everest

Virtual Research Environment
Data Management Plan

Workpackage	**WP6**	**VRE deployment, maintenance and operations**	
Task	6.3	VRE population	
Author (s)	Rosemarie Leone		ESA
	Fulvio Marelli		T2
Reviewer (s)	Cristiano Silvagni		ESA
Approver (s)	Ugo Di Giammatteo		ACS
Authorizer	Mirko Albani		ESA
Document Identifier	EVER-EST DEL WP6-D6.4		
Dissemination Level	Public		
Status	Approved by the EC		
Version	1.1		
Date of Issue	14 December 2018		

Document Log

Date	Author	Changes	Version	Status
15/07/2016	Rosemarie Leone		v0.1	Draft to discussed with VRCs
01/09/2016	Rosemarie Leone	Include comments from VRCs	v0.2	Draft to be circulated for comment
30/09/2016	Rosemarie Leone/Fulvio Marelli	Revision of scope discussed during Edinburgh Plenary meeting	v0.3	Draft to be circulated to WP Leaders
14/10/2016	Rosemarie Leone/ Fulvio Marelli	Comments addressed by WP Leaders implemented	v0.4	Draft to be circulate to VRCs for review
14/10/2016	Rosemarie Leone/ Fulvio Marelli	Comments from VRCs implemented	v0.5	Draft to be circulated to formal reviewer
05/11/2016	Rosemarie Leone/ Fulvio Marelli	Comments from formal reviewer implemented	v1.0	Draft to be approved by the EC
14/12/2018	Cristiano Silvagni	Document status	v1.1	Approved by the EC

Table of Contents

Abstract

This document presents the EVER-EST project Data Management Plan (DMP). This is the first version of the DMP and will be updated over the course of the project whenever significant changes arise as ruled by AD[2]. The document provides a description of the data summary, the scope of the DMP and the details of the management plan for Findable, Accessible, Interoperable and Reusable (FAIR) Data, as defined at this stage of the EVER-EST project.

Definitions and Acronyms

Acronym	Description
CFSP	Common Foreign and Security Policy
DHA	Daily Hazard Assessment
DMP	Data Management Plan
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable and Reusable
GEO	Group on Earth Observations
GSNL	Geohazard Supersites and Natural Laboratories
HIM	Hazard Impact Model
IPR	Intellectual Property Rights
MSFD	Marine Strategy Framework Directive (MSFD).
NHP	Natural Hazards Partnership
RO	Research Objects
VRC	Virtual Research Community
VRE	Virtual Research Environment
WF	Work Flow

Applicable Documents

Document ID	Document Title
1	Amendment to Grant Agreement 674907
2	Guidelines on Findable, Accessible, Interoperable and Reusable (FAIR) Data Management in Horizon 2020
3	D3.1. Deliverable: EVER-EST VRE Detailed Definition of Use Cases 1.1
4	D4.1. Deliverable: EVER-EST VRE Research Objects in Earth Science version 1.1
5	D4.2. Deliverable: Research Objects Models and Support Technology in Earth Sciences Version 1.0
6	D5.1 EVER-EST VRE Infrastructure and Services Design version 1.1

1 EVER-EST Data Management Plan

1.1 Document scope

This document presents the EVER-EST project Data Management Plan, describing how EVER-EST Virtual Research Communities data is made Findable, Accessible, Interoperable and Reusable (FAIR), taking into account the VRCs specific requirements in relation to openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security and long term preservation needs.

The EVER-EST DMP has been added to list of the EVER-EST deliverables following the amendment of the Grant Agreement that was signed on the 16th of August 2016 to provide a sound data management plan as this is an essential part of the Earth Science research life cycle and best practices.

1.2 Document structure

The overall structure of this document is based on the guidelines on FAIR Data Management in Horizon 2020 (Version 3.0 dated 26 July 2016) and the new DMP template included in the latter guideline. The DMP will be updated over the course of the project whenever significant changes arise (as foreseen by the guidelines and in line with the periodic evaluation/assessment reviews of the EVER-EST project).

2 Data Management Plan Components

2.1 Data management plan data summary

EVER-EST will provide earth scientists with the means to seamlessly manage both the data involved in their computationally intensive disciplines and the scientific methods applied in their observations and modelling, which lead to the specific results that need to be attributable, validated and shared within the community e.g. in the form of scholarly communications. Such data management capabilities will be augmented with the models, techniques and tools necessary for the preservation of scientific methods and their implementation in computational forms such as scientific workflows, which are increasingly used in the Earth Science domain. The scientific community involves multi-disciplinary scientists in all Earth Science disciplines and policy impact areas. Policy makers are responsible for defining the main Earth health indicators, disaster risk management actions and investments.

For each of the EVER-EST communities the relevant data summary will be provided in the following subchapter highlighting:

- The purpose of the data collection/generation for the specific community;
- Types and formats of data that are generated and/or collected by the community;
- Existing data re-use and how the data are re-used;
- Origin of the data;
- Expected size of the data;
- The 'data utility'

A detailed description of the pre-selected communities input data need and generated out put data is provided by AD[3].

Sea Monitoring Virtual Research Community

The Sea Monitoring VRC focuses on finding new ways to measure the quality of the maritime environment and it is quite wide and heterogeneous, consisting of multi-disciplinary scientists such as biologists, geologists, oceanographers and GIS experts, as well as agencies and authorities (e.g. ARPA or the Italian Ministry of Environment). The scientific community has the main role of assessing the best criteria and indicators for defining the Good Environmental Status descriptors defined by the Marine Strategy Framework Directive (MSFD). The indicator derivation process includes the following:

- **Datasets:** Raster data for seafloor bathymetry, backscatter and hydrodynamic models, vector data for the coral occurrences; jellyfisch occurences from citizen science monitoring (video, photo, reports, social media information); Mediterranean sea physical and biogeochemical variables from satellite data platform (Copernicus, http://marine.copernicus.eu/services-portfolio/access-to-products/, aviso, http://www.aviso.altimetry.fr/en/data/products.html, ecc.) Posidonia meadows habitat mapping.
- **Software:** ArcGIS tool for deriving environmental variables and geospatial/ statistical analysis, .xls matrix, R, Maxent.
- **Documents:** Published abstract, PPT presentation of the models, MSFD document on habitat extent and invasive species distribution, previous paper on habitat suitability models.

GeoHazards Supersites Virtual Research Community

The Geohazard Supersites and Natural Laboratories (GSNL) is a collaborative initiative supported by GEO (Group on Earth Observations) within the Disasters Resilience Benefit Area. The goal of GSNL is to facilitate a global

collaboration between Geohazard monitoring agencies, satellite data providers and the Geohazard scientific community to improve scientific understanding of the processes causing geological disasters and better estimate geological hazards. The Geohazards presently addressed in the GSNL initiative are all hazards linked to earthquakes and volcanic eruptions (e.g. seismic shaking, ground deformation, seismically triggered landslides, ash fall, pyroclastic flow, lava flow). The monitoring of these hazards is done via Permanent Supersites, which deal with prevention activities (i.e. science to support seismic and volcanic hazard assessment), and Event Supersites, which have a limited duration and are dedicated to intensive scientific research on specific eruptions or earthquakes. In EVER-EST, the activity of the Geohazard VRC is focused on Permanent volcanic Supersites (Mount Etna, Islandic volcanoes, Campi Flegrei/Vesuvio). The main activities of this VRC need the following resources:

- **Datasets:** geophysical parameters describing seismic and volcanic processes and phenomena (e.g. ground displacement and velocity, gas composition, atmospheric water content, ash particle density, etc.), SAR and optical satellite data (e.g. Sentinel1 & 2, COSMO-SkyMed , TerraSAR X, Radarsat 2, ALOS 2, MODIS, MSG, Pleiades, etc.), GPS data.
- **Software/Models:** Scientific modeling codes used to simulate the effects of the phenomena and processes. They are used to generate space/time representations of geophysical phenomena (e.g. measures of surface deformation, models of ash dispersal, models of the magmatic reservoir). Commercial image analysis software for SAR and optical data (SARSCAPE). Commercial software for data analysis (Matlab, ENVI/IDL, Fortran, Python, etc.). Geographic Information System software (ArcGis).
- **Documents:** Publications on journals or conference proceedings, validation reports, reports on research results, Research Objects including scientific results, workflows, bibliography, topical discussions, etc.

Land Monitoring Virtual Research Community

The European Union Satellite Centre (SatCen) represents, in the framework of EVER-EST and in line with the Secure Societies Horizon 2020 Societal Challenge, the stakeholders involved in the decision-making process of the EU in the field of the Common Foreign and Security Policy (CFSP).

Land Monitoring is key in providing useful information to those entities that have to:

- Make informed decisions referred to the monitoring of urban, build-up and natural environments;
- Identify certain features and anomalies or changes over areas of interest as well as of natural resources;
- Monitor features/changes condition and exploitation to address related environmental, scientific, humanitarian, health, political and security issues as well as to adopt sustainable management practices.

Thus the Land Monitoring community can be described as composed by institutional and operational entities as well as by scientific and research entities, potentially having different final goals but using the same space assets and similar services/techniques.

The Land Monitoring VRC data generation process includes:

- **Datasets**: Satellite images (e.g. Sentinel 1 and other data from the Copernicus programme and third party missions), other geotagged data (structured and unstructured) coming from social, commercial, open and other sources (e.g. social media information and newsfeed);
- **Software**: Data ingestion tools (from catalogues as the ESA Sentinels Scientific Hub); pre-processing and processing tools (e.g. calibration, co-registration, change detection) from open software (e.g. SNAP), open libraries (e.g. GDAL) and custom developed algorithms (mainly written in Java); these tools might be readapted to be used in the frame of EVER-EST project;
- **Documents**: Documentation on the data (e.g. Sentinels' guidebooks or data provenance) and the (pre-) processing algorithms ingested (e.g. reference papers) as well as validation procedures and reports (e.g. description of possible methods to validate the whole processing chain).

Natural Hazards Virtual Research Community

The Natural Hazards Partnership (NHP) is a group of 17 collaborating public sector organisations comprising government departments, agencies and research organisations. The NHP provides a mechanism for providing co-ordinated advice to government and those agencies responsible for civil contingency and emergency response during natural hazard events.

The NHP provides daily assessments of hazard status via the Daily Hazard Assessment (DHA) to the UK responder and resilience communities, pre-prepared science notes providing descriptions of all relevant UK hazards and input to the National Risk Assessment. In addition, the NHP has set up a Hazard Impact Model (HIM) group tasked with modelling the impact of a range of UK hazards within a common framework and operational delivery of the model outputs. Initially they are concentrating on modelling the impact of 3 key hazards – surface water flooding, land instability and high winds – on people, their communities and key assets such as road, rail and utility networks. The partners share scientific expertise, data and knowledge on hydrological modelling, meteorology, engineering geology, GIS and data delivery and modelling of socio-economic impacts.

The HIM data generation process includes:

- **Dataset:** Impact Library, a repository of pre-calculated impact data for each HIM, a surface water flooding hazard footprint generated, using the G2G modelling process, in ASCII grid format, county level reporting areas generated by the flood forecasting centre in ESRI shapefiles;
- **Software/Methods:** R and Python scripting languages used for modelling impacts of hazards based on hazard footprint data and the impact library; ArcGIS geoprocessing tools for generation of polygonised impact outputs.
- **Documentation:** impact results that require summary and presentation to end users, including an interpretation of the risk when forecast data used in initial stages of the modelling. Guidelines on running hazard impact modelling scenarios and schematic descriptions of the hazard impact modelling workflows. Hazard Impact Framework report enabling standards across different hazard scenarios. Related conference presentations, papers and proceedings as well as peer review papers authored by NHP partners and their individual institutions.

2.2 Data management plan scope

Earth Science communities using EVER-EST infrastructure during the research life cycle generate scientific peer-reviewed publications for which open access obligation in Horizon 2020 apply. The underlying research data and products within the scope of this data management plan are heterogeneous as summarized in the previous chapter and can be grouped in:

- Research data collected or processed/generated as part of the VRC research life cycle, intermediate products, as preliminarily identified in [AD4] and summarized in chapter 2.1;
- Research objects.

3 Findable, Accessible, Interoperable and Reusable (FAIR) Data

The research object concepts, technologies and methodologies enable the vision for 'FAIR' Findable, Accessible, Interoperable and Re-usable data management practices while supporting VRCs specific requirements in relation to both openness and protection of scientific information, commercialisation and IPR, privacy concerns, security and long term preservation needs.

The research object paradigm, life cycle model and technology support FAIR data management recommendations related to sharing documentation/communication of scientific knowledge as well the reproducibility of scientific results including:

- Documenting best practices (WFs, analysis methods, monitoring methods, etc.).
- Providing long term preservation of scientific knowledge (how data are analysed, how results are validated, etc.)
- Providing long term preservation of end-user stories (demonstrating scientist-end-user interactions), also for public dissemination.
- Executing of "standard" workflows for data analysis/modeling in order to validate results and generate "standard" products (e.g. deformation maps) as mass products.
- Testing algorithms and data, either modifying the workflow to execute new analysis methods/models on the same dataset, or executing the original workflow on different datasets;
- Supporting long term data series and historical science based on past observations and the validation of models with actual data

Research Objects for EVER-EST VRC can encapsulate the following data/product information.

- **Workflows:** High level flowchart and formal workflow descriptors (e.g. Taverna bundles). Also, metadata such as text files describing the general workflow, including all information needed by scientists to choose this workflow for other use cases (assumptions, usage issues, etc.)
- **Documentation:** ranging from scientific papers, bibliography, user manuals to impact results, report, etc.
- **Data:** Input data (for processing and for validation), output data (intermediate non-validated and final validated) and a report on use case data and results.
- **Processing components:** Software, web services, configuration setup, hardware requirements.
- **Products**: results obtained using workflow-centric RO or external processing tools. These results may be preliminary or not yet published, but need to be encapsulated in RO for scientific purposes or for risk management purposes. Usually correlated by explicative text files.

At this stage of the project, for the scope of this data management plan the following RO types as described in [AD4 and AD5] have been identified:

- **Workflow-Centric RO**: contain a workflow, whether a Taverna WF bundle or just an executable code and/or a Fortran, Matlab, etc. source code, executable not only on the VRE.
- **Data-Centric RO**: contains reference to a dataset or observation (normally many of them). Depending on the scope it may be static or be a live RO to which further data are added periodically.
- **Research Product Centric RO**: It contains the (normally validated) results of one or more processing runs (e.g. a workflow for source modeling). It could contain instead the result of qualitative interpretations (e.g. a map of geomorphological features).

In addition, the following RO type, not under the scope of this DMP has been identified:

- Documentation and bibliographic Research Objects.

3.1 Making data findable, including provisions for metadata

EVER-EST includes activities aiming on definition and harmonization of metadata for the VRE as part of the RO model definition. The detailed description of these activities can be found in [AD4, AD5]. This work is intended for harmonization, in the course of the project, of the data and research object produced using the VRE and the VRCs communities have already started more and more to benefit of the internal training-by-doing, generating and using ROs. VRCs taking part of the project might have their own community-specific metadata schemes. However, the overall aim of the EVER-EST data management policy at the start of the project was to encourage the use of the latter schemes and documentation methods, meanwhile progressing on the harmonization of the metadata and ontologies taking into account the specific needs of the VRCs. Use of suitable international standards (e.g. INSPIRE directive, RDA Metadata standard directory, metadata standard for long term data preservation) have been assessed. Data produced and used during the project will be identifiable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers), Registry of Research Data Repositories and repositories like Zenodo, OpenAIRE and CERN are currently under assessment.

3.2 Making data openly accessible

All EVER-EST project results that are open to use for any purpose will be appropriately licensed using open licensing policy (e.g. Creative Commons 4.0BY or similar). Unless required by the Consortium Agreement or VRC specific IPRs, all EVER-EST data products openly accessible will be discoverable (i.e. via metadata harvesting access) in reasonable time after data collection and/or generation. Default time for this is 6 (six) months from the end of the result generation. It is to be noted that the data provided by the VRCs may not be fully open, depending on the specific license and conditions of use of the input data. This may apply for instance to some satellite data (e.g. COSMO-SkyMed or Radarsat 2) or to some in situ datasets.

For many datasets produced, the storage and access management will be implemented using the Research Objects environment and, EVER-EST VRE and VRC repositories, if applicable. Access will be provided to the Commission officials and their appointed reviewers. Access to IPR sensitive data will be adequately controlled. The detailed description of data access infrastructure, data set and RO catalogues is provided by [AD4, AD5, AD6].

3.3 Making data interoperable

As part of the project objectives, work is on going to assess the interoperability of VRCs research data and research object. Metadata vocabularies, standards and commonly used ontologies are being assessed to facilitate inter-disciplinary cross-fertilization of results. At this stage of the project, the research object model has been updated and extended as follows:

- Included the required vocabulary terms for describing geographic and time information, data access policies and intellectual property.
- Updated and aligned the research object core ontology and required extensions with the latest model of the Annotation Ontology, called Open Annotation Ontology (and since July 2016 Web Annotation Ontology, W3C Candidate Recommendation).
- Cleaned and properly annotated all the ontologies with provenance and metadata information.
- Adaptation and integration of existing Earth Observation metadata specifications.

3.4 Increase data re-use

Each VRC is currently assessing how to license data to permit the widest reuse possible and clearly identify any requirements for data embargo and length of time for which the data will remain usable if applicable. Data quality assurance processes are being implemented within the Research Object embedded checklist.

3.5 Allocation of resources, long term data preservation

Each community is responsible for the VRCs specific data storage requirements. The EVER-EST project will provide services for data set storage sharing and backup as described in [AD6]. Data selected for long term preservation will be included in the VRC specific long-term preservation requirements. In the data preservation decision the following aspects will be considered: 1) Re-usability of the data (including metadata), 2) needed resources for long term storage (size, access), 3) expected storage period, 4) possibility of external data storage using non- project related repositories. Data set storage, curation and maintenance costs during the project life time are valid EVER-EST costs. The long term resources needed for long term preservation and storage will be considered in the sustainability plan. To be noted that the adoption of the research object paradigm includes additional metadata in the form of checklists that monitor and diagnose potential decay derived e.g. from issues with the availability or accessibility of the data due to platform downtime or data format changes, either as a fork at the VRE or as a reference to the original dataset at the side of the data provider.

3.6 Data security

Data recovery, secure storage and transfer of sensitive data are being addressed at architectural design level [AD11] and will be described in detail in the next release of the plan. Basic access control to the content of the research object, particularly by third parties accessing the research object is currently under implementation.

3.7 Ethical aspects

As stated in the Grant Agreement, data sets collected or generated in EVER-EST do not have ethic aspects concerns.

3.8 Other: Licensing and IPR

Ownership of the data and results produced throughout the project activities is defined in the Consortium Agreement and by the VRCs specific IPRs regulations. The following requirements on functionalities related both to the research object paradigm and impacting in EVER-EST architecture design, are under implementation:

- Citation and attribution: sharing of data and methods, particularly at a point in time before an actual paper is published by a team of scientists to assure that data and methods are fully referentially, e.g. as a research object with its own DOI.
- Licensing mechanisms: allow scientists to define the terms in which their research objects can be used. This would allow creating confidence on the research object and establishing etiquette for acknowledgement that would support the previous point.