WFCatalog: a catalogue for seismological waveform data

Luca Trani[a,1,2,*], Mathijs Koymans[a,1], Malcolm Atkinson[b,2], Reinoud Sleeman[a,1], Rosa Filgueira[c,3]

[a]*Utrechtseweg 297, 3731 GA, De Bilt, The Netherlands*
[b]*Informatics Forum, University of Edinburgh, Edinburgh, EH8 9AB, UK*
[c]*The Lyell Centre, Research Avenue South, Edinburgh, EH14 4AP, UK*

Abstract

This paper reports advances in seismic waveform description and discovery leading to a new seismological service and presents the key steps in its design, implementation and adoption. This service, named WFCatalog, which stands for waveform catalogue, accommodates features of seismological waveform data. Therefore, it meets the need for seismologists to be able to select waveform data based on seismic waveform features as well as sensor geolocations and temporal specifications. We describe the collaborative design methods and the technical solution showing the central role of seismic feature catalogues in framing the technical and operational delivery of the new service. Also, we provide an overview of the complex environment wherein this endeavour is scoped and the related challenges discussed. As multi-disciplinary, multi-organisational and global collaboration is necessary to address today's challenges, canonical representations can provide a focus for collaboration and conceptual tools for agreeing directions. Such collaborations can be fostered and formalised by rallying intellectual effort into the design of novel scientific catalogues and the services that support them. This work offers an example of the benefits generated by involving cross-disciplinary skills (*e.g.* data and domain expertise) from the early stages of design, and by sustaining the engagement with the target community throughout the delivery and deployment process.

Keywords: waveform catalogue, metadata, seismological waveform data, data quality

1. Introduction

This paper reports advances in seismic waveform description and discovery leading to a new seismological service and presents the key steps in its design, implementation and adoption. This service, named WFCatalog, which stands for waveform catalogue, accommodates features of seismological waveform data.

In recent years seismology has experienced a paradigm shift accompanied by major innovations and changes. Seismology has become a data-intensive science where the increasing abundance of data plays a crucial role. This change carries inevitable consequences and affects the way seismologists pursue their research. Network operators, data producers and data centres are equally impacted by this revolution. The role of data centres is changing dramatically, moving from being "simple" data repositories to providers of advanced data services, *e.g.* for data and metadata curation, data exploration and access, analysis and processing. Connection and engagement with user communities has helped steer this

*Corresponding author
Email address: trani@knmi.nl (Luca Trani)
[1]Department of R&D Seismology and Acoustics, Royal Netherlands Meteorological Institute (KNMI)
[2]School of Informatics, University of Edinburgh
[3]British Geological Survey, Edinburgh

transition. The availability of easily accessible data and derived products increases the demand on data centres to provide better and more efficient services for their users. Feedback from user communities influences the design of data centres' technical and organisational architectures.

Our contribution is driven by user demand, existing limitations in current seismic waveform data descriptions and the consequent shortcomings of the paradigms of discovery and access. These limitations provided the motivation for improving the interaction mechanisms between users and seismological data centres. This paper presents a novel approach to seismic waveform description which is central to the enhancement of seismic discovery and access services – we describe a concrete technical solution which has been implemented and is being deployed in the major European seismological data centres federated in the European Integrated Data Archive (EIDA)[4].

The paper is organised as follows: Section 2 describes motivation and context; Section 3 illustrates the methodology adopted and details of the architecture; Section 4 describes the challenges encountered; Section 5 presents related work; Section 6 discusses results and application scenarios, Section 7 outlines conclusions and future directions.

2. Motivation and context

A typical modern seismic station provides continuous, 3-component recordings of ground motion that are typically between 1 and 100 samples per second. A seismic network comprises a number of geographically distributed seismic stations, from which the data streams usually are transmitted in real-time to a data centre. Here, data are archived, processed and analysed by seismologists to extract seismological information (*e.g.* earthquake location and sub-surface structure).

Seismic waveforms are the "primary" data and the seed that yields a multitude of higher-order derived products, thus they should be treated as first class citizens in seismological data centres. Observatories and Research Facilities for European Seismology[5] is the organisation that coordinates the seismic waveform data acquisition and provisioning in Europe. Under the aegis of ORFEUS, EIDA provides a framework to define and share policies for seismic waveform data acquisition, curation and access.

Refining and improving data services according to users' requirements is a major task of EIDA, which requires a deep understanding of and engagement with the user community. The requirements of this community are continuously evolving, thus presenting new challenges to data and service providers. Methods and data analysis techniques have an impact on data management and contribute to pushing the limits of existing infrastructures. For instance, data intensive techniques, like cross correlation of accumulated datasets (Galea et al., 2013; Addair et al., 2014), require the efficient management of and provisioning for large volumes of data.

Typically an analysis workflow starts with *data acquisition*. This time and resource consuming step entails users' interaction with one or several data centres. Data centres usually offer several methods and tools to support users' data acquisition providing discovery and access to their data holdings. These tools are continuously improved and have gone through substantial enhancements, for instance moving from email based tools (*e.g.* BreqFAST) to web services (*e.g.* FDSN web services[6]). The latter enable machine-to-machine communication which is a fundamental requirement to achieve automated workflows. Nowadays many scientific methods in seismology are encapsulated and formalised as workflows, drawing on standard libraries for data handling and transformation (Krischer et al., 2015; Filguiera et al., 2014; Atkinson et al., 2015). The automatic enactment of such workflows

[4] www.orfeus-eu.org/data/eida
[5] www.orfeus-eu.org
[6] www.fdsn.org/webservices

poses additional requirements on the data services, such as managing their rapid burst of requests for data access, distributing resources and responses according to agreed policy and automatically maintaining usage and accounting records to justify resources and support planning.

The paradigm underpinning the request of seismological waveform data has remained almost identical for many years leveraging well-known and common query parameters – including sensor (*e.g.* network, station and channel) and temporal descriptions (*e.g.* start-time and end-time). This set of parameters is well-known among seismologists and satisfies the requirements of several use cases. Nevertheless the current data services suffer from important drawbacks *e.g.* the lack of a mechanism to check the availability of a certain dataset, or lack of an overview of the content of seismic streams. In most of the current seismological data services seismic waveforms are treated as *opaque* objects, meaning that very little information is exposed about their actual content. Direct consequences of such shortcomings are: 1. increased load at users' sites, in terms of data volume and CPU usage; 2. higher rates of request misses; and 3. higher rates of unusable data downloads.

Leveraging on users' requirements we reduce some of these shortcomings in the current seismic waveform description, discovery and access methods.

3. Methods

At the foundation of this effort there is the concept of a catalogue. This catalogue organises and conveys information embedded in continuous seismic streams, *i.e.*, it is a *seismic waveform feature* catalogue. Catalogues are commonly used in seismology *e.g.* to collect and distribute seismic events (Godey et al., 2013), historical earthquakes information (Albini et al., 2013), strong motion parameters (Cauzzi et al., 2016), *etc*. To the best of our knowledge, the description and discovery of seismic waveforms in terms of their content has not been addressed so far.

Building and populating such a catalogue requires a good understanding and knowledge of the seismologists' practices and the common patterns of seismic data analysis.

Without direct access to such features users would have to compute them on datasets downloaded as opaque objects, risking that unwanted characteristics would lead to data disposal. Potentially this situation may result in a vicious circle with a conspicuous waste of resources. Instead we pursue a *virtuous circle* with an efficient use of resources which is a fundamental requirement of any data-driven science. The key to invert this cycle has been identifying a number of tasks and operations of general concern and moving them from *users' sites* to *data centres' sites*. Moving repeated resource consuming tasks into data centres provides several advantages: 1. it reduces users' resource consumption supporting more efficient use of resources at data centres; 2. it leads to a *canonical* definition and representation of seismic waveform features; and 3. it supports and enhances data discovery and access services making them tailored to users' requirements. For instance, a data centre can cache the results of common operations, thus amortising the computational costs over many users. Also, data centres can tune and optimise the performance of such computations and develop the necessary expertise. This can be seen as a *delegation of responsibilities* from the users to the data centres that must deliver: **trust** and **reliability**, and provide **verifiable** and **guaranteed results**.

In the subsequent sections we present details of the `WFCatalog`'s operations, data model and architecture.

3.1. `WFCatalog` *operations*

`WFCatalog` supports several operations:

1. computation, collection, ingestion of metadata
2. stewardship of metadata – update, delete, versioning

3. query functionalities
4. metadata publication
5. data access based on queries over the metadata

Metadata computation, collection and ingestion (1) are core functionalities provided by `WFCatalog`. The computation of metadata is performed close to the related data archive, and requires direct access to the raw seismological data. Computation can be scheduled according to a configurable frequency – this feature provides flexibility and allows us to meet the related policies within the federation. The management of metadata (2) must reflect the chosen policies and the data lifecycle.

`WFCatalog` provides *readonly* capabilities to the users. Metadata ingestion and update are delegated to data centres' operators. This choice reflects the idea that data centres are responsible for the curation of their data holdings, which includes the generation and curation of the related metadata. Metadata may have different versions identified by a timestamp and a version number. At present querying the catalogue for specific metadata versions or performing timestamped queries is not supported – the most recent version of the metadata is provided by default. This behaviour will be extended in future releases in order to facilitate reproducibility. Different query patterns (3) are currently implemented. *Multifaceted* queries spanning across multiple parameters are supported, including: temporal constraints, stream specifications (network, station, channel, location id, *etc*), quality parameters and continuous segments (see Table 1). *Multisite* queries are not supported because data centres should expose only the information, data and data products which they are responsible for. Metadata publication (4) is essential when considering cross-disciplinary science. Adopting standards to publish datasets enables easier discovery and interoperation in broader contexts. `WFCatalog` supports the usage of *Persistent Identifiers*, which entails a commitment to guarantee access to metadata even beyond the data lifespan. `WFCatalog` improves discovery and access (5) to waveform data. At present direct access to data is not provided, but it can be enabled in combination with data access services *e.g.* `fdsnws-dataselect`[7]. The partial API compatibility allows the sharing of queries across services. In a future release persistent identifiers pointing to the data objects (*e.g.* EPIC Handle[8]) will be embedded in the responses from `WFCatalog`.

3.2. Data model

Improving the description and representation of seismic waveform is a major goal of this effort. Such a representation should be: 1. recognised and shared; 2. flexible and extensible; and 3. lightweight and suitable for machine-to-machine communication. The interoperation with broader multidisciplinary environments demands clear formats and well-defined interfaces. Our solution has been designed with these general requirements in mind; we also address *attribution*, *citation* and *reproducibility*.

3.2.1. Data quality metrics

We perform the qualification of seismic waveform according to well-defined and agreed *data quality* metrics, which can be derived from seismic waveforms. The selection of the metrics is not a trivial task and it has been accomplished in successive steps involving several stakeholders. Besides the purely technical issues there are other relevant aspects to consider. A major hurdle is the difficulty to find a common, meaningful and shared way to define *data quality*. The interpretation of data quality is often subjective and varies significantly from case to case. Metrics should span a broad set of use cases and target different users. The selection process was initiated and carried out in the context of EU FP7 project

[7] `www.fdsn.org/webservices`
[8] `www.pidconsortium.eu`

Parameter	Type	Description
network	string	network code
station	string	station code
channel	string	channel code
location	string	channel location identifier
starttime	ISO8601	start time of the selection
endtime	ISO8601	end time of the selection
format	string	specify the output format (default JSON)
include	string	specify the level of detail of the results, *e.g.* $include = sample$
granularity	string	define the desired level of granularity, *e.g.* day
minimumlength	float	limit results to continuous data segments of a specified minimum length in seconds
longestonly	boolean	limit results to the longest continuous segment per channel
csegments	boolean	include information about continuous segments
[metric_filter]	metric dependant	select streams that satisfy a filter on a specific metric value for any metric defined in table 2, *e.g.* $sample\_max\_lt = 10$ & $sample\_max\_gt = 3$

Table 1: Query parameters currently supported by `WFCatalog`

NERA[9] (Sleeman, 2014a,b). This delivered a coherent preliminary set of data quality metrics, which have been further developed and endorsed by EIDA data centres. Subsequently, a broader community has been involved by targeting the International Federation of Digital Seismograph Networks (FDSN). That discussion is currently ongoing and a core set of metrics and their associated definitions has been identified. Consensus and shared definitions of such metrics are fundamental requirements to ensure compatibility, exchange and comparison of results across different systems. The list of data quality metrics adopted in the current version of `WFCatalog` is provided in Table 2. For a more complete overview we refer to the `WFCatalog` specification[10].

3.2.2. WFMetadata schema

`WFCatalog` publishes metadata according to the Waveform Metadata (`WFMetadata`) JSON schema[11]. This schema is novel and represents (seismic) waveform metadata including data quality metrics and additional features as shown in Table 3. An important feature is the possibility to extend its applicability beyond seismological waveforms *e.g.* infrasound time series data. The `WFMetadata` schema sets the basis to become a standard way to represent and exchange seismic waveform metadata, thus filling a gap in the current seismological metadata offerings. Noteworthy is the support for *Persistent Identifiers* which coupled to versioning and information about the producer, foster proper attribution, citation and reproducibility.

3.3. Architecture

`WFCatalog`'s architecture is modular and is composed of the following main elements: data analysis and metrics computation module, metadata store and web service API. (Fig. 1)

3.3.1. Data Analysis and Metadata Computation module (DAMC)

Waveform data analysis and data quality metrics computation are core functions. They yield the features extracted from waveform data which are then stored and made accessible

[9] `www.nera-eu.org`
[10] `www.orfeus-eu.org/data/eida/eidaws/wfcatalog`
[11] `github.com/EIDA/wfcatalog/blob/master/wf_metadata_schema.json`

Sample metrics	
num_samples	sample_max
sample_min	sample_mean
sample_median	sample_stdev
sample_rms	sample_lower_quartile
sample_upper_quartile	num_gaps
num_overlaps	max_gap
max_overlap	sum_gaps
sum_overlaps	percent_availability
MiniSEED header metrics	
encoding	num_records
quality	record_length
sample_rate	timing_correction
timing_quality_mean	timing_quality_median
timing_quality_lower_quartile	timing_quality_upper_quartile
timing_quality_max	timing_quality_min
data_quality_flags	
amplifier_saturation	digitizer_clipping
spikes	glitches
missing_padded_data	telemetry_sync_error
digital_filter_charging	suspect_time_tag
activity_flags	
calibration_signal	time_correction_applied
event_begin	event_end
positive_leap	negative_leap
event_in_progress	
io_and_clock_flags	
station_volume	long_record_read
short_record_read	start_time_series
end_time_series	clock_locked

Table 2: Data quality metrics currently implemented in `WFCatalog`

Feature name	Description
wfmetadata_id	identifier of the returned metadata document
producer: data centre, agent, date creation	producer of the metadata document
waveform_type	type of the related waveform, *e.g.* seismic and infrasound
waveform_format	format of the related waveform, *e.g.* MiniSEED
version	progressive number indicating the document version

Table 3: `WFCatalog` additional features

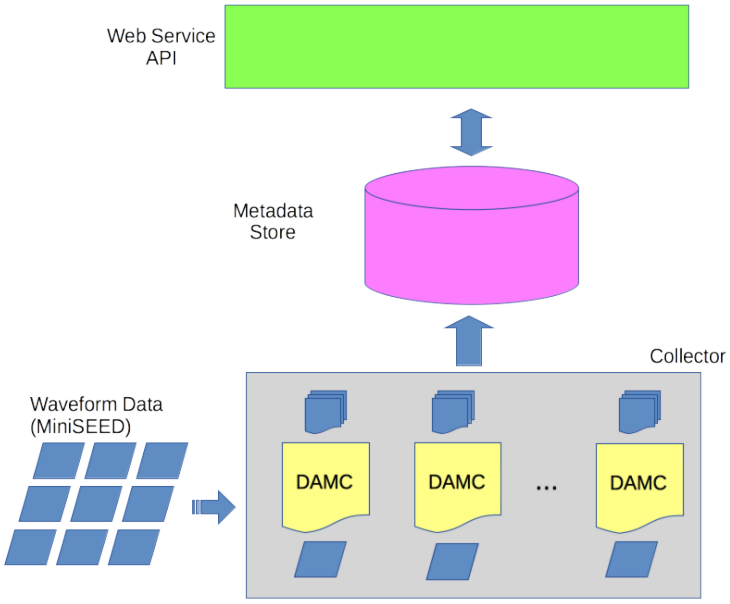


Figure 1: `WFCatalog` architecture overview – Seismic streams encoded in MiniSEED feed the `Collector`. This component performs parallel processing of seismic streams instantiating multiple Data Analysis and Metadata Computation modules (DAMCs). DAMC's implementation builds on a popular seismological software library: ObsPy[13]. Each DAMC extracts features and metadata from seismic streams and populates the Metadata Store. The `WFCatalog` web service API provides programmatic access to the metadata stored in the database.

to users. The DAMC implements these operations complying with specified and agreed metric definitions. Our strategy has been to decouple the definition of the features from their implementation. As a consequence each data centre has the freedom to implement their own DAMCs as long as they comply with the agreed definitions. We provide a reference implementation which is adopted across EIDA data centres. This implementation builds on top of a popular community-driven Python library, namely ObsPy (Krischer et al., 2015), and it takes MiniSEED (Ahern et al., 2009) data as input. We chose MiniSEED as it is by far the most used format in EIDA data centres and for compatibility with the `fdsnws-dataselect` web service. However, `WFCatalog` and its metadata model are not bound to any specific data format. Liaising with the ObsPy developers and the user community we developed extensions and made an additional module available[12]. The inclusion of the DAMC code in a software library widely used by seismologists has been a strategic choice with various advantages: 1. it establishes a direct communication channel with the user community; 2. it involves the user community in its design, maintenance and evolution; 3. it enables users to have the same functionalities available at their sites; and 4. it builds consensus and promotes adoption.

The DAMC is configurable and integrated in the `WFCatalog` ingestion process, namely `WFCatalog Collector`. At this stage features are computed with a *daily* granularity. The DAMC, and the `WFCatalog`, have been designed to scale in terms of new features and/or additional time granularities. The ingestion into the database is performed by run-

[13] github.com/obspy

[12] `docs.obspy.org/master/packages/autogen/obspy.signal.quality_control.html`

ning multiple DAMC processes in parallel managed by the `Collector`. We store hash signatures which can be used to trigger re-computation of the features when changes occur in the data files.

3.3.2. Metadata Store

The features extracted from seismic waveform data require the support of a suitable database infrastructure offering: scalability, performance and optimisation in terms of storage space, query functionality and query time. We decided to benchmark several technologies before opting for a solution. The final choice has been driven by pragmatic aspects, and it might evolve over time as the architecture design is technology independent. Our evaluation considered the following factors: maturity, language support, availability of connectors and software libraries, scalability and extensibility. A complete technology review is out of the scope of this paper, nevertheless, it is worth mentioning the systems we evaluated, namely: MySQL, MonetDB (MonetDB BV, 2013), Cassandra (Apache Software Foundation, 2013a), CouchDB (Apache Software Foundation, 2013b) and MongoDB (MongoDB, Inc., 2016). Of particular interest was the experience with MonetDB (Ivanova et al., 2013a,b). This technology, when further developed and refined, has the potential to provide functionalities hardly achievable with the other candidates. However, at the time of our choice this technology was not considered stable enough for production and we opted for another DBMS: MongoDB. MongoDB is a very popular document store which provides native scalability and its internal model is flexible and allows for extensions.

In the current setting the database hosts two collections: one holds the features computed on daily files whereas the other holds the features about the continuous segments contained in a specific day with start-time and end-time of each segment. Therefore, we are able to provide a detailed description of the availability of data in each waveform stream. Moreover, the loose coupling of the collections allows for extensions that include additional features and time granularities, *e.g.* hourly.

3.3.3. Web API

The web API facilitates the interaction with third-party software and users. This component has to promote usability, support diverse use cases, address the evolving nature of the user community's requirements and allow for extensibility – ideally it should be possible to add features and modify the current query patterns according to new scientific methods whilst maintaining consistency.

The design of the API of the `WFCatalog` has been an iterative, collaborative work involving several stakeholders including data-centre operators, developers and seismologists. The participation of several actors from the early stages of the design contributed useful perspectives and requirements.

The discussion was triggered by a prototype showing the potential capabilities, this prototype has been refined incrementally during further stages. One of the requirements was to allow compatibility with existing service standards (*e.g.* FDSN). This reduces the learning curve and facilitates the uptake of a new service. It also enables users and data curators to retain the value of prior investments in methods, workflows, code and working practices. We were able to fulfil this backward compatibility constraint only partially.

Table 4 summarises the available methods, for an extensive description we refer to the published web API[14].

[14] `www.orfeus-eu.org/eidaws/wfcatalog/alpha/application.wadl`
`geofon.gfz-potsdam.de/eidaws/wfcatalog/alpha/application.wadl`
`eida.bgr.de/wfcatalog/alpha/application.wadl`
`catalog.data.ingv.it/wfcatalog/1/application.wadl`
`eida.gein.noa.gr/eidaws/wfcatalog/alpha/application.wadl`

Method	Description
query	enables metadata queries with the supported parameters
version	returns the version of the web service
application.wadl	returns the WADL document describing the service

Table 4: `WFCatalog` webservice API methods

4. Challenges

In the previous sections we described how we implemented the `WFCatalog`, interpreting users' requirements and translating them into a concrete architecture. In the subsequent sections we introduce the challenges encountered during the design and construction process. Recognising the main challenges and their implications can provide a better understanding of the complexity of the environment in which this work is framed. These challenges can be divided in two sub-categories: *socio-political* and *technical*.

4.1. Socio-political challenges

Seismology has a long tradition of global collaboration and data sharing, as well as knowledge and experience about definition, design and implementation of data models, formats, services and tools. Consequently the maturity of the community is reflected and formalised in a number of international coordination and collaboration frameworks at global and European scale *e.g.* IASPEI[15], FDSN[16], ESC[17], ORFEUS[18]. The role of such organisational bodies is fundamental to guarantee authoritativeness, trust, acceptance and adoption on the form of shared services and data they deliver. Alongside the official formalised contexts, there often exist community-driven efforts, which may have an equally large impact. These initiatives can be powerful and direct vehicles to reach out to large and broad communities outside the formal schemes. Identifying the key players and stakeholders of a specific community is essential when designing innovative services for such a community. We addressed a mix of official and de-facto processes in order to facilitate the definition and uptake of the `WFCatalog`. We targeted FDSN, ORFEUS and EIDA as formal frameworks and ObsPy as community-driven effort.

The European seismological landscape has a distributed organisation with responsibilities shared across a number of recognised data centres. This organisation has historical and cultural roots but it is also a design choice to address the evolving data challenges. An example is the official establishment of EIDA within ORFEUS in 2013. Previously the ORFEUS Data Centre (ODC) was the centralised European data archive. The newly constituted federated structure responds better to modern challenges but it requires well-defined, shared agreements and a common vision.

Another important aspect is understanding the users, their requirements, the set of tools and methods they use and the limitations of these tools. In the seismological domain there are a number of well-known and widespread tools, libraries, methods and data exchange standards. Seismologists exploit such common building blocks by applying customisations and extending them with new methods. However, the sharing of methods and algorithms for data analysis and processing, in the form of workflows, is quite new to the community and gained popularity only recently supported by initiatives such as the VERCE project (Atkinson et al., 2015). Customisation may inhibit the adoption of standard representations.

Seismologists are accustomed to delegating data management operations to data centres whereas processing and analysis remain the users' focus. A reason for this may be the novelty of method sharing and lack of "control" when delegating operations.

[15] `www.iaspei.org`
[16] `www.fdsn.org`
[17] `www.esc-web.org`
[18] `www.orfeus-eu.org`

An important lesson learned is that technical changes ought to be supported and sustained by appropriate organisational frameworks. These frameworks can provide the context and vocabulary to steer collaborative discussions, pooling insights and efforts thereby accelerating convergence on solutions. They offer trusted environments that facilitate technology uptake and long-term sustainability. `WFCatalog` is the result of a collaborative work initiated within EIDA that provided a proper organisational framework to exchange ideas, requirements and define strategies and policies. These elements are equally important because a catalogue is not just a piece of software – a fundamental component is related to the authoritativeness of the information therein maintained and offered to the users. Clear and well-defined policies to building, operating, revising and decommissioning such catalogues are key elements. The combination of policies, software and communication providing high compatibility across a federation like EIDA, allowed us to reach the highest level possible of agreement among partners, whereas such consistency was not replicated at the FDSN. FDSN provides a broad platform to coordinate, discuss, promote and exchange ideas, nevertheless, it has a looser coupling among participants which is reflected in a slower pace to forming global agreements. An additional factor in the agreement forming is the level of commitment which can vary depending on different priorities, available resources, *etc*. In our case the clear engagement of most of the contributing partners in an overarching research infrastructure for solid-Earth science, namely the European Plate Observing System[19], constituted an accelerating factor. Inevitably the boundary conditions provided by EPOS influenced the timeline ensuring a rapid convergence towards a common goal. Therefore the catalysing role of projects and research infrastructures should not be underestimated.

Within large collaborations a major challenge is the multiplicity of factors that ought to be synchronised and aligned for a common purpose. Communication, engagement and commitment to key roles and of representatives are essential. In order to foster these activities technical architectures need to reflect the complexity of the surrounding environment and offer intellectual ramps (Atkinson et al., 2010).

4.2. Technical challenges

The computation of the quality metrics presented several challenges. In order to align the theoretical definitions with the computation, we had to overcome several issues mainly introduced by the SEED (Ahern et al., 2009) data format and by the data archival system. Adaptations allowed us to obtain thorough and accurate results conforming with the definitions. As the system has to cope with the steady growth of the data and the consequent increase of the metadata volume, scalability is essential. The approach and the technology adopted enable us to deal effectively with these issues.

The chosen data model guarantees the flexibility and extensibility required to address the expansion of the set of metrics and features. Another critical aspect concerning the metrics is the granularity, that is the time window over which the metrics are calculated. As previously mentioned we chose to compute the metrics on daily intervals. This choice is a tradeoff between meaningfulness and pragmatism. The alternatives are fixed time granularity of a different length and dynamic computation tailored on users' requests. The latter represents the ideal solution. For instance, scientists performing analysis on long period signals might be interested in metrics computed and aggregated on a yearly basis. Unfortunately the dynamic solution is also the most expensive from the computational point of view. Also, it requires proper technological support not easily achievable with most DBMSs. We experimented with this approach with MonetDB but for the reasons previously mentioned, maturity and support, we decided to move towards a less advanced but more stable solution. We adopted a fixed granularity but as a mitigating factor we designed the system to accommodate multiple independent granularities.

[19] `www.epos-ip.org`

Another challenge regards the performance of the metrics computation. This aspect influences the database update policies. Ideally a user may want the metrics and the raw data available simultaneously which means near real time. However, processing in near real time the incoming data of thousands of waveform streams can be expensive, especially considering the limited capacity of some data centres. We optimised the DAMC in order to speed up the critical operations. As ObsPy is predominantly a Python framework, in an initial phase some operations proved to be slow and we switched to native C implementation for the critical methods in order to achieve better results. This optimisation provided a gain of a factor of 10 on the most compute-intensive methods.

5. Related work

The design, development and deployment of `WFCatalog` was influenced by many aspects of contemporary research including attempts to assess seismic waveform data quality.

Data quality has been a debated topic in seismology for a long time. A number of tools and software packages have been produced addressing data quality. An example of such software is PQLX (McNamara and Boaz, 2006) which provides a graphical user interface on top of a MySQL database containing probability density functions (PDF) of power spectral densities (PSD). PQLX has been widely used and it became a de facto standard for certain metrics.

Another application is the Data Quality Analyzer (DQA) (Ringler et al., 2015). That application, developed at USGS, follows an approach similar to `WFCatalog` to present data quality metrics and facilitate the assessment of the quality of seismic stations. However, DQA's approach is focused primarily on stations whereas `WFCatalog` is waveform data centric. Although DQA computes and stores data quality metrics in a central PostgreSQL DBMS it does not expose them as metadata, thus not enabling machine-to-machine communication. DQA has a rich web interface with diverse visualisations.

The above products focus purely on data quality metrics mainly addressing the diagnostics of seismological stations. They do not aim at extending the description of waveform data and they do not provide such metrics as a service. They are valuable tools in the context of their application but they do not address the broader scope. Moreover each solution adopts a different set of metrics and definitions.

The Modular Utility for Statistical Knowledge Gathering (MUSTANG)[20] has a web service interface providing programmatic access to a number of quality metrics in different formats and a data browser for visualising such metrics. Our work is an attempt to homogenise the metrics across different systems. In particular we identified a set of metrics shared with MUSTANG as a common base for discussions at FDSN.

6. Results and Discussion

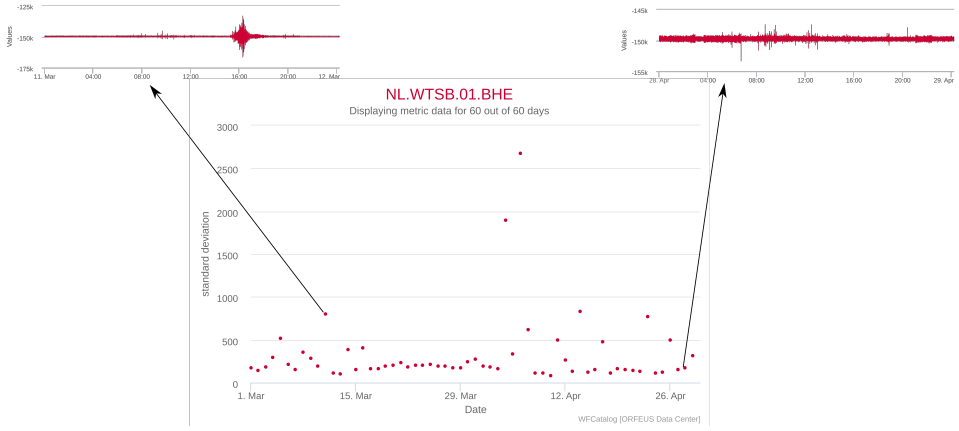
`WFCatalog` can be used to assess seismic waveform data quality, in this way data centres have a powerful tool to evaluate and present the quality of their data holdings. Similarly, network operators have an effective instrument that offers them immediate feedback about the status of their sensors, thus helping them addressing potential issues and delivering better quality data. By offering a catalogue which contains metadata that users would otherwise compute on downloaded data, `WFCatalog` provides major savings: 1. overall computation is reduced because computation results are reused; 2. access to primary data that then proves unusable is avoided, with a substantial network traffic reduction; 3. users do not need to perform quality analyses before they use the data but they still can if they have additional criteria; and 4. gradually the standards for data quality will emerge leading to more consistent science.

[20] `service.iris.edu/mustang`

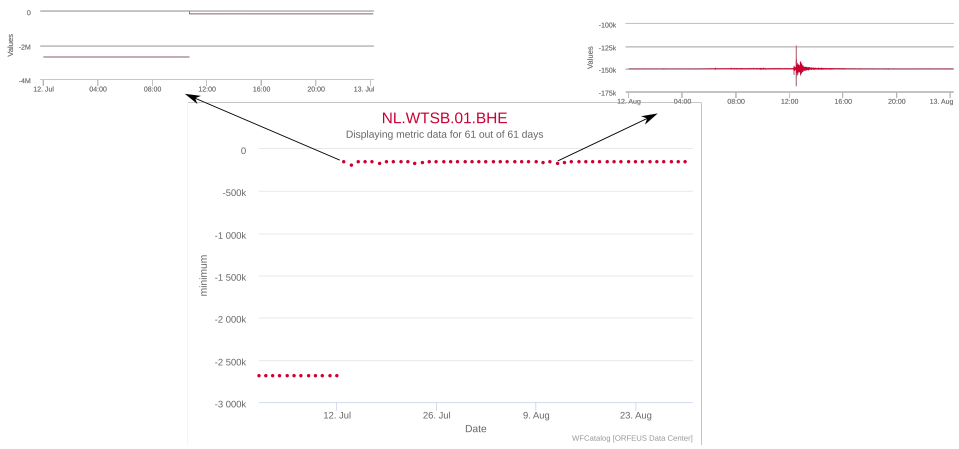
The benefits of `WFCatalog` are its data model, its exchange format, `WFMetadata` schema, and the programmatic access to a standardised set of predefined features. `WFMetadata` schema provides a canonical representation of waveform data metadata which helps establish trustworthy communication in a federated environment. The `WFCatalog` constitutes an important addition which combines with other components and services to provide substantial advantages in seismic waveform discovery and access. It helps to steer the discovery process filtering the results tailored by user's requirements about waveform data content. An example of such interaction and service composition is the combination of `WFCatalog` (discovery) and `fdsnws-dataselect` (access).

Service composition is one possible application of `WFCatalog`, another application is visualisation. At ODC we developed web interfaces fed by `WFCatalog` in order to check the availability of datasets and visualise multiple data quality metrics. Fig. 3 illustrates an example of a visual interface that can be enabled on top of `WFCatalog`. This interface provides a visual inspection of the available seismic waveform streams with a detailed overview of the continuous segments contained in a daily stream.

Figure 2 shows another interface which can be used to browse graphically through data quality metrics. This figure shows multiple metrics computed for different days. This interactive tool allows seismologists to spot possible issues with the underlying data – by clicking on a specific point it is possible to drill down to a preview of the underlying data.



(a) Showing the standard deviation of the raw data. A standard deviation higher than average may indicate the detection of an event during that day (left). Lower standard deviations are representative of ambient noise (right).



(b) Showing the minimum value of all samples for each daily granule. A sudden jump (left) in the minimum, maximum or mean may indicate an offset in the waveform baseline. A small dip in the minimum (right) may be a feature introduced by an event.

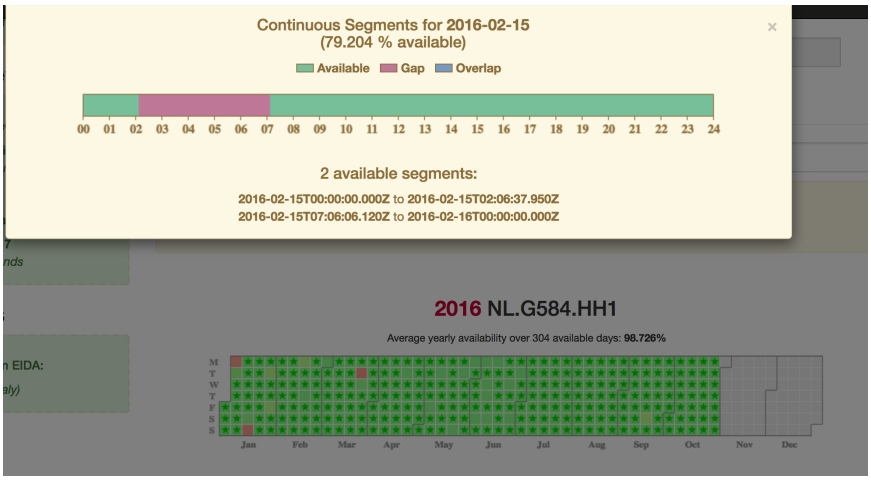


Figure 3: Data availability visualisation – this graphical interface allows users to browse through a daily calendar and view data availability. Green and red colours represent high and low availability, respectively. A tile marked with a star indicates full channel availability. Non-continuous days can be clicked to investigate the available data segments for that day[21].

(c) Showing the maximum sample value of the daily granules. Abnormally high maximum or minimum values are indicative of spikes in the data (left and right).

Figure 3: Data metric visualisation – this graphical interface illustrates a collection of sample metrics for each day in the requested time window.

6.1. Evaluation

We evaluated different aspects of `WFCatalog` discussed below.

6.1.1. Ingestion statistics

Processing of data progresses at roughly 50 - 60 gigabytes an hour running four parallel ingestion processes (DAMCs) on a quad-core machine[22] excluding network latency. The

[21] Source: `www.orfeus-eu.org/data/odc/quality/availability`
[22] Intel(R) Xeon(R) CPU E3-1225 v3 @ 3.20GHz

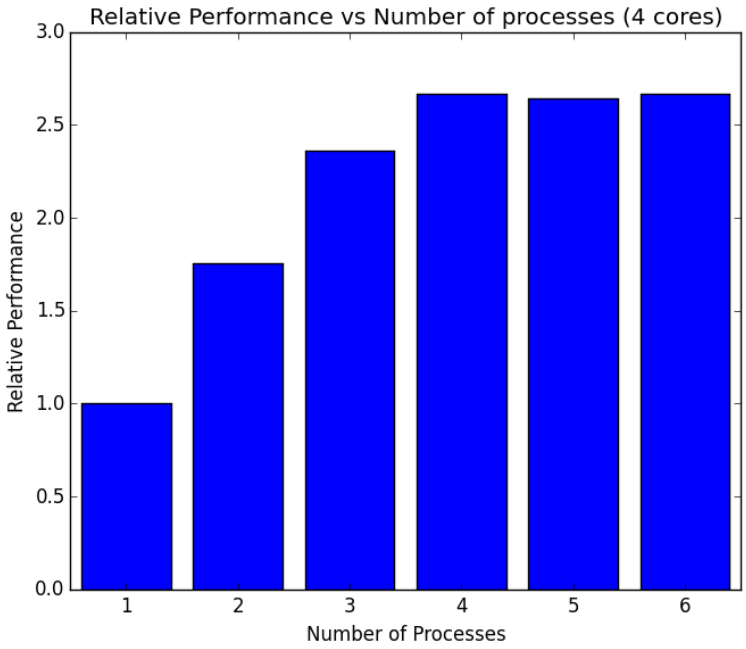


Figure 4: The figure illustrates the behaviour of the `WFCatalog` ingestion process. The relative performance $RelPerf = \frac{T_{1proc} * N_{proc}}{T_{Nproc}}$ is measured in terms of time.

relative performance for each additional process is illustrated in figure 4. The relative performance ($RelPerf$) is obtained as follows: $RelPerf = \frac{T_{1proc} * N_{proc}}{T_{Nproc}}$ where T_{1proc} is the time required by one reference process, T_{Nproc} is the average time required by N processes and N_{proc} is the number of processes. The time required for processing increases roughly linearly with increasing sample rate of the waveform, that is initially limited by the overhead of the calculation. On average, we observed that metric calculation on 24 hours of data takes roughly 2.5 seconds for data with 100 Hz sample rate, 1 second for 40 Hz, and 0.05 seconds for 1 Hz. Memory consumption presents a limiting factor in running multiple parallel processes for high sample rates because waveforms are required to be read into memory in their entirety for metric calculation.

6.1.2. Metadata store statistics

At present the `WFCatalog` at ODC has information on roughly 4 million daily streams accounting for a total of 400 million continuous segments. The storage size of the metadata of the daily streams, including indexed fields is 1.22 GB using the WiredTiger storage engine available in MongoDB. The metadata about continuous segments accounts for the most significant usage of disk space with a total of 85 GB. The amount of data that is represented by this metadata is roughly 15 TB distributed through 4 million daily waveform files. The storage size comes down to a compressed 315 bytes for each daily stream and 83 bytes for each additional continuous segment. Poor waveform data including many gaps may consist of up to 500,000 individual traces and will be a strain on the database. Future limits on the minimum length for a continuous segment may be set to prevent explosive growth of the database.

6.1.3. Benefits for the users

We investigated the advantages provided by `WFCatalog` for improving data discovery. Because at the time of writing we could not measure the operational usage, we opted

for a semi-simulated scenario. We analysed a sample of real queries (ca. 400,000) submitted by users to `fdsnws-dataselect`, which is currently the most used service to retrieve seismic waveform data. Users can submit time constrained queries attempting to get the desired data streams. However, data delivery is not guaranteed because no *a priori* information about data availability is provided by this service. `WFCatalog` can be used to get the availability information. We show that by exploiting `WFCatalog` we are able to improve data retrieval and reduce the number of requests which would deliver unusable data. For 'usable' data we mean requested time windows without gaps. By consulting several users we can consider this as a likely situation. We submitted the same users' queries to `WFCatalog` with the option to include gaps information in the requested time window. Figure 5 shows the distribution of the requested time windows which have been analysed. We notice that the majority of the requests are clustered into three main groups: *small range*, *medium range*, *large range*. These groups represent the most popular use cases, which we addressed in our analysis. We compared the responses of the queries with the expected criteria (continuous data), results are shown in figure 6. The percentages indicate the relative gain: $\frac{requests\_with\_gaps}{total\_number\_of\_requests} * 100$. In general there is a substantial improvement in the delivery of correct results as `WFCatalog` informs the users in advance about the time windows that should be discarded without attempting to download them. As expected the benefits increase on larger time windows because there the probability to have a gap is higher.

Therefore by interacting with `WFCatalog` before posing the actual data request, users save time and resources avoiding unnecessary downloads of discontinuous data streams.

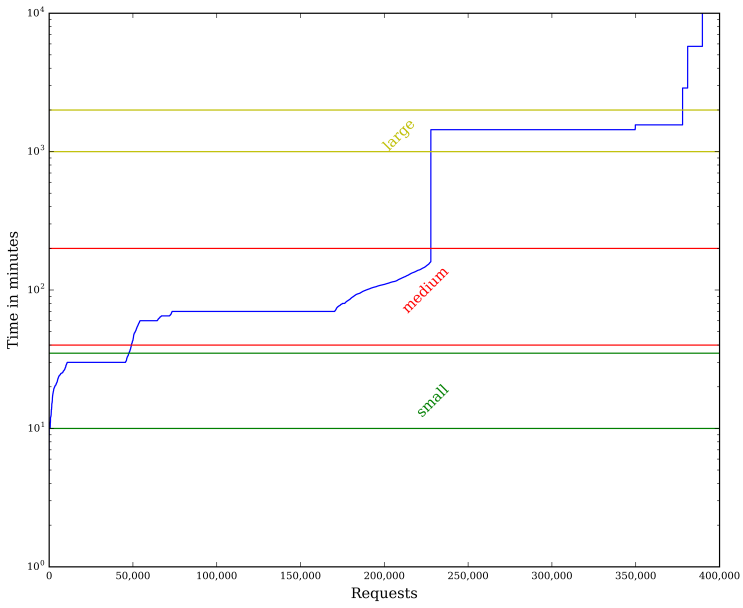


Figure 5: Requested time window distribution – the figure represents a sample of real users' requests submitted to `fdsnws-dataselect`. The highlighted regions show the most popular time window lengths.

7. Conclusions and future work

We presented a novel service which is currently being deployed across the major European seismological data centres of EIDA. `WFCatalog` will enrich the portfolio of tools

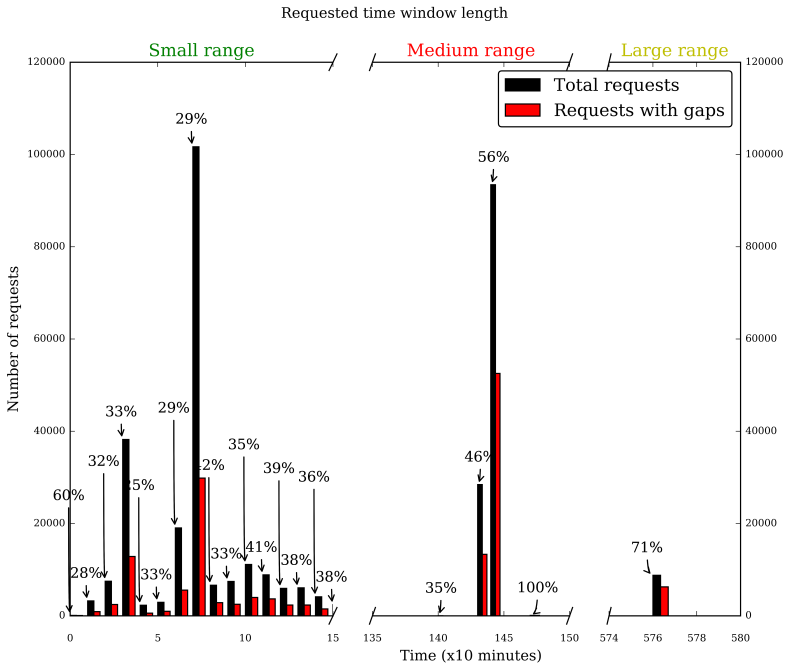


Figure 6: Improvement in data delivery: avoiding 'gappy' data – the figure shows the potential gain that can be obtained with `WFCatalog` filtering out time windows with gaps. The percentages indicate the relative gain: $\frac{requests\_with\_gaps}{total\_number\_of\_requests} * 100$

and services for seismology providing clear advantages for the discovery and access of seismic waveform data. The information provided in a machine-readable form will foster automated workflows and improve the data acquisition process. `WFCatalog` with its WFMetadata schema set the basis for a standardised way to exchange seismic waveform metadata and for a canonical representation of quality metrics and data features. The current schema will be maintained and supported by a large community in ORFEUS – this can ensure long term sustainability. Moreover, the continuous interaction with the users will guarantee extensions in order to address new use cases and scenarios. One such extension is for instance the integration of Power Spectral Density functions which is planned. The interoperability with broader communities beyond seismology is another aspect which will be improved, by enriching the published metadata including persistent identifiers and Dublin Core[23]. `WFCatalog` is a step towards making seismic waveform datasets FAIR – 'Findable', 'Accessible', 'Interoperable' and 'Reusable' – (Wilkinson et al., 2016).

Acknowledgements

We thank the EIDA team for their precious help and support. In particular Javier Quinteros and Andres Heinloo from GFZ Potsdam who provided valuable inputs in shaping the web service API. Also, we thank the ObsPy team and in particular Lion Krischer for their support developing the DAMC and the WFMetadata schema. We thank two anonymous reviewers for their constructive comments and suggestions which helped us to improve the manuscript. Trani, Koymans and Sleeman were supported by the EU project EPOS-IP

[23] `dublincore.org`

No. 676564[24]. Atkinson was supported by the EU project ENVRI^{plus} No. 654182[25].

Bibliography

M. Galea, A. Rietbrock, A. Spinuso, L. Trani, Data-Intensive Seismology: Research Horizons, John Wiley & Sons, Inc., ISBN 9781118540343, 353–376, doi:\bibinfo{doi}{10.1002/9781118540343.ch17}, URL http://dx.doi.org/10.1002/9781118540343.ch17, 2013.

T. G. Addair, D. A. Dodge, W. R. Walter, S. D. Ruppert, Large-scale seismic signal analysis with Hadoop, Computers and Geosciences 66 (2014) 145–154, ISSN 00983004, doi:\bibinfo{doi}{10.1016/j.cageo.2014.01.014}, URL http://dx.doi.org/10.1016/j.cageo.2014.01.014.

L. Krischer, T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, J. Wassermann, ObsPy : a bridge for seismology into the scientific Python ecosystem, Computational Science & Discovery 8 (1) (2015) 1–17, ISSN 1749-4699, doi:\bibinfo{doi}{10.1088/1749-4699/8/1/014003}, URL http://dx.doi.org/10.1088/1749-4699/8/1/014003.

R. Filguiera, I. Klampanos, A. Krause, M. David, A. Moreno, M. Atkinson, Dispel4Py: A Python Framework for Data-intensive Scientific Computing, in: Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems, DISCS '14, IEEE Press, Piscataway, NJ, USA, ISBN 978-1-4799-7038-4, 9–16, doi:\bibinfo{doi}{10.1109/DISCS.2014.12}, URL http://dx.doi.org/10.1109/DISCS.2014.12, 2014.

M. Atkinson, M. Carpene, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garth, A. Gemund, H. Igel, I. Klampanos, A. Krause, L. Krischer, S. H. Leong, F. Magnoni, J. Matser, A. Michelini, A. Rietbrock, H. Schwichtenberg, A. Spinuso, J.-P. Vilotte, VERCE Delivers a Productive E-science Environment for Seismology Research, 2015 IEEE 11th International Conference on e-Science (e-Science) 00 (undefined) (2015) 224–236, doi:\bibinfo{doi}{doi.ieeecomputersociety.org/10.1109/eScience.2015.38}.

S. Godey, R. Bossu, J. Guilbert, Improving the mediterranean seismicity picture thanks to international collaborations, Physics and Chemistry of the Earth 63 (2013) 3–11, ISSN 14747065, doi:\bibinfo{doi}{10.1016/j.pce.2013.04.012}, URL http://dx.doi.org/10.1016/j.pce.2013.04.012.

P. Albini, R. M. Musson, A. A. G. Capera, M. Locati, A. Rovida, M. Stucchi, D. Viganò, Global Historical Earthquake Archive and Catalogue (1000-1903), Tech. Rep., doi:\bibinfo{doi}{10.13117/GEM.GEGD.TR2013.01}, 2013.

C. Cauzzi, R. Sleeman, J. Clinton, J. D. Ballesta, O. Galanis, P. Kästli, Introducing the European Rapid Raw Strong-Motion Database, Seismological Research Letters 87 (4) (2016) 977–986, ISSN 0895-0695, doi:\bibinfo{doi}{10.1785/0220150271}, URL http://srl.geoscienceworld.org/content/87/4/977.

R. Sleeman, Automatic data QC and distribution statistics for data providers, D2.3, Report, KNMI, URL http://www.orfeus-eu.org/organization/projects/NERA/Deliverables/NERA_D2.3.pdf, 2014a.

[24] www.epos-ip.org
[25] www.envri.eu

R. Sleeman, Data quality improvement statistics, D2.4, Report, KNMI, URL www.orfeus-eu.org/organization/projects/NERA/Deliverables/NERA_D2.4.pdf, 2014b.

T. Ahern, R. Casey, D. Barnes, R. Benson, T. Knight, C. Trabant, SEED Reference Manual, IRIS, 2009.

MonetDB BV, MonetDB, URL www.monetdb.org, 2013.

Apache Software Foundation, Apache Cassandra, URL http://cassandra.apache.org, 2013a.

Apache Software Foundation, Apache CouchDB, URL http://couchdb.apache.org, 2013b.

MongoDB, Inc., MongoDB, URL www.mongodb.com, 2016.

M. Ivanova, Y. Kargin, M. Kersten, S. Manegold, Y. Zhang, M. Datcu, D. E. Molina, Data vaults: a Database Welcome to Scientific File Repositories, Proceedings of the 25th International Conference on Scientific and Statistical Database Management - SSDBM (2013a) 1doi:\bibinfo{doi}{10.1145/2484838.2484876}.

M. Ivanova, M. Kersten, S. Manegold, Y. Kargin, Data Vaults : Database Technology, Computing in Science & Engineering 15 (3) (2013b) 32–42, doi:\bibinfo{doi}{10.1109/MCSE.2013.17}.

M. Atkinson, D. De Roure, J. Van Hemert, D. Michaelides, Shaping ramps for data-intensive research, UK e-Science All Hands Meeting 2010 (2010) 1–3URL http://eprints.soton.ac.uk/271235/.

D. E. McNamara, R. I. Boaz, PQLX: A Software Tool to Evaluate Seismic Station Performance, AGU Fall Meeting Abstracts .

A. T. Ringler, M. T. Hagerty, J. Holland, A. Gonzales, L. S. Gee, J. D. Edwards, D. Wilson, A. M. Baker, The data quality analyzer: A quality control program for seismic data, Computers and Geosciences 76 (2015) 96–111, ISSN 00983004, doi:\bibinfo{doi}{10.1016/j.cageo.2014.12.006}, URL http://dx.doi.org/10.1016/j.cageo.2014.12.006.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. a.C 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. a. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018, ISSN 2052-4463, doi:\bibinfo{doi}{10.1038/sdata.2016.18}, URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4792175{\&}tool=pmcentrez{\&}rendertype=abstract.