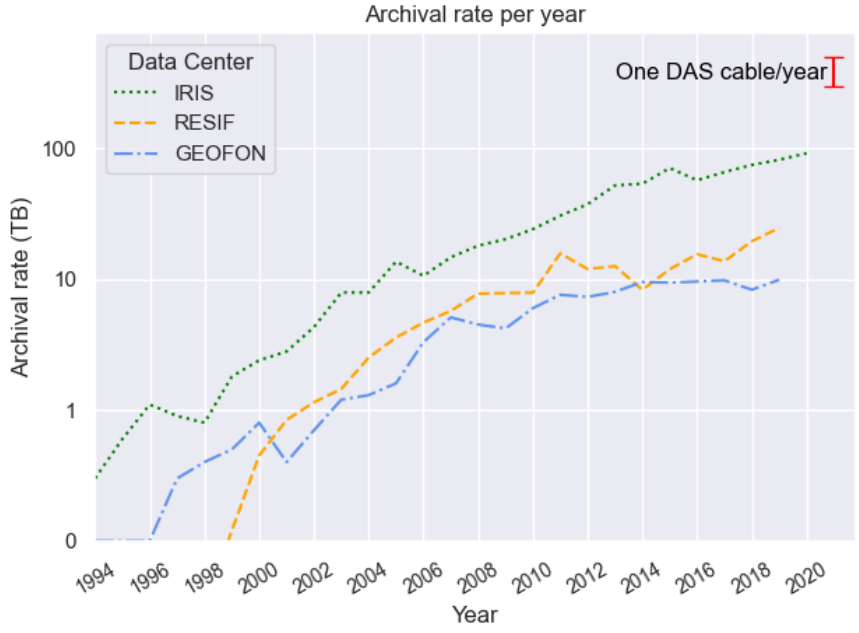
DAS data and products management policies within Geo-INQUIRE

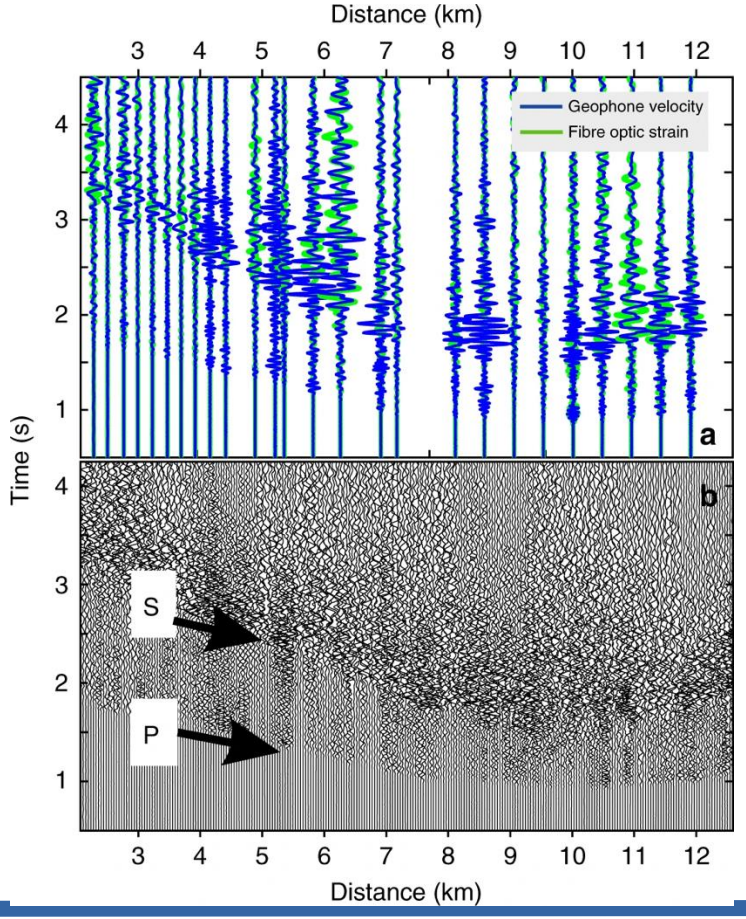
Current status and outlook

Geo-INQUIRE Seminars – 5th December 2024

Javier Quinteros on behalf of GEOFON and members of the Geo-INQUIRE DAS Group

Archival Rate vs. Resolution





Top: A simple DAS experiment can generate more data than the one archived in a whole year at any data centre in the world. Right: Comparison of the resolution between broad band sensors and a DAS interrogator.

Quinteros et al. (SRL, 2021); Jousset et al. (Nature, 2018)

Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives

Javier Quinteros[*1], Jerry A. Carter[2], Jonathan Schaeffer[3], Chad Trabant[2], and Helle A. Pedersen[3,4]

Abstract

New data acquisition techniques are generating data at much finer temporal and spatial resolution, compared to traditional seismic experiments. This is a challenge for data centers and users. As the amount of data potentially flowing into data centers increases by one or two orders of magnitude, data management challenges are found throughout all stages of the data flow.

The Incorporated Research Institutions for Seismology—Réseau sismologique et géodésique français and GEOForschungsNetz data centers—carried out a survey and conducted interviews of users working with very large datasets to understand their needs and expectations. One of the conclusions is that existing data formats and ser-

Quinteros et al. (SRL, 2021) doi:1.1785/0220200390

Roadmap for DAS standardization

- Metadata: community and seismological
- Data format(s)
- Long-term archival
- Data provisioning
- Real-time transmission
- Processing
- Ethical issues, or related to privacy and security

Issues to be addressed / Roadmap

- A basic Channel Naming convention has been suggested some years ago, but has some limitations.
  - Network: Fiber/cable
  - Station: DAS Channel
  - Location: empty
  - Channel: HSF (e.g. for 100/200 Hz).
- What do we do with experiments with more than one fiber?
- Could we find a solution with the new Source Identifiers approved by FDSN?
- Would this be a triggering factor for adoption within the community?
- Check technical completeness of the JSON representation and define a mapping to StationXML.
- FDSN WS capable of providing metadata in this new format.

Current problems: Data formats

Proprietary formats

- TDMS (Silixa)
- HDF5 (OptoDAS, Silixa v2, others)

Community:

- Seg-Y (some manufacturers)

Other solutions:

- Ad-hoc user-tailored formats (usually HDF5-based)
- miniSEED

Candidates to be the next ‘de-facto’ (?) standard:

- Something based on HDF5
  - Known in the community
  - Not well suited for multithread/multiprocess
- Zarr
  - RW multithread/multiprocess
  - Cloud is supported
- TileDB
  - Full multi-threaded implementation
  - Different storage solutions supported natively
  - Versioning

Data Management

- Until new standards are developed and adopted the community needs a seamless way to integrate the DAS datasets using current seismic standard formats (e.g. miniseed for data, StationXML for metadata).

- Strategy to standardize these datasets by downsampling them and creating a basic standard metadata (StationXML) mapped from raw data and extra information provided by the PI. The result is ready to be archived in a standard way.

Toward a Metadata Standard for Distributed Acoustic Sensing (DAS) Data Collection

Voon Hui Lai[*1], Kathleen M. Hodgkinson[2], Robert W. Porritt[2], and Robert Mellors[3]

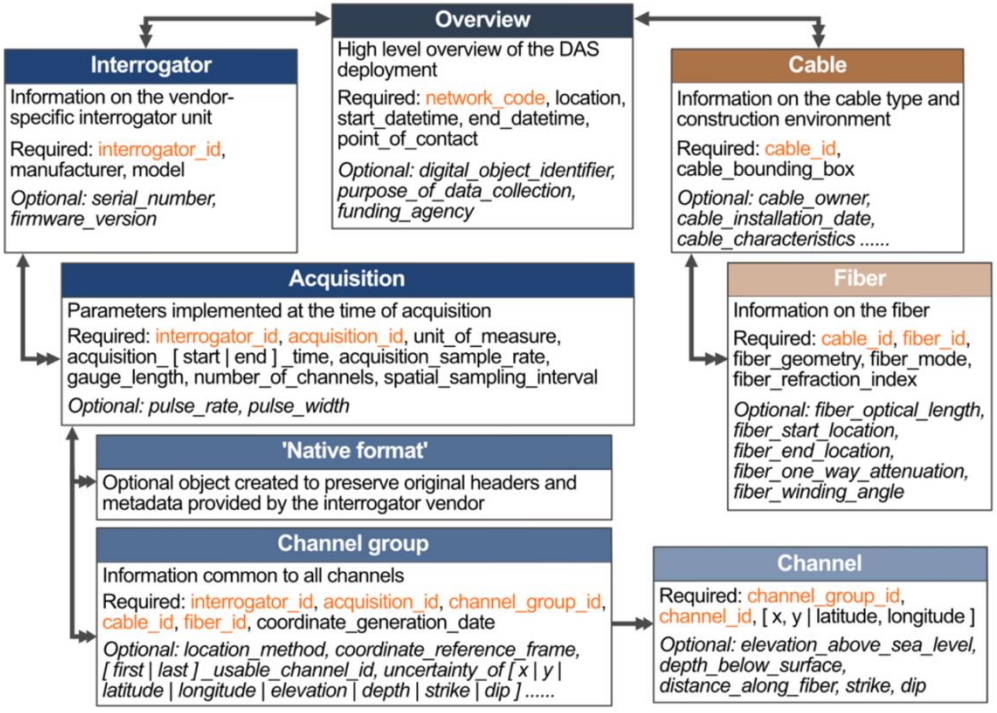
Abstract

With increasing geophysical applications using distributed acoustic sensing (DAS) technology, there is a need to implement a metadata standard specifically for DAS to facilitate the integration of DAS measurements across experiments and increase reusability. We propose a metadata standard intended primarily for the DAS research community, which fully describes the five key components of a DAS experiment: (1) interrogator; (2) data acquisition; (3) channels; (4) cable; and (5) fiber. The proposed metadata schema, which is the overall structure of the metadata, is hierarchical based, with a parent "overview" metadata block describing the experiment, and two main child

Voon Hui Lai, et al. (2024), doi:10.1785/0220230325

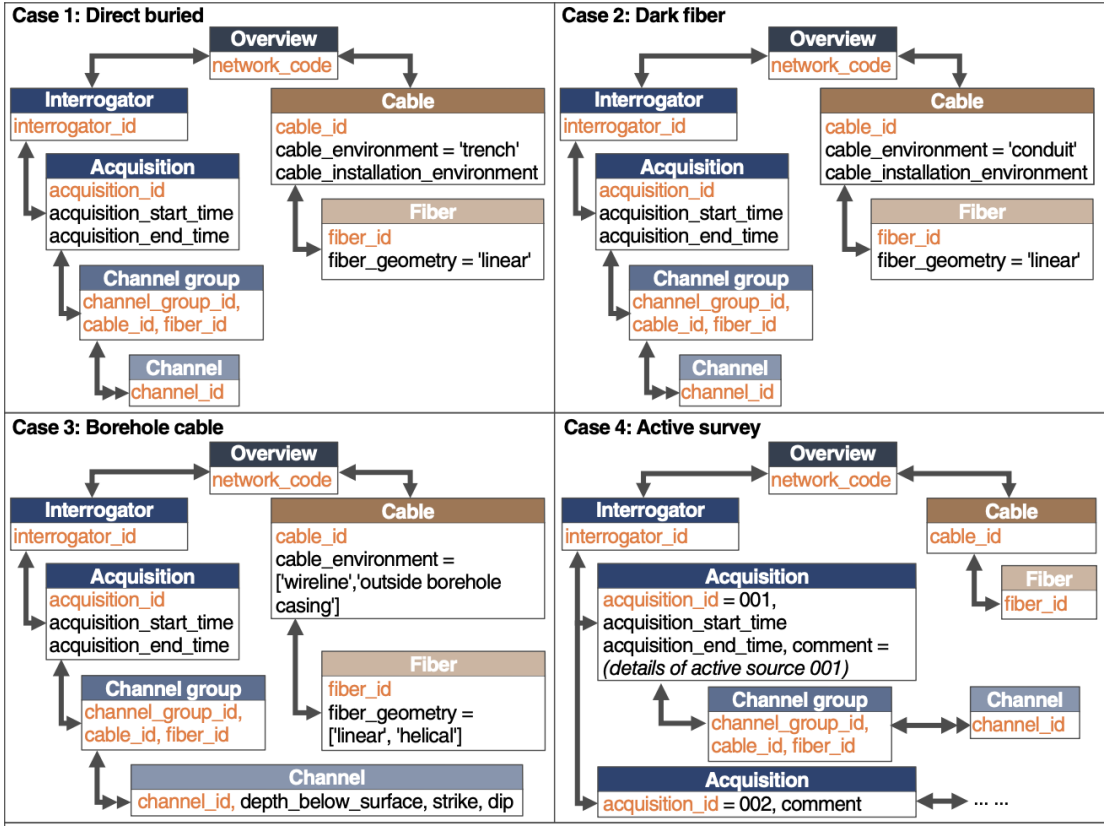
DAS Metadata

- DAS-RCN Data Management Working Group proposed a starting point for a common DAS metadata standard for archival purposes and to guide data collection at experiments.
- The specification was published after 2 years of discussion within the community.
- Orfeus and EarthScope did a first evaluation of the schema and met the authors to prepare a proposal for FDSN evaluation and adoption.



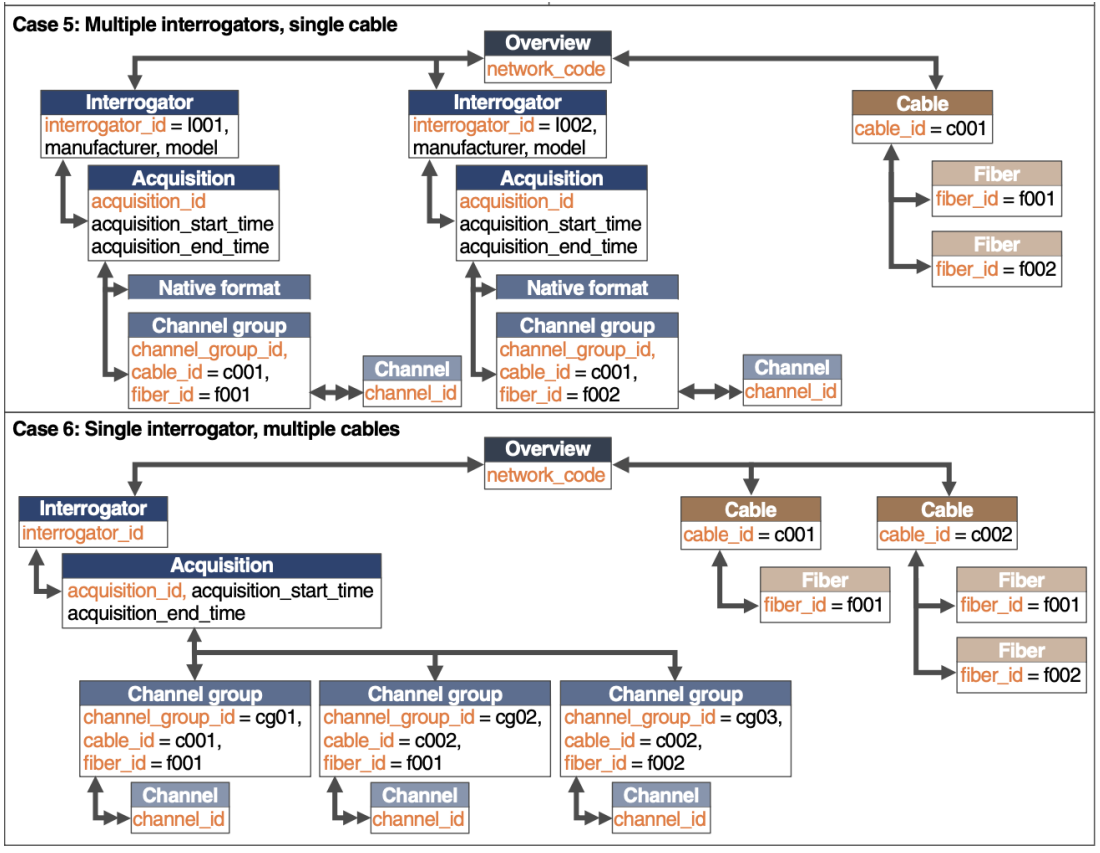
Voon Hui Lai, et al. (2024), doi:10.1785/0220230325

DAS Metadata



Voon Hui Lai, et al. (2024), doi:10.1785/0220230325

DAS Metadata



Voon Hui Lai, et al. (2024), doi:10.1785/0220230325

Metadata schema as FDSN standard

- Meeting with authors some weeks ago
- Proposal to FDSN presenting their work before end of 2024
- First version includes what is present in the paper
- Formal JSON schema including with very minor improvements
- (Hopefully) Approved and adopted as a standard by WG2
- FDSN WG2 takes over the work in a public repository
- Authors remain as members of future Review Teams
- Next release will include a “Processing” section
- Build (or adapt) our software ecosystem on top of that

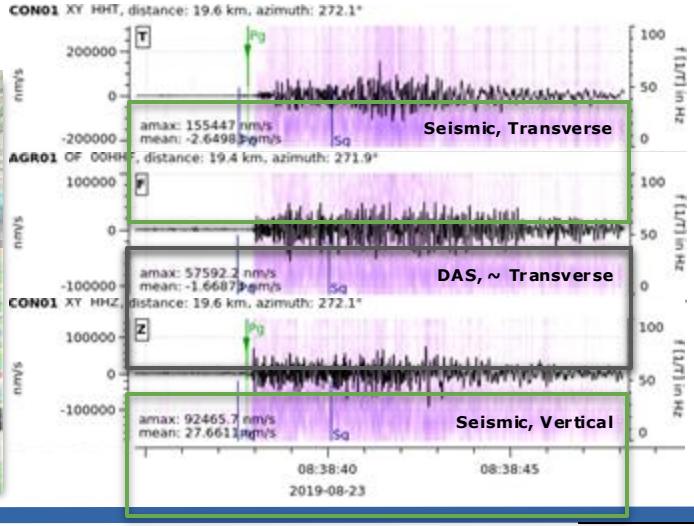
Seismological Metadata

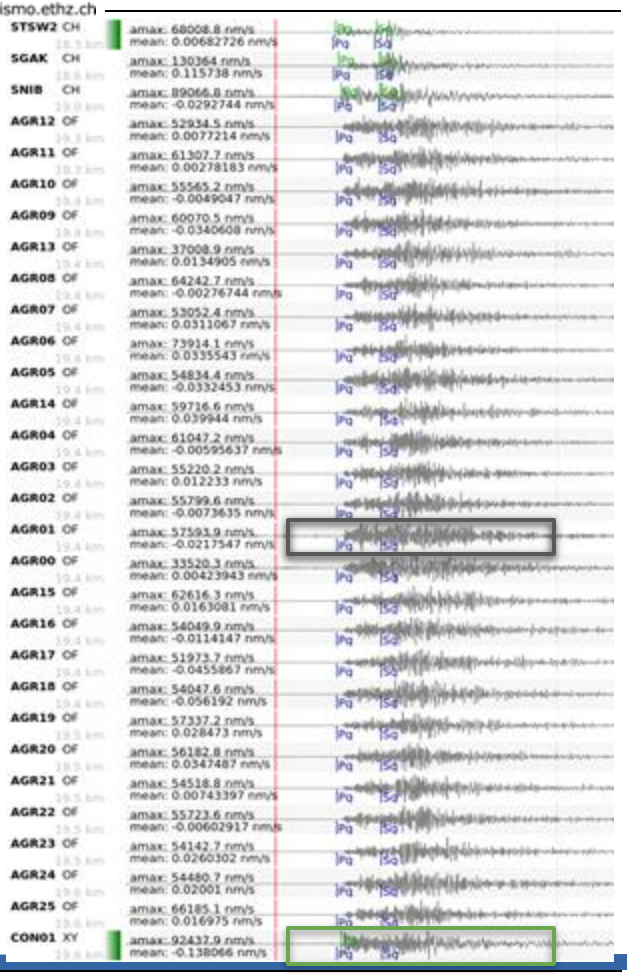
Fred Massin, Pascal Edme, John Clinton (SED-ETH)

Combine event-based data from DAS and seismic network using manual earthquake analysis software.

The goal is to get familiar with DAS data/quality and encourage wider usage.

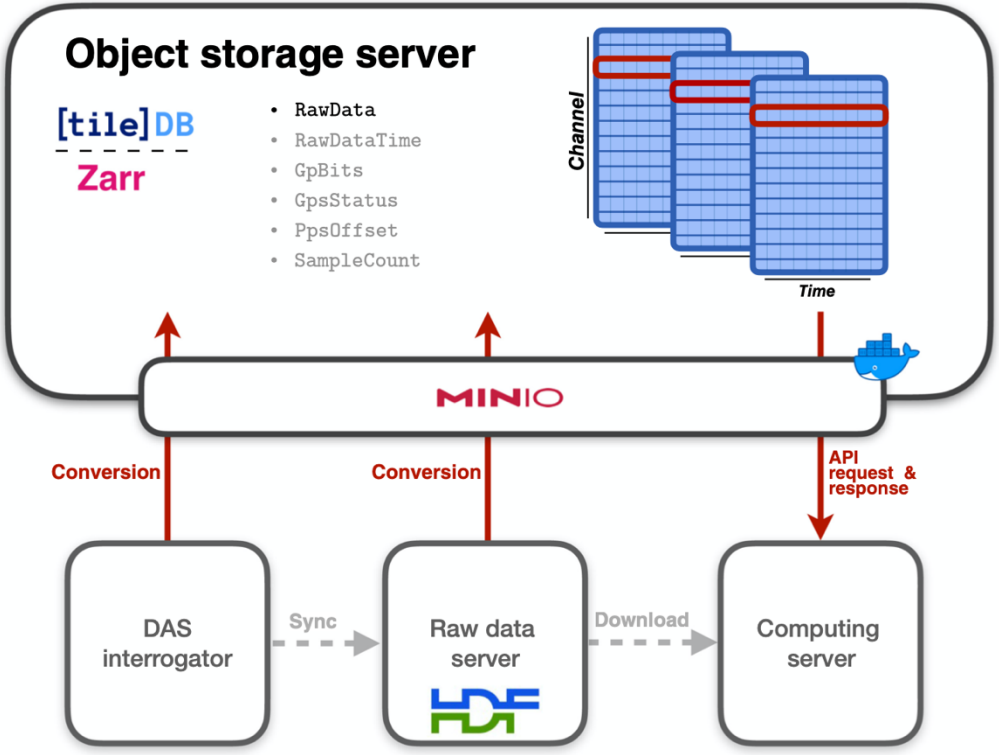
- Pre-processing (*DAS scientist, bespoke processing*):
  - Spatial coherence analysis and stacking
  - Spatial integration to velocity
- Conversion (*can be easily automated*):
  - Strain-rate and velocity timestamping and conversion to mseed
  - Fiber setup data conversion to FDSNxml & SeisComP bindings.
- Open in SeisComP for manual analysis:
  - ***scolv***: pick, location & magnitude





Data formats

- HDF5 can dramatically slow down DAS data sharing in modern distributed computing environments.
- Ni et al. show a promising proof-of-concept based on TileDB, S3 and MinIO.
- TileDB natively parallelizes I/O.
- Parallelize I/O with MPI for TileDB scales very well up to 16 concurrent processes.



Ni, Y. et al. (2023) SRL, doi:10.1785/0220230172

EIDA Guidelines for derived products from DAS experiments

First version agreed with Geo-I partners

Topics included and suggestions:

- Data Management
- How to subsample
- Channel Naming
- Miniseed technicalities
- New DAS metadata (Voon Hui Lai et al, 2024)
- DataCite metadata

Introduction

EIDA nodes are seismological data centres that host hundreds of seismic networks that can be easily discovered by the community, improving the impact of their research. The data provisioning for the community (or project members in case of embargoed data) follows all the international standards from the seismological community and is reachable by most of the software clients available.

Revisions

Version	Collaborators	
1.0	Javier Quinteros	Angelo Strollo
	Frederick Massin	Christopher Wollin
Date	Pascal Edme	Veronica Rodriguez
5th Nov 2024	Philippe Kaestli	John Clinton
	Gilda Currenti	Christos Evangelidis
	Peter Danecek	Michelle Prestifilippo
	Diane Rivet	Shane Murphy
	Jonathan Schaeffer	Jan Michalek

https://orfeus.readthedocs.io/en/latest/das_guidelines.html

Data Management

- Web page of a DAS dataset (Potsdam, Global DAS month) standardized and archived in downsampled form.
- Same standardization for similar datasets from now on.

Network code: 3U

Extended Network Information for network 3U [hide/show]

Creator(s): Wollin, Christopher [a]; Ehsaninezhad, Leila [a]; Hart, Johannes [a]; Rodríguez Tribaldos, Verónica [a]; Krawczyk, Charlotte M. [a]

[a] GFZ German Research Centre for Geosciences, Potsdam, Germany

Title: Global DAS Month 2023, Teleseismic Event Recordings, Potsdam Fiber

Publisher: GFZ Data Services

Resource Type: Other/Seismic Network

Network DOI: doi:10.5880/GFZ.2.2.2023.001

Related Reference(s):

1. Object Storage with raw data and documentation. WWW
2. Wuestefeld, A.; Spica, Z. J.; Aderhold, K.; Huang, H.; Ma, K.; Lai, V. H.; Miller, M.; Urmantseva, L.; Zapf, D.; Bowden, D. C.; Edme, P.; Kiers, T.; Rinaldi, A. P.; Tuinstra, K.; Jestin, C.; Diaz-Meza, S.; Jousset, P.; Wollin, C.; Ugalde, A.; Ruiz Barajas, S.; Gaite, B.; Currenti, G.; Prestifilippo, M.; and Araki, E.; Tonegawa, T.; de Ridder, S.; and Nowacki, A.; and Lindner, F.; and Schoenball, M.; Wetter, C.; Zhu, H.; Baird, A. F.; Rørstadbotnen, R. A.; Ajo-Franklin, J.; Ma, Y.; Abbott, R. E.; Hodgkinson, K. M.; Porritt, R. W.; Stanciu, C.; Podrasky, A.; Hill, D.; Biondi, B.; Yuan, S.; Luo, B.; Nikitin, S.; Morten, J. P.; Dumitru, V.; Lienhart, W.; Cunningham, E.; Wang, H. (2023). The Global DAS Month of February 2023. Seismological Research Letters. doi:10.1785/0220230180 DOI
3. GLOBUS Server (DAS): folder overview with raw data (registration required). WWW

[+]

Place(s): Study area southwest of Potsdam (Germany)

Subject(s)[a]: Monitoring system ; fibre optics

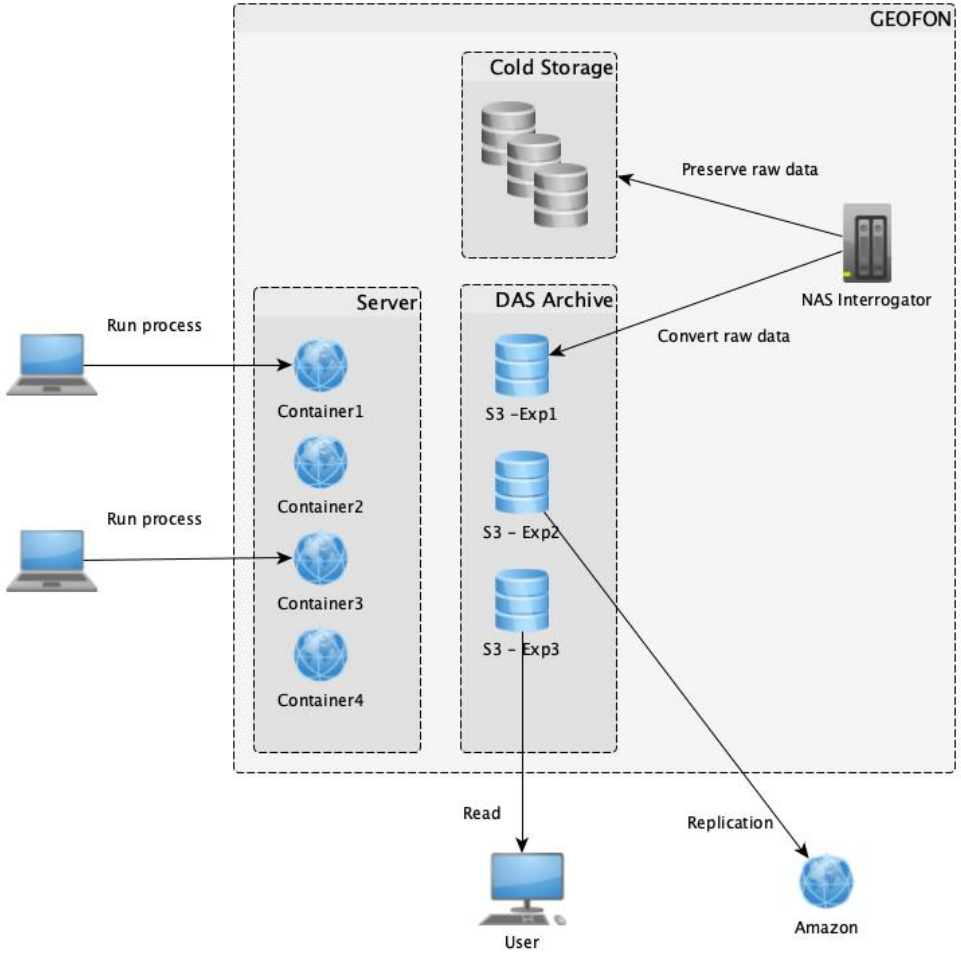
GCMD keyword(s): Geophysical stations/networks | Solid earth

Funding Reference(s): Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum GFZ

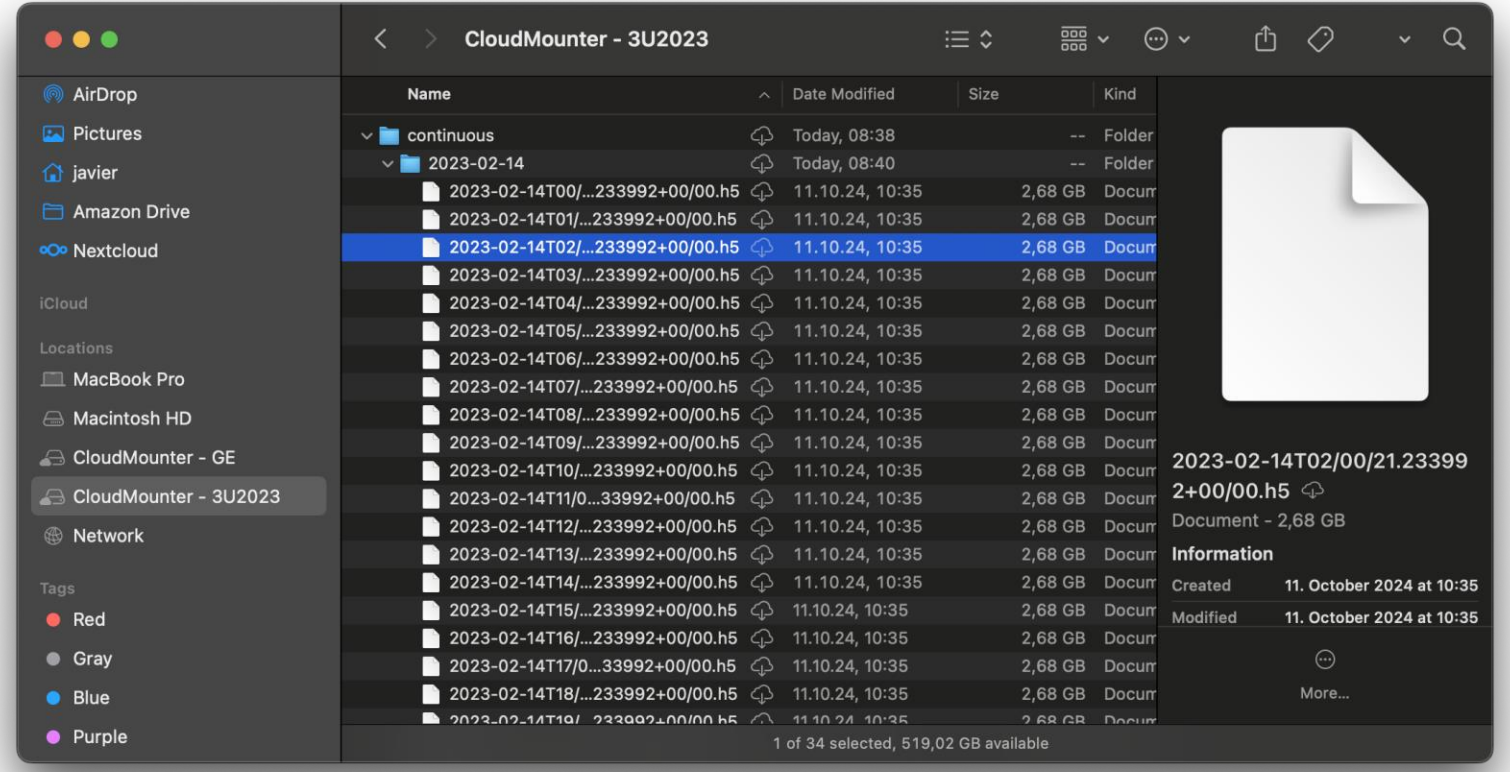
(2023): Global DAS Month 2023, Teleseismic Event Recordings, Potsdam Fiber. GFZ Data Services. Other/Seismic Network. doi:10.5880/GFZ.2.2.2023.001.

Our vision

- Raw data preserved in cold storage (tapes?).
- Full resolution data staged on S3 buckets.
- Open to users and ready to synchronize with other providers.
- Some users could run their codes in containers with fast (local) access to data.
- We have an example you can try:
  - 3U-2023. Check our landing page of the network.

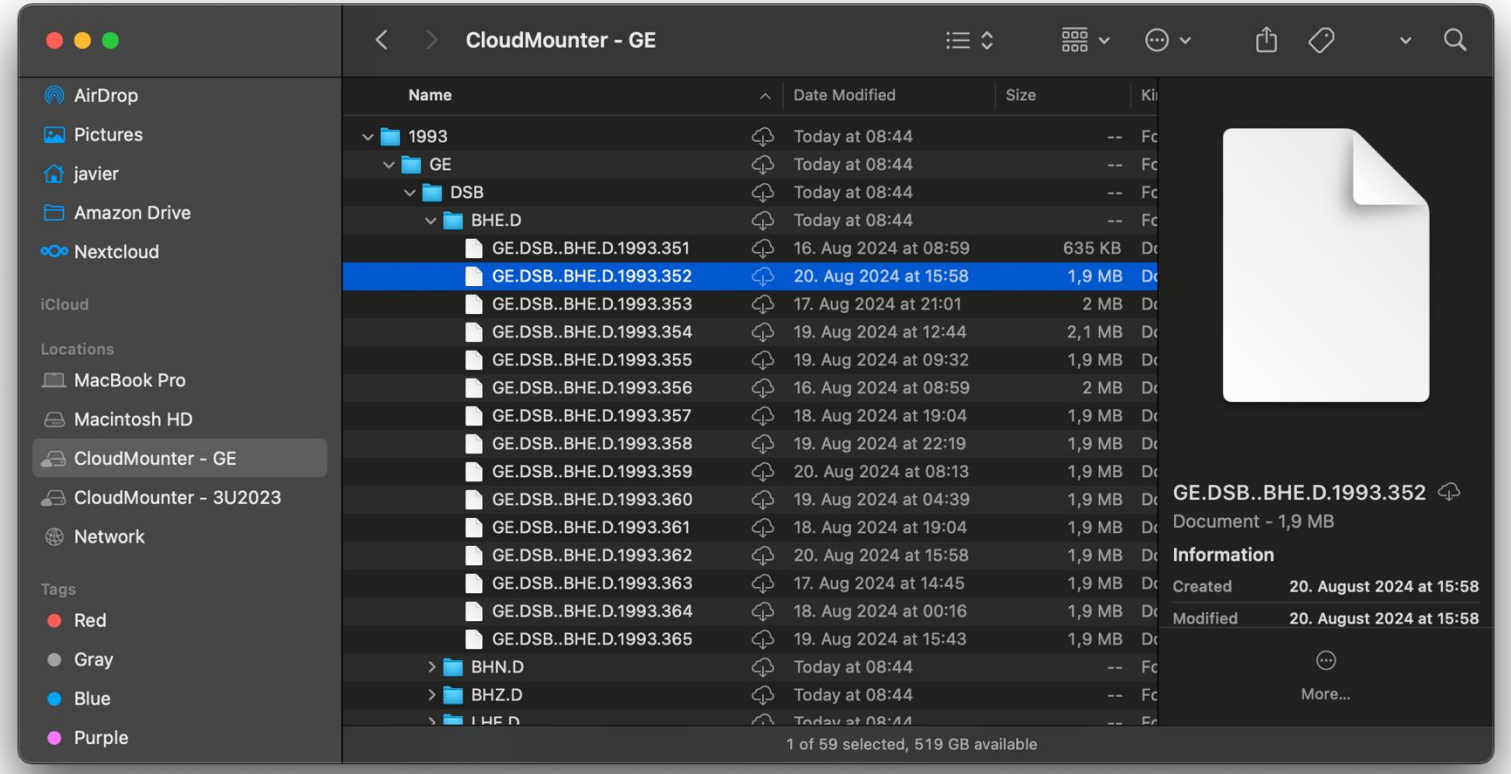


S3 Bucket with DAS data



Endpoint: https://s3.gfz-potsdam.de/ Bucket: gc.3u2023 Size: 468 GB. Objects: 195

S3 Bucket with Miniseed data

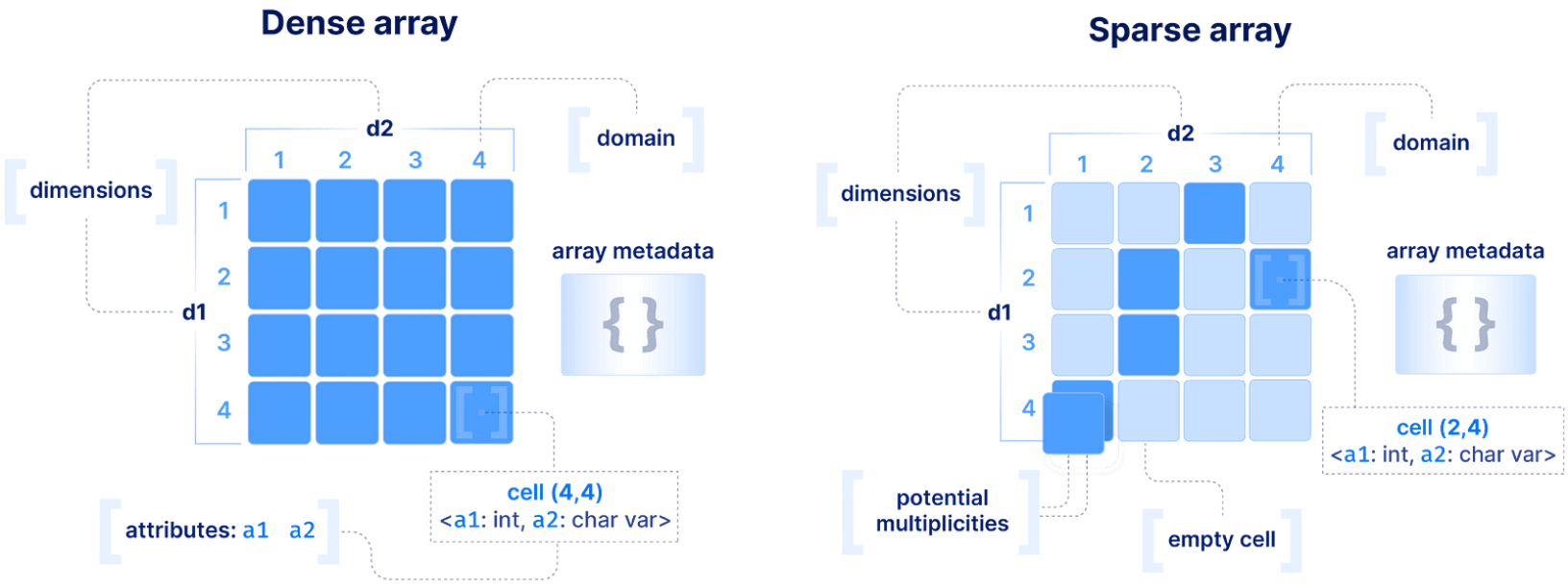


Endpoint: https://s3.gfz-potsdam.de/ Bucket: gc.ge Size: 18 TB. Objects: 17.355.337

TileDB

- Support for both dense and sparse arrays
- Support for dataframes and key-value stores
- Optimized for object stores (AWS S3, Google Cloud Storage, Azure Blob Storage)
- Chunked (tiled) arrays
- Tiling and compression
- Parallel IO
- Data versioning (rapid updates, time traveling)
- Groups
- Arbitrary metadata
- APIs from most typical programming languages

TileDB

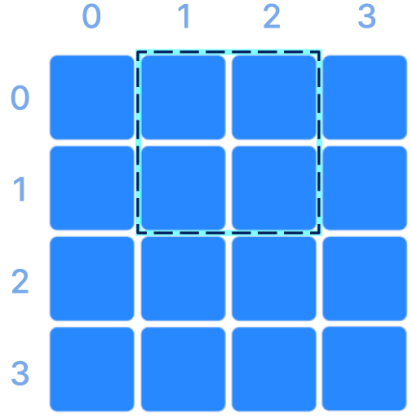

Dense array
d2
1 2 3 4
dimensions
domain
1
2
d1
3
4
array metadata
{ }
cell (4,4)
<a1: int, a2: char var>
attributes: a1 a2
Sparse array
d2
1 2 3 4
dimensions
domain
1
2
d1
3
4
array metadata
{ }
cell (2,4)
<a1: int, a2: char var>
potential multiplicities
empty cell

TileDB

Slicing

```
A [0:2, 1:3]
```

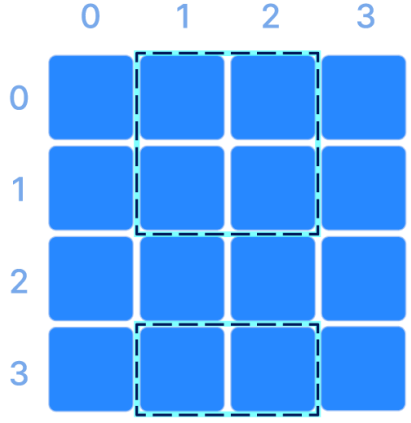
```
SELECT attr FROM A
WHERE d1>=0 AND d1<=1 AND
      d2>=1 AND d2<=2
```



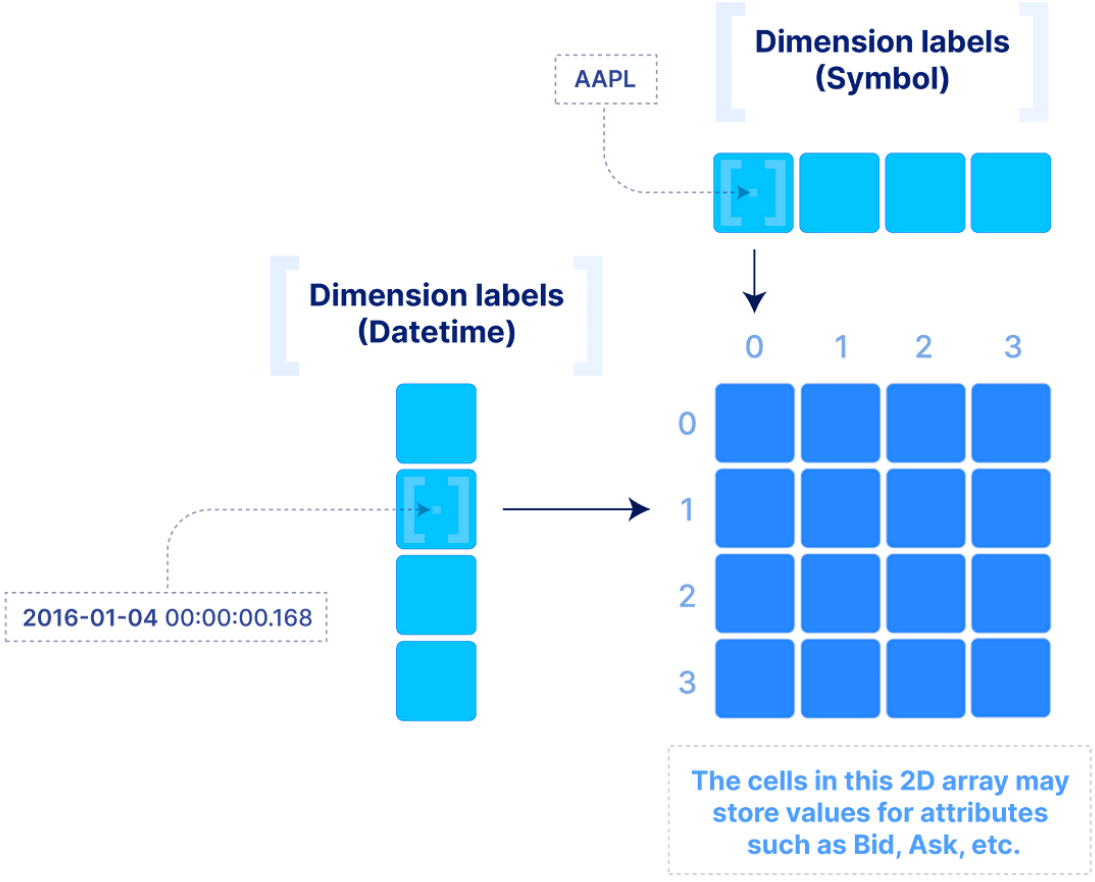
Multi-range Slicing

```
A [[0,1,3], 1:3]
```

```
SELECT attr FROM A
WHERE ((d1>=0 AND d1<=1) OR
       d1==3) AND
      d2>=1 AND d2<=2
```



TileDB

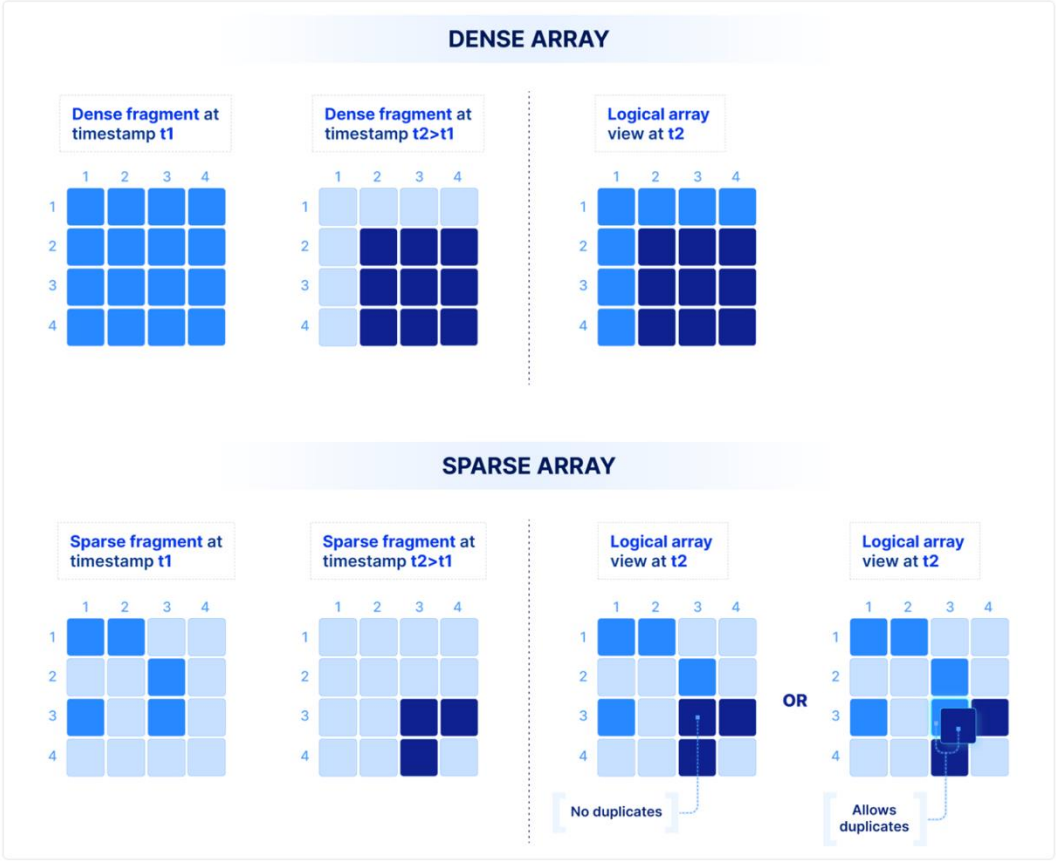

Dimension labels (Symbol)
AAPL
Dimension labels (Datetime)
0 1 2 3
0
1
2
3
2016-01-04 00:00:00.168
The cells in this 2D array may store values for attributes such as Bid, Ask, etc.

TileDB

```
my_array                                # array directory
├── ...
├── __fragments
│   └── <timestamped_name>              # fragment directory
│       ├── __fragment_metadata.tdb     # fragment metadata
│       ├── a0.tdb                      # fixed-sized attribute
│       ├── a1.tdb                      # var-sized attribute (offsets)
│       ├── a1_var.tdb                  # var-sized attribute (values)
│       ├── ...
│       ├── a2_validity.tdb             # validity of fixed- or var-sized attribute
│       ├── ...
│       ├── d0.tdb                      # fixed-sized dimension
│       ├── d1.tdb                      # var-sized dimension (offsets)
│       ├── d1_var.tdb                  # var-sized dimension (values)
│       └── ...
└── ...
```

Copy

TileDB


DENSE ARRAY
Dense fragment at timestamp t1
Dense fragment at timestamp t2>t1
Logical array view at t2
SPARSE ARRAY
Sparse fragment at timestamp t1
Sparse fragment at timestamp t2>t1
Logical array view at t2
Logical array view at t2
OR
No duplicates
Allows duplicates

AI/ML (yes... always, everywhere)

What do we actually need to be fully ready for AI/ML?

- Common standard format and fully homogeneous across data centres (and datasets).
- Be ready to support parallelization and scalability.
- Think in advance about privacy and security issues.
- Licenses! Yes, believe it or not…
  - We need to know if we can use some data for training and under which conditions.

Conclusions

- Different lines to work on:
  - Metadata format
  - Data format
  - Data provisioning
  - Real-time transmission
- Discussion within FDSN about the ”easy” topics:
  - Metadata format
  - Mapping to StationXML. Both will coexist for some time.
  - Channel Naming
- More complex issues:
  - Real-time transmission
  - Data provisioning via S3 buckets
  - Standard format for TileDB? Or standard interface (API)?
  - Abandon the synchronous behaviour?
  - Computation on top-of-the-data
- If the approach succeeds we can expect a slow migration of standard seismic data to this new solution.

Thanks for your attention!