Seismological Research Letters

The Italian Node of the European Integrated Data Archive

--Manuscript Draft--

Manuscript Number:	SRL-D-20-00409R2
Full Title:	The Italian Node of the European Integrated Data Archive
Article Type:	Focus Section - European Seismic Networks
Corresponding Author:	Peter Danecek Istituto Nazionale di Geofisica e Vulcanologia Rome, Rome ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Istituto Nazionale di Geofisica e Vulcanologia
Corresponding Author's Secondary Institution:	
First Author:	Peter Danecek
First Author Secondary Information:	
Order of Authors:	Peter Danecek
	Stefano Pintore
	Salvatore Mazza
	Alfonso Mandiello
	Massimo Fares
	Emiliano Della Bina
	Ivano Carluccio
	Diego Franceschi
	Milena Moretti
	Valentino Lauciani
	Matteo Quintiliani
	Alberto Michelini
Order of Authors Secondary Information:	
Manuscript Region of Origin:	ITALY
Suggested Reviewers:	Davide Scafidi davide.scafidi@unige.it
	Peggy Hellweg peggy@seismo.berkeley.edu
	Chad Trabant chad@iris.washington.edu
	Petr Kolinsky Has provided a first review with useful suggestions.
Opposed Reviewers:	

The Italian Node of the European Integrated Data Archive

Danecek P., Pintore S., Mazza S., Mandiello A., Fares M., Carluccio I., Della Bina E., Franceschi D., M. Moretti, Lauciani V., Quintiliani M., Michelini A

Affiliation:

Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Roma, Italy

Corresponding author:

Peter Danecek,

Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti,

Via di Vigna Murata, 605, 00143 Roma, Italy

peter.danecek@ingv.it

Abstract

The Orfeus European Integrated Data Archive (EIDA) provides a federated approach to the dissemination of seismological waveform data and ensures access to 12 regional seismological data centres – the EIDA nodes. The Istituto Nazionale di Geofisica e Vulcanologia (INGV), is one of the founding partners of this EIDA federation and manages the EIDA data distribution node in Italy. INGV has actively managed the smaller MedNet archive since 1990 and adopted a more comprehensive and systematic approach to seismological data archiving since 2007. The Italian EIDA node data archive currently totals 90 TBytes of waveform data available for download, originating from 25 networks and 974 stations, provided by INGV, MedNet or contributed by various partner institutions. Geographically it covers mainly Italy and some stations from the Mediterranean region. The archive is currently growing at a rate of approximately 11 TB/year.

INGV recently strengthened its data management capabilities, resources and infrastructure in order to effectively respond to growing scale of station inventory, archive and volumes of delivered data, and to acknowledge increasing attention towards open data sharing, appropriate attribution and FAIR principles, as well as higher demands on data quality and expectations of the scientific user community. To this end, it established a dedicated internal unit in charge of all relevant activities related to the Italian EIDA node. In this paper we address key aspects of EIDA node in Italy, such as evolution and status of the seismological waveform archive, and we describe the technical, organisational and

operational set-up of data and service management. We also outline ongoing activities and future evolutions aiming to further increase the quality of services, data availability, data and metadata quality, resilience and sustainability.

Introduction

The hosting Institution: INGV

The Istituto Nazionale di Geofisica e Vulcanologia (INGV) was founded in 1999 by integrating the former Istituto Nazione di Geofisica (ING), the Volcano observatories at Naples (Vesuvius Observatory) and Catania (Etna Observatory), and other laboratories and institutes. This merger resulted in the biggest non-university research institution with a focus on Geophysics and Earth Sciences in Italy, and one of the largest institutions of its kind in Europe and worldwide. Besides its vast range of research activities, INGV has an explicit mandate to provide seismic surveillance and data sharing services established by the instituting law and its statute (Decreto Legislativo N. 381/1999). It is an integral component of the civil protection system in Italy, both as a centre of expertise on seismic, volcanic and tsunami risks, and as provider of surveillance, monitoring or alert services for seismic, volcanic and tsunami hazard through agreements with the Civil Protection Department (Dipartimento di Protezione Civile, DPC).

In order to fulfil these objectives, INGV operates a permanent seismic network: The Italian National Seismic Network (Rete Sismica Nazionale - RSN, INGV Seismological Data Centre, 2006; Michelini et al. 2016) has its origins as pure surveillance infrastructure, but evolved due to technological updates and improved instrumentation into a more versatile research infrastructure. In contrast, the Mediterranean Very Broadband Seismographic Network (MedNet, MedNet Project Partner Institutions, 1990; Pondrelli et al. 2019) was initially established as a pure research infrastructure operated offline that gained relevance to seismic surveillance when real-time data transmission was added. Other national institutions contribute to the coverage of the territory and data is exchanged with neighbouring countries. Moreover, INGV operates three operational rooms with staff on duty twenty-four-seven in Rome, Naples and Catania, and a monitoring service for induced seismicity, run at INGV Bologna, for the Ministry of Economic Development (Ministero dello sviluppo economico). The earthquake surveillance system is described in further detail in Margheriti L. et al. (2021, same volume) and for more details on the tsunami warning system refer to Amato A. et al. (2021, same volume).

Optimizing resources: Combining surveillance and research

The control room in Rome provides primarily seismic surveillance at the National level and constitutes a natural aggregation point where a substantial part of seismological waveform data is received in real time from stations across Italy, adjacent countries and the Mediterranean region. It is therefore an obvious choice to locate the National seismological waveform archive close to this seismic surveillance facility. In fact, for many years the waveform archive and associated services were managed jointly with the surveillance services (Mazza et al. 2012). This approach and development were realized within the European Network of Research Infrastructures for European Seismology (NERIES) and the follow-up project Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation (NERA 2010-2014) (https://www.orfeus-eu.org/organization/projects/). The same project ultimately led to the formal establishment of the Orfeus European Integrated Data Archive (EIDA) working group (now a service management committee within ORFEUS), a federated approach to seismological data distribution in which INGV participated as founding partner with a primary node (Strollo et al. 2021, same volume). On the other hand, the standardised EIDA data distribution services started to support surveillance procedures and products.

Drawbacks of a combined approach

Combining surveillance, data archiving and distribution services had advantages in terms of cost and resource efficiency, as it allowed for sharing of the underlying infrastructure and effort, like networking, hosting, implementation, developments and management. The drawback, however, was that in a resources-limited environment dominated by the requirements, urgencies and importance of a very critical, always-available twenty-four-seven agency service of public interest, it occurred that rather specific tasks and developments with no immediate relevance to the surveillance services suffered from a lack of adequate attention and resource allocation. In order to overcome this limitation and to acknowledge the increasing attention towards open data sharing, appropriate attribution and FAIR principles for scientific data management (Wilkinson et al. 2016), the constantly growing scale of the station inventory, of the archive, of the volumes of delivered data, as well as higher expectations with respect to data, metadata and service quality from the scientific user community, it became necessary to provide additional resources. As a result, in 2019 a dedicated INGV internal team for data archive and data services management was established. Its main responsibilities largely coincide with all activities relevant to the EIDA federation system and its Italian primary node: managing the computational infrastructure of the data centre for the surveillance room and for the data archive; data acquisition from data providers; data archiving and curation; metadata management; sharing and distribution of data. While retaining largely the synergies of an integrated approach for data acquisition, data and metadata management, surveillance and data distribution services, the scope is to ensure adequate exploitation of the data assets and attention to the evolution and development of further data related services.

In this paper we describe key aspects of the EIDA node in Italy, providing details of the archive, covering the organizational and technical setup, infrastructure, data acquisition, operational procedures, challenges, project activities and future plans.

Seismological waveform archive

Early stages

Today's seismological waveform archive at INGV originates from the establishment of the MedNet Data Centre. The very first MedNet very broadband stations with high dynamic range and native digital recording have been installed since 1988 in Italy and in the EuroMed region. In fact, the first continuous waveform data available at the Italian EIDA node are from the MedNet network and for 15 years (1990-2004) MedNet stations remain the only available data in the archive. Initially, these stations were operated purely offline and data was recorded on magnetic tape. The MedNet Data Centre had among others the laborious task of handling these tape records by downloading the data therein, converting, processing, verifying and validating the data, and managing the archived data (Beranzoli et al. 1993). The data (modest amounts by today's standards) were provided upon request by copying them on magnetic media (or other storage media) and sending them by regular mail.

At the same time, the National Seismic Network (RSN) consisted of approximately 100 analogue mostly vertical short period stations which were used primarily for seismic surveillance and monitoring purposes. The stations transmitted their analogue signals to the surveillance operations room via telephone lines or via radio transmission, which limited their dynamic range and signal quality. These signals were continuously recorded only on paper rotating drums. Since 1984 these analogue signals were also digitised for automatic processing by a quasi-real-time system on a Digital VAX microcomputer (Console R., Taccetti Q. 1986). Although the digitalised event waveforms have been partly preserved, the lack of detail and quality of instrumentation information hindered the inclusion into the EIDA archive.

The Digital Broadband National Network

From 2001 several projects targeted the modernization and digitalization of the RSN which resulted in improved geographic coverage, improved data quality and rapid automatic real-time monitoring products (Amato et. al 2006). Unfortunately, at the very beginning of the network upgrade, the quality of station information was not given equal attention owing to sometimes casual reporting and updating of the station metadata. In addition, the use of a proprietary protocol and data format became an obstacle to systematic data archiving. Data from the upgraded seismic stations have become regularly available only since 2005, extending the acquisition and archiving procedures already in place for MedNet data. Successively, a steadily growing number of digital stations was integrated into a modern data archiving and distribution system. This system based on miniSEED format (SEED 2012) and Seedlink protocol (Weber et al. 2007; Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences and GEMPA GmbH 2008), now a de-facto standard in seismology, was developed in the context of nationally funded projects and the European project Network of Research Infrastructures for European Seismology (NERIES 2006-2010). Since 2008 data from other collaborating institutions have been acquired and archived, notably from neighbouring countries and from local networks in northern Italy (e.g. FDSN network codes: GU, NI, RF, SI; see Table 1 for more details). Data from the later networks are now provided through the Italian EIDA node.

Current archive content

At present the Italian node EIDA waveform archive totals at 90 TB of data available for download, originating from 25 networks (including FDSN network TV, a collection of temporary deployments) and 974 stations. To a large extent these data are provided by the INGV managed networks, permanent or temporary, but a considerable portion of the data holdings are nowadays contributed by other Italian institutions and universities, and are provided to EIDA users via the Italian node. We also receive data from neighbouring countries, which are processed and stored for seismic monitoring purposes, but only data from the Albanian network (FDSN network code AC) are currently made available to the EIDA federation. Figure 1 provides an overview of the geographical coverage of all permanent stations present in the archive of the Italian EIDA node. Most of the other seismic network data utilized to accomplish national monitoring and surveillance services (i.e., mainly the stations near the borders of Italy in the neighbouring countries) are available via other nodes of the EIDA federation, in particular RESIF, ETH, ODC, GFZ, NOA (see Data and Resources), with the exception of data from the Italian Strong Motion Network (Rete Accelerometrica Nazionale, RAN, FDSN network code IT). The RAN data are not yet available within EIDA. We summarise in Table 1 all data providers which contribute data to the Italian EIDA node and the respective networks. The relative proportion to the overall data holdings is shown in Figure 2 and the year-by-year evolution of the archive is illustrated in Figure 3.

The archive at the time of writing grows at a rate of approximately 11 TB/year and the station inventory currently declares 532 permanent stations as active, of which 122 are contributed by various partner institutions and 410 are permanent stations managed by INGV or MedNet (The reported numbers are valid for end of October 2020). The predominant portion of these active stations are equipped with velocity sensors (475 stations), and out of these nearly 80% consist of broadband stations (372 broadband vs. 103 short period stations). The prevailing sensor type is the Nanometrics Trillium 40s, due to its large number at INGV's National Seismic Network. Strong motion sensors are presently co-located at somewhat less than half of the seismometer stations (195 stations), mainly in combination with broadband sensors (156 stations), and a few stations are equipped with a strong motion sensor only (57 stations). The geographical distribution of the various sensor types is provided in Figure 4. The archive up to date holds relatively few data from temporary deployments by INGV (34 deployments). Some of these (23 deployments) are currently inserted with permanent network codes TV or IV for historical reasons, which however is expected to change in future. Most temporary deployments have been installed in the framework of Sismiko rapid deployments service (Moretti et al. 2016). Two temporary datasets are part of larger deployments with data distributed across several nodes of the EIDA federation: The Argostoli basin experiment performed within the NERA project and the AlpArray experiment (AlpArray Seismic Network 2015). The data archived at our EIDA node from the modest number of 16 AlpArray stations still account for approximately 26 % of the temporary data volume. The total number of temporary stations is approximately 357, and out of these currently only 16 temporary stations are declared active and functioning. At the time of writing no other types of sensors (rotation, infrasound, OBS) are included in the archive, though plans and discussions suggest that such instruments will be added in future.

Infrastructure and technical setup

Physical and virtual infrastructure

Currently the hosting hardware and infrastructure for the Italian EIDA node is located in Rome at the data centre shared with the seismic surveillance system. In this way we retain synergies, between the teams devoted to data archiving and seismic surveillance. The hosting hardware comprises a limited number of bare metal servers, for all I/O intensive workloads (notably for the central database management systems and real-time data streaming) and a virtualization infrastructure. The bare metal servers realize a twofold data acquisition chain in an active-active configuration and the primary MySQL database server in high availability. The virtualization infrastructure is split in a main and a backup VMWare cluster and hosts approximately 100 of virtual machines. This virtualization infrastructure is planned to be superseded by an OpenStack-based in-house cloud distributed among three different locations. A new server cluster has been installed in order to deploy this solution. The archive is held in a network attached storage backed up by a twin hardware. This data centre has a dedicated point of access to the Internet, which is separated from the rest of the facility of INGV's head office in Rome.

Metadata management

One of the more challenging tasks in the data centre is to keep track of all the stations and the related information regarding setup, configuration and status of the instrumentation and transmission. This is particularly crucial in an environment where a very large number of persons is involved in the overall process of maintaining and managing stations, transmission paths based on heterogeneous technologies, data acquisition, metadata and instrumentation information, as well as their use in the context of a surveillance service. In order to deal with the increasing complexity and scale of stations, operations, continuous changes and updates at INGV's data centre already back in 2007 SeisNet/SeisFace was developed and introduced (Pintore et al. 2012). This highly integrated database and management tool combines a series of aspects of seismic network management into one single web interface, and introduces an abstraction from the technical configuration and implementation of the acquisition system. In this way it becomes possible to divide the management efforts among a larger group of people with heterogeneous technical skills. In particular, SeisFace integrates the management of:

1. General information about individual stations (logistics, contacts, technical details and transmission paths);
2. Instrument templates with nominal instrument and response information;
3. Instrumental metadata, response information and acquisition configuration of specific stations;
4. Transmission protocol and connection details for system configuration;
5. Configuration and flags for usage in the surveillance system or public access;
6. Monitoring of station status, dataflow and latencies;
7. Tracking of seismic station and transmission issues (faults and repairs).

SeisFace has some provisions to manage the workflow for insertion or updating of stations information and metadata, which is initiated by a request for modification. Subsequently, this request is revised for correctness and accepted, or rejected and reiterated. This protocol and review process by a person which acts as a sort of *gatekeeper* is particularly important at INGV because possible errors would have a direct impact on the surveillance services and products (e.g., incorrect sensor metadata would result in wrong ground motions estimates inserted into the maps of ground shaking routinely calculated by INGV). The information and data inserted through the SeisFace interface are inserted into the SeisNet database hosted on the shared DBMS infrastructure. This database is queried by our FDSN compliant station webservice developed in-house (see Data and Resources).

Dataflow

In our organizational setup, the seismological data centre at INGV acts as an intermediary collecting and aggregating data from various seismic network operators, the data providers (including INGV's networks and networks provided by partner institutions) and in turn provides services and data to the National seismic surveillance services as well as to the scientific user community. The importance and relevance of the surveillance services at INGV implies particular attention, high service level and privileged access.

The sketch in Figure 5, illustrates the simplified dataflow in the seismological data centre. Configuration of data acquisition is edited in SeisFace and inserted in the underlying database SeisNet. This information is then used to configure the various components of all acquisition systems. The real-time data acquisition is handled by two different protocols: while SeedLink is now dominating as the principal protocol in our data centre, we still have a series of stations with legacy Nanometrics instrumentation which transmit data via the proprietary protocol to multiple instances of NaqsServer. There are other protocols in use in various networks that are not visible to our data centre, as they are handled and converted to SeedLink by the respective data providers. All data, from intermediate servers or directly from single stations, is aggregated at a central pair of two independent SeedLink servers for redundancy. All data acquired at the Nanometrics servers are also passed to these two Seedlink servers using the in-house developed plug-in nmxptool (Quintiliani 2007). In order to allow for retransmission of lost data and gap filling, the data stream to the redundant SeedLink servers may be delayed by up to 10 minutes. Therefore, these Naqs streams are also passed directly and without this delay to the real-time processing system of the surveillance services. The real-time data streams are further relayed to two other hosts via SeedLink for data distribution. One dedicated host (discovery) is used to exchange data with partner institutions, the other data stream ensures that also real-time data is available in EIDA through FDSNWS dataselect webservice (see Data and Resources).

The real-time seismic waveform data from the stations are received and stored in miniSEED format and SeisComP data structure (SDS; Weber et al. 2007; Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences and gempa GmbH 2008) as two principal and independent master copies. This redundancy is necessary for sufficient resilience of the overall system, and to allow for maintenance and reconfiguration of these central servers. Seismic waveform data archiving occurs by an explicit and automated process which starts after midnight when the daily miniSEED files stored in the SDS directory structure have been completed and closed. Before archiving the two independent copies are merged. If the data from a station are incomplete and the station is appropriately configured for this purpose, an attempt to recover the missing portion of the data from the local copy at the station is triggered. This attempt to recover data may be repeated during a limited period within the following days. The consolidated version of the single miniSEED files is copied to the waveform archive that is hosted on our dedicated storage cluster infrastructure. The archive data is then subject to further processing and analysis in order to collect the necessary data for the EIDAWS WFcatalog (Trani et al. 2017), for SQLX (see Data and Resources) and other in-house scripts for quality control purposes.

Technical details of the EIDA service implementation

A dedicated host and instance of SeisComP is then used exclusively for FDSNWS dataselect service. FDSNWS station is based on an in-house implementation of this standard service which accesses directly our SeisNet database. This development was motivated by the need to ensure correctness and consistency of the inventory and instrument response information provided to the surveillance system and to the scientific user community. The MongoDB document store database for WFcatalog, the related EIDAWS wfcatalog webservice, the EIDAWS routing services (Quinteros, J. 2017) and further smaller internal microservices are deployed as docker containers (see Data and Resources) on dedicated virtual hosts deployed on the virtualization infrastructure. All public EIDA webservices are accessed through one unique nginx (see Data and Resources) reverse proxy. The surveillance system utilises the same standardised webservices to access inventory and station response information for configuration. It also uses the EIDA webservices to retrieve waveform data for analysis and to generate products after the initial detection and localization.

Activities and future evolution

The new organization setup implemented 2019 and described above has provided us with the opportunity to address some challenges or issues present at the Italian EIDA node with more focus and intensity. Various of these activities go back well before 2019 but have been intensified since.

Hosting infrastructure maintenance and renewal

Renewal and maintenance of the service hosting infrastructure is a continuous ongoing process, and staggered in time for the various hardware components reaching their programmed end of lifetime. At the time of writing, apart from rather modest renewal of obsolete server hardware, a fundamental innovation and scaling of the virtual hosting ecosystem occurs by moving away from a moderate scale commercial solution towards an in-house cloud solution at multiple premises based on OpenStack (see Data and Resources). The scope of this solution is not only increased scale, redundancy and flexibility, but also possibility to integrate more heterogeneous hardware and, above all, to significantly improve resilience and disaster recovery including relocation of operations to different physical locations. Another important activity regards the renewal of the storage cluster. The renewal is planned for next year and the project also includes the replication to a second premise in Naples with the same target to improve resilience and disaster recovery. We are currently in the process of procuring a replacement solution for hardware and software. The project also includes the addition of a *computational archive*, a cluster and software platform which combines storage and computational resources in order to improve data access and analysis capabilities to our archives.

Metadata and data curation

Investments into infrastructure and connectivity, as well as more resources, attention and effort, led already to a noticeable improvement of EIDA service quality. The rapidly growing volume of requested data, however, requires constant evolution of the system, monitoring and elimination of new bottlenecks. Considerable effort goes into the verification and improvement of data and metadata quality. The curation of data and metadata is an ongoing process, often triggered by the need to solve specific issues which are reported or discovered, or in other occasions due to systematic analysis. Given the tight relation to the surveillance system most attention is dedicated to problems related directly to data and metadata which are effectively used in real-time earthquake detection and analysis processing. The currently established procedures, in general, ensure that metadata of stations that are newly inserted or updated are formally correct, would not create obvious issues during processing, and that data streaming is working correctly. The continuous monitoring, usage and feedback from the surveillance staff itself also provides an important element of quality control. On the other hand, the requirement to provide new station data quickly or to keep station data available, occasionally prevented a more careful analysis or even prompted “preliminary” insertion without subsequent corrections. More subtle issues regarding metadata, data quality, historical data or issues introduced during later modifications of the inventory or archive, though, may not always have received sufficient attention due to resource shortage.

Operational guidelines and best practices

Since several years we have been actively analysing and addressing data and metadata quality issues. Though we were able to intensify this effort significantly over the last year, we also realise that while on the level of the data centre we have some possibilities to detect issues, it is much less straightforward to act on them and to come up with an adequate solution. Important information can be obtained only in collaboration with the data providers. The data centre simply is not the instance to act on specific data quality issues. In particular, waveform data quality is often determined by the station instrumentation itself, e.g. compromised by deterioration of the hardware. In other cases, the quality of deployment does not match instrument performance. The logical consequence is that our effort as data centre needs to focus on acting earlier in the process by establishing well-defined operational guidelines and best-practice protocols which significantly limit the introduction of any quality problems into the archive. Such protocols, however, require a wider consensus and acceptance of established rules and procedures, and the need to include operations well beyond the data centre and the EIDA node itself. While improving archive quality will keep us busy, the intensified effort and dialogue has already noticeable effects, as illustrated in Figure 6.

A related aspect regards the establishment of well-described relationships, updated agreements with our partner institutions to clarify the details of collaboration. Their data have been provided primarily for real-time surveillance and monitoring purposes and most data providers agreed to make their data available also to EIDA. Closer collaboration and shared ownership of the data and metadata holdings of the archive will provide reciprocal benefits. More intense participation and support by data providers help the EIDA node to manage the growing scale of the archive, resolve problems and increase the data and metadata quality. This is particularly true for the data holdings from the past which require additional curation effort and are not well covered by real-time data provision only. For the data provider it helps to improve visibility and to ensure more complete and correct information on data provenience, data attribution and licensing, of which they require better control. Several initiatives seek to improve on this collaborative aspect, among others supported by Italian EPOS activities.

Recent or ongoing technical developments strongly reflect the activities on archive quality, intensifying collaborations with data providers and data attribution. Above all the newly developed website, apart from addressing better visibility of the EIDA node itself, also targets at providing improved visibility of institutions contributing their data to the Italian node, and ensuring correct data attribution. Future releases of the website should also include additional information and tools to support the assessment of station status metadata correctness. (see Data and Resources).

Next generation tools for metadata management

SeisFace has been a major step forward in inventory and data centre management, when it was introduced and it remains a central component of data centre operations. Despite all advantages of its integration, the tight relations to its database structure and initial monolithic design now make it rather difficult to adapt or extend the functionality for new requirements or to modern formats, above all to StationXML. New developments and functionality are also becoming increasingly important, in order to support the implementation of updated operational procedures and protocols, to improve structured interaction with various actors and data providers, e.g. by providing for an enhanced and fine-grained collaborative issue tracking system. Moreover, its scalability does not match the massive growth of the station inventory, the archive and executed operations. We are therefore designing and developing a new modular version of this instrument. The different functionalities of SeisFace are mapped to separate modules, and implemented, where possible, by leveraging existing open source solutions (configured or adapted for our purpose). These are complemented with in-house developments. In particular, we are using Tuleap to substitute and extend tracking of issues and network modification requests; adopting eXist-db to store the station inventory natively as StationXML documents; and exploring yasmine as stationXML template and metadata editor (see Data and Resources). Microservices and web APIs serve as abstraction layers, which support the creation of an integrated environment and user experience similar to SeisFace, and which allow for stepwise migration towards the new tools. The intention of this design is to facilitate future evolution by replacing only single modules.

Quality-driven waveform data ingestion

Systematic reanalysis of the archive holdings has provided evidence that invalid or incorrect data have been introduced into the archive in the past. Some of these problems are recurrent and could be caught automatically. Though the currently deployed archiving routines already perform a series of tasks to enhance data quality (merge, refetch), to verify and validate data before archiving, and to process archived data for quality analysis, they lack an overarching framework which could help to manage, modify and monitor the rather complex workflow. Therefore, an important objective and area of activity is the redesign and reimplementation of waveform completion, quality control, archiving and analysis process. The plan is to integrate all processing tasks into a rule-based workflow management system. This way adding or modifying tasks, managing conditions, checking the status of files and processes, or understanding which tasks have been executed, should be simplified, and manual laborious and error-prone operations significantly reduced. Once the workflow framework is sufficiently mature it could be shared with other data centres.

Discussion and Conclusion

For many years the seismological waveform data archive and Italian EIDA node were managed in the context of the National surveillance service at INGV, where data from the vast majority of Italian seismic stations arrive in real time. Despite the obvious opportunities this combination offers, we also had to recognize that more specific aspects and tasks of data archive and EIDA service management had suffered due to the overwhelming importance of the surveillance service.

This problem has been addressed with the addition of new resources and a new organizational setup which resulted in the division of tasks and focus into two dedicated teams with distinct focus, while retaining a shared computational and networking infrastructure. This allows to address some of the long-standing issues of the Italian EIDA node. The increased focus and effort on EIDA related activities, along with upgrades on the data infrastructure and network connectivity, had already their positive impact on EIDA service quality and some metadata and data issues. We also have to acknowledge that the data centre itself has only a very limited impact on waveform data quality, but can produce quality metrics and provide prompt feedback. Significant improvements of the quality of seismic waveform data, and to some extent also metadata quality, will only be achieved if data providers manage to contribute to this important objective, and above all if sufficient resources become available for the renewal of the seismic network hardware.

To improve the interaction with data providers, we are in the process of developing guidelines, protocols and minimal standards for station management. These documents describe or extend current practice, provide major detail and should help to reduce errors and effort. Moreover, to facilitate collaboration, we are revising our management system and we are developing new tools to provide more useful information to data providers or to enable them to contribute to metadata and data management.

Data and Resources

The seismological waveform data and metadata described in this work are available from the EIDA, either from the Italian node (usually identified by the acronym INGV) or through federated services. Table 1 provides a complete list of networks available from the Italian EIDA node and the respective data providers. The following EIDA nodes (https://www.orfeus-eu.org/data/eida/nodes/) are identified by their acronyms in the text:

- ETHZ: ETH Zurich Schweizerischer Erdbebendienst (SED)
- GFZ: Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences

- NOA: National Observatory of Athens
- ODC: ORFEUS Data Center
- RESIF: Reseau Sismologique & Geodesique Francais

The EIDA organization, the federated services for data access are described in Strollo et al. (2021). For information about the Italian Strong Motion Network (Rete Acceleromentrica Nazionale, RAN) and data access see: http://ran.protezionecivile.it/EN/index.php, Presidency of Counsil of Ministers - Civil Protection Department (1972).

In this work we used or referenced information from various online resources as listed below. Last access occurred in October 2020.

- Orfeus and EIDA federation: http://www.orfeus-eu.org/data/eida/; https://www.orfeus-eu.org/organization/projects/
- Italian EIDA node: http://www.eida.ingv.it
- Documentation about usage of FDSNWS station, FDSNWS dataselect, EIDAWS routing, EIDAWS WFCatalog, EIDA federator: http://www.orfeus-eu.org/data/eida/webservices/
- FDSN web service specifications: https://www.fdsn.org/webservices/

Various open-source software products are key to the operation of the data centre and EIDA node, or the software is explicitly mentioned in the text:

- SeisComP: Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences and gempa GmbH (2008)

- Nmxptool: Quintiliani (2007); https://gitlab.rm.ingv.it/matteo.quintiliani/nmxptool
- MySQL: https://www.mysql.com
- Xataface: http://xataface.com
- SQLX: https://sqlx.science
- nginx: https://www.nginx.com
- OpenStack: https://www.openstack.org
- Docker: https://www.docker.com
- Tuleap: www.tuleap.org
- eXist-db: http://www.exist-db.org
- yasmine: https://github.com/iris-edu-int/yasmine-stationxml-editor

Acknowledgments

Many colleagues and partners have contributed, through various intermediate achievements and steps, to the creation of the seismological data centre at INGV and the Italian EIDA node of today. Above all, the immense effort by network operators, at INGV and at all partner institutions, providing data to the EIDA node, made it possible to create the impressive seismological data asset that is now available to the scientific community. Colleagues from the surveillance services are constantly monitoring the status of stations, networks and other components, or provide feedback and hints useful to improve the quality of seismic data, the waveform archive and station inventory. Feedback from the scientific community is equally important. We would like to thank Petr Kolínský, Lars

Ottemöller and an anonymous reviewer for their feedback and valuable suggestions.

Funding by European projects NERIES and NERA significantly contributed to the establishment of a modern data archiving and data distribution system at INGV's seismological data centre and created the base for the participation in EIDA. The establishment of the EPOS infrastructure and associated projects (EPOS-PP, EPOS-IP) have fostered awareness of the value of proper data management and sharing culture. Furthermore, activities and infrastructure of the Italian EIDA node received institutional funds and support from various national projects of the Italian ministry of Research, e.g. EPOS Italy, PON GRINT, FISR SOIR. The continued funding provided by the Italian Civil Protection Department (Dipartimento della Protezione Civile, DPC, Presidenza del Consiglio dei Ministri) helps to improve longer-term sustainability of the archive and the EIDA node.

References

AlpArray Seismic Network. (2015). AlpArray Seismic Network (AASN) temporary component. AlpArray Working Group. doi: 10.12686/ALPARRAY/Z3_2015.

Amato, A, L. Badiali, M. Cattaneo, A. Delladio, F. Doumaz, F. Mele (2006). The real-time earthquake monitoring system in Italy, Géosciences - Révue du BRGM, no. 6 (September 2006).

Amato A., Avallone A., Basili R., Bernardi F., Brizuela B., Graziani L., Herrero A., Lorenzino M.C., Lorito S., Mele F.M., Michelini A., Piatanesi A., Pintore S., Romano F., Selva J., Stramondo S., Tonini R., Volpe M. (2021). From seismic monitoring to tsunami warning in the Mediterranean, Seismol. Res. Lett. submitted for this Focus section.

Beranzoli L., D. Giardini and N. A. Pino (1993). Seismogram processing at Mednet, Computers and Geosciences 19, 2, 167-174.

Console R., Q. Taccetti (1986). Telemetry for the Italian Seismological Network. 1st Meteosat DCP Users. Conference. ESA. Lisbon, January 1986.

European Gravitational Observatory (2019). Virgo Interferometric Antenna for Gravitational Waves Detection. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/VR.

Geological Survey-Provincia Autonoma Di Trento (1981). Trentino Seismic Network. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/ST.

Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences and GEMPA GmbH (2008). The SeisComP seismological software package. GFZ Data Services.

doi:10.5880/GFZ.2.4.2020.003.
INGV Seismological Data Centre (2006). Rete Sismica Nazionale (RSN). Istituto Nazionale di Geofisica e Vulcanologia (INGV), Italy. doi: 10.13127/SD/X0FXNH7QFY.
Institute of Geosciences, Energy, Water and Environment (2002). Albanian Seismological Network. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/AC.
Margheriti L., C. Nostro, O. Cocina, M. Castellano, M. Moretti, V. Lauciani, M. Quintiliani, A. Bono, F. M. Mele, S. Pintore, P. Montalto, R. Peluso, G. Scarpato, S. Rao, S. Alparone, S. Di Prima, M. Orazi, A. Piersanti, G. Cecere, M. Cattaneo, A. Vicari, V. Sepe, Ch. Bignami, L. Valoroso, M. Aliotta, A. Azzarone, P. Baccheschi, A. Benincasa, F. Bernardi, I. Carluccio, E. Casarotti, C. Cassisi, B. Castello, F. Cirilli, M. D'Agostino, C. D'Ambrosio, W. De Cesare, E. Della Bina, A. Di Filippo, R. Di Stefano, L. Faenza, L. Falco, M. Fares, P. Ficeli, D. Latorre, M. C. Lorenzino, A. Mandiello, A. Marchetti, S. Mazza, A. Michelini, A. Nardi, M. Pastori, M. Pignone, M. Prestifilippo, P. Ricciolino, G. Sensale, L. Scognamiglio, G. Selvaggi, O. Torrisi, F. Zanolin, A. Amato, F. Bianco, S. Branca, E. Privitera and S. Stramondo (2021). Seismic Surveillance and Earthquakes Monitoring in Italy. Seismol. Res. Lett. In preparation for this Focus section.
Mazza, S., A. Basili, A. Bono, V. Lauciani, A. Mandiello, C. Marcocci, F. Mele, S. Pintore, M. Quintiliani, L. Scognamiglio, and G. Selvaggi (2012). AIDA – Seismic data acquisition, processing, storage and distribution at the National Earthquake Center, INGV ANNALS OF GEOPHYSICS, 55, 4, 2012; doi: 10.4401/ag-6145.

MedNet Project Partner Institutions (1990). Mediterranean Very Broadband Seismographic Network (MedNet). Istituto Nazionale di Geofisica e Vulcanologia (INGV). doi: 10.13127/SD/FBBBTDTD6Q.

Michelini, A., L. Margheriti, M. Cattaneo, G. Cecere, G. D'Anna, A. Delladio, M. Moretti, S. Pintore, A. Amato, A. Basili, A. Bono, P. Casale, P. Danecek, M. Demartin, L. Faenza, V. Lauciani, A. G. Mandiello, A. Marchetti, C. Marcocci, S. Mazza, F. M. Mele, A. Nardi, C. Nostro, M. Pignone, M. Quintiliani, S. Rao, L. Scognamiglio, and G. Selvaggi (2016). The Italian National Seismic Network and the earthquake and tsunami monitoring and surveillance systems, Adv. Geosci., 43, 31–38, doi: 10.5194/adgeo-43-31-2016.

Moretti, M., L. Margheriti, and A. Govoni (2016). Rapid response to earthquake emergencies in Italy: temporary seismic network coordinated deployments in the last five years. in: D'Amico S. (Ed): Earthquakes and their impacts on Society, Springer.

OGS (Istituto Nazionale Di Oceanografia E Di Geofisica Sperimentale) and University of Trieste (2002). North-East Italy Broadband Network. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/NI.

OGS (Istituto Nazionale Di Oceanografia E Di Geofisica Sperimentale) (2016). North-East Italy Seismic Network. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/OX.

Pintore S., C. Marcocci, A. Bono, V. Lauciani, and M. Quintiliani (2012). Seisface: interfaccia di gestione delle informazioni della rete sismica nazionale centralizzata, Rapporti Tecnici INGV, 218, ISSN 2039-7941, http://istituto.ingv.it/en/le-collane-editoriali-ingv/rapporti-tecnici-ingv.html.

Pondrelli, S., F. Di Luccio, L. Scognamiglio, I. Molinari, S. Salimbeni, A. D'Alessandro, and P. Danecek (2019). The First Very Broadband Mediterranean Network: 30 Yr of Data and Seismological Research, Seismol. Res. Lett. 91(2A): 787–802. doi: 10.1785/0220190195.

Presidency of Counsil of Ministers-Civil Protection Department (1972). Italian Strong Motion Network. Presidency of Counsil of Ministers - Civil Protection Department. doi: 10.7914/SN/IT.

Quinteros, J. (2017): Routing Service: A data centre federation for the seismological community. GFZ German Research Centre for Geosciences. doi: 10.5880/gfz.2.4.2017.001.

Quintiliani, M. (2007). libnmxp e nmxptool: software Open-Source per trasmissioni dati sismici Nanometrics, Rapporti Tecnici INGV, 51, ISSN 2039-7941, http://istituto.ingv.it/en/le-collane-editoriali-ingv/rapporti-tecnici-ingv.html.

SEED (2012). SEED Reference Manual. Standard for the exchange of earthquake data, SEED format version 2.4. International Federation of Digital Seismograph Networks Incorporated Research Institutions for Seismology (IRIS), USGS, http://fdsn.org/pdf/SEEDManual_V2.4.pdf.

Strollo A., D. Cambaz, J. Clinton, P. Danecek, Ch. P. Evangelidis, A. Marmureanu, L. Ottemöller, H. Pedersen, R. Sleeman, K. Stammler, D. Armbruster, J. Bienkowski, K. Boukouras, P. L. Evans, M. Fares, C. Neagoe, S. Heimers, A. Heinloo, M. Hoffmann, P. Kaestli, V. Lauciani, J. Michalek, E. O. Muhire, M. Ozer, L. Palangeanu, C. Pardo, Javier Quinteros, M. Quintiliani, J. A. Jara Salvador, J.

Schaeffer, A. Schloemer, and N. Triantafyllis (2021). EIDA: the European Integrated Data Archive and service infrastructure within ORFEUS. Seismol. Res. Lett. in preparation for this Focus section.

Trani, L., M. Koymans, M. Atkinson, R. Sleeman, and R. Filgueira (2017). WFCatalog: A catalogue for seismological waveform data, Comput. Geosci., doi: 10.1016/j.cageo.2017.06.008.

University of Bari “Aldo Moro.” (2013). OTRIONS, Seismic networks of Gargano Area (Italy). International Federation of Digital Seismograph Networks. doi: 10.7914/SN/OT.

University of Genova (1967). Regional Seismic Network of North Western Italy. International Federation of Digital Seismograph Networks. doi:10.7914/SN/GU.

University of Trieste (1993). Friuli Venezia Giulia Accelerometric Network. International Federation of Digital Seismograph Networks. doi: 10.7914/SN/RF.

Weber, B., J. Becker, W. Hanka, A. Heinloo, M. Hoffmann, T. Kraft, D. Pahlke, J. Reinhardt, J. Saul, and H. Thoms (2007). SeisComP3 - automatic and interactive real time data processing. Geophysical Research Abstracts, Vol. 9, 09219, 2007.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 doi: 10.1038/sdata.2016.18.

List of mailing addresses

Peter Danecek, peter.danecek@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Pintore Stefano, stefano.pintore@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Mazza Salvatore, salvatore.mazza@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Mandiello Alfonso, alfonso.mandiello@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Fares Massimo, massimo.fares@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Carluccio Ivano ivano.carluccio@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Della Bina Emiliano, emiliano.dellabina@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Franceschi Diego, diego.franceschi@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Milena Moretti, milena.moretti@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Lauciani Valentino, valentino.lauciani@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Quintiliani Matteo, matteo.quintiliani@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

Michelini Alberto, alberto.michelini@ingv.it, Via di Vigna Murata 605, 00143 Rome, Italy

List of Figure Captions

Figure 1: Map of the geographical coverage with the permanent stations archived at the Italian EIDA node. Some MedNet stations beyond the map are omitted, see Pondrelli et al. (2019) for specific details on this network.

Figure 2: Overview of archive composition of the total data holdings (left), relative shares of data from contributed networks (centre) and temporary networks (right). The network TV is a collection of several temporary deployments.

Figure 3: Evolution of the yearly volume increment of waveform data holdings at the Italian EIDA node. Before 2005 only data from the MedNet network was archived. In late 2007 a modern archiving system was introduced. The volume of temporary networks (orange) is relatively low at the Italian EIDA node, while contributed data (yellow) gained increasing relevance.

Figure 4: Map illustrating the geographical distribution of different seismic sensor types of all currently active permanent seismic stations.

Figure 5: Schema (simplified) of the dataflow at the INGV data centre, the redundant layout of critical systems is omitted. All waveform data from the stations are aggregated at the principal SeedLink server, relayed to selected components in real time and archived daily by a dedicated process. See text for detailed discussion.

Figure 6: The plots show relative shares of seismic stations classified by their overall data availability in 2019 and 2020. The comparison of data availability between years 2019 and 2020 shows a noticeable improvement of this metrics. In particular, the share of stations with data availability above 99% increased from 26% to 50% and shares with low data availability (<95%) decreased. This improvement is the result of combined effort by the data centre and by data providers.

Tables

Table 1: Data providers and networks archived at the Italian EIDA node

Data provider	Network code (FDSN)	Network	Citation
Istituto Nazionale di Geofisica e Vulcanologia	MN	Mediterranean Very Broadband Seismographic Network	MedNet Project Partner Institutions (1990)
	IV	Italian Seismic Network	INGV Seismological Data Centre (2006)
	2009/4A	Emersito Seismic Network for Site Effect Studies in L'Aquila town, Central Italy	
	2014/5J	The Sardinia Passive Array experiment	
	2016/XO	EMERSITO Seismic Network, 2016 Central Italy	
	2016/3A	Seismic Microzonation Network, 2016 Central Italy	
	2017/ZM	Seismic Emergency for Ischia by Sismiko	
	2018/YD	Seismic Emergency for Molise-Italy by Sismiko	
	2020/XK	Temporary network seismic in Central Calabria	
	2020/X3	INGV SISMIKO emergency seismic network for Salemi-Italy	
	2011/4C	NERA-JRA1 Argostoli basin experiment, Greece	
	2015/Z3	AlpArray Seismic Network (AASN) temporary component	AlpArray Seismic Network. (2015).

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Centro Ricerche Sismologiche	OX	North-East Italy Seismic Network	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (2016)
	NI	North-East Italy Broadband Network	Istituto Nazionale Di Oceanografia e di Geofisica Sperimentale and University of Trieste. (2002)
Università degli studi di Genova, Dip. di Scienze della Terra dell'Ambiente e della Vita	GU	Regional Seismic Network of North Western Italy	University of Genova (1967)
Università degli studi di Trieste, Dip. di Scienze della Terra	NI	North-East Italy Broadband Network	Istituto Nazionale Di Oceanografia e di Geofisica Sperimentale and University of Trieste. (2002)
	RF	Friuli Venezia Giulia Accelerometric Network, Italy	University of Trieste (1993)
Università degli Studi di Napoli Federico II, Dipartimento di Fisica, RISSC-Lab	IX	Irpinia Seismic Network, Italy	
Università degli Studi di Bari Aldo Moro, Dip. Scienze della Terra e Geoambientali	OT	OTRIONS Local Seismic Network, Apulia, Italy	University of Bari “Aldo Moro.” (2013)
Università degli Studi della Basilicata	BA	Universita della Basilicata Seismic Network, Italy	
Provincia autonoma di Bolzano	SI	Sudtirol Network, Italy	
Provincia autonoma di Trento	ST	Trentino Seismic Network, Italy	Geological Survey-Provincia Autonoma di Trento (1981)
Institute of Geosciences, Energy, Water and Environment (IGEWE), Polytechnic University of Tirana	AC	Albanian Seismic Network	Institute of Geosciences, Energy, Water and Environment (2002)
European Gravitational Observatory	VR	Virgo Interferometric Antenna for Gravitational Waves Detection	European Gravitational Observatory (2019)

Figures

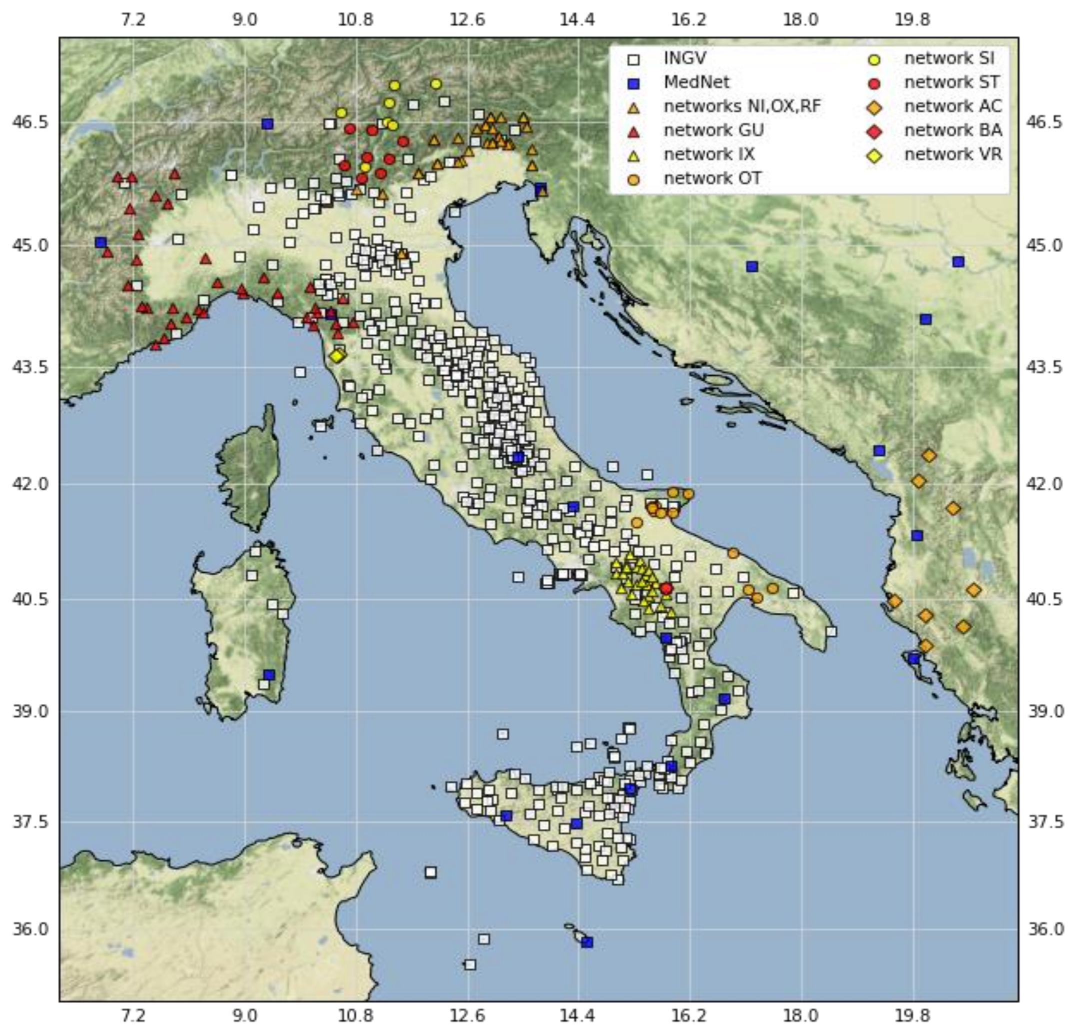


Figure 1: Map of the geographical coverage with the permanent stations archived at the Italian EIDA node. Some MedNet stations beyond the map are omitted, see Pondrelli et al. (2019) for specific details on this network.

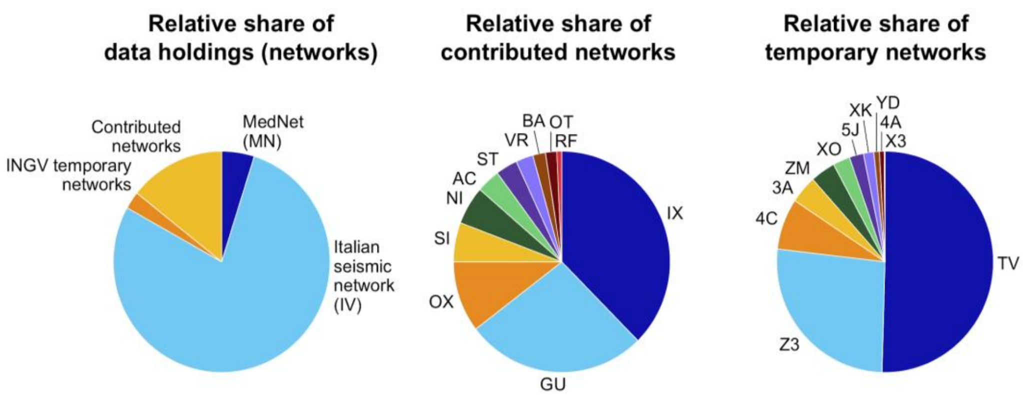


Figure 2: Overview of archive composition of the total data holdings (left), relative shares of data from contributed networks (centre) and temporary networks (right). The network TV is a collection of several temporary deployments.

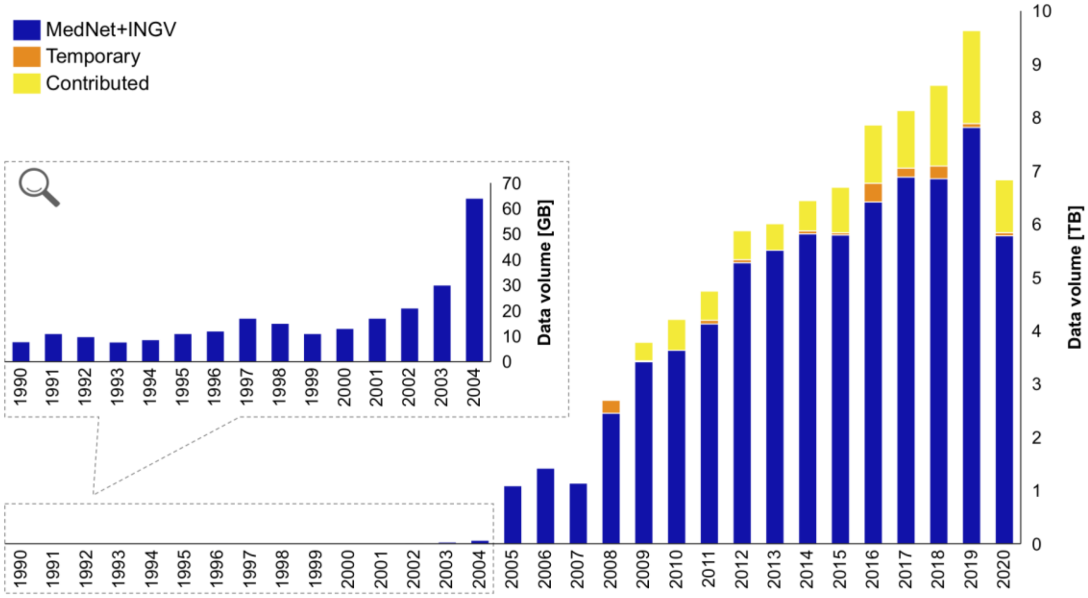


Figure 3: Evolution of the yearly volume increment of waveform data holdings at the Italian EIDA node. Before 2005 only data from the MedNet network was archived. In late 2007 a modern archiving system was introduced. The volume of temporary networks (orange) is relatively low at the Italian EIDA node, while contributed data (yellow) gained increasing relevance.

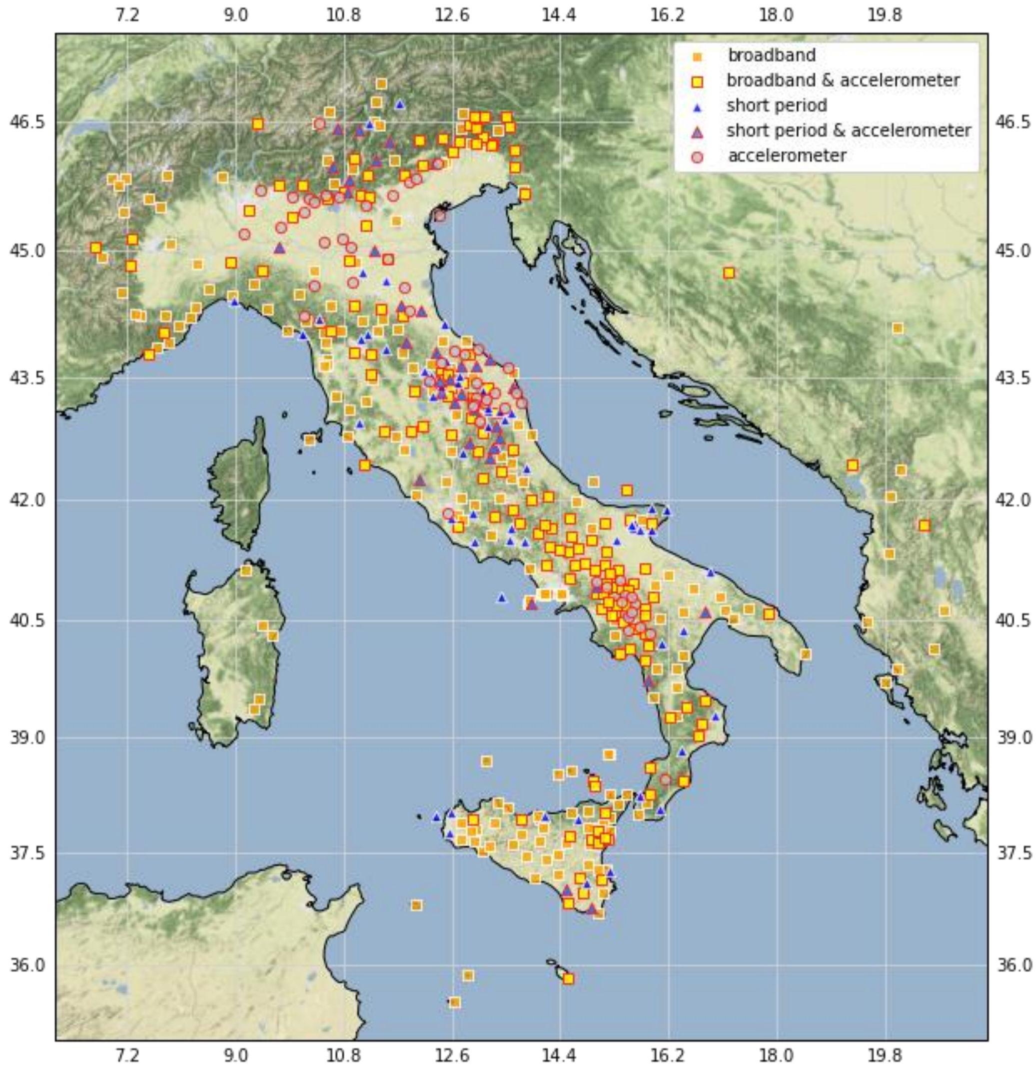


Figure 4: Map illustrating the geographical distribution of different seismic sensor types of all currently active permanent seismic stations.

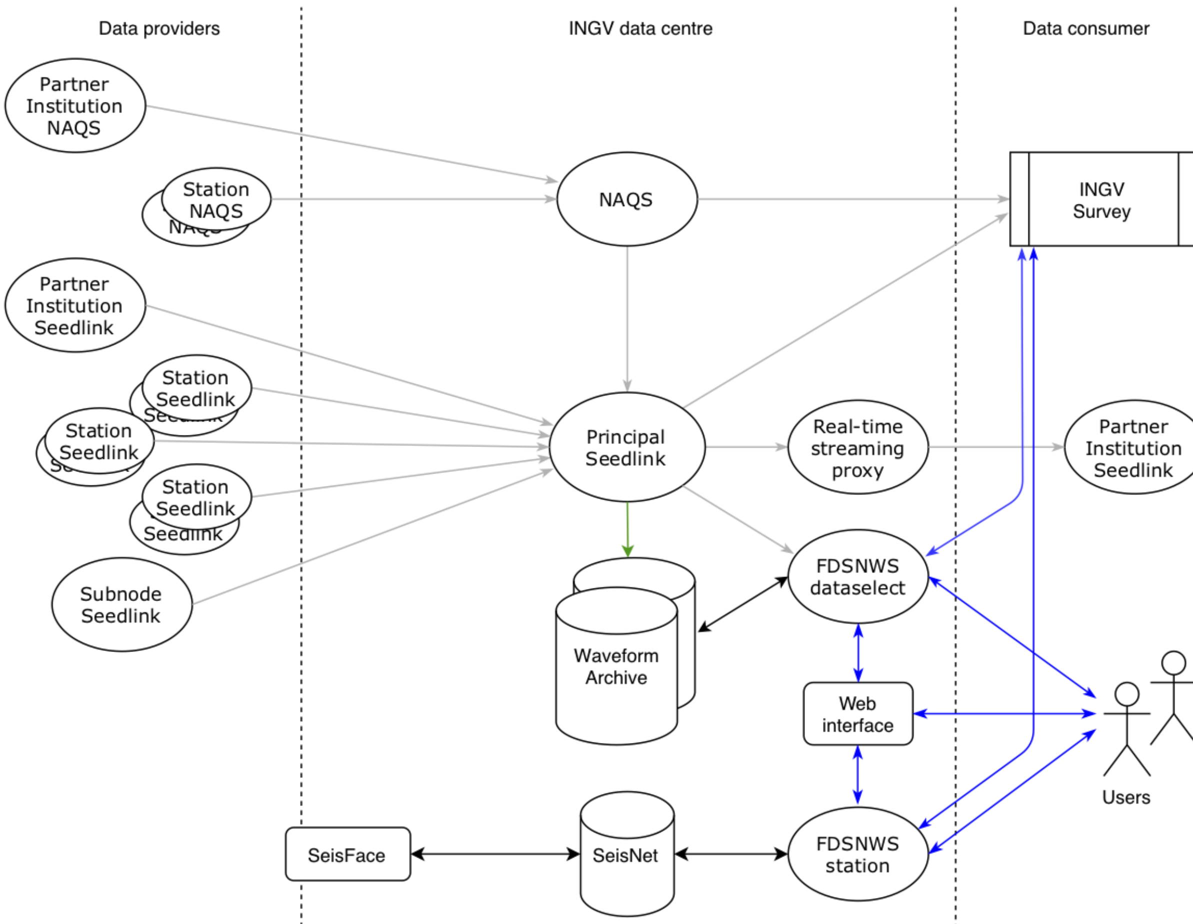


Figure 5: Schema (simplified) of the dataflow at the INGV data centre, the redundant layout of critical systems is omitted. All waveform data from the stations are aggregated at the principal SeedLink server, relayed to selected components in real time and archived daily by a dedicated process. See text for detailed discussion.

Figure 6: The plots show relative shares of seismic stations classified by their overall data availability in 2019 and 2020. The comparison of data availability between years 2019 and 2020 shows a noticeable improvement of this metrics. In particular, the share of stations with data availability above 99% increased from 26% to 50% and shares with low data availability (<95%) decreased. This improvement is the result of combined effort by the data centre and by data providers.

Figure 1

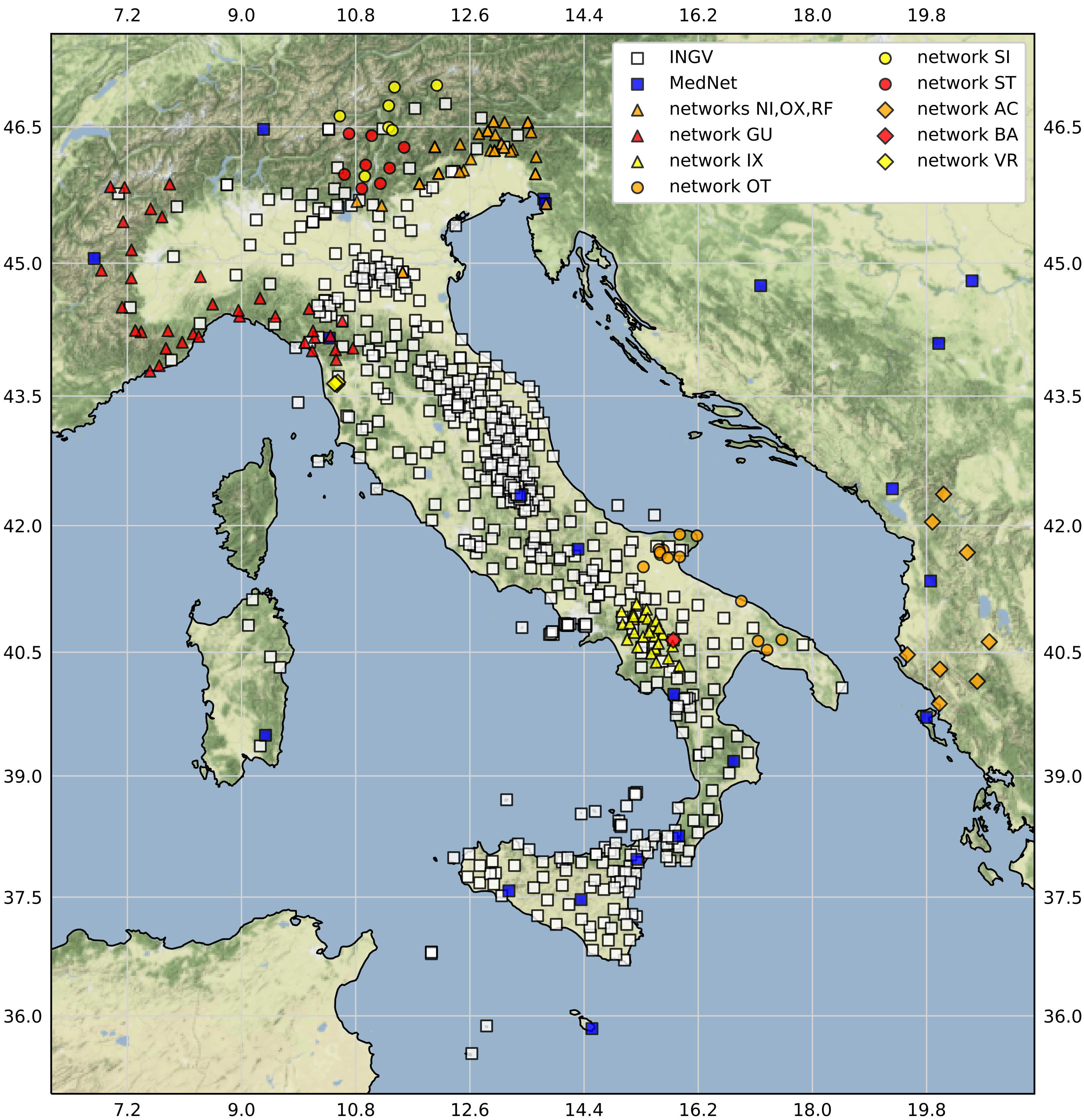

INGV
MedNet
networks NI,OX,RF
network GU
network IX
network OT
network SI
network ST
network AC
network BA
network VR
7.2
9.0
10.8
12.6
14.4
16.2
18.0
19.8
46.5
45.0
43.5
42.0
40.5
39.0
37.5
36.0

Figure 2

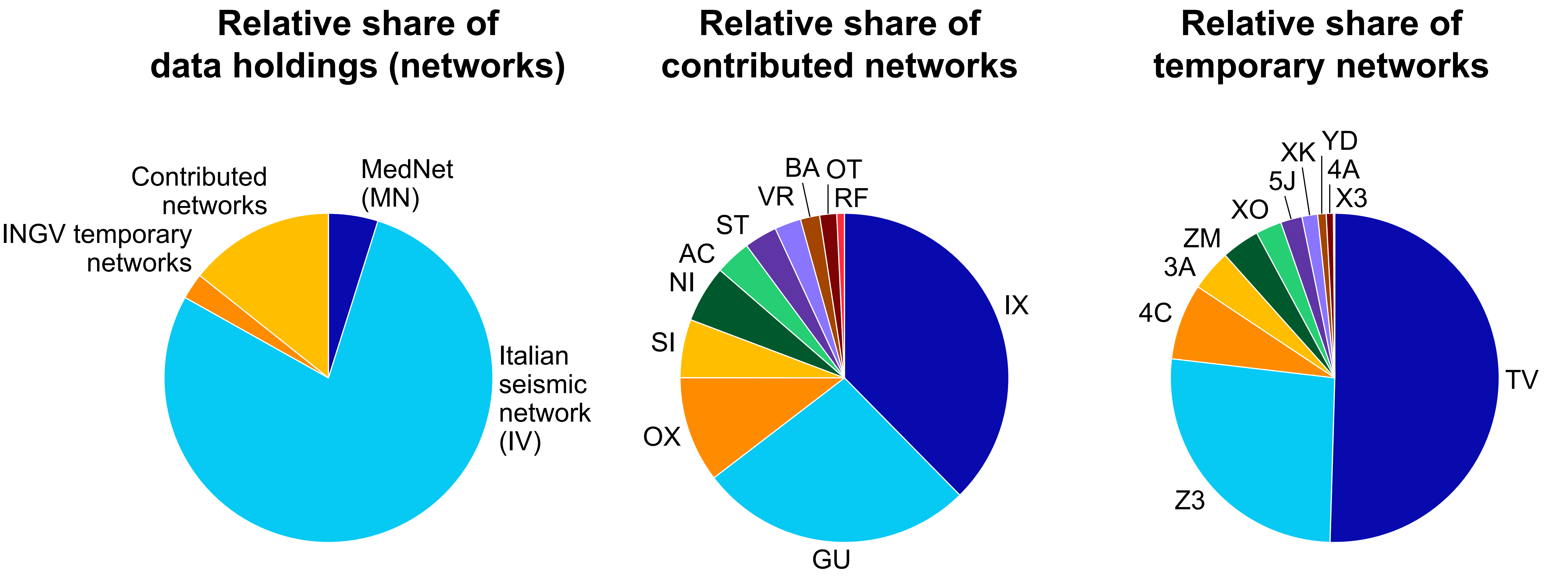
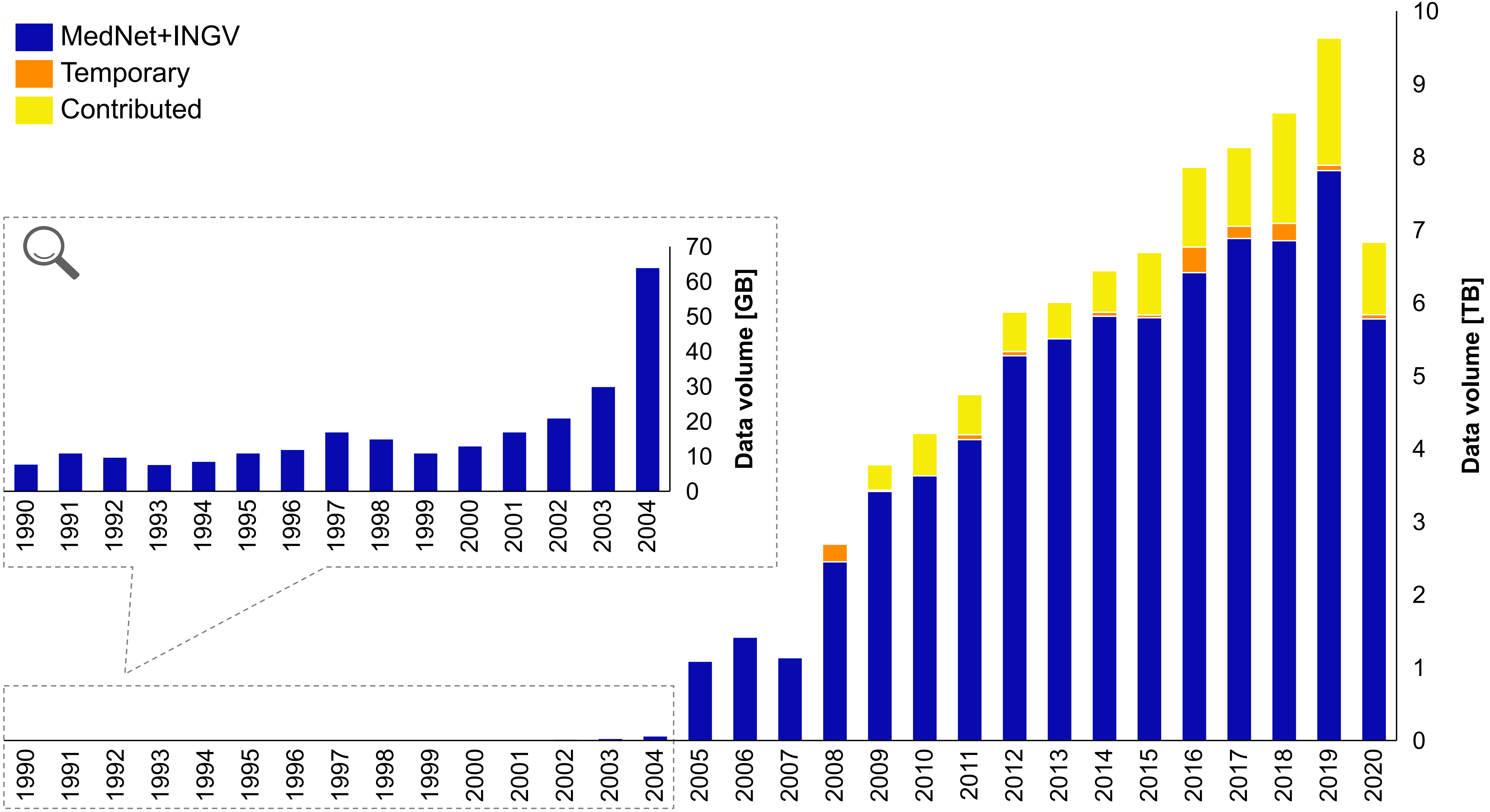
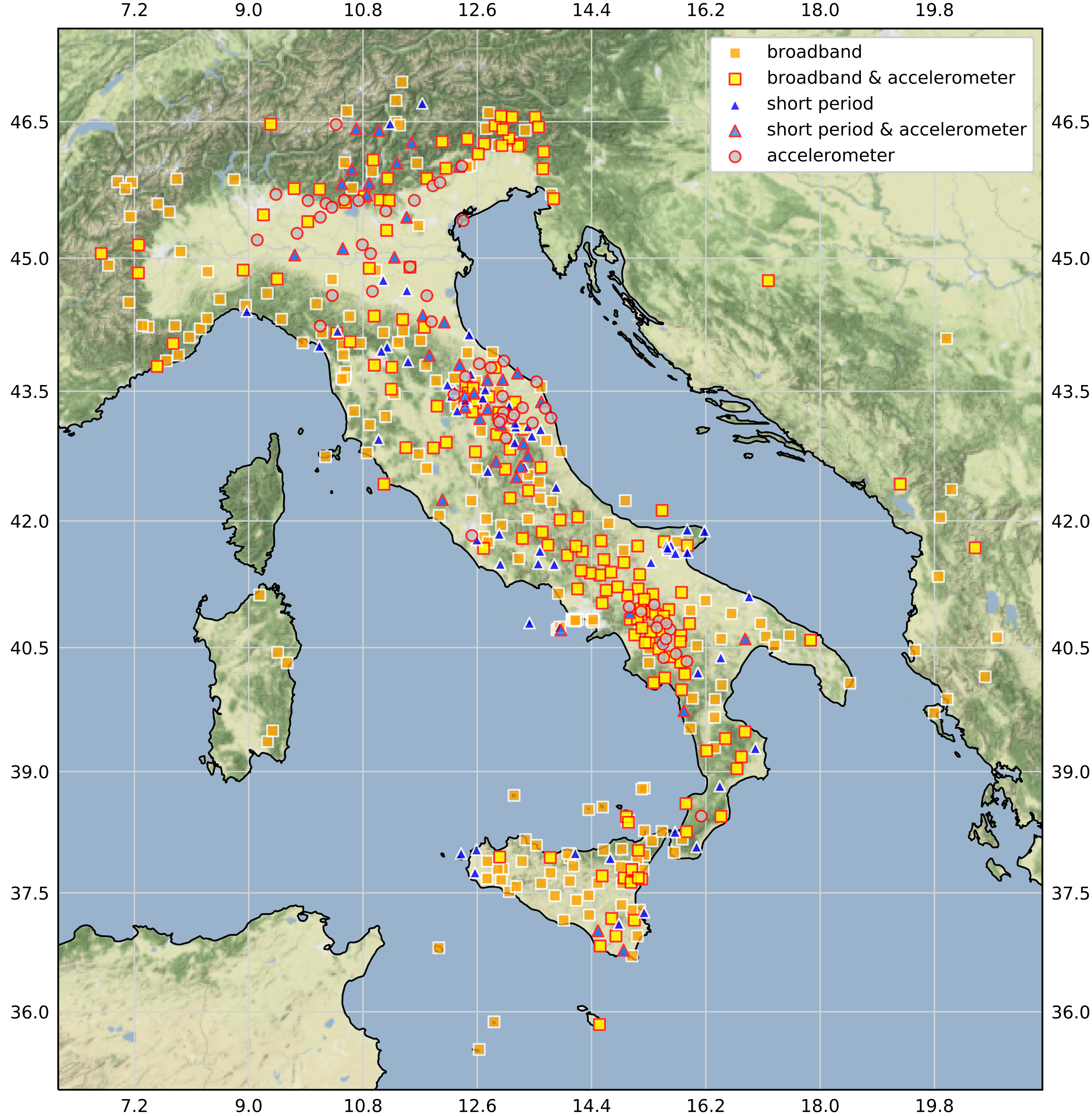

Relative share of data holdings (networks)
MedNet (MN)
Contributed networks
INGV temporary networks
Italian seismic network (IV)
Relative share of contributed networks
BA
OT
VR
RF
ST
AC
NI
IX
SI
OX
GU
Relative share of temporary networks
YD
XK
4A
5J
X3
XO
ZM
3A
4C
TV
Z3

Figure 3



MedNet+INGV
Temporary
Contributed
Data volume [GB]
0
10
20
30
40
50
60
70
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
Data volume [TB]
0
1
2
3
4
5
6
7
8
9
10
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Figure 4



broadband
broadband & accelerometer
short period
short period & accelerometer
accelerometer
7.2
9.0
10.8
12.6
14.4
16.2
18.0
19.8
46.5
45.0
43.5
42.0
40.5
39.0
37.5
36.0

Figure 5

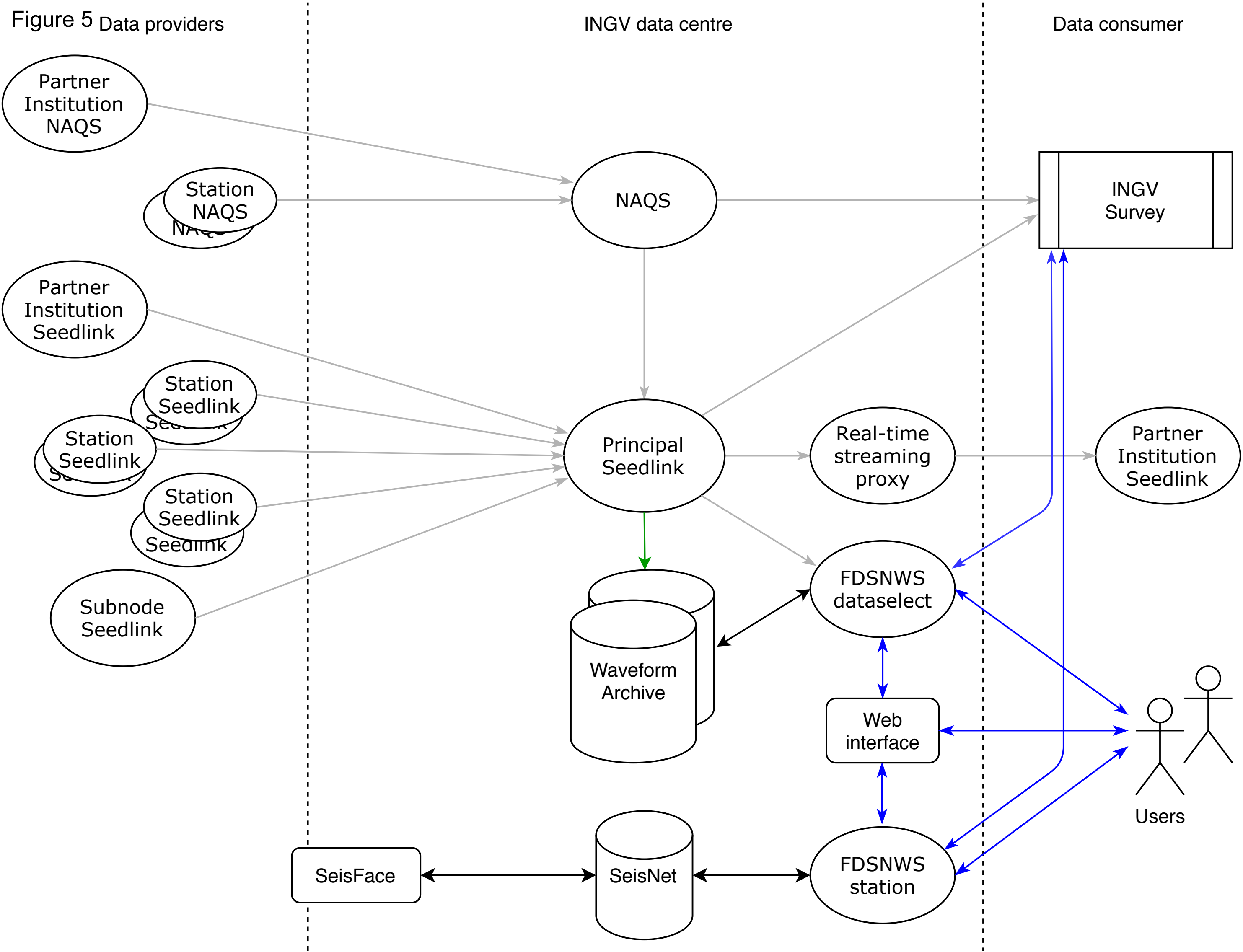

Data providers
INGV data centre
Data consumer
Partner Institution NAQS
Station NAQS
NAQS
INGV Survey
Partner Institution Seedlink
Station Seedlink
Station Seedlink
Station Seedlink
Subnode Seedlink
Principal Seedlink
Real-time streaming proxy
Partner Institution Seedlink
FDSNWS dataselect
Waveform Archive
Web interface
Users
SeisFace
SeisNet
FDSNWS station

Figure 6

Availability network IV MN 2019
availability<90
availability>=99
8.5%
25.8%
90<=availability<95
15.6%
15.2%
95<=availability<97
35.0%
97<=availability<99
Availability network IV MN 2020
availability<90
90<=availability<95
5.6%
7.4%
95<=availability<97
10.4%
availability>=99
50.0%
26.7%
97<=availability<99