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

Data Management Plan

Tamás Kazinczy
NIIF

VI-SEEM Regional Climate Training
11-13 Oct 2017

The VI-SEEM project initiative is co-funded by the European Commission under the H2020 Research Infrastructures contract

Overview

- Data Management Plan (DMP) in general
- VI-SEEM DMP

DMP in general [1]

- Why do we need one?
- What is the aim of DMP?
- What else is required by EC?

DMP in general [2]

- All EU projects are required to have one
- DMP tells all the details about data in a project
  - **What** to collect / generate
  - **How** to handle / exploit / publish / curate / preserve
  - **Where** to store
  - **Which** formats / standards to use
  - and many more...
- DMP is a living document
  - updated as needed
- EC requirements: see later

Current structure of VI-SEEM DMP

- Data description
- Data collection and documentation
- FAIR Data
- Data set identifiers, level of access and preservation

Data description

- Levels of data (based on purpose)
- Possible users
- Levels of access for users
- General Metadata
- SC specific data
- Cross-disciplinary data

Levels of data by purpose

- Scientific data (A)
  - experimental, observational, computational
  - Corresponding SWs and workflows
  - Documentation on how to access data and results
- Publications and other materials (B)
  - Conference papers
  - Additional documentation that puts results in context
- Simplified data formats for immediate reuse (C)
  - Visualization for educational/outreach purposes
  - Theory interpretations

Possible users for those levels

- VI-SEEM collaborators, SC members
  - Would benefit from all categories (A, B, C)
- SCs worldwide
  - Would benefit from all categories (A, B, C)
- Public at large
  - Would benefit from publications and simplified data formats (B, C)

Levels of access for users

- Non-registered users
  - Publications
  - Outreach material

  (whatever is available)
- Registered users
  - Browse or search datasets
  - Get access to data
- VI-SEEM Contributors
  - Upload data sets
  - Get access to VI-SEEM only data

General Metadata [1]

- Metadata: data about your data
- Concept: identify common elements
- Currently only three of them are mandatory
  - Title
  - Identifier
  - Primary Contact

  (this may not be the case for all services)

General Metadata [2]

- Current elements:
  - Title
  - Identifier
  - Primary Contact
  - Description
  - Keywords
  - Creator
  - Community
  - Discipline

General Metadata [3]

- Current elements (continued):
  - Publisher
  - Public Since
  - Language
  - Format
  - Temporal Coverage
  - Spatial Coverage
  - Technical Contact

SC specific data example: Climate [1]

- Model simulation data
  - metadata is important → **reproducibility**
  - Also needed:
    - input parameters
    - source code
  - Time matters → **version control** is also needed
- Observational data
  - diverse sources
  - scattered
  - most of the data are **freely accessible** though

SC specific data example: Climate [2]

- Metadata
  - follows Climate and Forecast (CF) Metadata Conventions
    - self-describing metadata
    - easily readable
    - simple
  - Climate metadata in VI-SEEM
    - on the way of adopting CF 1.7 specifications

Cross-disciplinary data

- Bringing different SCs together; this may
  - be a **mutual interest** in some data
  - be a basis for **new research efforts**
  - help **data reuse**

    (e.g. output of a climate research could serve as a parameter for a life science research)

- Identification is in progress

- DMP is to be updated accordingly

Data collection and documentation

- Data quality control and assurance
- Quality Assurance / Quality Check officers
- Requirements for Data Contributors

Data quality control and assurance [1]

- QA is about processes
- QC is about checking for errors
- E.g. as part of our QA process
  - Dataset providers are required to fill in a form before uploading datasets
- There is a team of QA/QC officers in VI-SEEM
  - They will assure that QC was performed for datasets

Data quality control and assurance [2]

- If an issue arise for a specific QC
  - Case escalated to SC leader
- Method of selecting data for initial / later checks will be defined by SC leaders
- SC leaders approve the form and the QC done
  - It is then when data becomes available
- Periodic checks for QC processes and data selection method will be done by WP5 leader and will work together with SC leaders if issues arise

QA / QC officers [1]

Country	QA/QC Officer	Contact details
Armenia	Artashes Mirzoyan	amirzoyan@sci.am
Bosnia and Herzegovina	Vladimir Risojevic	vladimir.risojevic@etf.unibl.org
Bulgaria	Vladimir Dimitrov	vgd@acad.bg
Cyprus	Panayiotis Charalambous	ps.charalambous@cyi.ac.cy
Egypt	Mohammed Elfarargy	Mohammed.Elfarargy@bibalex.org
FYR of Macedonia	Sonja Filiposka	sonja.filiposka@finki.ukim.mk
Georgia	Ramaz Kvatadze	ramaz@grena.ge
Greece	Ioannis Liabotis	iliaboti@admin.grnet.gr

QA / QC officers [2]

Country	QA/QC Officer	Contact details
Hungary	Lajos Bálint	lajos.balint@niif.hu
Israel	Zivan Yoash	zivany@mail.iucc.ac.il
Jordan	Mostafa Zoubi	mostafa.zoubi@sesame.org.jo
Moldova	Mihail Matenco	mihail.matenco@renam.md
Montenegro	Lidija Milosavljevic	lidija@ac.me
Romania	Marian Neagul	marian.neagul@e-uvt.ro
Serbia	Petar Jovanovic	petarj@ipb.ac.rs

Requirements for Data Contributors

- Well documented **process**
  to be used to perform **quality check**
  on the **dataset**

- First QC is to be presented to QA/QC officer in charge
  - Later checks might not require that

- There are some more (discussed later)

FAIR Data

- What is the FAIR principle?

  VI-SEEM **Data Services** are aimed at **providing** its **capabilities** so that the various **datasets** of the different communities are:

  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable

Data access and sharing

- Findability
  - VI-SEEM PID Service
  - VI-SEEM Data Discovery Service
  - Metadata should be searchable at all times
- Access to data
  - Open / restricted / closed
  - “as open as possible and as close as necessary”
  - Reasons for restriction are to be well justified
- Sensible data is a special matter (see later)

Data storage and preservation

- Levels of preservation
  - Short term (< 6 months)
  - Medium term

    (~1 year, within duration of a computational project)
  - Long term (within lifetime of VI-SEEM)
- VI-SEEM Data Services for the levels above
  - Simple Storage (short)
  - Repository, Archival (medium and long)
- Metadata - standards, common schema and mapping
  - Moving target

Security, privacy and ethical aspects

- Handling personal data requires special attention
- GDPR (General Data Protection Regulation) goes live next May
- Additional requirements for Data Contributors
  - Best would be if datasets already comply with the regulation at time of ingestion
  - This means e.g.
    - Data minimization
    - Pseudonymization
  - Else a preprocessing step is required (implementation **and** documentation)

Data set identifiers, level of access and preservation [1]

- Based on survey made by WP5
- These informations are placed into the DMP as well
- Describes
  - Application
  - Regional community dataset
  - Level of preservation
  - Level of user access
  - Data type / Format

Data set identifiers, level of access and preservation [2]

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
ENB-RCM	Rainfall records at rain gauge stations in the Eastern Nile Basin	Medium term	Restricted	Observational, Simulation / NetCDF

End

Thank you!