# TCS AH ENODE

## HORIZON 2020 DMP

### 1. DATA SUMMARY

*State the purpose of the data collection/generation*

Anthropogenic hazard episodes/data are stored in two eNode's located in France and Poland. Each eNode stores various data types related to anthropogenic hazard research purposes. Within EPOS-IP project data are organized in exceptional datasets, called 'episodes', which comprehensively describe a geophysical process; induced or triggered by human technological activity, posing hazard for populations, infrastructure and the environment. At least 20 episodes from EU, UK, USA and Vietnam should be implemented within the project.

*Explain the relation to the objectives of the project*

eNode's provide access to advanced Research Infrastructures to the Anthropogenic Hazards Community. This objective is implemented as a web platform which offers access to various datasets related to selected anthropogenic seismicity cases. The relevant seismic and non-seismic data are gathered in the so-called episodes of anthropogenic seismicity. From the architecture point of view IS-EPOS is built of two major components: the AH local data centres (Polish eNode – CIBIS and French eNode – CDGP), gathering episodes and their multidisciplinary data and the IT platform binding together access to the episodes' data, software, applications and documents. The eNode's integrates currently episodes linked to:

- ✓ Underground hard rock and coal mining
- ✓ Salt solution mining
- ✓ Hydro energy production
- ✓ Geothermal energy production
- ✓ Conventional and unconventional hydrocarbon production
- ✓ Underground gas storage
- ✓ Wastewater injection

*Specify the types and formats of data generated/collected*

The data formats are based on EPOS guidelines. Here, are the current data formats supported by EPOS-IP WP14:

| Data category | Standard format |
|---|---|
| Seismic / ground motion catalogue | Mat, QuakeML |
| Seismic signals (seismogram / accelerogram) | miniSEED / SEED |
| Seismic / ground motion network | SeisComp inventory xml |
| Water quality, air quality, industrial data and geodata* | GDF (mat), xlsx, csv, txt |
| Other geodata | geotiff / shapefile |
| Satellite data | proposed: GDF (mat), shapefile, geotiff |
| Documents | pdf / graphic and video formats / presentation formats / other formats |

Seismic catalogues in QuakeML are widely spread exchange standard in the seismological community. It is recommended by EPOS-IP seismological TCS.
The seismic signals and waveform format is maintained by the International Federation of Digital Seismograph Networks and documented in the SEED Manual.
Seismic and ground motion networks are saved in standard SeisComP inventory xml files. SeisComP is a seismological software for data acquisition, processing, distribution and interactive analysis that has been developed

by the GEOFON Program at Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences and gempa GmbH.
Geodata, which is frequently available in the form of maps is stored as geographically orientated data in geotiff (raster) or shapefile (vector) formats.
All the Generic Data Format (GDF) and catalogs are in .mat files with defined structures and documentation available via eNodes or IS-EPOS platform.
GDF, miniSEED/SEED, inventory XML and .mat catalog are the working formats of the IS-EPOS platform, which is the main sharing point of data.
Detailed description of the format used by the IS-EPOS platform is documented and available in documents repository of the IS-EPOS platform (https://tcs.ah-epos.eu/eprints/1675).

*Specify if existing data is being re-used (if any)*

Six episodes were collected within national project IS-EPOS next six were integrated within SHEER project and are available for the other on-going projects (eg. EPOS IP, S4CE, SERA, …) and all registers.
IS-EPOS episodes are:

1. BOBREK (continued in EPOS-IP project)
2. CZORSZTYN (continued in EPOS-IP project)
3. GROSS SCHOENEBECK (continued in SHEER and EPOS-IP projects)
4. LGCD
5. SONG TRANH (continued in EPOS-IP project)
6. USCB (continued in EPOS-IP project)

SHEER episodes are:

1. GRONINGEN
2. LUBOCINO
3. OKLAHOMA (continued in SHEER and EPOS-IP projects)
4. PREESE HALL (continued in SHEER and EPOS-IP projects)
5. THE GEYSERS
6. WYSIN

EPOS-IP episodes are:

1. 1993 SOULTZ-SOUS-FORETS STIMULATION
2. 2000 SOULTZ-SOUS-FORETS STIMULATION
3. 2003 SOULTZ-SOUS-FORETS STIMULATION
4. ASFORDBY
5. GAZLI (restricted data access)
6. GISOS-Cerville
7. LACQ GAS FIELD (restricted data access)

8. MONTEYNARD
9. NORTHWICH
10. PREESALL MINE
11. PYHASALMI MINE (restricted data access)
12. STARFISH (restricted data access)
13. THE GEYSERS Prati 9 and Prati 29 cluster
14. THORESBY COLLIERY
15. VAL D'AGRI (restricted data access)
16. VAL D'AGRI FIELD (restricted data access)
17. VOUGLANS

S4CE episodes are:

1. CARBFIX (restricted data access)
2. ST. GALLEN (restricted data access)

POL-VIET episode is:

1. LAI CHAU (restricted data access)

BOIS episode is:

1. BOGDANKA (restricted data access)

*Specify the origin of the data*

Data sets are provided by several Institutions and Industrial partners worldwide: EU, UK and Vietnam (state on December 2017). All data providers have long-lasting experience in anthropogenic hazard research and assessment.

*State the expected size of the data (if known)*

Average size of the data related with episode is 33 GB, therefore the capacity of the data of forseen at least 21 episodes is at least 690GB.

*Outline the data utility: to whom will it be useful*

Several user types of the data can be distinguished:
The principal users of AH data sets is individual **researcher (R)**, who is authorized by the platform/eNode admin upon his/her own request. This class membership is resolved on the basis of applicant's affiliation.
The next is Internal **Project Participant (PP)**. NODE is going to be an infrastructural pillar, enabling anthropogenic hazard community data access and storage to carry on collaborative research within EPOS IP. The most often PP is a researcher and then has also all the powers of R. If not, the powers of PP against the resources of eNode arise from belonging of the user to a specific users' class.
**External Project Participant (EPP).** A member of a research project, whose coordinator agreed with eNode management special conditions concerning the use of its infrastructural resources for the project purposes. EPP, registered and authorized upon request of the project coordinator, obtains access to restricted resources of the

project, located on eNode. The most often EPP is a researcher and then has also all the powers of R. If not, the powers of EPP against the other resources of eNode arise from belonging to a class of users.
**Representative of Industry Partner (IP).** The Anthropogenic Hazards Community - industry partnership concept responds to the needs to raise contacts between science and industry to the level, at which the two parties mutually impact one another. In the framework of this partnership, Industry Partner makes a part of its infrastructures, either hard (instruments, monitoring networks) or soft (data, including relevant operational), available for research through eNode. It is worth to note that research does not need the most recent data, which can be sensitive for Industry Partner. The comprehensiveness of the studied case description with data is that, which is essential. The eNode management, from its side, implements the mechanisms, which prevent violation of rights of Industry Partners, including intentional or unintentional abuse of information that can be related to the Partner. Furthermore, Industry Partners provide advice on the potential usefulness and practical applicability of the solutions meant to be used in practice. In return, Industry Partners obtain possibility to convey to research community wider problems of their interest. The partnership has the form of a signed agreement between the two parties IP is authorized by eNode admin upon request of the management of the Industry Partner.
**Representative of Institutional User (IU).** Various interactions between Anthropogenic Hazards Community and either academic of non-academic organizations/institutions are envisioned. In every case its specific form, appropriate for the Institutional User's needs and eNode management strategy will be determined in a signed agreement between the two parties. The agreement will define the access extent to the resources of the NODE. IU is authorized by eNode admin upon request of the management of the Institutional User.
IS-EPOS platform as well as the EPOS ICS website will be main a data sharing points, both will have an access to eNode data specified within the EPOS IP project agreement and access rules, which can not be contradictory to above but may include access to open data for **Trainer/Trainee (T)** from Institutions, which will decide to use platform for training/education, will agree the use of IS-EPOS resources for these functions in accordance with IS-EPOS management.
**Public (P).** A registered but not authorized visitor of the IS-EPOS platform. This user has access only to the preview of the open access data stored on eNode.

## 2. FAIR DATA

### 2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA [FAIR DATA]

*Outline the discoverability of data (metadata provision)*

Metadata are prepared according to the guidelines of the EPOS Programme. Various metadata fields are required depending on the object type (episode, directory or file). Values of metadata are inherited down through the structure of data in the episode. For example, it is enough to set the value of metadata field 'episode name' for episode and then all directories and files belonging to this episode will have the same value of field 'episode name' as this episode. The same rule applies to directories. Of course it is possible to change inherited metadata value if needed. Metadata set for each file is saved in e-nodes in xml or equivalent exchange format.

*Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?*

Metadata are provided for data search and discovery as well as the keywords taking into account the anthropogenic hazard inducing factor, data type, file type etc. In case of sharing through the virtual portals, the unique link to data is also provided. Every available episode will have DOI registered as data set in DOI registry.

*Outline naming conventions used*

Naming conventions within the episode are related to the episode name and standard discipline name of the particular data types accepted within the AH and EPOS community.

*Outline the approach towards search keyword*

Genral search keywords are mainly related to the episode name, technology inducing anthropogenic hazard, type of the data and geographical location of the episode. Detailed search can be also done with time of the episode and related hazard occurence as well as by the file extension, methodology of the data collection and file name.

*Outline the approach for clear versioning*

Once the episode is created and published on www sharing platforms it can only have updates of the new data, which is denoted in data description and files metadata. In case of some fundamental change it has to be denoted in metadata and data description.

*Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how*

Metadata is prepared according to EPOS rules, following the CERIF naming and scheme convention.

**2.2 MAKING DATA OPENLY ACCESSIBLE [FAIR DATA]**

*Specify which data will be made openly available? If some data is kept closed provide rationale for doing so*

Data stored on eNode's may be of a different kind and different projects. In case of the H2020 projects it is forseen to make all the data open, if the data provider do not specify other access rules. The following statuses shown below take into account, that in some cases data providers such as industrial partners may not be willing to open all the technological data for public or the project managers keep data embargoed until the end of the project. In case of any project storing the data on eNode's the status of the data will be set individually according to the data provider – eNode agreement. Below are eNode's statuses for data:

| Level of confidentiality | Restrictions | Data type |
|---|---|---|
| **Confidentiality level 0 : Visualisation of metadata** | No restriction. Access to non sensible data : All users have access to these data. | Public data |
| **Confidentiality level 1 : Public data** | Data available to any person with an EPOS account (need of traceability) | Public data with traceability Direct download |
| **Confidentiality level 2 : Dissemination to academic field with justification** | Data for specialized researchers from the academic domain (researchers and project students). Before downloading the data, the user needs to give a reason and the name of the project he is working on. The owner of the data needs to give his accord. | Data available to the academic field with justification. Free data with a standard contract. Depending on the sensibility of data, the data can be directly downloaded |
| **Confidentiality level 3: case by case** | Restricted data: the user needs to contact the owner of the data in order to have an access. A specific contract needs to be signed by the two parties. | Restricted data (Possible remuneration needed) |

*Specify how the data will be made available*

The IS-EPOS platform (https://tcs.ah-epos.eu) and EPOS IP ICS platform are the main data sharing points. The direct acces to the data via eNode webpage is also forseen in case of the individual request.

*Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?*

Data will be accesible by webpages under https protocol, all the data is possible to visualise on the IS-EPOS platform or with use of standard software used in AH community.

*Specify where the data and associated metadata, documentation and code are deposited*

Metadata is stored on eNode and are synchronized with IS-EPOS platform and further with EPOS ICS platform, which allows to search the data through metadata and keywords on eNode storages. Detailed description of the formats and metadata used by the IS-EPOS platform is stored on eNode and available in documents repository of the IS-EPOS platform.

*Specify how access will be provided in case there are any restrictions*

The access is after signin in to the sharing points, according to the user roles and data access rules descibed in paragraph 1 (Data summary) and above in this paragraph.

## 2.3 MAKING DATA INTEROPERABLE [FAIR DATA]

*Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability*

The documentation of data formats is prepared and accessible at the level of data access. The metadata is implemented so that all information needed to read, use and interpret data in the future is available.

*Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?*

The metadata and documentation of data formats is prepared and accessible at the level of data access. More information can be added through documentation. All the vocabularies and keywords are in accordance with the general EPOS rules following CERIF standards.

## 2.4 INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES) [FAIR DATA]

*Specify how the data will be licenced to permit the widest reuse possible*

The licenses are defined in individual data owner/provider and eNodes agreements.

*Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed*

In case of the H2020 projects it is forseen to make all the data open, if the data provider do not specify other access rules or embargo period. The following take into account, that in some cases data providers such as industrial partners may not be willing to open all the technological sensitive data for public or the project managers keep data embargoed until the end of the project to guarantee the project members the priority in using the data to fulfill project goals. It is recommended to end the embargo period when the project ends.

*Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why*

In case of H2020 projects related data, the use of Creative Commons 4.0 International CC BY NC is recommended and all data after the end of the project or after embargo period end will be available according to the access rules described in 2.2, mainly they will be openly available registerd in EPOS.

*Describe data quality assurance processes*

**Data Quality workflow:**
Data upload: Raw data are uploaded to proper ‘buffer’ subdirectory or a quality server by data provider. Each data provider has an access to dedicated buffer directory in eNode ‘Administrator’ for this episode and its ‘Control group’ in eNode are assigned.

- ✓ QC1: Administrator sets new data as a new task in database task management system. (20%).
- ✓ Data conversion and validation: Data are verified, converted (if needed) and homogenized by people assigned by Administrator.
- ✓ QC2: The completeness and quality of prepared data are checked (50% of workflow).
- ✓ Data transfer, metadata preparation and publication: All files are described with sets of metadata prepared according to eNode rules. Standard metadata are checked.
- ✓ Data can be published. Metadata are visible for users.
- ✓ QC3: Sets and accepts data as correct (100%).
- ✓ Data is published.

*Specify the length of time for which the data will remain re-usable*

The long-term preservation plan is based on EPOS programme. Data will be stored after the EPOS IP project ends. Other national and international funding through the projects are also planned as well. Therfore the data reuse is forseen in any projects aiming in anthropogenic hazard. Curently data availability is guaranteed to 2020, according to the EPOS IP and national funding obligations, but it is forseen to be continued beyond 2025.

## 3. ALLOCATION OF RESOURCES

*Estimate the costs for making your data FAIR. Describe how you intend to cover these costs*

The main cost is the quality control and IT staff. The servers, hard drive, other technical infrastructures and work stations for QC and IT staff are funded by eNode hosting institution – EOST and IG PAS.
The main funding is currently from EPOS-IP. In the future, further funding is expected within the follow-up of EPOS. Additionally, national funding is expected as well. The securing of funding is the responsibility of eNode's managers. Current annual cost estimates of e-nodes within EPOS:

| | **Cost of IG PAS Node** | **Cost of EOST Node** |
|---|---|---|
| Servers (storage of the data) | 68 500,00 € | |
| QC group and IT | 120 000,00 € | 132 000,00 € |
| Place for QC group and node – operational cost | 60 000,00 € | 58 000,00 € |
| **TOTAL** | 248 500,00 € | 190 000,00 € |

Any storage of additional data should be supported by funding secured from the project or institution acquiring the data.

*Clearly identify responsibilities for data management in your project*

eNode manager is responsible for the operation of the eNode and implementation of the provider-eNode agreements as well as supervision of the QC. The IT staff is responsible for the servers performance and network security. QC group is responsible for the seting metadata and its validation, correctness of data formats and publication of the data through sharing points (IS-EPOS platform). The providers are responsible for the content of the data.

*Describe costs and potential value of long term preservation*

According to the above table cost of long term preservation is mainly the servers (they need to be replaced after the machine reach usage time) and operational costs of eNode with limited IT staff. Which means that server costs anually will be the same as in above table and operational costs will also stay the same as well.

## 4. DATA SECURITY

*Address data recovery as well as secure storage and transfer of sensitive data*

The main storage is on eNode's servers with daily or weekly backups. An external offline backup is planned annually.
The data access and management in storage spaces are available only for authorised eNode personel. The sensitive data such as the names and Institutional adresses and emails are stored in database, but the acces is limited for administarators of eNode only. The personal adresses or phones etc. are not stored for users and administration of eNode.
Data providers are obligated to denote the sensitive data and define their access rules or embargo if needed. eNode management will secure the data according to the detailed provider – eNode agreement covering above mentioned issues.

## 5. ETHICAL ASPECTS

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

The eNode management implements the mechanisms, which prevent violation of rights of data providers, including intentional or unintentional abuse of information that can be related to the provider. Namely the mechanisms are the Authentication, Authorization Identification and tracability of the data users.

**6. OTHER**

*Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)*

Question not answered.