SERA
Seismology and Earthquake Engineering
Research Infrastructure Alliance for Europe

DELIVERABLE

D4.3 Best Practices Guidelines

Work package	WP4 Expanding access to the European seismic monitoring infrastructure
Lead	Deutsches GFZ Potsdam
Authors	Javier Quinteros (GFZ), Reinoud Sleeman (KNMI), Angelo Strollo (GFZ), Jarek Bienkowski (KNMI), Helle Pedersen (CNRS), Christos Evangelidis (NOA), John Clinton (ETH), Constantin Ionescu (INFP)
Reviewers	Jan Michalek (UIB)
Approval	Management Board
Status	Final
Dissemination level	Public
Delivery deadline	30.04.2020
Submission date	27.04.2020
Intranet path	DOCUMENTS/DELIVERABLES/ SERA_D4.3_Best_Practices_Guidelines.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730900.

Table of Contents

Summary 3
1 Introduction 4
2 General organization of EIDA 5
3 EIDA Nodes Operation 6
3.1 Services 6
3.1.1 FDSN Station WS 7
3.1.2 FDSN Dataselect 8
3.1.3 WFCatalog (Data Quality Parameters) 9
3.1.4 Routing Service 10
3.1.5 EIDA Authentication System (AAI) 11
3.1.6 Feedback and communication with users and developers 12
3.2 Data Centre registration at FDSN 12
4 Data Protection related issues (GDPR) 13
5 Integrating exotic data 13
6 Training and Outreach 14
Contact 15

Summary

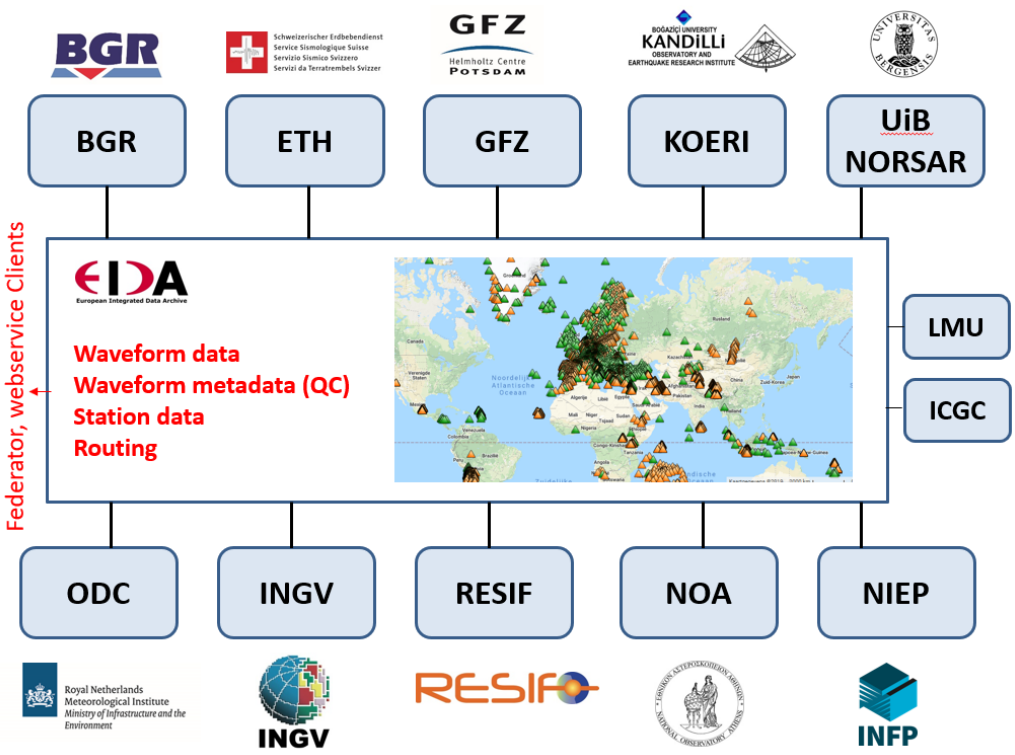
This report provides specific guidelines for the configuration and operation of services maintained at nodes in the European Integrated Data Archive (EIDA): fdsnws-station, fdsnws-dataselect, eidaws-wfcatalog, eidaws-routing, EIDA Authentication System and EIDA Issue Tracker. These guidelines address a) specific details with the aim to harmonize the operations across EIDA, b) procedures to update information in the global context (FDSN) and c) awareness of regulations concerning general data protection.

1 Introduction

The current expansion of the European infrastructure for seismological waveform data EIDA (European Integrated Data Archive) is driven by the increase in number of data, new data types and the user community. Examples that illustrate this expansion during SERA are: the addition of two new nodes into EIDA, the number of stations that distribute their data through EIDA increased with about 4500 to more than 11500, the amount of data exported by the ORFEUS Data Centre increased from 8 TB in 2017 to 27 TB in 2019, and the development and deployment of new services like the EIDA Federator, Authentication System and Issue Tracker.

To ensure long term sustainability of such a complex infrastructure it is essential to a) standardize as much as possible on formats and services for the users, b) standardize operational procedures as much as possible, and c) create and maintain up-to-date documentation.

This deliverable D4.3 accompanies deliverables D4.2 (Report on metadata challenges and proposed solutions), D4.4 (Metadata model standards) and D4.5 (EIDA Documentation System) and describes the current operational procedures at EIDA nodes based on EIDA standards and guidelines ("best practices") and how to implement these efficiently. Best practices are a set of guidelines that represent the most efficient course of action in a given situation. Therefore, these guidelines will become part of the EIDA Documentation system as a living document.



EIDA infrastructure: 12 federated data archives providing access to their data holdings through standardized services.

2 General organization of EIDA

EIDA, an initiative within ORFEUS, is a distributed federation of data centres established to securely archive seismic waveform data and metadata gathered by European research infrastructures and provide transparent access to data for the geosciences research communities. EIDA's organization and management is handled by the EIDA Management Board (EMB). The EIDA nodes are data centres that collect and archive data from seismic networks deploying broad-band sensors, short period sensors, accelerometers, infrasound sensors, and other geophysical instruments. In addition to the EMB, the EIDA Technical (ETC) works in close collaboration and under supervision of the EMB. The ETC is where all technical aspects of new developments and implementations within EIDA are discussed.

Each node nominates a person to represent it at each Board/Committee (EMB, ETC). In the case of the ETC, this person will be the technical contact for the node. Other persons could also be nominated as contacts to provide faster reactions.

Currently EIDA comprises 12 data archives, hosted by the following institutions/organizations:

ODC/KNMI: Orfeus Data Centre / Koninklijk Nederlands Meterologisch Instituut
GFZ: Deutsches GFZ Potsdam
RESIF: Reseau Sismologique & Geodesique Francais
ETH: Zurich Schweizerischer Erdbebendienst (SED)
INGV: Istituto Nazionale di Geofisica e Vulcanologia
INFP/NIEP: National Institute for Earth Physics
KOERI: Kandilli Observatory and Earthquake Research Institute
NOA: National Observatory of Athens
BGR: Bundesanstalt für Geowissenschaften und Rohstoffe
UiB / NORSAR: University of Bergen
LMU: Ludwig Maximilians Universität München
ICGC: Instituto Cartográfico y Geológico de Catalunia

EIDA is in close contact with the users and the research communities by two mechanisms: a) the ORFEUS User Advisory Group (UAG) with the task to review practice and services and to provide suggestions and recommendations for improving current services or implementing new services; b) through the EIDA Issue Tracker on GitHub.

3 EIDA Nodes Operation

EIDA provides a set of FDSN and EIDA standardized and EPOS compatible web services, which are most useful for mass downloads of data, metadata, and related seismic information, within automated seismic analysis workflows (e.g. regional/global tomography and receiver functions, regional seismicity, attenuation, ambient noise cross-correlations).

It is expected that these services are running and performing 24/7. In case of technical problems, or issues related to the quality of data or services, the technical contact of the EIDA node should react as soon as possible with a maximum delay of 48 hours. This includes also the issues assigned to them in any of the EIDA communication channels (e.g. Issue Tracker @Github, mailing lists).

As a general rule, all nodes should be running the latest stable versions of the services. However, it is of ultimate importance to coordinate the upgrade of the services between the nodes to avoid incompatibilities between the nodes. If possible, nodes should upgrade services and schemas in a reasonably short time window.

The technical setup of EIDA is basically a Federation of data and services. Its operation is mostly automatic and runs in the background in a transparent way for users, but sometimes also for the node operators. In order to keep incompatibilities and unwanted changes to a minimum, all operational changes should be not only done technically, but also **announced in advanced** to the EIDA Maintenance Group mailing list (eida_maint).

Details and resources like documentation for the operation of EIDA nodes, as well as for the integration of new nodes, can be found at the ORFEUS documentation repository[1] for public access and Redmine[2] for internal access, as well as in the Deliverable D4.5 of this same project.

In the next sub-sections we provide some specific guidelines for the configuration and operation of these services.

3.1 Services

Each of the data centres belonging to EIDA must run at least four services (http://www.orfeus-eu.org/data/eida/webservices):

- FDSN StationWS: provides metadata / inventory information,
- FDSN Dataselect: provides seismic waveforms,
- WFCatalog: provides quality parameters and availability of the waveforms,
- Auth: provides (temporary) credentials data from a digitally signed token (only mandatory for nodes hosting restricted data).

Other services are being provided centrally by EIDA:

- Routing Service: provides information/routes to services,
- EIDA Authentication System (AAI): provides digitally signed tokens to be used in EIDA services,
- Issue Tracker: official communication channel for user feedback on issues and problems related to EIDA data and services.

[1] https://orfeus-eu.readthedocs.io/
[2] https://dev.knmi.nl/projects/eida

- Federator: provides a single, unified access point to the waveform archives and the station and quality control information from the entire EIDA data holdings.

Despite that there are mechanisms in place for the monitoring of most of these services, each EIDA node is responsible for keeping their own services up and running 24/7. Centralized monitoring, like the one deployed at ETH to measure the performance of nodes[3], should also be reviewed daily to early detect potential problems.
This is also valid for the data and metadata archived at the data centres. Each node should thoroughly check the consistency of data and metadata to early detect problems on the gains, orientation, timing, or any other information needed to process the data. There is still an on-going discussion towards a unique tool to be run at each node, and currently this remains an individual decision of the data centre. The decision regarding the need of new services for EIDA, or the adoption of already existing services, is a task of the EMB. Once the decision was taken, the ETC should discuss about all technical aspects of the new development, or adoption.

3.1.1 FDSN Station WS

A big collaborative effort has been done during the last 6 years to identify seismic networks by means of persistent identifiers. These identifiers are meant to be used in the citation of scientific and public communications, so that data providers can be properly attributed. Available DOIs are already easily accessible for all networks through FDSN web page and several EIDA nodes. Users start to be aware of the necessity to acknowledge the use of seismic data by proper DOI citation. Therefore, data centres should adopt this as part of their usual data workflow.

Most of the details regarding the creation of a DOI for a seismic network can be found in the FDSN Recommendation for DOIs (2014)[4], as well as in the Deliverable D4.4 of this project (EIDA Metadata model standards).

With the new release of the StationXML schema v1.1 it is now possible to include the persistent identifier directly in the metadata (e.g. StationXML). Data centres should do this by including the tag “Identifier” under the “Network” tag in the following way:

```
<Network code="XX" startDate="1980-01-01T00:00:00" restrictedStatus="open">
<Identifier type="DOI">10.12345/SOMEID</Identifier>
…
</Network>
```

Declaration of a DOI identifying a seismic network.

Because persistent identifiers can only be provided by means of the new StationXML schema v1.1, data centres should migrate to that version as soon as possible, but in a coordinated manner with the other nodes to avoid multiplicity of versions.

All services have implicit or explicit limits in the amount of information which can be sent. As these limits can be specified in different units, or quite difficult to understand, or not even known by the users, an easy rule-of-thumb has been agreed. Data centres should adjust their configuration so that

[3] http://eida-webtests.seddbd.ethz.ch/en/home/
[4] http://www.fdsn.org/pdf/V1.0-21Jul2014-DOIFDSN.pdf

the *full inventory of each network* at "response" level can be downloaded. This provides a clear directive for developers in order to provide sustainable implementations.

3.1.2 FDSN Dataselect

One of the recommendations of the ORFEUS User Advisory Group (UAG) was to make available to EIDA users some simple scripts to discover, select and download data in a proper way. Therefore, users will avoid trivial errors and can adapt these scripts to their own needs.

In the EIDA official Github repository[5], we included a set of scripts as well as documentation with recommendations on tools supporting all (most of) EIDA features (e.g. Obspy, fdsnwsscripts).

Users can get a seamless first approach to data discovery and access by means of these resources. For most of the users this should be enough to satisfy their needs. More complex examples can be implemented but this can still be the best starting point to avoid usual mistakes.

All users who cannot find proper support in the available documentation, issue trackers, ORFEUS contact page and other communication options, should feel free (and are encouraged) to contact the ORFEUS UAG (current composition to be found at: http://www.orfeus-eu.org/organization/structure), not to solve a particular problem, but to help the UAG to steer the service developments that are required by the research community.

Regarding the configuration on the data centre side, limitations on the web server and other systems don't need to be identical, but preferably should be coherent and reasonable.

- Firewall policies managed by IT departments on which the nodes operator has no control must be avoided.
- The number of allowed connections could be between 30 and 100.
- Some limit on the number of connections per IP could be set (e.g. 10 to 30).
- If some of these limits are hit, the user should receive a 503 HTTP error. Then, the user should split the request in smaller pieces and try again.
- Number of samples per request should be at least 1.000 million.

Special attention should be payed to the case in which a user requests data and there is nothing to return (204 No Data). This operation can be very fast, and in the case that a user is running a code sequentially requesting data, it can lead to many tens (hundreds?) of valid requests per second. This workflow should not be interpreted as a Denial-Of-Service (DOS) attack, because a big part of valid requests received by EIDA nodes don't provide any data for various reasons related to protocol issues or data issues. Until we foster the adoption of services like WFCatalog or the new, standard availability-WS, this situation will happen very often.

The ultimate check for accessibility of data must be easily done by means of Obspy and the fdsnws_fetch client (from fdsnwsscripts package), which are the recommended clients in the EIDA documentation. Full instructions on how to do this can be found in README file of the EIDA repository[6].

[5] https://github.com/EIDA/userfeedback
[6] https://github.com/EIDA/userfeedback/blob/master/README.md

3.1.3 WFCatalog (Data Quality Parameters)

The WFCatalog web service provides detailed information on the contents of waveform data including quality control parameters. Information can be included on sample metrics, record header flags, and timing quality. The WFCatalog can serve as an index for data discovery as it has support for range filtering on all available metrics.

According to the ORFEUS UAG, the WFCatalog should be deployed and used on all existing and future EIDA nodes. EIDA nodes should ensure that the WFCatalog is always up-to-date and displays the most recently added/changed data. A fully federated and fully working WFCatalog is the base for any further development towards data quality selection and to quickly identify (meta)data issues.

Quality parameters should be calculated daily, approximately one day after the data files have been closed, with a maximum delay of 36 hours. At that moment one can consider that its data are complete. Processing of daily waveform data with a small delay (~1-2 days) ensures an optimum between data completeness and quick availability of the quality parameters. At the same time the computational resources needed to process the data are kept constrained. Current standard granularity for the quality parameters is one day. Smaller granularities are technically supported but will increase resources significantly.

A list of the metrics calculated from the seismic waveforms can be seen in the following table:

QUALITY METRIC	DESCRIPTION
MAX_GAP	Maximum gap length in seconds
MAX_OVERLAP	Maximum overlap length in seconds
NUM_GAPS	Number of gaps
NUM_OVERLAPS	Number of overlaps
NUM_SAMPLES	Number of samples
PERCENT_AVAILABILITY	Percentage of available data
SAMPLE_MAX	Maximum sample value
SAMPLE_MIN	Minimum sample value
SAMPLE_MEAN	Mean sample value
SAMPLE_RMS	Quadratic mean of samples
SAMPLE_STDEV	Standard deviation of samples
SAMPLE_LOWER_QUARTILE	25th percentile of samples
SAMPLE_MEDIAN	50th percentile of samples
SAMPLE_UPPER_QUARTILE	75th percentile of samples
SUM_GAPS	Sum of data gaps in seconds
SUM_OVERLAPS	Sum of data overlaps in seconds

Metrics calculated by the data centres.

3.1.4 Routing Service

EIDA offers FDSN web services for accessing their holdings. Depending on the type of service, these may only provide information about a reduced subset of all the available waveforms.

To assist users to locate data, we have designed a Routing Service, which runs centrally at ODC. This (meta)service is supposed to be queried by clients (or other services) in order to localize the address(es) where the desired information is provided.

The Routing Service must serve this information in order to help the development of smart clients and/or services of higher level, which can offer the user an integrated view of the entire EIDA, hiding the complexity of its internal structure. However, the Routing Service needs not to be aware of the extent of the content offered by each service, avoiding the need for a large synchronized database at any place.

Each data centre should provide a set of routes to declare its own data holdings (e.g. archived networks). These routes need to be imported during the synchronization process by the EIDA central service. Routes consist of a NSLC (Network, Station, Location, Channel) code, a time window and the URLs where services provide information related to this time series.

Once a day the central Routing Service merges the routes from all data centres and checks for consistency. In order to improve the workflow of this process, as well as the downstream services, data centres should provide as few routes as they can.

For instance, if the data centre provides access to all existing waveforms from a network (XX), just one route at the network level should be declared.

```
    <ns0:route networkCode="XX" stationCode="" locationCode="" streamCode="">
```

This reduces the number of routes managed by the service, but also the number of entries in the output to users. Therefore, it also reduces the number of requests that a user/smart-client will do after querying the Routing Service.

In case of a network that is distributed across different data centres the route can be defined at a finer level (e.g. station, channel). The aim is always to generate the minimum number of routes which describes the data holding at the data centre.

```
   <ns0:route networkCode="XX" stationCode="STA1" locationCode="" streamCode="">
…
   <ns0:route networkCode="XX" stationCode="STA2" locationCode="" streamCode="">
```

Like in the example above, if a network XX has 3 stations and only two of them (STA1, STA2) have been archived at your data centre one should find two routes, one for each station. The third one, for STA3, should be provided by the data centre hosting it. The three routes will be merged centrally during the synchronization.

The process of integrating the basic services of an EIDA node is reduced to the inclusion of an URL in the central Routing Service running at ODC. This URL should provide routes like defined above. There are two options on how to do this. The easiest one is to provide an URL to a static XML file with all this information. In some special cases, an EIDA node can also decide to run a Routing Service. This instance of the service can expose the list of routes to data from the node, so that the central service at ODC can harvest it. The node operator can check if this is working by calling the “localconfig” method of its

local (@node) Routing Service. Detailed instructions on how to use this method can be found in the section "Importing remote routes"[7] from the Routing Service documentation.

As it has been suggested for other services, despite any changes in routes will be automatically ingested into the EIDA system (and also to FDSN) and forward downstream, changes to the routes are expected to be announced in the eida_maint mailing list.

The full documentation for users, operators and developers can be found at https://routing.readthedocs.io .

3.1.5 EIDA Authentication System (AAI)

EIDA has worked for many years as a federated archive, allowing users to request data which are actually distributed among many data centres in a simple way. This works very well for open data. However, restricted data present some difficulties.

The FDSN web services use HTTP basic digest authentication, which is not the best technical decision for federations like EIDA, where we try to offer the user experience of a single European data centre, which in the background is formed by a federation of data centres.

From the point of view of the authorization the data centres have to manage their own access control lists. This works well only if an experiment is archived at only one data centre, because the access control list is stored next to the data, but in the case of collaborative experiments it implies again problems of synchronizing the access control lists between nodes. Moreover, the increasing number of examples in which institutions from different countries cooperate splitting the data from an experiment between different archives (e.g. networks NERA JRA1, Alparray) showed the need for a scalable solution for user management and access control to the data.

B2ACCESS[8] is the EUDAT[9] federated cross-infrastructure authorisation and authentication framework in Europe for user identification and community-defined access control enforcement. The B2ACCESS service managers are responsible for managing users and their assignments to groups for their specific service. They can also manage (edit, delete and modify) groups of their service.

B2ACCESS allows users to authenticate themselves using a variety of credentials. The most interesting way of authenticating (at least from the point of view of EIDA users) is to use the eduGAIN[10] backbone.

eduGAIN is an initiative with the aim of allowing users to authenticate (log in) at their home institutions and let them use services all around the world. The idea behind is analogue to the service that eduROAM provides. There are strict rules for an institution to be accepted as part of eduGAIN and these rules assure the quality of the credentials issued. Their members are mostly universities and research institutions.

Data centres should:

- Foster users to create a B2ACCESS account using their institutional credentials.
- Local B2ACCESS accounts should be used if the connection to the home institution of the user is not available.
- Manage the Access Control List (ACL) locally for all networks when their data are not distributed between data centres.

[7] https://routing.readthedocs.io/en/latest/userdoc.html#importing-remote-routes
[8] https://b2access.eudat.eu/
[9] https://eudat.eu/
[10] https://edugain.org

- If a network is distributed (e.g. Alparray), permission should be managed using the graphical interface of B2ACCESS. The token produced by the EAS will contain these permissions and users will be authorized (if proper) when querying restricted data. Local permissions should be configured with the name of the B2ACCESS group related to the network/experiment instead of the list of users.
- Foster users to request data with smart clients supporting routing and the token-based authorization: fdsnws_fetch as a command line tool (replacement for arclink_fetch) and Obspy as a python-based multipurpose solution.

3.1.6 Feedback and communication with users and developers

EIDA needs to early detect problems related to any service being provided or the quality of its data holdings (data/metadata). The **official communication channel** for this type of issues is the Issue Tracker at the EIDA repository at Github[11].

Problems with each EIDA service/client/product should be posted here by users and properly labelled. Issues will be classified and assigned to the operators/developers responsible for that within the same day.

The only way to improve the operations of our data centres is to receive continuous and constructive feedback. Whenever a user has a problem or detect an issue, the following options are given:

- Read very carefully the available documentation.
- Check for existing posts in the Issue Tracker[12]. It is possible that another user posted some similar problem and the answer is already available. Do not forget to check the **closed issues**, as these are the ones which already provide a solution.
- Post a new issue in the Issue Tracker. There are already templates defined to guide the user in order to provide all information needed and avoid extra requests and waiting for more feedback.

In the case that some changes should be applied to any system, or when a new system is developed, it is important to communicate this to the technical contacts (or developers) of the clients recommended by EIDA/ORFEUS. These are:

- fdsnwsscripts: geofon@gfz-potsdam.de, andres@gfz-potsdam.de
- Obspy: lion.krischer@erdw.ethz.ch, tobias.megies@geophysik.uni-muenchen.de
- Federator: kaestli@sed.ethz.ch, daniel.armbruster@sed.ethz.ch

3.2 Data Centre registration at FDSN

The FDSN approved in its Meeting in Montreal (July 2019) a formal procedure to register a Data Centre, solving most of the limitations related to the old (manually curated) list of data centres in a HTML page.

The new FDSN Data Centre Registry is designed to provide discovery of FDSN data centres (in machine-readable format), discovery of services offered by each data centre, and identification of priority (primary, secondary, …) for data sets offered by each centre.

The Registry consists of a central repository at http://www.fdsn.org . Each of the entries (a data centre) of the registry will be directly controlled by the person authorized at each data centre. The web pages

[11] https://github.com/EIDA/userfeedback/issues
[12] EIDA Issue Tracker is the official communication channel (https://github.com/EIDA/userfeedback/issues).

at the FDSN site related to the data centre will be based on the information declared on this registry. Also, a web service is available to provide basic operations on the registry.

If an operator wants to perform operations on the Data Centre Registry, it will need to use its FDSN Message Centre (mailing list) account. The creation of a new data centre is still a moderated process in order to avoid spurious creation of data centres.

Once the data centre exists, the user can update the entry in multiple ways. The easiest option is to use a simple web form available via the “Edit” button in the Data Centre page. Something more advanced, and that it can be also automated, is to POST an update through the web service.

The preferred one and recommended for EIDA nodes is the third one, the regular harvesting of the entry from a supplied URL. This URL will be the official EIDA Routing Service. The functionality of the Routing Service was extended to be able to “export automatically” this curated information to the FDSN Data Centre Registry.

It could be normal that some errors will be shown with inconsistencies between the harvested data and the data provided by other data centres (not from EIDA). Consistency between EIDA nodes is being checked by the Routing Service, but with other data centres (e.g. IRIS) there is no previous check.

If there are inconsistencies, take note of all of them and which data centre is the one responsible for that. Please, get in touch with the technical contact of the other data centre and agree the priorities assigned to each route by each of you. You can always check the FDSN pages for a particular network and verify who is responsible for that route.

4 Data Protection related issues (GDPR)

Each node must provide daily summary of usage without sensitive/private data. The future Logging System is currently being designed and developed, so no technical details in the format of best practices can be given at this point. However, node operators should be very careful of the new European regulations (GDPR, https://gdpr-info.eu) regarding Data Protection.

Each data centre must provide a “Data Protection” text on their web sites, specifying in detail their internal policies related to data such as logs, user information, emails, etc. This text must declare explicitly how long each piece of data will be stored and, if it is the case, with whom these data will be shared with.

5 Integrating exotic data

In the context of this project, the Deliverable D4.2 provides an extensive and detailed list of recommendations for OBS, Infrastructure monitoring, and other types of exotic (or non-classical seismic) data. All data centres are expected to follow the recommendations published in that document. For instance, what is suggested for time correction for OBS data: shift the data to closest time and change quality flag to ”Q” (page 19).

However, some widely used implementations of the FDSN Dataselect WS (i.e. SeisComP3) do not support more than one version of the stream (they have the same NSLC code and differ only in the header flag). Therefore, a recommended practice is to provide via WS the best quality data (Q) following the recommendation of the Deliverable D4.2, and to be ready to provide the unmodified version of the data by some other service if requested by users.

6 Training and Outreach

ORFEUS is usually offering sessions on “Data Availability, Data Access, Data Quality” as well as “Later developments” at the Annual ORFEUS Workshops to enable discussions between the infrastructure and the scientific communities on these topics. The attendance to these workshops, as well as the participation of a critical mass of community members in their preparation, or formal communication channels with the ORFEUS Management related to it, is needed in order to satisfy the needs of the community.

Contact

Project lead	ETH Zürich
Project coordinator	Prof. Dr. Domenico Giardini
Project manager	Dr. Kauzar Saleh
Project office	ETH Department of Earth Sciences

Sonneggstrasse 5, NO H-floor, CH-8092 Zürich

sera_office@erdw.ethz.ch

+41 44 632 9690

Project website www.sera-eu.org