Data Management Plan all’ONT
Gruppo EIDA-UF3

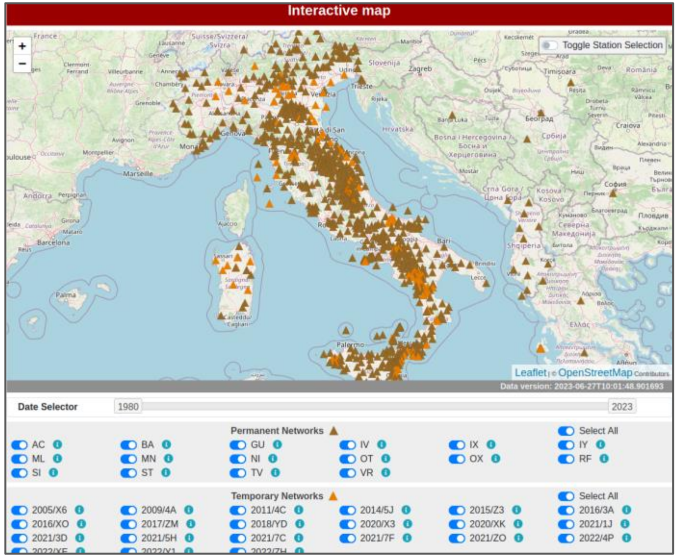
Carluccio I., Daneck P., Della Bina E., Fares M., Franceschi D., Mandiello A. , Maniscalco M., Mazza S.,Pintore

Keywords

Policy dati INGV
Open Data
Open Access
Trusted-Archive
FAIR data

EIDA Italian Node

The EIDA Italia node is the INGV managed datacenter that stores and distributes seismic waveforms collected by the INGV seismic networks, and other Italian and some foreign data providers. The list of institutions that contribute with their seismic networks data to EIDA Italia can be found at www.eida.ingv.it site. Data from other countries in the Euro-Mediterranean area are collected directly by INGV and partners by the MedNet network.



Data version: 2023-06-27T10:01:48.901693

Date Selector 1980 2023

Permanent Networks

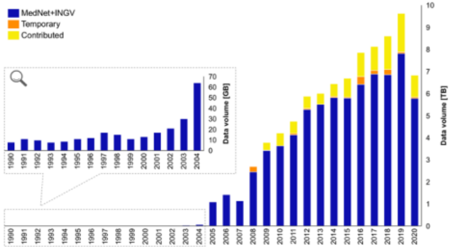
Select All

AC, BA, GU, IV, IX, IY
ML, MN, NI, OT, OX, RF
SI, ST, TV, VR

Temporary Networks

Select All

2005/X6, 2009/4A, 2011/4C, 2014/5J, 2015/Z3, 2016/3A
2016/XO, 2017/ZM, 2018/YD, 2020/X3, 2020/XK, 2021/1J
2021/3D, 2021/5H, 2021/7C, 2021/7F, 2021/ZO, 2022/4P
2022/XE, 2022/Y1, 2022/ZH

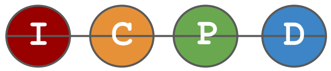


EIDA-ITALIA:

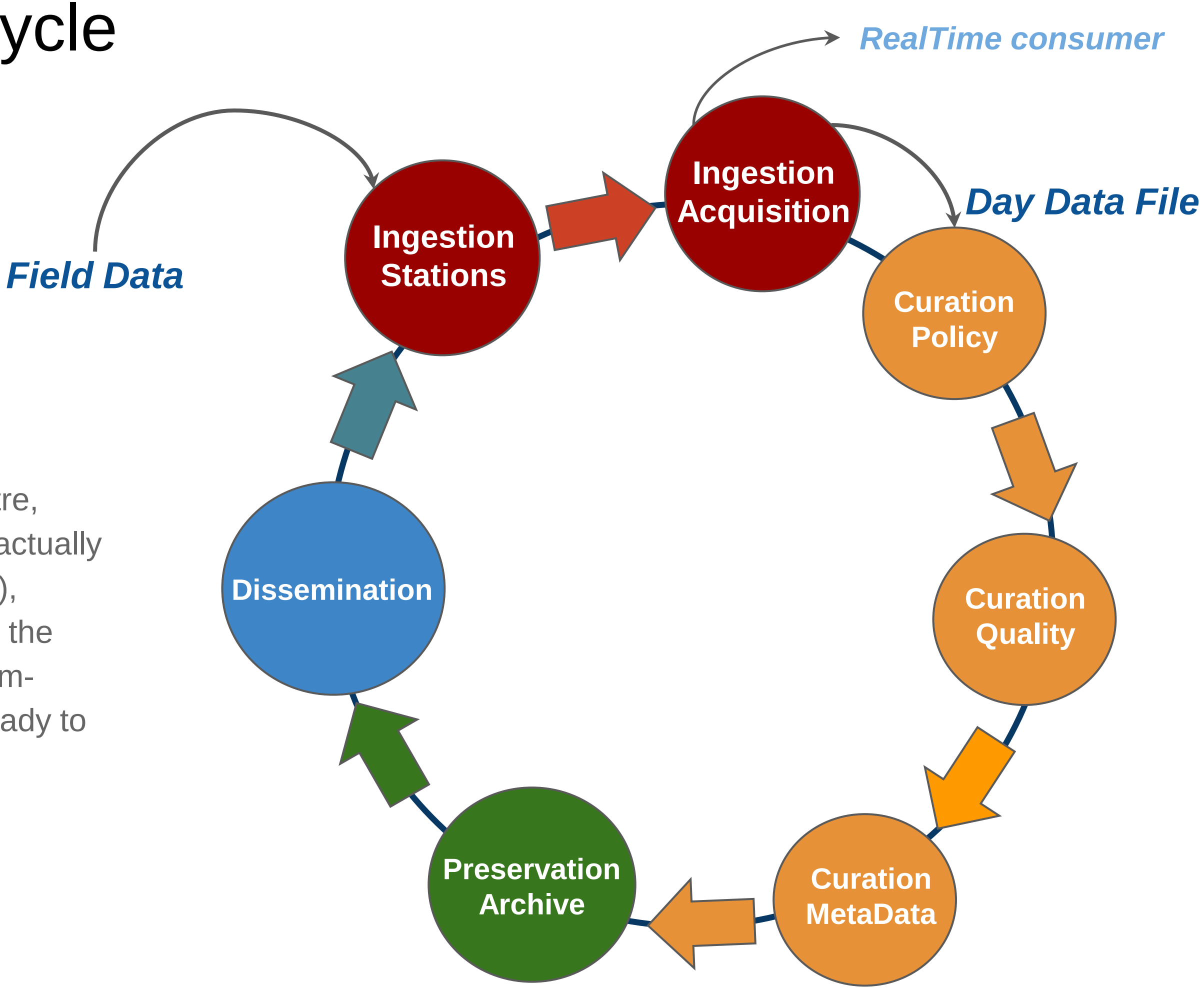
- 15 data providers
- 16 permanent networks
- 21 temporary deployments
- 1000+ stations
- 11k channel-epoch
- 120+ TB waveform data files

Seismic Data LifeCycle

Data Centre point of view

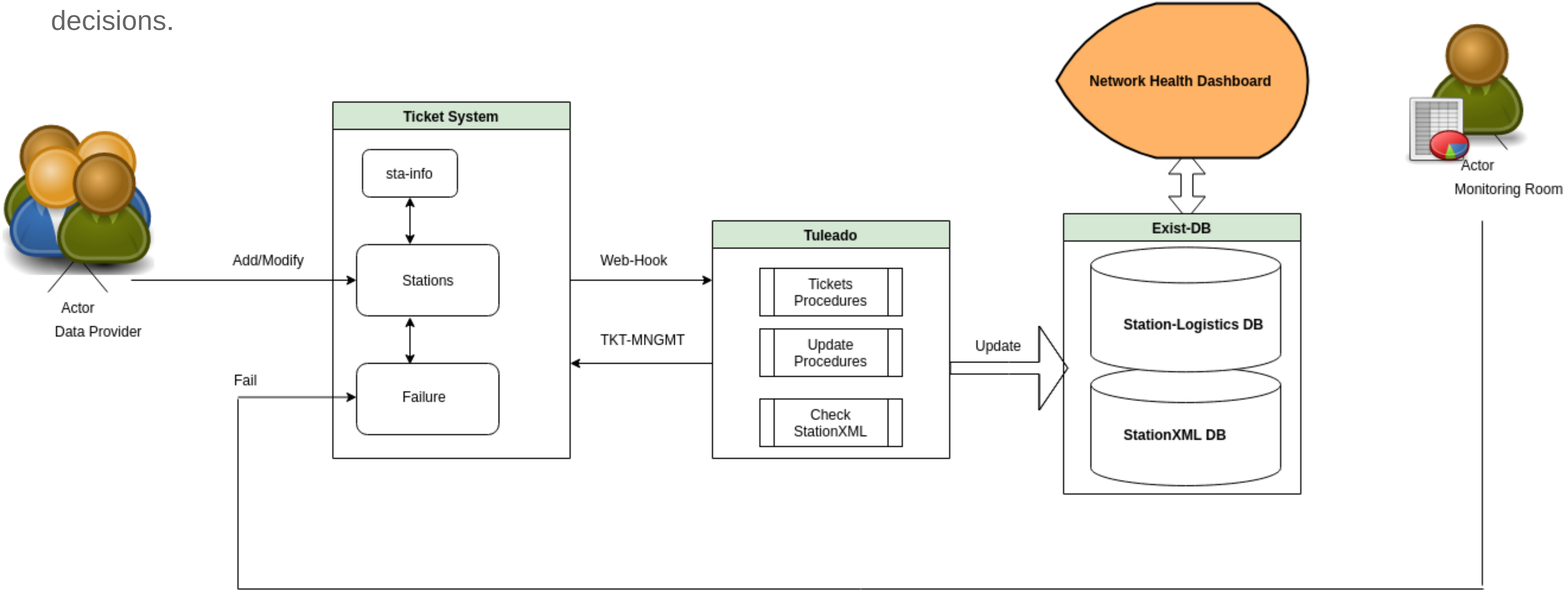


Our data life cycle, as Seismic data centre, start from initial generation: the station (actually we consider stations as our data-source), hence via the ingestion procedure, go to the curation phase, then will go into long-term-preservation phase; there data will be ready to be disseminate and re-use.



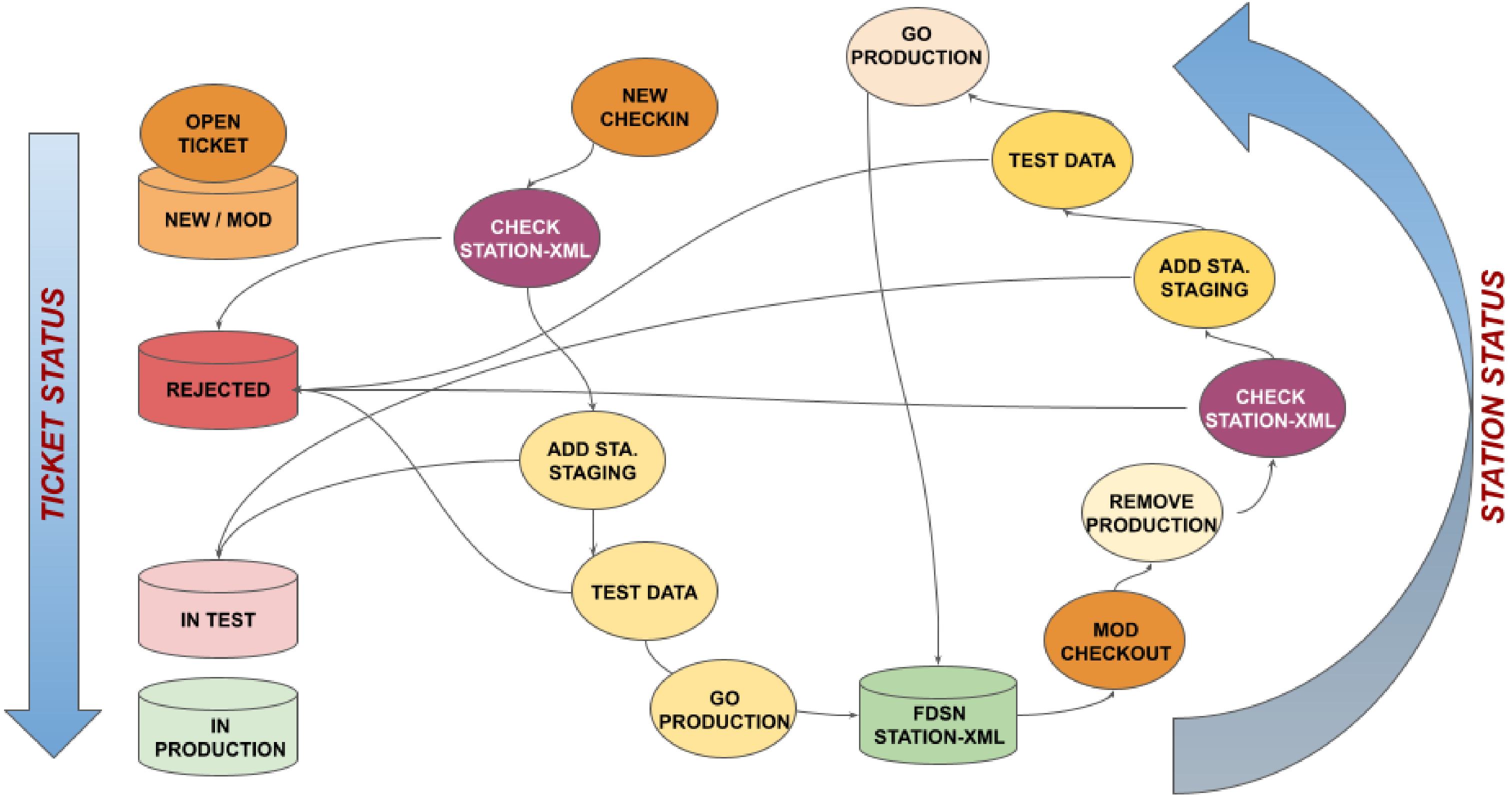
Data Ingestion Station management system

Our system has a modular design, by leveraging existing open-source solutions i.e. as a customized implementation of a ticketing system, a GUI leverages Grafana creating a dashboard to present an overview of the system status of all components. Ticket status drives the station process through all of our systems (from acquisition to distribution). All these steps are managed through *Tuleado* who is able to follow some policies and make some decisions.



Data Ingestion Station management model

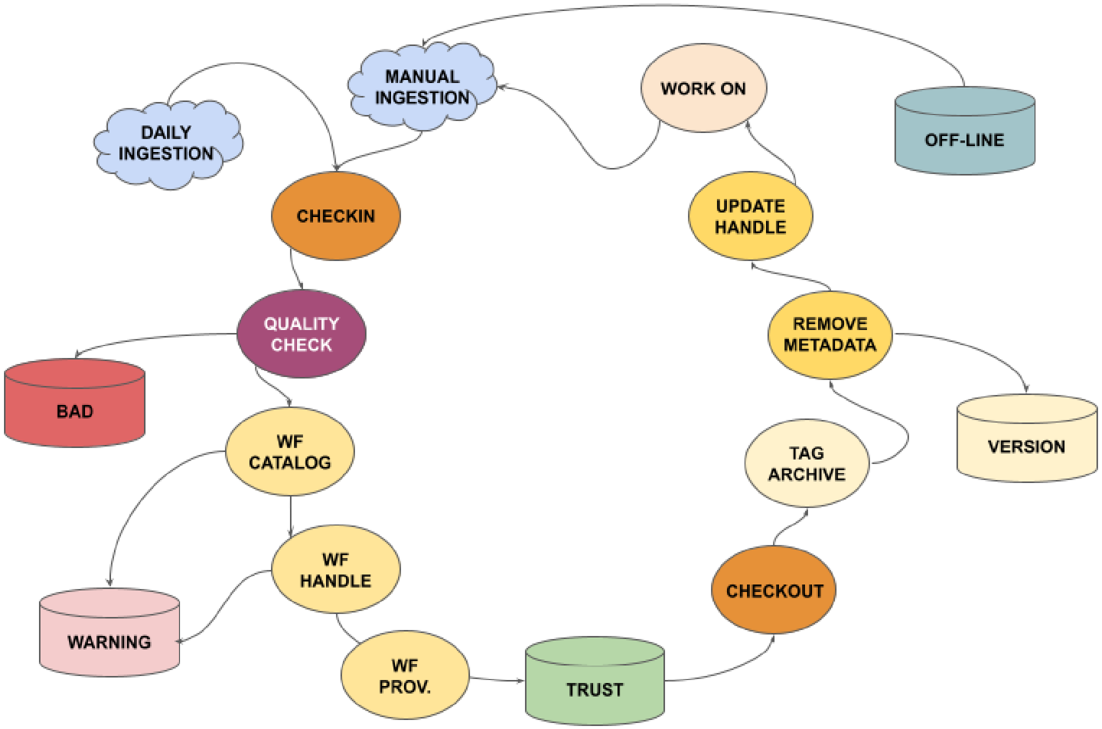
Our station management model is based on a tight coupling between a ticket system and ours acquisition/dissemination systems; avoiding as much as possible human intervention in back office. To do this a Business Automation is implemented via an in house developed software agent called “*Tuleado*”.



Data Curation Policies & Quality

We have deployed an "**Policy Enforcement Point**" in order to execute our policies, expandable as needed, on data file.

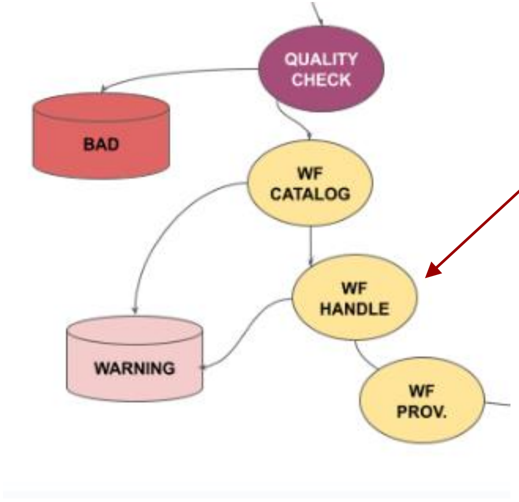
E.g. our 'checkin' policy do this: after pass sanity check and quality assessment, data file is directed to extract community metadata and metrics, then is turned from data file into digital object, lastly others metadata are applied.



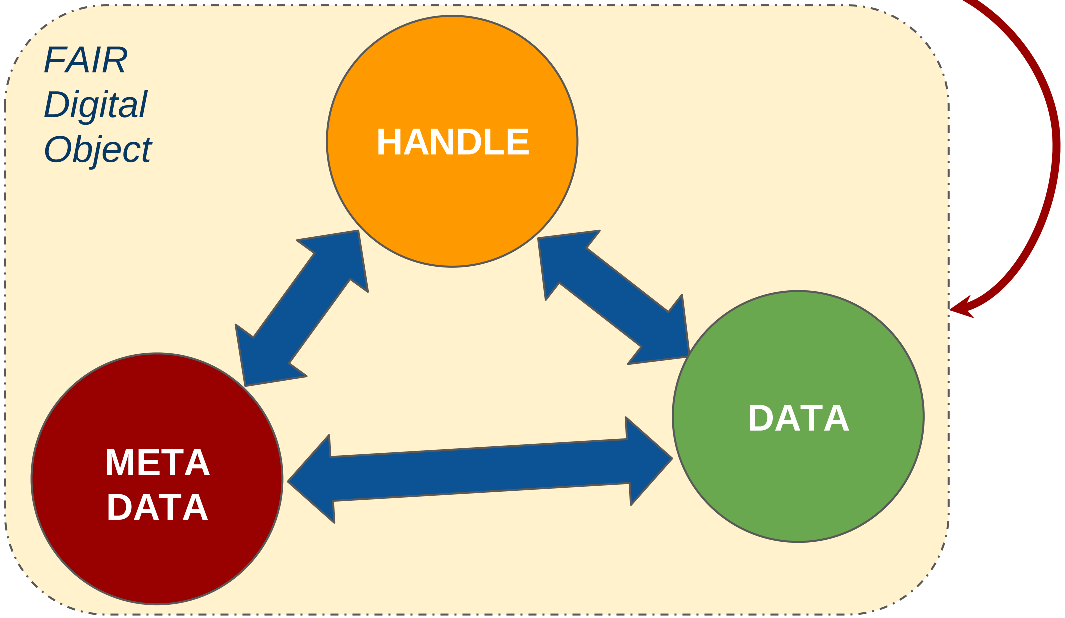
- Check-in
- Quality Checks
- Extract Meta
- Mint a PID
- PUT IN Trust Archive

- Check-out
- Remove Meta
- PULL OUT Trust Archive
- Version Update

Data Curation Data File + unique identifier = Digital Object

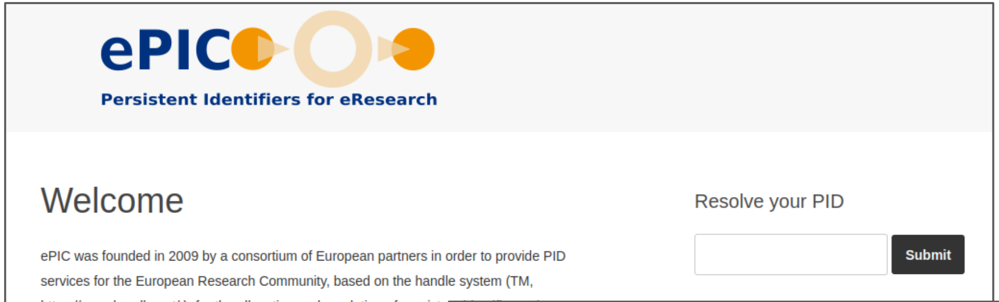


We add a Unique Persistent Identifier to all Data Files that have passed quality check & extracted Metadata:
Data File become Digital Object



Data Dissemination Handle resolver: a Digital Object approach

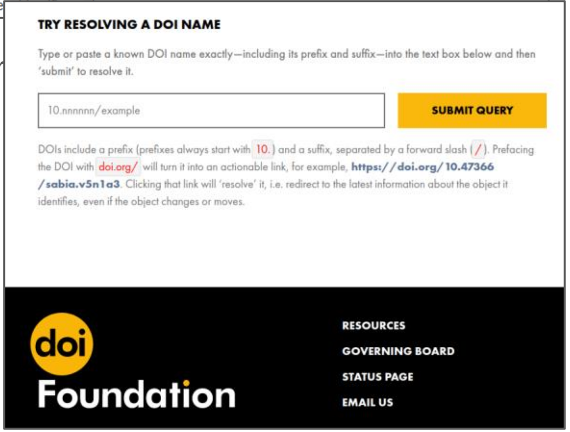
- *Provides uniform ‘information’ interface*
- *Easy to use*
- *DO Citation (ePIC-PID)*
- *Availability ensured to decades*
- *Hides system details*



Welcome

ePIC was founded in 2009 by a consortium of European partners in order to provide PID services for the European Research Community, based on the handle system (TM, https://www.handle.net/), for the allocation and resolution of persiste

Resolve your PID


TRY RESOLVING A DOI NAME

Type or paste a known DOI name exactly—including its prefix and suffix—into the text box below and then ‘submit’ to resolve it.

DOIs include a prefix (prefixes always start with 10.) and a suffix, separated by a forward slash (/). Prefacing the DOI with doi.org/ will turn it into an actionable link, for example, **https://doi.org/10.47366/sabia.v5n1a3**. Clicking that link will ‘resolve’ it, i.e. redirect to the latest information about the object it identifies, even if the object changes or moves.

https://www.doi.org/11099/11ed-996f-0242ac120005 :: latest ver

http://www.handle.net/11099/11ed-996f-0242ac120005 :: latest version

11099/11ed-996f-0242ac120005#**version=1** :: specific version

11099/11ed-996f-0242ac120005#**metadata** :: DC metadata (WF-HANDLE)

11099/11ed-996f-0242ac120005#**provenance** :: provenance (WF-PROV)

11099/11ed-996f-0242ac120005#**document** :: doc of file (Human readable)

11099/wf-search#lat; lon; rad; dstart; dend; :: list of PID in that area at given time-window

Data Preservation Long term preservation: Storage - Tape backup

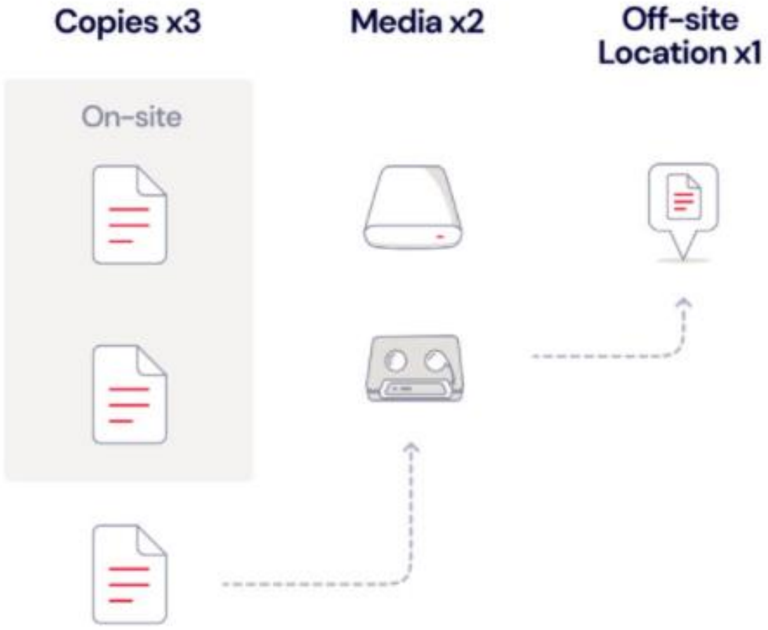
Our policy for long-term-preservation is to have 3 copies: in local storage, then make a replication at a new storage archive in Naples and a backup in a tape library, compliant to 3-2-1 backup policy.
A test suite on data backupped is performed on regular base

***Long Term**: Hardware and file systems are upgraded on regularly base in order to allow readable over time (by now last 3 decades are available).*

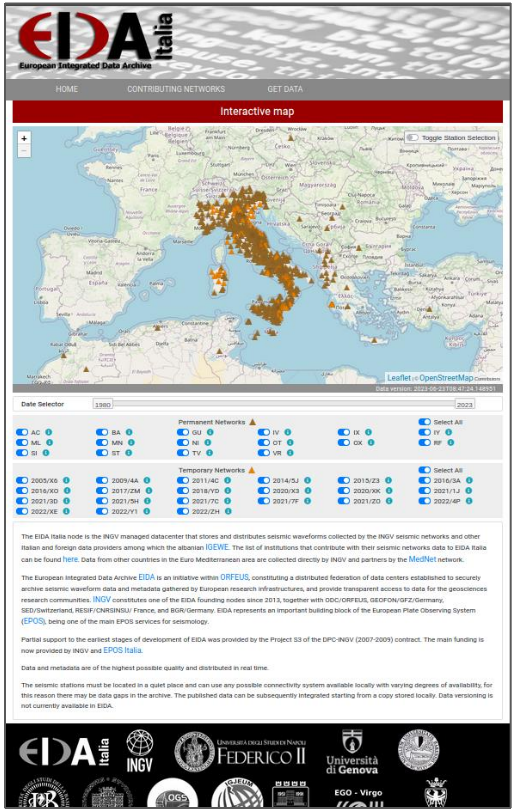
Follow 3-2-1 Backup rules coined by Peter Krogh

- **Have at least 3 copies of your data**
- **Store the copies on 2 different media**
- **Keep it safe with 1 backup copy off-site**

Tape Backup are tested cyclically on monthly base



Data Dissemination WEBsite, Web-services and so on..



EIDA Italia
European Integrated Data Archive
HOME
CONTRIBUTING NETWORKS
GET DATA
Interactive map
Toggle Station Selection
Leaflet | © OpenStreetMap contributors
Data version: 2023-06-23T08:47:24.148951

Date Selector 1980 2023

		Permanent Networks			● Select All
● AC	● BA	● GU	● IV	● IX	● IY
● ML	● MN	● NI	● OT	● OX	● RF
● SI	● ST	● TV	● VR		

		Temporary Networks			● Select All
● 2005/X6	● 2009/4A	● 2011/4C	● 2014/5J	● 2015/Z3	● 2016/3A
● 2016/XO	● 2017/ZM	● 2018/YD	● 2020/X3	● 2020/XK	● 2021/1J
● 2021/3D	● 2021/5H	● 2021/7C	● 2021/7F	● 2021/ZO	● 2022/4P
● 2022/XE	● 2022/Y1	● 2022/ZH			

The EIDA Italia node is the INGV managed datacenter that stores and distributes seismic waveforms collected by the INGV seismic networks and other Italian and foreign data providers among which the albanian IGEWE. The list of institutions that contribute with their seismic networks data to EIDA Italia can be found here. Data from other countries in the Euro Mediterranean area are collected directly by INGV and partners by the MedNet network.

The European Integrated Data Archive EIDA is an initiative within ORFEUS, constituting a distributed federation of data centers 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. INGV constitutes one of the EIDA founding nodes since 2013, together with ODC/ORFEUS, GEOFON/GFZ/Germany, SED/Switzerland, RESIF/CNRSINSU/ France, and BGR/Germany. EIDA represents an important building block of the European Plate Observing System (EPOS), being one of the main EPOS services for seismology.

Partial support to the earliest stages of development of EIDA was provided by the Project S3 of the DPC-INGV (2007-2009) contract. The main funding is now provided by INGV and EPOS Italia.

Data and metadata are of the highest possible quality and distributed in real time.

The seismic stations must be located in a quiet place and can use any possible connectivity system available locally with varying degrees of availability, for this reason there may be data gaps in the archive. The published data can be subsequently integrated starting from a copy stored locally. Data versioning is not currently available in EIDA.

After all these previous steps we are ready to disseminate data through web services and website.

Web site:

- EIDA Italia (https://eida.ingv.it/): frontend for browsing and downloading Italian Stations data and metadata belonging to Italian EIDA node.

Web Services:

- FDSN StationXML: nosql xml based system (eXist-db) provides APIs for download stations metadata based on FDSN standard.
- FDSN Dataselect: SeisComP based system provides APIs for download waveforms data in mseed format.
- FDSN availability based on mongo database.
- EIDA WS
  - Routing: find correct url to download across EIDA federation
  - WF-Catalog: data metrics; data quality; data availability

Data Dissemination Computational Archive: processing as a service

To serve users we are adding significant computational resources and an adequate processing and analysis framework combining Apache Spark, and ObsPy, creating a "computational archive" where storage resources and computational resources converge.

Worker resource (aggregated)

CPU: 384 cores (8x2x24); AMD Epyc 7352, 2.3 GHz base clock

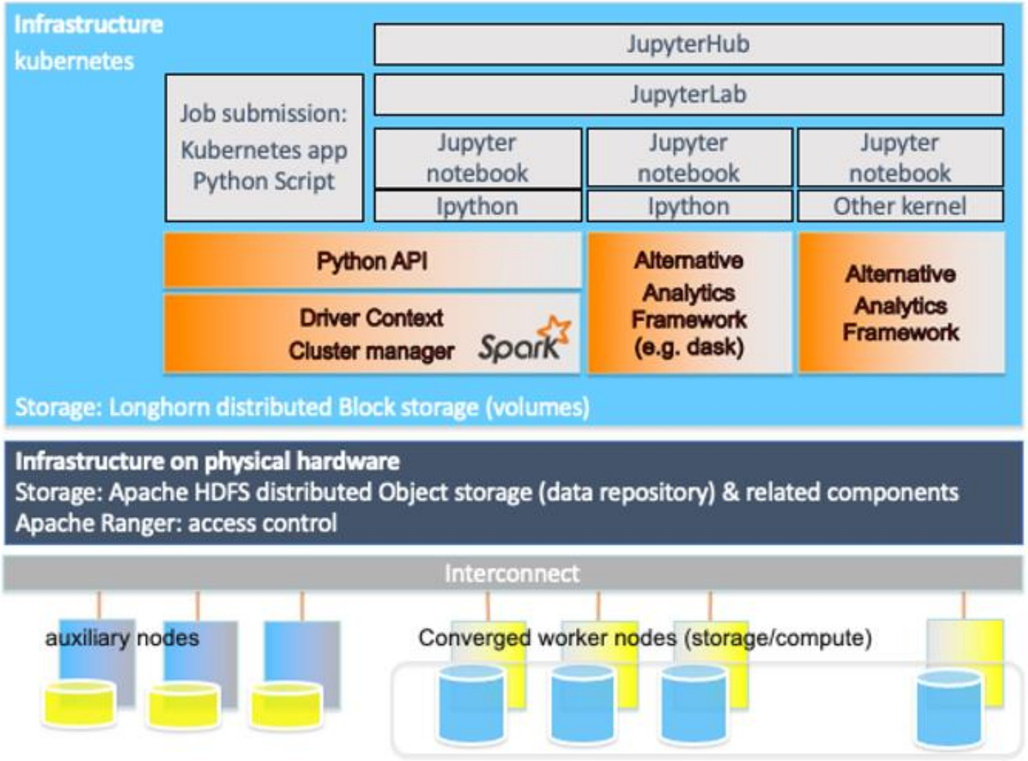
RAM: 8 TB; DDR4

Mem bandwidth: 204.8 GB/s /CPU

Storage resources (aggregated)

SAS HDD: 1440 TiB (8x10x18 GiB)

SAS SSD: 61,440 GiB (8x2x3,840 GiB)


Infrastructure
kubernetes
JupyterHub
JupyterLab
Job submission:
Kubernetes app
Python Script
Jupyter notebook
Ipython
Jupyter notebook
Ipython
Jupyter notebook
Other kernel
Python API
Driver Context
Cluster manager
Spark
Alternative Analytics Framework (e.g. dask)
Alternative Analytics Framework
Storage: Longhorn distributed Block storage (volumes)
Infrastructure on physical hardware
Storage: Apache HDFS distributed Object storage (data repository) & related components
Apache Ranger: access control
Interconnect
auxiliary nodes
Converged worker nodes (storage/compute)

Sostenibilità

Fondi Istituzionali

Allegato A DPC WP 3

JRU EPOS-Italia OS1

Progetti: Centro Italia DL-50 OR AA

Cosa manca

Una politica dei dati in streaming

Repository INGV dei Dati repo.data.ingv.it

Grazie