Plan Overview

A Data Management Plan created using DMPonline

Title: European Databases of Seismogenic Faults: EDSF Installation

Creator: Roberto Basili

Principal Investigator: Roberto Basili, Roberto Vallone

Data Manager: Roberto Vallone

Project Administrator: Roberto Basili

Affiliation: Other

Funder: European Commission

Template: Horizon Europe Template

ORCID iD: 0000-0002-1213-0828

ORCID iD: 0000-0003-1208-9412

Project abstract:

The European Databases of Seismogenic Faults (EDSF) installation operates under the auspices of the EPOS TCS-Seismology work program, particularly those of theEFEHR Consortium, and considers the principles expressed by theEPOS Data Policy. EDSF distributes services for data about seismogenic faulting proposed by the scientific community or solicited to the scientific community or stemming from project partnerships that involved the use or development of the EDSF installation itself.

ID: 109454

Start date: 01-10-2015

Last modified: 20-06-2023

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

European Databases of Seismogenic Faults: EDSF Installation

Data Summary

Will you re-use any existing data and what will you re-use it for?

The compilation of seismogenic fault datasets exploits the wealth of information available from Earth Science studies, including, but not limited to, earthquake geology, seismology, seismotectonics, and geodynamics. The main purpose of such datasets is the geometric reconstruction of potential earthquake sources and the estimation of their activity rates. Re-used data mainly come from the scientific literature.

The generated datasets are integrated data products coming from complex analyses or community-shared data harmonization.

What types and formats of data will the project generate or re-use?

Generated and re-used data are most often geospatial data providing the location of potential seismogenic faults, their geometry, and their behavior. Parameters detailing both geometry and behavior are linked to the spatial data in the form of tabulated attributes.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

Data generation or re-use serves the creation of harmonized datasets of faults deemed capable of generating earthquakes above a pre-defined magnitude. These datasets are commonly used as input to analyze earthquake hazards, such as the hazards posed by ground shaking, tsunami, fault displacement, and several earthquake secondary effects (e.g., liquefaction and landslides). Altogether, they provide the basis to perform multi-hazard risk analyses. In addition, these datasets provide insights into the seismotectonics of large regions, thereby fostering research studies to understand the deformation mechanism of the Earth's outer shell.

What is the expected size of the data that you intend to generate or re-use?

The volume of the data is in the order of 1 GB, including all the processed data and the secondary outputs (documentation, journal articles, reports). Storage, access, and preservation do not imply additional costs than those already envisaged. Transferring data does not pose a challenge to the EDSF installation system.

What is the origin/provenance of the data, either generated or re-used?

The generated and re-used data come from scientific studies.

To whom might your data be useful ('data utility'), outside your project?

Datasets on seismogenic faulting are commonly used in earthquake hazard analyses, geodynamic modeling, and Earth sciences studies. Civil protection authorities use these datasets for the prevention and preparedness for natural hazards. These datasets are frequently used by geologists, engineers, and other professionals in science and engineering. The general public often consults these datasets on the occasion of earthquakes to understand the geological structure that generated them.

FAIR data

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

A persistent identifier must identify each and every dataset published on the EDSF installation. The EDSF Core Team recommends the use of a DOI minted by a recognized organization (e.g., DataCite), which also provides access to the relevant metadata.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Provisions for metadata include:

- metadata offered with the DOI as required by DataCite;
- metadata offered through the INGV Open Data Portal when the owner is INGV;
- metadata offered through the standard OGC protocol CSW.
- EPOS-DCAT-AP when the dataset is mapped in the EPOS ICS-C portal;
- INSPIRE when the dataset is mapped in the Italian "Repertorio Nazionale dei Dati Territoriali."

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Search keywords are provided in the metadata section of web services (WMS, WFS), in the metadata catalog, and in the DOI metadata. They include specific terms related to seismology and earthquake geology to ensure data discovery through web services and search engines. The same list of keywords is also included in the EDSF portal to promote Search Engine Optimization (SEO) common criteria.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

When the owner is INGV, the metadata are offered through the INGV Open Data Portal. Other cases will be evaluated.

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

Data are deposited in two servers owned by INGV, installed in two different institutional premises for security reasons.

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Different storage solutions were evaluated, and others are still under evaluation. The current solution was chosen as the best solution in terms of cost/benefits.

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

The repository only distributes datasets for which a DOI has been assigned. The recommended organization to mint DOIs is DataCite.

2.2. Making data accessible - Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All data is made openly accessible.

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

No embargo is adopted.

2.2. Making data accessible - Data:
Will the data be accessible through a free and standardized access protocol?

All the datasets distributed by the EDSF installation are accessible through free, standard, protocols ensured by web services of the Open Geospatial Consortium, or freely downloadable from the EDSF portal.

2.2. Making data accessible - Data:
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

No restriction on the use of the data exists, apart from giving the appropriate credit to the data creator.

2.2. Making data accessible - Data:
How will the identity of the person accessing the data be ascertained?

The identity of the person accessing the data is not ascertained.

2.2. Making data accessible - Data:
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

There is no need for a data access committee.

2.2. Making data accessible - Metadata:
Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Metadata is openly available and contains enough information (direct link) to enable the user to access the data.

2.2. Making data accessible - Metadata:
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Both data and metadata stored in the INGV repositories will remain available indefinitely.

2.2. Making data accessible - Metadata:
Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

The data are available in several, widely-used, cross-platform, GIS formats. Neither specific documentation nor specific GIS software for their use is needed. A brief tutorial on how to load the WMS/WFS services on the QGIS software (a free and open-source cross-platform desktop GIS) is offered on the EDSF portal. Training on how to access and use the data can be occasionally organized.

2.3. Making data interoperable:
What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

To guarantee interoperability between datasets made available through the EDSF portal and other spatial data, the standard OGC protocols WMS and WFS are adopted.

Metadata is published through the standard OGC protocol CSW.

Also, the availability of the EDSF datasets as popular formats downloadable files (GeoJSON files, ESRI shapefiles, MapInfo Tables, KML), facilitates users who need to combine EDSF datasets with other geographically referenced data in desktop GIS.

2.3. Making data interoperable:

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

There is no standard vocabulary for this type of data. However, the most common definition of the relevant scientific community is used as much as possible.

2.3. Making data interoperable:

Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1] ***A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)***

All published datasets must include qualified references to the broadest level possible.

2.4. Increase data re-use:

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

All datasets published in the EDSF installation must have comprehensive documentation addressing the data structure, the definition of variables, and the units of measurement. Ideally, the documentation is published in peer-review journals.

2.4. Increase data re-use:

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

All datasets published in the EDSF installation are made freely available and licensed under theCreative Commons 4.0 CC-BY license as prescribed by the EPOS Data Policy.

2.4. Increase data re-use:

Will the data produced in the project be useable by third parties, in particular after the end of the project?

All datasets published in the EDSF installation are freely usable by anyone.

2.4. Increase data re-use:

Will the provenance of the data be thoroughly documented using the appropriate standards?

The documentation and metadata of each dataset recognize the data provenance through proper citation of the source of information using the formats usually accepted by the relevant scientific community.

2.4. Increase data re-use:

Describe all relevant data quality assurance processes.

The EDSF Core Team carries out quality control of the distributed data according to a multi-step workflow described in thedata quality assurance document available in the documentation section of the EDSF portal.

2.4. Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

The authors of each dataset published on the EDSF installation participate in research projects and other initiatives using their respective datasets. Depending on the scale of the initiative, the participants estimate the allocation of resources. The EDSF Core Team assists in these initiatives to ensure the correct interpretation of the distributed data and help avoid misuse, giving special attention to research outputs coming from actionable science, such as earthquake hazards and risk products.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

When other research outputs are generated (only digital outputs are envisaged), the management of those outputs is shared with their creators.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

When other research outputs are generated, compliance with the FAIR principles shall be managed by their creators.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

The maintenance of the EDSF installation is presently estimated to be in the order of 100 k€/year.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Storage, archiving, re-use, and security costs are partly covered by EPOS and INGV institutional funding. When additional resources are necessary, they are sought through project funding.

Who will be responsible for data management in your project?

The EDSF Core Team is responsible for the data management of all web services published on the EDSF installation. Each published dataset has its own data manager(s) specified in their respective web pages, websites, or metadata.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The data will be retained indefinitely. Long-term preservation is ensured by storing data on the INGV IT infrastructure.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Data are backed up every night in separate storage through custom shell scripts. Recovery procedures, implemented with the same tools, were tested. A semi-automatic recovery solution is under evaluation as a commercial cloud storage solution for backups.

The operating system and software installed on the servers that store data are updated nightly to ensure security bug fixes are installed as soon as possible.

SSL transfer for HTTP (HTTPS) was implemented and is chosen per default for all hosted services.

The installation is actively monitored both at the Virtual Machine level by the hypervisor through which they are managed (vCPU, network traffic, storage, etc.) and at the Operating System level by the Nagios monitoring application specifically implemented in a virtual machine (processes, CPU load, disk partitions, logged users, services and websites). In case of anomalous behavior, Nagios sends an alert to the system administrator (email and instant message).

No sensitive data are stored.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There is no ethical reason that could impact data distribution and sharing. A disclaimer is associated with each dataset to remove legal liability from the data owner and the data publisher. Users are also cautioned to consider carefully the nature of the datasets before using them for decisions that concern personal or public safety or in relation to business involving substantial financial or operational consequences.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

No personal data is collected or distributed by the EDSF installation.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

There is no plan to use another data management procedure for the EDSF Installation. Additional DMPs using the DCC template will be adopted for specific datasets distributed by the installation.