the use of data in VI-SEEM, an imperative role considering the amount of data planned to be collected through the project. Following the Data Management Plan guidelines, the VI-SEEM community benefits by taking advantage of quality assurance and control, searchability, and increased usability of the data. In addition, the Intellectual Property Rights and conformance to legal and ethics issues are ensured. The VI-SEEM technical coordinator bears the overall responsibility for providing support and guidance for the implementation of the policies outlined in the Data Management Plan. The WP5 leader will monitor and update the Data Management Plan if needed, while the QA/QC officers will be responsible for quality control and assurance, in collaboration with the WP5 leader and the SC leaders. The individual publishers of the data in the VI-SEEM VRE have the responsibility of the data they are publishing.

1. Introduction

VI-SEEM brings together three scientific communities by providing the infrastructure to create Virtual Research Environments (VREs). These VREs provide the framework where common data and analysis tools can be shared and accessed by the members of the collaboration, but also serve educational and outreach activities targeting the public at large. The data span the three scientific communities: climate, cultural heritage and life sciences and consist of different formats, while initially being located at different sites in the Southeast Europe and the Eastern Mediterranean (SEEM) region.

This document describes the Data Management Plans for VI-SEEM, specifically the description of the data, the policies for collection, access and sharing of the data, the infrastructure for storage and plans for preservation, the processes for quality control and assurance, and the responsibilities regarding the governance of the data. Tables showing the specific datasets to be collected and their level of access and preservation are included in the document.

The aim of this Data Management Plan is to provide a thorough description of the policies for data management during the VI-SEEM project. Since research data forms the basis of the VI-SEEM project, effective data management policies are needed to ensure the verification and reuse of research results, the interoperability of heterogeneous datasets, the correct treatment of privacy issues, as well as the sustainable storage of the datasets. The intermediate objectives are: a) to set up the procedures for contributors of data to be able to share their data, while taking into account any restrictions on sharing that may be required, b) to set up the procedures for quality control and assurance of the data, in order for the data to be useful to the users, c) to enable registered users of VI-SEEM to search for and find relevant data to support their scientific or educational activities, d) to set up the procedures for developing ontologies suitable to address the issue of interoperability of datasets from cross-disciplines, e) to establish the policies for security and preservation of the data. The Data Management Plan provides to data providers a set of policies that have to be applied for each of the data sets to be published and managed in the context of the VI-SEEM project. Data set providers will apply the relevant policies based on the nature and characteristics of their specific data sets. Possible additional policies might need to be introduced in cases where the newly produced and collected data sets, products of the open calls, will have different requirements than the existing ones.

2. Data Management Plan

2.1. Data description

The data collected and processed within the VI-SEEM project can be broadly grouped in three categories, each corresponding to the three scientific communities (SC):

- Climate,
- Digital Cultural Heritage and
- Life Sciences.

Depending on the expected usage of each category of data, the data can be divided in three types:

a) *Scientific data* - includes experimental, observational, computational and/or curated data, including the methodology to acquire it, i.e. the software, analysis workflows and documentation to generate, collect, analyze and access the data and related published results.

b) *Publications* - including peer-reviewed journal and conference proceedings, as well as any additional documentation that is necessary to understand the data related procedures and that puts the results in context.

c) *Simplified data formats* - for immediate re-use, such as visualization for education and outreach purposes, as well as the theory interpretations and educational analyses.

Accepted standard encoding for data is UTF-8.

Users of the categories of data described above are classified as follows, according to their intended use of the data:

1. VI-SEEM collaborators, members of the relevant SCs that require access to the data for analysis, research and production of new scientific results. Such users would require access and would benefit from all three Levels described above (Levels a/b/c)
2. Worldwide scientific communities that require data for further analysis including qualitative and quantitative, comparison with other data sets and in general any activities related to their research and the production of new scientific results. Worldwide scientific community is not restricted to the academic world but also involves SMEs, spinoffs as well as NGOs carrying out research related to the three communities. Such users would require access and would benefit from all three Levels described above (Levels a/b/c).
3. The public-at-large, including outreach coordinators, educators and students. Such users would benefit from publications, simplified data formats as well as visualizations for education and outreach purposes (Levels b/c).

According to their access rights, users are classified in the following categories, with increasing access rights:

1. Non-registered users: Would be able to access publications and educational and outreach material, when these are available.

2. Registered users: Would be able to browse collections of scientific data, search for studies, and access and analyze data in the relevant scientific community.
3. VI-SEEM Contributors: Would be able to upload datasets in the relevant SC, as well as access data shared only within VI-SEEM.
4. VI-SEEM Administrators: Would be able to perform user and access management.

The process of authentication and authorization of users, is controlled by the WP3 leader in accordance to D3.1 [3]. Registered VI-SEEM Contributors will be all scientists who are granted resources through the Integration Phases as well as open calls for access. In the following subsections, the data to be collected by the VI-SEEM Scientific Communities is described, followed by useful definitions.

2.1.1. Structure of the general metadata

Although different communities have different ways of describing their research data, there is a common set of metadata shared by all and this set represents the core of our General Metadata. The issue of deriving a common “language” in order to make data sets from different communities interoperable is discussed in detail further on in subsection 2.1.5.2

The structure of this core is intentionally flat as it is essential to avoid increasing complexity of handling metadata (to reduce the burden of both data set producers and consumers).

Core metadata consists of the following elements ('*****' means mandatory):

- Title*****
- Description
- Identifier***** (source URL/URN or PID or DOI; VI-SEEM PID preferred)
- Creator
- Community (climate, digital cultural heritage or life sciences)
- Discipline
- Publisher (for citation purposes)
- Public Since (year and month from which publicly available)
- Language
- Format
- Keywords (for content description)
- Temporal Coverage (period of time the research data resource is related to)
- Spatial Coverage (geolocation the research data resource is related to)
- Primary Contact***** (primary responsible for the data set)
- Technical Contact
- Date
- Type
- Rights

This set of general metadata may be revised over time.

Additional, critical pieces of information, as these derive from the characteristics of the heterogeneous data sets about the metadata will also be included in subsection 2.1.5.2.

2.1.2. Climate data

2.1.2.1 Description of climate data

The climate modelling and weather forecasting communities use mainly two types of data: 1) simulation and 2) observational data. Both kinds cover a wide range in terms of volumes and licensing. In the following two subsections a description of these two types of data is provided.

2.1.2.1.1 Model simulation data

In HPC applications, the climate community users store the full output only temporarily and post process the data remotely with only the final results being transferred and stored by the user. While not all model output data can be kept permanently, it is very important to keep all metadata of the simulation to enable reproduction of the simulation if needed. For this, all input parameters, and the source code need to be stored and made accessible under case-specific licenses. The user communities require a version control system for source code and storage so that model input and output can be used by multiple users. This prevents resource-expensive re-runs, allows analysis of simulation ensembles, and promotes data re-use. For instance, the output of global weather forecasts is used to set the boundary conditions of high-resolution regional models for downscaling. Final simulation model output is uploaded to central, internationally managed data repositories such as the CORDEX initiative.

2.1.2.1.2 Observational data

Observational data is widely used in the community. Observations include weather station measurements and remote sensing data such as satellite images. The sources and locations of this data are very diverse. Some international initiatives exist to collect observation data centrally, but significant amounts of data are scattered and self-hosted by the data sources. Most data sources are freely accessible for non-commercial research and education. The data integration in VI-SEEM will only focus on the data freely available for general research organizations.

2.1.2.2 Climate metadata structure

In principle most VISEEM climate data will adopt, comply with, and conform to the Climate and Forecast (CF) Metadata Conventions (http://cfconventions.org). The CF conventions are increasingly gaining acceptance and have been adopted by a number of projects and groups as a primary standard. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.

Principles of CF include self-describing data (no external tables needed for understanding); metadata equally readable by humans and software; minimum redundancy and maximum simplicity; and a development process focusing on existing needs. The primary focus of the conventions is to enable applications to locate, both spatially and temporally, data contained in conforming files. The second main purpose is to provide a definitive description of the data in each variable to help users of data from different sources decide which quantities are comparable.

The currently-released CF 1.7 specification is to be adopted, as this may provide better explanations or context, or more advanced capabilities. But generally these specifications do not conflict with CF 1.6, a previously released version, so version 1.6 may also be followed.

The CF conventions generalize and extend the COARDS conventions (named for their sponsor, the Cooperative Ocean/Atmosphere Research Data Service). The extensions include metadata that provides a precise definition of each variable via specification of a standard name, describes the vertical locations corresponding to dimensionless vertical coordinate values, and provides the spatial coordinates of non-rectilinear gridded data. Since climate and forecast data are often not simply representative of points in space/time, other extensions provide for the description of coordinate intervals or multidimensional cells and indicate how a data value is representative of an interval or cell. This standard also relaxes the COARDS constraints on dimension.

The conventions for CF metadata are designed to promote the processing and sharing of files created with the NetCDF API. Therefore we will be referring to the CF convention as NetCDF CF. In addition to NetCDF CF metadata schema, other schemas are also used, although they are not so common. These schemas include the Dublin Core, WRF model schema, Proprietary OpenFOAM schema, GeoTIFF and GRIB. All the different formats used in Climate applications are provided in Table 1.

2.1.3. Digital cultural heritage data

2.1.3.1 Description of digital cultural heritage data

In the field of Digital Cultural Heritage the data produced and used include the following types of data: 1) Modules and Components data, 2) Documentation and Analysis, 3) Formal Knowledge Representation data, 4) 3D Visualization and Analysis Data. In the following four subsections a description of these four categories is provided.

2.1.3.1.1 Modules and components data

One of the aims is to produce modularized components that enable simple use of VI-SEEM scientific application environments for individual operations, like a geolocation service for mapping semantically organised data, such as ARCHES, docker containers for every module that can be shared and/or reused through the VI-SEEM Clowder, for example - 3D model cleaning, generating and streaming - as well as learned image representations, image search engines and trained classifiers. These include, among others, various components that can be used in OCR, Web UI and databases of complex data sets - like handwritten Arabic, Hebrew and Karamanlidika texts - for searching and identifying the phonetic varieties of the indexed lexemes, as well as the finding of grammatical and formational suffixes.

2.1.3.1.2 Documentation and analysis data

This category covers documentation and analysis of material properties of structures, works of art and artifacts as well as investigating the cultural heritage objects production, use and history of development through quantitative - chemical and physical - analysis and digital methods including remote sensing and geo-surveying. Examples of this include web UI for new tools used for modelling of geoelectrical tomographic data, subsurface reconstruction and imaging. It also includes workflows and pipelines related to artifact/object digitization, interactive 3D museum tours, databases of surveyed materials, objects and buildings, technical datasets of analyzed materials, consultation services of technical details as well as data analysis methodologies that involve regional datasets like MEGA Jordan GPS geo-referenced data. This category includes online visualization tools for x3D, 3D PDF, RTI ptm and Giga-pan image files, as well as Microsoft Word and PDF files.

2.1.3.1.3 Formal knowledge representation data

Formal knowledge representation relates to how we build explanations in humanities research, how we formulate declarative statements, which are their components and how such process can be formally expressed within the digital domain, in particular in datasets of thousands of digitized books from the Banatica database and the Museum of the Republic of Srpska collections, Ptolemaic inscriptions and The “Aharoni” Online Digitized Collection, the National Natural History Collections at the Hebrew University of Jerusalem. The main components of this process are: ontology engineering (CRM), including development of knowledge repositories, causality and formal reasoning, expressing uncertainty and data transparency.

2.1.3.1.4 3D Visualization and analysis data

This research focuses on developing innovative 3D-based methods and techniques for data acquisition, processing, simulation and analysis. Research focuses on scientific visualization of built heritage and geometric analysis of data, simulation of crowds and space occupation and development of virtual museums. Main components are: simulation algorithms, 3D documentation, shape analysis, modeling and real-time rendering, interactive big data and large scale 3D reconstruction models and point-clouds (from Structure-from-Motion photogrammetric techniques), as well as Unity3D mobile apps for geolocation of digital assets, convolutional neural networks (HDF5 format usable for deep learning libraries, codes and workflows),

workflows and training material for machine learning applications in remote sensing (HRRS Image Classification of SAT-4 and SAT-6 datasets) and geophysics, like Evaluation of Convnets for Large-Scale Scene Classification From High-Resolution Remote Sensing Images, and for large image-set to be automatically matched with a given ortho-photo and geo-referenced in a completely automated way. This category involves online visualization viewers for x3D, 3D PDF, simulation optimization codes, 3D model files, Unity3D plugins and scenes, raster images and text files.

2.1.3.2 Digital cultural heritage metadata structure

The VI-SEEM Cultural Heritage metadata mostly follow the CIDOC-CRM RDF (STARC Dioptra), ARC2 triple store, ISBD–M, and UNIMARC (BVL) standards (https://www.loc.gov/marc/bibliographic/), respectively. Metadata will mostly be generated from the operation of Digital Libraries, the application of semantic referencing and annotation, users' annotations of digitized artifacts and reconstructed historical objects, as well as from the publication of databases and use of OCR tools. The Dioptra metadata schema is based on previous research, taking into consideration LIDO (http://www.lido-schema.org/schema/v1.0/lido-v1.0-schema- listing.html) and CARARE metadata schemas (http://www.carare.eu/swe/Resources/CARARE-Documentation/CARAREmetadata-Schema) and having at its base CIDOC-CRM as a reference model (Le Boeuf et al. 2012). The sister schema, CARARE metadata schema, was proposed and it is used within digital libraries communities. -It is used in digital libraries initiatives aimed at the aggregation and description of cultural heritage items (archaeology, cultural heritage and their digital surrogates). The Dioptra metadata schema guarantees the sustainability of the content, an open access to the digital resources archived in a repository and the management of various datasets, allowing also data interoperability. Its structure allows us to archive and to retrieve three-dimensional models, activities and all the decisions taken during the creation process of the digital data. In fact, the novelty of this metadata schema respect to other standards and metadata schemas is the possibility to register and trace the information regarding the digital provenance of 2D and 3D objects. Metadata standardization e.g., Dublin Core and derived/related standards, XML, is important as it will allow for their mapping, e.g., Open CV and MINT, and interoperability across platforms, e.g., Spark SQL, ASCII.

Metadata of the Digital Cultural Heritage community will aim to be freely exchangeable and open, except for cases of copyrighted material - in particular rare books and unpublished Ptolemaic inscriptions.

2.1.4. Life sciences data

2.1.4.1 Description of life sciences data

The Life Sciences Scientific Community focuses on five scientific areas: 1) Modeling and Molecular Dynamics (MD) study of important drug targets, 2) Computer-aided drug design, 3) Analysis of Next Generation DNA sequencing data, 4) Synchrotron data analysis, and 5) Image processing for biological applications. In the following five subsections we provide a short description for each scientific area:

2.1.4.1.1 Modeling and molecular dynamics (MD) study of important drug targets data

Molecular dynamics (MD) is a computer simulation method for studying the physical movements of protein in real time and space. The process of simulation of MD is the process of integration of Newton's second law of motion equations. Coordinates of atoms and velocities are estimated through a fixed time step, an integration step. The output of MD simulations is trajectories representing snapshots of evolution of the protein system and appropriate values of time, energy (for example van der Waals), applied force, temperature of the system *etc*. Trajectories are sequential snapshots of simulated molecular system, which represents atomic coordinates at specific time periods, stored in a binary format. To obtain information from such files, special programs and information processing techniques are applied. Several programs are available for numerical calculations with the MD method, including the most popular GROMACS, CHARMM, NAMD, and AMBER, which produce the binary trajectory files .trr, .dcd, and .netCDF that contain all the coordinates, velocities, forces and energies of the simulation. Other data formats may also contain text files that are used to prepare and run the simulations, such as input files, biomolecule coordinate files (stored usually in .pdb format), output log files.

2.1.4.1.2 Computer-aided drug design data

Drug design is the inventive process of finding new medications based on the knowledge of a biological target. The drug is most commonly an organic small molecule that activates or inhibits the function of a biomolecule such as a protein, which in turn results in a therapeutic benefit to the patient. Computer-aided drug design (CADD) involves the design of molecules that are complementary in shape and charge to the biomolecular target with which they interact and therefore will bind to it aided by the use of computational methods. Datasets in CADD may contain Ligand (candidate drugs) Libraries for Computer-Aided Drug Design. Ligands are stored in multiple formats such as SMILES (Simplified Molecular Input Line System), which was developed as an unambiguous and reproducible method for computationally representing molecules, InChIKey (International Chemical Identifier Key), which was released in 2005 as an open source structure representation algorithm that is meant to unify searches across multiple chemical data bases using modern internet search engines, and other common ligand file formats such as .pdb, .mol, .sdf, .mol2. CADD may also contain protein target databases as therapeutic targets, which are produced by methods of structural biology (NMR, X-rays, neutron scattering) and are usually stored in Cartesian coordinates such as .pdb, .mol, .sdf, .mol2 files. Finally, CADD outputs are stored as a set of ligands (candidate drugs), which are predicted to activate or inhibit the protein of interest; these are stored in the above-mentioned ligand formats, and contain also predicted free energy of binding to the protein of interest in kcal/mol.

2.1.4.1.3 Analysis of next generation DNA sequencing data

Next-generation sequencing (NGS) is the sequencing of DNA and RNA with new technologies. NGS can generate billions of short reads for each sample and processing of the raw reads will add more information. Various file formats have been introduced/developed in order to store and manipulate this information. The most commonly used are the Raw Sequence, which contains the different nucleotides of a DNA strand, FASTA format, which provides one line of description, then the sequence, and GenBank record, which contains background information such as the source of the biological molecule (what organism) and the scientists who discovered the sequence, the various molecular features of the sequence and some of the biological activity. NGS file formats include FASTQ, FASTA, SAM/BAM, and VCF that are commonly used in analysis of next-generation sequencing DNA data. The FASTA format, generally indicated with the suffix .fa or .fasta, is a straightforward, human readable format. Normally, each file consists of a set of sequences, where each sequence is represented by a one line header, starting with the '>' character, followed by the corresponding nucleotide sequence, in multiple lines of regular width (generally 60 or 80 characters wide). In practice, some tools may produce a sequence with a header and a single long line of sequence. The FASTQ is a text file format (human readable) that provides 4 lines of data per sequence (sequence identifier, the sequence, comments, quality scores). FASTQ format is commonly used to store sequencing reads, in particular from Illumina and Ion Torrent platforms. SFF is a binary file format used to encode sequencing reads from the 454 platform. SAM/BAM File formats are used to encode short reads alignment. The Variant Call Format (VCF) is a specification used in bioinformatics for storing gene sequence variations.

2.1.4.1.4 Synchrotron data

Proteins can be visualized using experimental methods of structural biology such as X-ray crystallography. Out of over 75 000 crystal structures of macromolecules deposited in the Protein Data Bank (PDB) by 2012, about 72% were solved with the use of synchrotron radiation. The proportion of structure depositions based on synchrotron data has steadily increased since the first experiments were carried out at Stanford and accounts for about 85% of all depositions in the last 3–4 years. Each spot on a protein is a diffracted X-ray beam, which emerged from the crystal and was registered by the X-ray detector. Thousands of diffraction spots need to be collected to solve a protein structure. Using crystallographic terminology, this process is called X-ray data collection. Detectors, which are necessary for recording diffraction images during crystallographic data collection, have been developed for both in-house sources and synchrotron beamlines. The data format commonly collected at beamlines is CCD image file(s) (typically .osc or .adsc).

2.1.4.1.5 Biomedical image data

The major file formats currently used in biomedical imaging: Neuroimaging Informatics Technology Initiative (Nifti), Minc, and Digital Imaging and Communications in Medicine (DICOM). The format that we use in VI-SEEM for biomedical images is DICOM which is a standard for storing and transmitting medical images enabling the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware, and picture archiving and communication systems (PACS) from multiple manufacturers. It has been widely adopted by hospitals, and is making inroads into smaller applications like dentists' and doctors' offices. DICOM files can be exchanged between two entities that are capable of receiving image and patient data in DICOM format. The different devices come with DICOM Conformance Statements which clearly state which DICOM classes they support, and the standard includes a file format definition and a network communications protocol that uses TCP/IP to communicate between systems. DICOM is used worldwide to store, exchange, and transmit medical images. DICOM has been central to the development of modern radiological imaging: DICOM incorporates standards for imaging modalities such as radiography, ultrasonography, computed tomography (CT), magnetic resonance imaging (MRI), and radiation therapy. DICOM includes protocols for image exchange (e.g., via portable media such as DVDs), image compression, 3D visualization, image presentation, and results reporting.

2.1.4.2 Life sciences metadata structure

Metadata for Life Sciences depend strongly on the research area. Hence, this section is organized according to the thematic areas of Life Sciences.

Molecular Dynamics (MD) simulations metadata: Metadata associated with MD simulations are: dataset description (description/aim of the study, references), system description (reference experimental structure, organism, relevant sequence modifications, relevant local structures, cofactors), system simulation conditions (force field type and version, simulation length, temperature, solvent and ions, charge settings, added salt, type of trajectory, number of frames, time per frame, timestep), and preliminary analyses performed (RMSD, ratios of gyration, number of hydrogen bonds, etc).

Computer-aided drug design: Computer-aided drug design (CADD) metadata are associated with (a) the proteins of pharmacological interest used in CADD, (b) the initial ligand libraries used to target the protein of pharmacological interest, (c) the output ligands after processing the ligand libraries with CADD. Metadata concerning the protein of interest are associated with (i) its crystal structure and are described below in the section “Synchrotron metadata”, (ii) position of hydrogens placed in the protein, (iii) tautomerization and ionization states generated for the protein of interest, (iv) placement of important water molecules for drug design, (v) binding site descriptors. Metadata concerning the ligands to be considered for CADD or CADD output ligands are associated with ligands descriptors that are output after CADD within the ligand structure file or as separate text files. These are molecular weight, hydrogen bond donors and acceptors, lipophilicy (logP), solubility (logS), cell permability (CaCos2), polar surface area (PSA), and other physicochemical properties. Other metadata include ligand steric clashes with the protein, potential toxicity, clusters of output CADD ligands according to their physicochemical properties providing representative compounds for each cluster.

Next Generation DNA Sequencing metadata: Metadata associated with NGS refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. Information is supplied by completing all fields of a metadata template spreadsheet. Guidelines on the content of each field are provided within a spreadsheet such as the one provided by GEO (NIH) https://www.ncbi.nlm.nih.gov/geo/info/seq.html#metadata. The metadata associated with NGS samples are information about species, application, file locations, sample labels, lab name, library type, batch information and sample groups (samples which are forming groups to be compared).

Synchrotron metadata: The metadata of raw diffraction images stored are normally contained in the image header. However, not all image formats are equally informative or entries are missing. Ideally, metadata should comprise the following: identification of the image format, number of pixels, pixel sizes, byte-storage architecture, baseline offset and handling of overflows, information on the corrections that are applied (dark current, distortion correction, non-uniformity correction), detector gain, goniometer axes orientations and rotation directions, and information on the experiment such as exposure time, number of repeats, oscillation axis and range, wavelength used, beam polarization, detector position (or beam position), offsets, incident beam flux, byte-storage architecture. Because of the various detector formats, the unifying CBF/imgCIF format was developed. It provides a metadata structure in which all of the metadata can be found in one place. It consists of an ASCII imgCIF header and binary (CBF) or ASCII-based encoded data blocks. The binary format is reasonably space-efficient owing to the use of compression algorithms such as byte_offset compression, and it is useful for large images and for data transfer between collaborating groups. Three categories of data exist, ARRAY data, AXIS data and DIFFRN data, allowing a unique definition of how to interpret the data, and no prior knowledge would be required if all data items were filled in. This is often not the case, however; for example, PILATUS detector image files contain all relevant metadata in just a small comment line block, the so-called miniCBF format.

Biomedical images: Biomedical image metadata is typically stored at the beginning of the file as a header and contains at least the image matrix dimensions, the spatial resolution, the pixel depth, and the photometric interpretation. Thanks to metadata, a software application is able to recognize and correctly open an image in a supported file format simply by a double-click or dragging the image icon onto the icon of the application. In the case of medical images, metadata have a wider role due to the nature of the images itself. Images coming from diagnostic modalities typically have information about how the image was produced. For example, a magnetic resonance image will have parameters related to the pulse sequence used, e.g., timing information, flip angle, number of acquisitions, etc. A nuclear medicine image like a PET image will have information about the radiopharmaceutical injected and the weight of the patient. These data allows software like OsiriX to on-the-fly convert pixel values in standardized uptake values (SUV) without the need to really write SUV values into the file. Post-processing file formats have a terser metadata section that essentially describes the pixel data. The different content in the metadata is the main difference between the images produced by a diagnostic modality and post-processed images. Metadata are a powerful tool to annotate and exploit image-related information for clinical and research purposes and to organize and retrieve into archives images and associated data.

Regarding the "DICOM Network" which was developed in Moldova, the XML metadata is removed from DICOM file on the import and is inserted to the Database. This makes possible the following functionalities and features: 1) Size of DICOM image is reduced and in case we have multiple slices with the same information this approach removes the repeated data. 2) It is easier to impersonate the data. 3) The date in database is enriched and complemented by the data from HIS (Hospital information system). 4) It is possible to launch a search by metadata.

2.1.5. Cross-disciplinary data

2.1.5.1 <u>Description of cross-disciplinary data</u>

The scope of VI-SEEM is to provide services for the scientific disciplines of Climate, Digital Cultural Heritage as well as Life Sciences. Although high emphasis is given in applications focusing on the three communities, one of the important goals of the project is to enhance the cross-disciplinary synergy between the different communities. To this end, the updated version of the VRE portal provides a section dedicated to the cross-disciplinary character of the VI-SEEM project, namely, it provides access to the following entries:

- Data Visualization
- Simulation Data
- Data Analytics and Processing
- Geographic Datasets Description
- Analytical Studies and Portfolio
- Source Code

Data Visualization: Data Visualization will be providing the output of applications which involve visualization techniques such as those providing 3D reconstructed objects produced in the scope of Digital Cultural Heritage, which at the same time could be used in applications of life sciences (for instance bioarchaeology) and Climate (for instance in visualizing the outcome of climate effects such as winds and dust storms, as well as the role climate changes played in human societies, population movements in the past, etc). Hence, data visualization section will be a selection of datasets describing visualizations which could be used by more than one discipline.

Simulation Data: Simulation Data would provide datasets that have been produced by performing the computer simulations and which can be used by anyone of the three scientific communities. An example for such dataset would be Atmospheric Chemistry data produced via simulations in the framework of Climatology. These could also be used in Digital Cultural Heritage applications in identifying the corrosion of archeological monuments due to air pollution (and its impact on the conservation of built heritage), as well as in Life Sciences in order to identify the effects of air pollution and air quality in biological aspects.

Data analytics and processing: Data analytics and processing section will be providing all the associated tools which can be used for the examination of raw datasets with the purpose of finding patterns and drawing conclusions about the hidden information behind these datasets. Such tools will included algorithmic and mechanical processes which can be used to enlighten the insight of the data. Example of these tools would be workflows and machine learning codes making use of Convnets which can be used by researchers working in one of the three scientific communities.

Geographic Datasets Description: Geographic Datasets Description is the cross-disciplinary section that provides Geographic Information System (GIS) mapping datasets. These consist mostly of spatial or geographic data. Spatial analysis with geographical information systems is used extensively in the research fields of climatology (hydrological modeling, cartographic modeling), digital cultural heritage and in life sciences. Such datasets have already been identified and are ready to be uploaded in the virtual research environment. An example of this kind of datasets is drone mapper imagery provided by the First Montenegrin Centre of Excellence BIO-ICT. The drone mapper is an open source GIS and image processing tool which deals with images recorded by drones. BIO-ICT implemented this tool and uses it in processing of vineyard and sea images. It can be implemented in our project in all three communities, installed on large project infrastructure and can be useful to all project participants. Processed images will be prepared as datasets for further analysis.

Analytical Studies and Portfolio: Analytical Studies and Portfolio section provides studies based on analytical approaches such as physical processes as well as chemical processes which can provide information on the materials’ structure and chemical properties of substances. One example is chemical analysis studies as well as material analysis investigations. This kind of analytical works can be used broadly in the fields of digital cultural heritage (chemical properties of structures) as well as in the life sciences community.

Source Code: Source Code could provide access to the part of the code repository which focuses on codes which can be used by different communities. Such codes include, for instance, quantum-chemistry and chemical-kinetics related codes which can be used by both communities of climate and life sciences. In addition this section can include codes and scripts associated with visualization tools (ParaView for instance) to be used by the digital cultural heritage as well as climate scientific communities.

WP5 leader facilitated the identification of possible applications that could potentially enrich the cross-disciplinary fields of the VI-SEEM VRE. So far three such applications have been identified and one has already been implemented. These applications involve the synergy of Digital Cultural Heritage and Climate community as well as aerial imagery and photography. As of now, VI-SEEM integrated codes developed in the context of the CyI-Illinois RIVEEL3D (Real-time Immersive 3D Virtual Environment for Education and Learning) research activity which have been implemented and is accessible in the VI-SEEM CLOWDER. CyI-Illinois RIVEEL3D is developed in collaboration with the National Centre for Supercomputing Applications at the University of Illinois at Urbana-Champaign, and was used as host virtual environment by VI-SEEM for the testing of crowd simulation optimization algorithms for the study and analysis of historical sites in Eastern Mediterranean. CyI-Illinois RIVEEL3D contributes to the on-going study of the Green Line of Nicosia in Cyprus that still divides the city, and proposes the analysis of the use of public spaces in contested urban environments. The current stage of integration involves a Unity plugin for the mobile interaction with geo-located assets of digital cultural heritage, as well as visualization of the impact of climate change-induced extremes, like flooding, heat waves and dust storms, on the built environment of historic cities. RiVEEL3D can be accessed by clicking here; this requires to get authorization to use the VI-SEEM Clowder.

2.1.5.2 Cross-disciplinary metadata

Dealing with the cross-disciplinary character of VI-SEEM introduces a rather challenging task, namely, how to treat the metadata structure in order to provide interoperability between datasets from different disciplines and enhance the synergy and effectiveness of collaboration and cooperation of different scientific sectors. Hence we need to provide a structure which will enable our data sets to be easily found, accessed, retrieved as well as provide their long term preservation (in the context of cross-disciplinary nature).

As this was demonstrated in previous sections, datasets from different communities are described by different metadata schemas and conventions. One way of dealing with the lack of internal standards and internal communities' coherence and cohesion of current metadata effort is to evolve an ontology-based metadata approach.

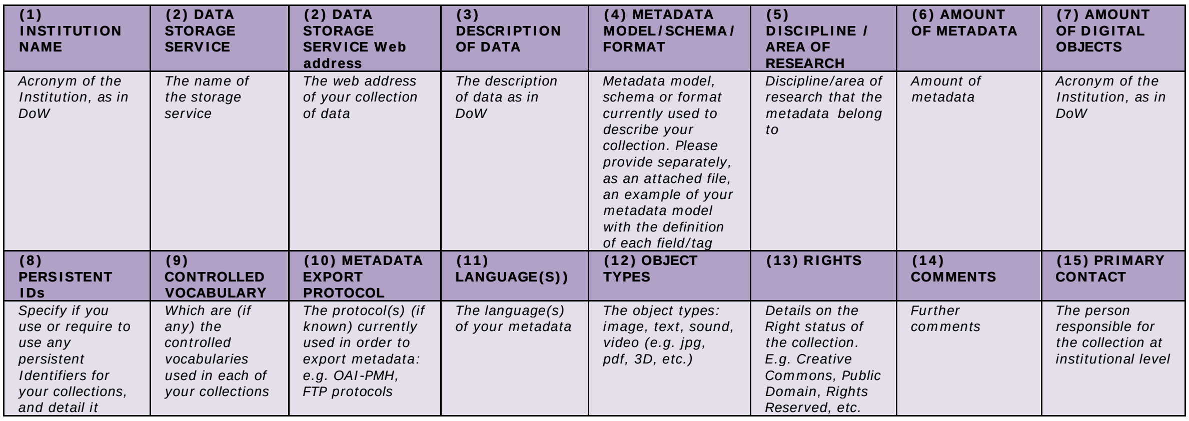
Hence, there is a necessity of developing *ontologies* which represent knowledge as a set of concepts within the domains and the relationships between those concepts, as well as metadata schemas. In other words, sets of metadata elements designed for a specific purpose, such as describing a particular type of information resource.

Practically, we should develop - or adapt an existing one - a framework which will provide answers to following:

1. Metadata schemas between different scientific communities, and even between different scientific applications within the same scientific community, use different structures/schemas. Therefore, a problem of inconsistency rises.
2. A framework requires to trace the information about the pipeline for creation of the digital resource from the acquisition to publication.

3. A framework requires to provide a detailed description of all the intermediate passages.
4. Such framework requires to preserve the digital content of the datasets, providing long term accessibility, content retrieval, data interpretation, data transparency, re-use of the data.

This framework requires to adopt a strategy that first requires to understand the diversity of the provided information. To address the above questions WP5 took the initiative and performed a Data Management Plan survey. The questions addressed in the survey are included in Appendix 1; in addition a picture of the survey as this was distributed to the developers contributing to the integration phases is provided in Figure 1.



(1) INSTITUTION NAME	**(2) DATA STORAGE SERVICE**	**(2) DATA STORAGE SERVICE Web address**	**(3) DESCRIPTION OF DATA**	**(4) METADATA MODEL/SCHEMA/ FORMAT**	**(5) DISCIPLINE / AREA OF RESEARCH**	**(6) AMOUNT OF METADATA**	**(7) AMOUNT OF DIGITAL OBJECTS**
Acronym of the Institution, as in DoW	*The name of the storage service*	*The web address of your collection of data*	*The description of data as in DoW*	*Metadata model, schema or format currently used to describe your collection. Please provide separately, as an attached file, an example of your metadata model with the definition of each field/tag*	*Discipline/area of research that the metadata belong to*	*Amount of metadata*	*Acronym of the Institution, as in DoW*
(8) PERSISTENT IDs	**(9) CONTROLLED VOCABULARY**	**(10) METADATA EXPORT PROTOCOL**	**(11) LANGUAGE(S))**	**(12) OBJECT TYPES**	**(13) RIGHTS**	**(14) COMMENTS**	**(15) PRIMARY CONTACT**
Specify if you use or require to use any persistent Identifiers for your collections, and detail it	*Which are (if any) the controlled vocabularies used in each of your collections*	*The protocol(s) (if known) currently used in order to export metadata: e.g. OAI-PMH, FTP protocols*	*The language(s) of your metadata*	*The object types: image, text, sound, video (e.g. jpg, pdf, 3D, etc.)*	*Details on the Right status of the collection. E.g. Creative Commons, Public Domain, Rights Reserved, etc.*	*Further comments*	*The person responsible for the collection at institutional level*

Figure 1. The data management survey as this was distributed to the developers of the VI-SEEM application

After collecting the contributions from the partners we analyzed the questions and provided directions regarding the progress of the development of the framework to address the cross-disciplinary metadata structure. The critical questions that will shape the framework on which our ontology will be based are:

- Metadata Model/Schema/Format. A short description for the metadata formats is provided in Appendix 2.
- Persistent IDs, namely whether a dataset is using or requires to use any persistent Identifiers.
- Controlled Vocabulary, namely which are the controlled vocabularies used in each collection of datasets.
- Export Protocol, namely the protocol(s) (if any) currently used in order to export metadata: e.g. OAI-PMH, FTP protocols.

The results of the survey for the research communities of Climate, Life Sciences, and Digital Cultural Heritage are provided in Table 1, Table 2 and Table 3 respectively.

Metadata Schema	Persistent identifiers	Controlled vocabulary	Export protocol
NetCDF CF (8)	N/A (12)	N/A (15)	OPeNDAP; KML-based export (4)
Dublin Core (1)	URI (2)		OAI-PMH (1)
WRF model (2)	EPIC PIDs (1)		FTP / SFTP (4)
Proprietary OpenFOAM (1)			N/A (5)
GeoTIFF (1)			SCP (1)
No metadata (1)			
GRIB (1)			

Table 1. The output of the data management survey for the climate scientific community

Metadata Schema	Persistent identifiers	Controlled vocabulary	Export protocol
.SDF format (1)	N/A (11)	N/A (12)	OPeNDAP (5)
TXT file (3)	EPIC (1)	List of terms (1)	N/A (4)
.DCD / .XTC (1)	Record identifier (1)		FTP (1)
No metadata (3)			SFTP (1)
CDISC / ODM (1)			XML-based export (2)
Gro / pdb format (2)			
NAMD, Spa, PDB, MAP, RasMol (1)			
DICOM (1)			

Table 2. The output of the data management survey for the life science scientific community

Metadata Schema	Persistent identifiers	Controlled vocabulary	Export protocol
UNIMARC / MARC 21 (1)	N/A (9)	UNIMARC / MARC 21 (1)	http (1)
Html format (1)	Record identifier (1)	N/A (7)	XML-based export (4)
Text format (2)		User defined (2)	OAI-PMH (1)
Dublin Core (1)			N/A (3)
N/A (4)			OPeNDAP (1)
CIDOC-CRM, RAW xml (1)			

Table 3. The output of the data management survey for the digital cultural heritage scientific community

The above tables reveal a number of issues regarding the derivation of an ontology capable of governing the interoperability of data sets between cross-disciplines. Namely:

- It appears that some data sets come without metadata nor a particular metadata schema. This provides inconsistencies in the way one handles different metadata formats.
- There are similarities between the Climate and Life Sciences metadata formats in the sense that most of the metadata schemas are produced by software, while in the case of Digital Cultural Heritage metadata structure requires human input (html, text formats).
- Only some collections of datasets have Persistent identifiers.
- Only a few collections of datasets use controlled vocabulary.
- Different methodologies of exporting protocols are being used.

To address the above challenges we will adopt the following actions. First, we provide for each discipline a common metadata structure and, secondly, we will focus on these structures in order to provide the common cross-disciplinary metadata structure. This, of course, is part of an ongoing work and since the Data Management Plan is a “live document” it will be updated accordingly.

Nevertheless, a good starting point to create a metadata schema is to first build a framework based on the General Metadata provided in Section 2.1.1 as well as to include information derived from the executed survey. This information involves the following:

- 3D data
- Site location – area
- Standards (e.g. WGS84)
- Mapping process

At the moment the above plan will provide the cross-disciplinary metadata information. This will be shortly updated in order to include more aspects of the data management survey. The collection of the results from the survey provides a metadata ingestion plan. This document is useful for tracing the activities of content preparation, content aggregation, data publication etc.

To transform heterogeneous metadata into one interoperable metadata standard we are planning to use MINT. MINT is an open source, web based platform for Metadata Interoperability. It has been successfully used in more than 15 Europeana feeder projects involving 300 cultural organizations and 500 users. In addition, more than 6.000.000 metadata records have been produced and published.

2.1.6. Definitions

Definitions that are used along the Data Management Plan include:

1. File format: Standard way of encoding and storing digital data.
2. Metadata: Data that provides information about the data, such as tags for better search and retrieval, context about data collection and data formats.
3. Ontology: a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.

4. Simulation data: Data as a result of modeling and simulation, such as climate projections.
5. Experimental data: Data as a result of experimental processes, such as biological data.
6. Observational data: Data as a result of observational surveys, such as climate data.
7. Image data: Data that consist mainly of images, such as medical and archeological artefacts.
8. Text data: Data that consist mainly of text, such as books and archeological inscriptions and papyri.

2.2. Data collection and documentation

The data is collected as a part of the VI-SEEM project, according to the following procedure. Only collaborators, registered at the level of contributor, will be eligible to upload data to the VI-SEEM VRE. Upon submission of the data, a standard form describing the data will be filled by the contributor and submitted alongside the data. The form will include, for each dataset, at minimum:

1. Data description.
2. The scientific community of the data.
3. File format.
4. Metadata.
5. Any pre-processing that was applied to the data.
6. Documentation that accompanies the data, to provide context about how the data was created, authors and their contributions.
7. How the data can be used (level of access).
8. How long the data needs to be preserved (level of preservation).
9. Future addition: Cross-Disciplinary metadata[1].

Clear directions for the procedure of data collection are shared with the contributors and are available on the website where data collection will take place. After being collected, the data will be controlled for quality by the QA/QC officer, as described in the following subsection.

2.2.1. Data quality control and assurance

Measures will be taken to assure quality of the uploaded dataset as a whole. As a minimum prerequisite, dataset providers are obliged to fill in the form described above as well as to provide a well-documented process that should be used to perform quality check on the dataset. Based on the latter, an initial quality check on selected data will be presented to the team of QA/QC Officers.

[1] As it has been discussed in the previous section our plan is to develop an interoperable metadata standard. This will be included in the future form of data collection.

Namely, each of the partners of VI-SEEM is required to have a delegate in the QA/QC Officers Team. The method of selecting data for later quality checks will be defined by SC leaders.

All this will ensure that data uploaded to VI-SEEM VRE (either to repositories or long term data archiving services) and available for sharing with the communities is of the satisfactory quality and that information about the data is readily available upon submission. In addition, this process will avoid the possibility of hosting and sharing illegal data. Legal and ethics issues are tackled in the next section on data access and sharing.

The data will not become available for users until the SC leader approves the associated form and the QA/QC officers verify the quality of the data. The WP5 leader will perform periodic biannual checks for the quality check processes, the data selection method and the associated information of the data, and will contact corresponding QA/QC officers, and SC leaders if issues arise.

2.2.2. Quality control and quality assurance officers

Each partner has assigned a Quality Assurance and Quality Control (QA/QC) officer. This decision was taken internally and the selected person is expected to have an extended knowledge on data management. Once the registered user uploads a dataset in the repository, which could be either the VI-SEEM repository or the VI-SEEM Clowder for Digital Cultural Heritage, the corresponding QA/QC officer as well as the SC and WP5 leaders will receive automatically a notification email assigning the Quality Control and Assurance of the data set to the QA/QC officer. WP5 leader will supervise whether the officer delivers the task in due time (approximately a week). If during the QA/QC procedure a problem arises, it is escalated to the Service Enabler, and if the problem still persists, then it is directed to the associated scientific leader. The names of the QA/QC officers are provided in Table 4. In addition we provide the contact details (emails) on the same table. Regarding the datasets that have already been uploaded in the VRE, QA/QC officers will be assigned the performance of quality assurance and quality control on these data sets.

Partner	QA/QC officer	Contact details
Greece	Ioannis Liabotis	iliaboti@admin.grnet.gr
Cyprus	Panayiotis Charalambous	ps.charalambous.@cyi.ac.cy
Bulgaria	Vladimir Dimitrov	vgd@acad.bg
Serbia	Petar Jovanovic	petarj@ipb.ac.rs
Hungary	Lajos Bálint	lajos.balint@niif.hu
Romania	Marian Neagul	marian.neagul@e-uvt.ro
Albania	Marin Aranitasi	maranitasi@fti.edu.al
Bosnia and Herzegovina	Vladimir Risojevic	vladimir.risojevic@etf.unibl.org
FYR of Macedonia	Sonja Filiposka	sonja.filiposka@finki.ukim.mk

Montenegro	Lidija Milosavljevic	lidija@ac.me
Moldova	Mihail Matenco	mihail.matenco@renam.md
Armenia	Artashes Mirzoyan	amirzoyan@sci.am
Georgia	Ramaz Kvatadze	ramaz@grena.ge
Egypt	Mohammed Elfarargy	Mohammed.Elfarargy@bibalex.org
Israel	Zivan Yoash	zivany@mail.iucc.ac.il
Jordan	Mostafa Zoubi	mostafa.zoubi@sesame.org.jo

Table 4. The QA/QC officers and their contact details for each partner

2.2.3. Quality control minimum requirements

A set of minimum requirements the datasets need to fulfill should be defined as guidelines for QA/QC officers in order to perform the Quality Assurance and Quality Control. The following bullet points provide the minimum criteria for a dataset to be validated:

- Is the dataset accessible?
  - Answer should be “yes”
- Is the dataset corrupted?
  - Answer should be “no”
- Is the particular dataset in accordance to DoW?
  - Answer should be “yes”
- Does the dataset have the right format?
  - Answer should be “yes”
- Is the dataset accompanied with a description?
  - Answer should be “yes”
- Does the dataset have the right description according to DoW?
  - Answer should be “yes”
- Does the submission provided the level of access?
  - Answer should be “yes”
- Does the submission define the level of preservation?
  - Answer should be “yes”
- Is the dataset accompanied with metadata?
  - Answer should be “yes”
- Is there documentation that accompanies the data, to provide context about how the data was created, authors and their contributions?
  - Answer should be “yes”
- Does the dataset have a PID?
  - Answer should be “yes”
- For future use only: Is the dataset accompanied with cross-discipline metadata?
  - Answer should be “yes”

The above points are the minimum requirements a dataset needs to obey to become accessible. In the future a set of minimum requirements for metadata based on the derived ontology will also be provided in the Data Management Plan.

2.3. FAIR Data

Data services of VI-SEEM are aimed at providing their capabilities according to the FAIR principle. This means that the various data sets of the different communities/disciplines shall be:

- Findable
- Accessible
- Interoperable
- Reusable

Efforts have been made to assure that data is handled by VI-SEEM Data Services in conformance with these requirements. These are presented in the following chapters.

2.3.1. Data access and sharing

Findability is assured by the VI-SEEM PID service which provides unique persistent identifiers for each dataset. The VI-SEEM Data Discovery Service (https://search.vi-seem.eu/dataset) could be used for searching for datasets by several criteria. Note that access to data may be limited but still, data shall be searchable by the means of accompanying metadata at all times.

Data collected through VI-SEEM will be governed by the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license by default, with some exceptions, as outlined in this document. This license "only allows users to download the work and share them with others, as long as they credit the authors, but they cannot change them in any way or use them commercially" [2]. For some datasets, an additional restriction will be instituted: an embargo on data access for scientific data that was directly collected or produced by a group, until that group has been provided with a reasonable time frame to study and publish the results, as determined by each group. Additional restrictions may be placed for ethical, or other reasons, as determined by the collaboration and reported in this document.

Data in the life sciences that might pertain to individual patients will be properly anonymized and only available in aggregate form to comply with ethical and legal directives set by the EU. Further ethical and legal constraints will be enforced where appropriate for the safety and privacy of subjects. Certain data is provided to the VI-SEEM collaborators from external agencies with the provision that it cannot be shared. In that case, the policies of the original data providers will be respected.

Educational and outreach material, as well as data for visualization will be available according to the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license, unless otherwise stated. This license "lets other remix, tweak, and build upon the work non-commercially, as long as they credit the authors and license their new creations under the identical terms" [2].

VI-SEEM collaborators will be encouraged to publish in open-access peer-reviewed journals and conferences, whenever possible, to ensure that fellow scientists and citizens have access to the results without limitations. This policy will not be strictly enforced. The scientific data is classified broadly in three levels of access:

1. Open: Governed by Creative Commons Attribution-NonCommercial-NoDerivs License, data is open for non-commercial use, as long as the authors are credited and the data cannot be changed.
2. Restricted: Based on the Creative Commons Attribution-NonCommercial-NoDerivs Licence, the dataset is available to be used from users and contributors of VI-SEEM platform with some restrictions. These restrictions may include an embargo for a time period after publication, and any other restrictions, as determined by the contributor of this dataset. These restrictions will be noted upon submission of the data.
3. Closed: Not available to any user other than the contributor of this dataset

According to the FAIR principle data should be “as open as possible and as close as necessary”.

The authors of any work that makes substantial use of the data must communicate with the data source prior to publication. The data will be quality controlled, but VI-SEEM does not accept any liability for the correctness and/or appropriate interpretation of the data, which is the sole responsibility of data users. When a contributor shares data, the contributor will electronically sign an agreement that specifies that data belong to the owner, are not stolen and can be used for the purposes they are shared for.

Data users must respect any restrictions on the use or reproduction of data. If a user encounters problems associated with the data, they can provide feedback to the WP4 leader through an email form in the webpage. The WP4 leader will then communicate with the contributor of the data to fix the reported issues. Such problems may include data corruption, inconsistent file formats, discrepancies between description and data. The authors of the data retain all Intellectual Property Rights.

2.3.2. Data storage and preservation

VI-SEEM is responsible for the preservation of data in the following three levels of preservation:

1. Short term (less than 6 months, available only within the duration of a computational project, scratch data).
2. Medium term (around 12 months, available only within the duration of a computational project).
3. Long term preservation of the data (36 months, within the duration of the project, catalogues / indexed searchable data).

Requirements for storage have been translated into a set of common services. These have been deployed along with the data repositories. Semantics and metadata associated with each dataset are stored in the VI-SEEM Repository Service. Each dataset requiring medium or long term preservation has a persistent identifier (PID) from the VI-SEEM PID Service to both assure sustainability and consistency of access and help data re-use.

For short term storage, users may choose the VI-SEEM Simple Storage Service. Note that for such short-lived data, no PID is required.

For medium and long term, VI-SEEM Repository Service is utilized coupled with the VI-SEEM Archival Service to improve data safety.

Archival and preservation of the collected data after the duration of the VI-SEEM project is not guaranteed. Contributors of data who desire to extend the duration of the preservation of their submitted data should communicate with their SC leader in a timely manner (before the end of the VI-SEEM project) to ensure that procedures are in place for the correct treatment of their data, in agreement with the infrastructure providers.

Metadata and other supplemental information will be included in the collection of the data. A common cross-disciplinary schema and mappings from/to accepted standards will be integrated. This is required to ensure interoperability.

2.3.2.1 Allocation of resources

Underlying data storage resources are integrated sequentially starting from the beginning of the project, and from project month 10 (July 2016) full capacity, 330 TB of disk and 510 TB of tape storage space, is made available to the scientific communities.

Provided resources and services overall are mainly used through the development access, as well as through calls for production use of resources and services. The VI-SEEM development access facilitates the development and integration of services by the selected collaborating user communities. In this process, applications are given access to necessary computational and storage resources for a six-month period, during which application developers are expected to develop and integrate relevant services. The calls for production use of resources and services target specific communities and research groups that have already began development of their projects. These calls are intended for mature projects, which require significant resources and services to realize their workplans. Therefore, significant utilization of VI-SEEM data storage comes from the calls for production use of resources and services, and an order-of-magnitude smaller utilization comes from the development access.

The VI-SEEM development access is divided into three integration phases, and three separate calls for production use of resources and services will be issued during the project lifetime. Up to now, three integration phases and one call have been issued. Thus, the provided infrastructure resources are utilized during:

- 1st integration phase from project month M6 to M11,
- 2nd integration phase from M9 to M19,
- 3rd integration phase from M21 to M26,
- 1st call for production use of resource from M18 to M30.

2.3.2.2 Security, privacy and ethical aspects

2.3.2.2.1 Security

There are two aspects of security in VI-SEEM infrastructure, the authentication and authorization, which handles users verification and access levels, and transport level security that secures data traveling over the network between the resources. On the authentication and authorization aspect, two types of resources can be discerned: web based and legacy resources.

Web based resources use VI-SEEM AAI for authenticating and authorizing their users. AAI provides federated access to web based services such as Simple Storage service, Repository service, Data Discovery service, and Persistent Identifier service.

In simple terms, in web based services, HTTPS protocol is used for secure communication between endpoints. HTTPS is the standard protocol for securing communication on the web. It is the usual HTTP which runs on top of encrypted sockets (SSL/TLS) on the transport layer of the network stack (TCP/IP). Its encryption algorithms use a long-term private and public keys to negotiate a shared session key which is then used for communication between client and server while the connection is active. The asymmetric keys are verified by trusted third party certificate authorities (CAs).

For legacy services, which are not directly integrated into the VI-SEEM AAI, access is handled through SSH, whose authentication and authorization rely on standard Unix mechanisms (of the underlying platform on which they run) to verify users and give them appropriate access.

SSH, with SCP and SFTP, is remote shell and file transfer protocol similar to SSL/TLS in that it uses the same encryption algorithms for encryption at the transport level. It has its own asymmetric key management which is used to authenticate users without using passwords. It provides access to Work Storage Space, Local Storage, Data Staging, and Data Analysis services.

Among the legacy services, iRODS also has its own internal authentication and authorization schemes and handles transport level security between its servers and clients via SSL/TLS. This is specific to iRODS because it has its own federated access mechanism to allow for cooperation and data sharing between multiple instances managed by different institutions.

2.3.2.2.2 Privacy

Special care must be taken when working with personal data especially when it is of sensitive nature. In general, data sources ingested into the VI-SEEM VRE – containing personal data – should already comply with requirements for minimization (process as few as possible and only when really necessary) and pseudonymization (replace most identifying parts of personal data with artificial identifiers).

Data set providers are required to indicate whether or not working with personal data. In addition, if the data source is not already pseudonymized then a preprocessing step to make it so is required to be implemented by the data source responsible.

This way, it could be assured that no data subject could be re-identified based on data sets ingested in the VI-SEEM VRE.

Compliance with the above requirements should be regularly audited for each of the dataset providers working with personal data.

Furthermore, important care is taken when working with clinical phenotypic datasets. To this purpose we have investigated REDCap as a platform for creating general data collection web applications which will take care of the anonymization requirements of these datasets. Suitability of its APIs and licensing were evaluated, and the license was obtained from Vanderbilt University, where the main group developing REDCap is from. Development of the biobank on the REDCap platform, that included creating database schema from IMGGE requirements. The specifications for the data were given as a set of excel spreadsheets in a semi-regular form. Since there were too many fields to be entered by hand without making a mistake, a conversion script was created to parse the spreadsheets and translate them into formal specification that REDCap could import as database schema. The service is online and it is available at https://biobank.ipb.ac.rs/.

2.3.2.2.3 Ethical Aspects

Regarding the analysis of Next Generation DNA sequencing data, VI-SEEM provides information about ethical, legal and societal issues for the biobanking community through its collaboration with BBMRI-ERIC. VI-SEEM is requesting compliance for collection and reuse of data (retrospective research) according to the current GDPR which will soon be replaced by the EU new GDPR to be released on 25 May 2018. VI-SEEM shares the GDPR Code of Conduct from BBMRI, which aims to contribute to the proper application of the regulation, taking into account the specific features of processing personal data in the area of health; clarify and specify certain rules of the GDPR for controllers who process personal data for purposes of scientific research in the area of health; help demonstrate compliance by controllers and processors with the regulation; help foster transparency and trust in the use of personal data in the area of health research (http://www.bbmri-eric.eu/BBMRI-ERIC/gdpr-code-of-conduct/). The revised version of the Council for International Organizations of Medical Sciences (CIOMS) ethical guidelines for human research is also available through BBMRI-ERIC (http://www.bbmri-eric.eu/news-events/revised-version-of-the-cioms-ethical-guidelines-for-human-research/).

Questions regarding the legal and ethical requirements of the personal data processing were discussed and highlighted in a workshop that took place on 20 June 2017 in Athens, Greece, organized by the Biomedical Research Foundation, Academy of Athens and BBMRI-ERIC. The workshop aimed to increase the awareness of the Greek audience on ethical & legal issues, within the context of biomedical research and more specifically biobanking on a national and regional level. In that context, the way that specific EU countries plan to implement the GDPR in their respective legislations were explored. Researchers, bioethicists, lawyers, students, as well as members of the general public who wished to get informed on ethical & legal issues concerning biobanking were invited to attend; the proceedings of the workshop are available here: http://www.bbmri-eric.eu/wp-content/uploads/Report-workshop-Athens-20-June-2017_FINAL.pdf

In order to ensure that all VI-SEEM data follow the EU guidelines and regulations regarding anonymization, sharing of data, retrospective research, and all issues pertaining to sensitive patient data, we have formed the VI-SEEM Life Sciences Scientific Advisory Board (SAB), which will ensure that all guidelines are being followed for VI-SEEM data. The SAB is comprised of Ms. Olga Tzortzatou, Lawyer, Expert in Bioethics and member of BBMRI-ERIC, Professor E. Dermitzakis, Department of Genetic Medicine and Development, University of Geneva, Switzerland and Director: Health 2030 Genome Center, Professor Leonidas Alexopoulos, Systems Bioengineering lab, National Technical University of Athens and Founder, ProtAtOnce, Professor Emilija Shukarova Stefanovska, Research Center for Genetic Engineering and Biotechnology “Georgi D. Efremov”, Macedonian Academy of Sciences and Arts. We are also in contact with Michaela Mayerhoff, Chief Policy Officer, Ethical, Legal and Social Issues (ELSI) Officer and Chief Coordination Officer, who is updating us on ethical aspects.

2.3.2.2.4 Other aspects – restriction in datasets

Apart from restrictions in access to data resulting from ethical issues affecting Life Science datasets other aspects also might confine the data set access. As can be seen in Table 6, Table 7 and Table 8, restricted access to data sets occur also for data sets in Climate as well as Digital Cultural Heritage.

Some datasets appear to be only available to a group of researchers while most probably, they can become available to other parties upon request. For instance, in Climate WRF-ARW datasets are only available to CYMET and designated CyI staff. Availability of data to other parties can be granted upon request and through CYMET.

Furthermore, a number of datasets are available only to the dedicated research group prior publication. Once the associated article is published they become publicly available without any restriction in the level of access. Examples of such datasets are those of RCM MENA-CORDEX and HIRECLIMS.

2.3.2.3 Available data services of VI-SEEM

VI-SEEM offers the following data services to serve the different requirements of the Data Management Plan for each of the communities and research groups participating in the project.

- VI-SEEM Simple Storage Service (VSS)

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.

- 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).

Clowder [ref] is a similar service (with enhanced functionality such as image viewing and processing) that is being used by the community of digital cultural heritage for storage and sharing of datasets.

- 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).

- 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.

- VI-SEEM Data Discovery Service (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 Data Analysis Service (VDAS)

VI-SEEM data analysis service (VDAS) provides the capability to carefully and efficiently investigate and analyze 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.

2.3.2.4 *Persistent identifiers policy*

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.

All datasets which are stored in the VI-SEEM repository service for the purpose of long term storage and data sharing get automatically a persistent identifier. VI-SEEM repository is using the single prefix 21.15102 for all datasets from all the different communities. Datasets are then assigned suffixes in the form of VISEEM-XXX where XXX is a serial number assigned to the datasets by the repository. An example of a dataset with a PID is the following:

Dataset name	Computer-aided drug design to predict binding poses and relative binding affinities for FXR ligands in the D3R Grand Challenge 2
PID	http://hdl.handle.net/21.15102/VISEEM-277

Table 5. An example of a dataset with a PID

As the VI-SEEM repository is being used not only to store community and application specific data but also to store publications, documents, video, images and presentation all different types of data stored by the VI-SEEM communities can acquire and use a persistent identifier.

2.3.3. Data governance

The VI-SEEM technical board will possess the overall responsibility, to ensure that the policy, as outlined in this document, is followed, and any issues that might arise with regards to this policy or its implementation will be dealt with by the responsible parties, or discussed by the VI-SEEM consortium when solutions are not clear. Possible issues include data corruption or inconsistency, security breaches and access to the data by registered users, non-registered users, and contributors.

The process of registration is controlled by the VI-SEEM SC leaders and administrators. Any member of the scientific community at large can be registered as a user, whereas members of the VI-SEEM collaboration can be registered as contributors, to be able to share their data.

Open publications and relevant educational and outreach material intended for educators and outreach coordinators can be accessed openly (without registering).

The responsibility for following the guidelines set by this Data Management Plan is shared across the VI-SEEM collaboration. WP3, WP4 and WP5 leaders are providing the implementation of the infrastructure supporting the data access, preservation and re-use. The contributors of the data, the respective SC leaders as well as the associated QC/QA officers are responsible for the data quality control and assurance as well as other processes that guard the integrity of the data. The SC leaders will review the needs for security and confidentiality of the data, and will convey clear directions to the team responsible for the infrastructure of the VI-SEEM project.

Data providers are responsible to ensure that the data they are storing and/or sharing abided to the policies specified in this document and the terms of use of each of the corresponding services. QA/AC officers are also responsible to supervise whether the above responsibilities are satisfied.

The policies of the Data Management Plan will be monitored and reviewed by the WP5 leader if deemed necessary, and this document will be updated accordingly. The Data Management Plan is a “live document” and therefore occasional of frequent updates are expected depending on the progress of the project. For example an updated version of the Data Management Plan is expected to provide progress on the creation of an ontology which will enable the interoperability of datasets produced under the scope of different scientific communities. In addition an update of the Data Management Plan will be provided along with the introduction of new datasets (for instance from the open calls) with different requirements than the currently known ones.

2.4.Data set identifiers, levels of access and preservation

The following tables (Table 5, Table 6 and Table 7) describes the data to be collected as part of the VI-SEEM project. The levels of preservation and access are included per dataset. There are three levels of preservation: Short term (less than 6 months), Medium term (12 months), and Long term (36 months or longer), and three levels of access/dissemination/user: Open (no constraints for non-commercial use), Restricted (embargoed for publication, or other restrictions), Closed (only available for contributors of data).

Additional data produced in parallel to the aforementioned, such as publications and educational and outreach material will be released as soon as they are available and accessible to all (when possible, for open-access publications) for the duration of the project and beyond (Long term and Open access).

Application Acronym	Regional Community Dataset	Level of preservation	Level of user access	Data Type/Format
WRF-ARW	Daily model output(raw output data as well as post-processed output)	Long term	Restricted	Simulation / NetCDF, Gif
VINE	Observation dataset on dust particles in ambient air available from Georgian National Environmental Agency	Long term	Open	Observational , Simulation data / NetCDF, Grib
RCM MENA-CORDEX	Gridded datasets of temperature and rainfall for the MENA, via the CORDEX data portals	Long term	Restricted	Simulation / NetCDF
HIRECLIMS	ROCADA (Romanian Climatic Dataset)	Long term	Restricted	Simulation / NetCDF
WRF-Chem (NOA)	WRF-Chem dust aerosol concentrations and various meteorological parameters	Long term	Restricted	Simulation / NetCDF
DREAMCLIMATE	Downscaled atmospheric-dust DREAM covering wide North Africa, Southern Europe and Middle East regions.	Long term	Open	Simulation / NetCDF, Grib
DRS-ACS	Characteristic constants describing the kinetics of atmospherically relevant processes and spectroscopic properties of the involved species.	Long term	Open	Simulation
OpenFOAM	Recorded meteorological parameters	Long term	Restricted	ASCII
Continuous_LTS	A data set of daily continuous land surface temperature at 1 km resolution	Long term	Open	Simulation/GeoTIFF

Table 6. Data set identifiers, levels of preservation and access levels for the climate community

Application Acronym	Regional Community Dataset	Level of preservation	Level of access	Data Type/Format
CSAD	Hundreds of RTi files of Ptolemaic inscriptions	Long term	Closed	Image / RTi files; txt
3DINV	Existing and new datasets of geoelectrical tomographic data collected from field campaigns.	Short term	Restricted	Zip file that includes the input parameter file (*.in) and the corresponding survey data file (*.d)
mGeoAI	A demo dataset will be used for demonstration purposes share with the VI-SEEM community for research purposes.	Short term	Closed	Image / JPG, TIF, GeoTIFF, PNG, txt
AutoGR	A demo dataset will be used for demonstration purposes share with the VI-SEEM community for research purposes.	Short term	Closed	Image / JPG, TIF, GeoTIFF, PNG, txt
3DGEOEX	Existing and new datasets of geoelectrical tomographic data collected from field campaigns.	Short term	Closed	Zip file that includes the input parameter file (*.in) and the corresponding survey data file (*.d)
ELKA	Word, .pdf files of Karamanlidika texts	Long term	Open	Text / Word, pdf
Manuscript	Datasets of digitized handwritten documents in Arabic or Hebrew	Long term	Restricted	Text / image / pdf
DCH	An Internet digitized catalog of “Aharoni Collection” providing access to 3D models of specimens for research and dissemination purposes.	Long term	Open	Image / 3D models, txt
PETRA	MEGA Jordan GPS/ geo-referenced data	Long term	Restricted	Image / txt; .r
BVL	Banatica database: 1000 books and 200 digitized books	Long term	Open	JPEG, PDF, txt
VirMuf	3D models, CH Multimedia content and geo-spatial data	Long term	Restricted	Image; video (.mov; .avi); 3D models; Unity scenes and files; .fbx; .txt
CHERE	3D model, Gigapixel image and Panoramic image libraries and frameworks for general purpose/application.	Long term	Open	Image / JPEG, TIFF; .txt
CH-CBIR	Aerial images with annotated land cover types. Learned image representations for imagery in the areas of cultural heritage and remote sensing	Long term	Open	Image / JPEG, TIFF; .txt
DataCrowds	Crowd simulation models that can be applied to different scenarios, such as evacuations, urban planning,	Long term	Open	Tracked trajectories of crowds in comma separated files

	virtual reality, entertainment, robotics, etc. Crowd data with videos and tracked trajectories with instructions on their formats.			(.csv) and videos of crowds (.mov; .avi); .txt
HaPPEn	Testing and implementing a set of tools to be used for image based 3D reconstruction. Assessment on performance and usability of the in the Digital Cultural Heritagecommunity will be realized, and benchmarks will be provided to the community.	Short term	Restricted / public reports	Dense point clouds / 3D models / images / .txt
MC4CH	Digital material for archaeological monuments	Long term	Open	Image / CIDOC-CRM

Table 7. Data set identifiers, levels of preservation and access level for the digital culture heritage community

Application Acronym	Regional Community Dataset	Level of preservation	Level of access	Data Type/Format
MD-Sim	MD trajectories of oncogenic proteins with mutations relevant to the SEEM area	Medium term	Open	Simulation
DICOM Network	Generalized statistical datasets. Patient dataset available after special permission or relevant anonymization of data.	Long term	Restricted	Image / XML, JSON
CNCADD	Produce and share parameter sets relevant to the community	Long term	Open	Simulation
PSOMI	Datasets with molecule synthesis results.	Medium term	Open	Experimental / PDB, GRO, NAMD, PSF, PDF
SQP-IRS	Biological dataset. Computational vs. Experimental database for proteins secondary structures. Crystallographic vs. Spectroscopic database for selected targeted proteins	Medium term	Restricted	Simulation and experimental
THERMOGENOME	Datasets with data for thermodynamic stability of RNA/DNA and DNA/DNA duplexes for all transcripts, exons, introns, 5-UTRs, 3-UTRs for Homo sapiens (human), A. thaliana, C. elegans, D. melanogaster, D. rerio.	Long term	Restricted	Experimental
MDSMS	The molecular dynamics simulation of mixed systems	Medium term	Open	Simulation / PDB, TRR, XTC, DCD

Table 8. Data set identifiers, levels of preservation and access levels for the life sciences community

3. Conclusions

This document constitutes a high-level Data Management Plan for the VI-SEEM project that defines the governance of all data sets to be provided in the scope of the project. The data sets include those already identified and presented in Table 1, as well as the new ones to be created in the duration of the project. The data to be collected is described in detail, and policies regarding the collection, quality control and assurance, access and sharing, storage and preservation, privacy and security and ethics and legal issues of the data are outlined.

This Data Management Plan serves to guide the use of data in VI-SEEM, an imperative role considering the amount of data planned to be collected and generated throughout the project. Following the Data Management Plan guidelines, the VI-SEEM community benefits by taking advantage of quality assurance and control, search ability, and increased usability of the data. In addition, the Intellectual Property Rights and conformance to legal and ethics issues are ensured.

The VI-SEEM technical board bears the overall responsibility for providing support and guidance for the implementation of the policies outlined in the Data Management Plan. As data is collected, the WP5 leader will update the Data Management Plan if deemed necessary. Users of the VI-SEEM platform share the responsibility to follow the guidelines as outlined in this document. The individual publishers of the data in the VI-SEEM VRE have the responsibility for the data they are publishing.

Appendix 1

Here we provide the survey which has been distributed to the scientific contacts for each application developed during the integration phases. These information enabled us to create the Data Management Plan.

1. Institution Name:…………………………………………………………………………………………….
   (*acronym of the Institution, as in DoW*)
2. Data Storage Service:……………………………………………………………………………………
   (*the name and the web address of your collection of data*)
3. Description of Data:…………………………………………………………………………………………….
   (*the description of data as in DoW*)
4. Metadata model/schema/format:……………………………………………………………………
   (*metadata model, schema or format currently used to describe your collection. Please provide separately, as an attached file, an example of your metadata model with the definition of each field/tag*)
5. Discipline/Area of Research:…………………………………………………………………………….
   (*Discipline/area of research that the metadata belong to*)
6. Amount of Data:…………………………………………………………………………………………………….
   (*the amount of content will be aggregated*)
7. Amount of digital objects:………………………………………………………………………………….
   (*the amount of digital objects linked to the metadata aggregated*)
8. Persistent IDs type:….……………………………………………………………………………………….
   (*specify if you use or require to use any persistent Identifiers for your collections, and detail it*)
9. Controlled vocabulary:…………………………………………………………………………………….
   (*which are (if any) the controlled vocabularies used in each of your collections*)
10. Protocol:………
    (the protocol(s) (if known) currently used in order to export metadata: e.g. OAI-PMH, FTP protocols)
11. Language:………
    (*the language(s) of your metadata*)
12. Object types:……………………………………………………………………………………………………
    (*the object types: image, text, sound, video (e.g. jpg, pdf, 3D, etc.).*)
13. Rights:………
    (*details on the Right status of the collection. E.g. Creative Commons, Public Domain, Rights Reserved, etc.*)
14. Comments:……….
    (*further comments*)
15. Primary Contact:………………………………………………………………………………………………….
    (*the person responsible for the collection at institutional level (name and e-mail address)*)
16. Technical Contact:………………………………………………………………………………………….
    (*the person responsible for the content aggregation process within the project, if different from the primary contact (name and e-mail address).*)

Appendix 2

Metadata Glossary: Here we provide a short description of the different Metadata schemas and formats used in the scientific communities of Climate, Life and Digital Cultural Heritage.

Schema/Format	Description
NetCDF CF	NetCDF Climate and Forecast (CF) metadata conventions is a metadata schema designed in such a way to promote the processing and sharing of files created with the NetCDF API. More information can be found in the following link: http://cfconventions.org/
Dublin Core	The Dublin Core Schema is a small set of vocabulary terms that can be used to describe web resources, as well as physical resources such as books or CDs, and objects like artworks. More information can be found in the following link: http://dublincore.org/
Proprietary OpenFOAM	Proprietary OpenFOAM metadata convention is a metadata schema build to promote the processing and sharing of datasets created with the OpenFOAM software. More information can be found in the following link: https://www.openfoam.com/
GeoTIFF	GeoTIFF is a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file. The potential additional information includes map projection, coordinate systems, ellipsoids, datums, and everything else necessary to establish the exact spatial reference for the file. More Information can be found at: https://trac.osgeo.org/geotiff
GRIB	GRIB metadata schema is used to promote the processing and sharing of files created with the COSMO model system http://www.cosmo-model.org/content/model/documentation/grib/grib_gribapi.htm
.SDF format	Structured Data Format (SDF) is a simple metadata schema used extensively in Life Sceinces. More information can be found in: http://dataprotocols.readthedocs.io/en/latest/simple-data-format.html
.DCD / .XTC	DCD and XTC metadata formats are simple metadata schemes which can be easily converted from one to another and they are used occasionally in the Life Sciences.
CDISC / ODM	CDISC ODM-XML is a vendor-neutral, platform-independent format for exchanging and archiving clinical and translational

	research data, along with their associated metadata. More information can be found at: https://www.cdisc.org/standards/transport/odm
Gro / pdb format	.gro and .pdb formats support the metadata which are suitable for the GROMACS and PYMOL software respectively. One can convert files from one format to the other by simply executing pdb2gmx/gmx2pdb. More information can be found at: http://manual.gromacs.org/current/online/gro.html and https://www.cgl.ucsf.edu/chimera/docs /UsersGuide/tutorials/framepdbintro.html
DICOM	A metadata schema used to promote the processing and sharing of files created within the scope of the DICOM network. More information can be found at: https://en.wikipedia.org/wiki/DICOM
UNIMARC / MARC 21	UNIMARC and MARC21 metadata standards are commonly used in Digital Cultural Heritage and can be easily converted from one to the other. More information can be found at: https://www.loc.gov/marc/bibliographic/