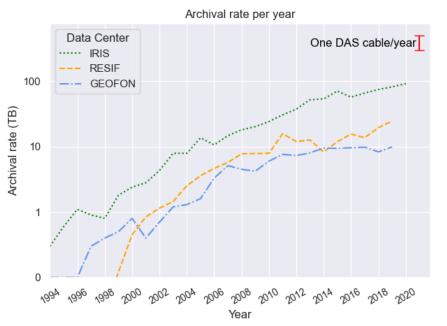
DAS data and products management policies within Geo-INQUIRE Current status and outlook

Geo-INQUIRE Seminars – 5th December 2024

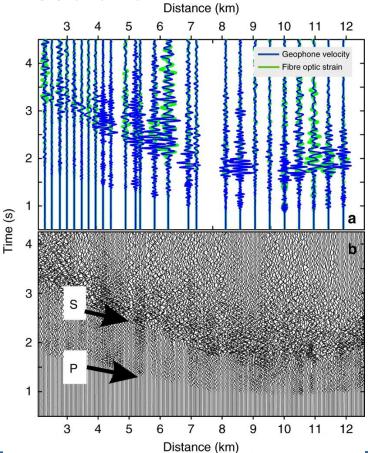
Javier Quinteros on behalf of GEOFON and members of the Geo-INQUIRE DAS Group



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. Carter², Jonathan Schaeffer³, Chad Trabant², 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-





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. Hodgkinson², Robert W. Porritt², and Robert Mellors³

Abstract

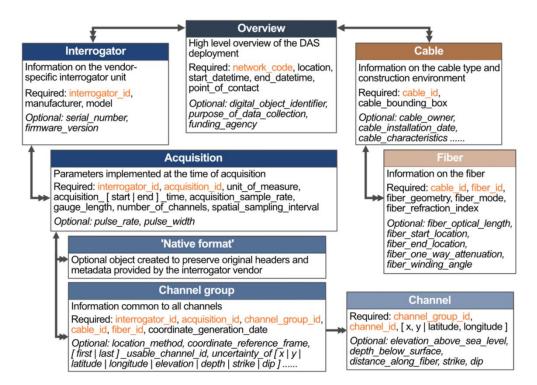
With increasing geophysical applications using distributed acoustic sensing (DAS) technology, there is a need to implement a metadata standard specifically for DAS to facilitate the integration of DAS measurements across experiments and increase reusability. We propose a metadata standard intended primarily 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

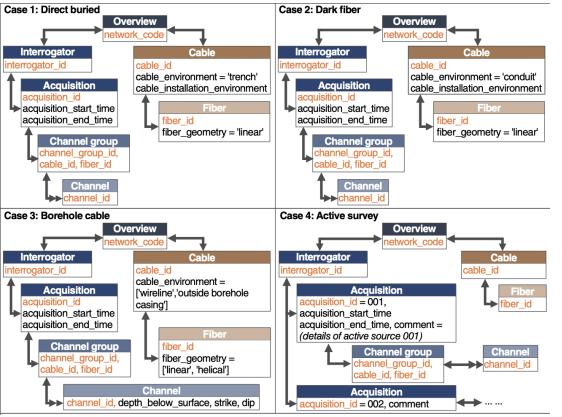
- DAS-RCN Data Management Working Group proposed a starting point for a common DAS metadata standard for archival purposes and to guide data collection at experiments.
- The specification was published after 2 years of discussion within the community.
- Orfeus and EarthScope did a first evaluation of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



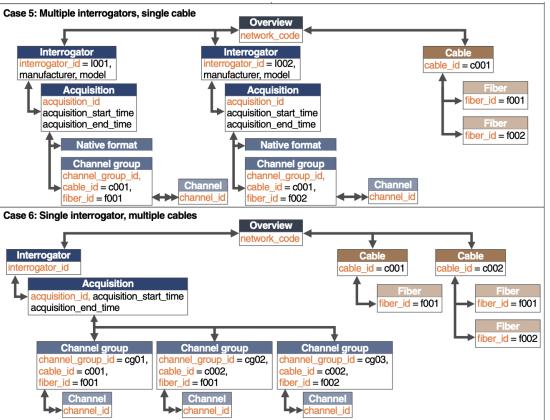
DAS Metadata



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)

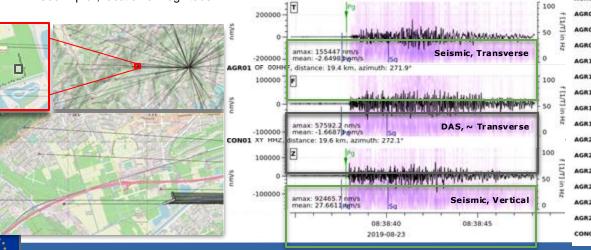
Combine event-based data from DAS and seismic network using manual earthquake analysis software.

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:

П

scolv: pick, location & magnitude



CON01 XY HHT, distance: 19.6 km, azimuth: 272.1

Velocity | Z comp (seismic) | BP 1-30Hz

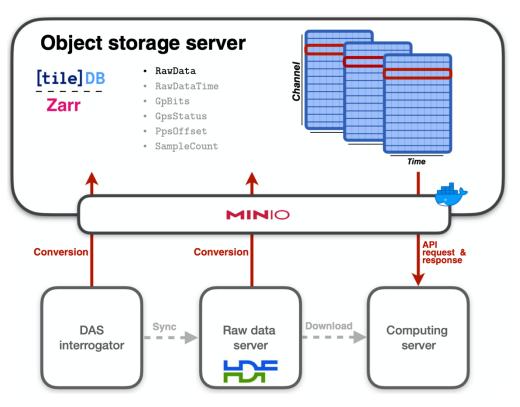
mo.et	hz.ch —				
STSW2	СН	amax: 68008.8 nm/s mean: 0.00682726 nm/		BQ	to lan
SGAK	CH	amax: 130364 nm/s mean: 0.115738 nm/s		- Barn	in the second
SNIB	СН	amax: 89066.8 nm/s mean: -0.0292744 nm/		Selv.	-kalon
AGR12		amax: 52934.5 nm/s mean: 0.0077214 nm/s		Po	Net M
AGR11		amax: 61307.7 nm/s mean: 0.00278183 nm		-	welling)
AGR10		amax: 55565.2 nm/s mean: -0.0049047 nm/	1	-	50
AGR09		amax: 60070.5 nm/s mean: -0.0340608 nm/		-	-
AGR13		amax: 37008.9 nm/s mean: 0.0134905 nm/s		Pa	anda
AGROS	OF	amax: 64242.7 nm/s mean: -0.00276744 nm		Pa	NUMBER
AGR07		amax: 53052.4 nm/s mean: 0.0311067 nm/s	-	-	wille
AGROG		amax: 73914.1 nm/s		Pg	ind inter
AGROS		mean: 0.0335543 nm/s amax: 54834.4 nm/s		P.G	EQ 1
AGR14		mean: -0.0332453 nm/ amax: 59716.6 nm/s		PQ	54
AGR04		mean: 0.039944 nm/s amax: 61047.2 nm/s	1	PQ	15:41
AGROS	OF	mean: -0.00595637 nm amax: 55220.2 nm/s		PQ 1	159 P
AGROZ	OF	mean: 0.012233 nm/s amax: 55799.6 nm/s		Pa	154
AGR01	OF	mean: -0.0073635 nm/ amax: 57593.9 nm/s		Pa	
AGROO	OF	mean: -0.0217547 nm/ amax: 33520.3 nm/s		10	153
	19.4 km	mean: 0.00423943 nm amax: 62616.3 nm/s	15	PQ'	153
AGR16	OF	mean: 0.0163081 nm/s amax: 54049.9 nm/s	1	Pq	153 110
	19.4 hm	mean: -0.0114147 nm/		Pq	15.2 M
	19.4 km	amax: 51973.7 nm/s mean: -0.0455867 nm/	-	Pa	154-11
	10.4 km	amax: 54047.6 nm/s mean: -0.056192 nm/s	1	Pa	154
	19.5 km	amax: 57337.2 nm/s mean: 0.028473 nm/s		Pg	Ba
	13.5.km	amax: 56182.8 nm/s mean: 0.0347487 nm/s		ra	153
	19.5. hm	amax: 54518.8 nm/s mean: 0.00743397 nm	1	Pa	15ari
	10.5 km	amax: 55723.6 nm/s mean: -0.00602917 nm		Pa	Bally
	11.5 km	amax: 54142.7 nm/s mean: 0.0260302 nm/s	3	Pg	时间
	13.8 km	amax: 54480.7 nm/s mean: 0.02001 nm/s	1	ra in	TSHIM
	19.8.km	amax: 66185.1 nm/s mean: 0.016975 nm/s		Po	*
CON01	XY	amax: 92437.9 nm/s mean: -0.138066 nm/s			Sla.

w seismo ethz ch

100	
Po	
Ing.	Alexandre and a second
(Alla)	with particular and the second
Po .	ange fillige have been as a second
IPa	Water State
IPg	erst philippi
1-a	to said of the family as a
IPa	
P q	hogginities and an and an
)Pg	ung Whele Ansama
lea'	He gali pinin in
Pq	Martin Martin Construction
Pa	nggilijijisisse name
JPg.	
Pa	Salating and a second s
Pin .	ALL STREAM PROPERTY AND A STREAM PROVIDENCE AND A ST
17.14	Line I al
	のないのでのないですのですのですのですのですのです。
	and the second
N 2 2	
Z Z Z Z	
Z Z Z Z	
2 2 2 2 2	
2 2 2 2 2 2 2	
2 2 2 2 2 2 2 2	
5 2 2 2 2 2 2 2 2 2	
2 2 2 2 2 2 2 2 3 2 2	
2 2 2 2 2 2 2 2 2 2 2	
2 2 2 2 2 2 2 2 2 2 2 2 2 2	
2 2 2 2 2 2 2 2 2 2 2 2 2 2	

Data formats

- HDF5 can dramatically slow down DAS data sharing in modern distributed computing environments.
- Ni et al. show a promising proofof-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

Revisions

First version agreed with Geo-I partners

Introduction

EIDA nodes are seismological data <u>centres</u> 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.

Topics included and suggestions:

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

Versio	n	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

Network code: 3U

D

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

Web page of a DAS dataset (Potsdam, Global DAS month) standardized and archived in downsampled form.

Same standardization for similar datasets from now on.

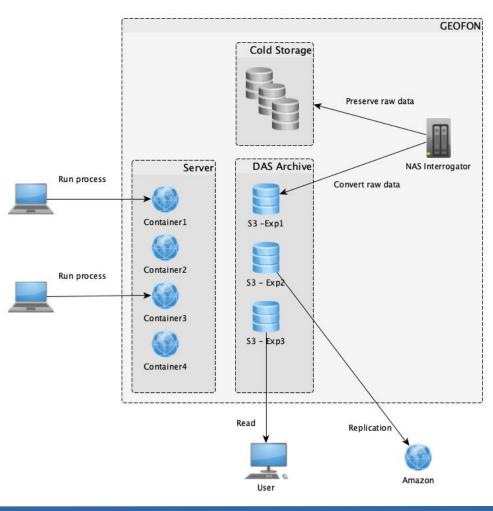
	(inde/snow)
Creator(s):	ा Wollin, Christopher ª; (в Ehsaninezhad, Leila ª; (в Hart, Johannes ª; в Rodríguez Tribaldos, Verónica ª; (в Krawczyk, Charlotte M. ª
	^a GFZ German Research Centre for Geosciences, Potsdam, Germany
Title:	
	Global DAS Month 2023, Teleseismic Event Recordings, Potsdam Fiber
Publisher:	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.
	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).
	[+]
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



alizar Varet

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

•••	< > CloudMounter - 3U2023			≔≎	000	~ @) ~	₾ ⊘		Q
ด AirDrop	Name		Date Modified	Size		Kind				
🖾 Pictures	✓	Ą	Today, 08:38			Folder				
iavier	v 🚞 2023-02-14	ф	Today, 08:40			Folder				
	2023-02-14T00/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
Amazon Drive	2023-02-14T01/233992+00/00.h	5 Q	11.10.24, 10:35	2,6	8 GB	Docum				
ᅇ Nextcloud	2023-02-14T02/233992+00/00.1	15 G	11.10.24, 10:35	2,6	8 GB	Docum				
	2023-02-14T03/233992+00/00.1	15 🗘	11.10.24, 10:35	2,6	8 GB	Docum				
	2023-02-14T04/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
	2023-02-14T05/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
	2023-02-14T06/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
MacBook Pro	2023-02-14T07/233992+00/00.H	15 \bigcirc	11.10.24, 10:35	2,6	8 GB	Docum				
👄 Macintosh HD	2023-02-14T08/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
	2023-02-14T09/233992+00/00.1	n5 🖓	11.10.24, 10:35	2,6	8 GB	Docum				
CloudMounter - GE	2023-02-14T10/233992+00/00.h	5 G	11.10.24, 10:35	2,6	8 GB	Docum	um 2023-02-14T02/00/21.			23399
🖂 CloudMounter - 3U2023	2023-02-14T11/033992+00/00.h	5 🖓	11.10.24, 10:35	2,6	8 GB	Docum	n 2+00/00.h5 🗇			
Network	2023-02-14T12/233992+00/00.h	5 Q	11.10.24, 10:35	2,6	8 GB	Docum	Documer	nt - 2,68 G	В	
	2023-02-14T13/233992+00/00.h	5 G	11.10.24, 10:35	2,6	8 GB	Docum	Informat	ion		
	2023-02-14T14/233992+00/00.h	5 G	11.10.24, 10:35	2,6	8 GB	Docum	Created	11. Oct	ober 2024 a	at 10:3
Red	2023-02-14T15/233992+00/00.h	5 🖓	11.10.24, 10:35	2,6	8 GB	Docum	Modified	11. Oct	ober 2024 a	at 10:3
	2023-02-14T16/233992+00/00.h	5	11.10.24, 10:35	2,6	8 GB	Docum				
Gray	2023-02-14T17/033992+00/00.h	5 🖓	11.10.24, 10:35	2,6	8 GB	Docum		\odot		
o Blue	2023-02-14T18/233992+00/00.H	5 Q	11.10.24, 10:35	2,6	8 GB	Docum		More		
- Durale	2023-02-14T1a/ 233aa2+00/00 H	5 1	11 10 2/ 10.35	26	RO R	Docum				
Purple		1	of 34 selected, 519	,02 GB available						

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



S3 Bucket with Miniseed data

•••	< > CloudMounter - GE		:≡ ≎)) 🖉	~	Q
ด AirDrop	Name		Date Modified	Size	Ki				
🖪 Pictures	~ 🚞 1993	Ą	Today at 08:44		Fc				
🏠 javier	~ 🛅 GE	ф	Today at 08:44		Fc				
	V 🚞 DSB	Ą	Today at 08:44		Fc				
🗎 Amazon Drive	v 🚞 BHE.D	Ą	Today at 08:44		Fc				1
ᅇ Nextcloud	GE.DSBBHE.D.1993.351	Ą	16. Aug 2024 at 08:59	635 KB	Do				
	GE.DSBBHE.D.1993.352		20. Aug 2024 at 15:58	1,9 MB	De				
	GE.DSBBHE.D.1993.353	Ą	17. Aug 2024 at 21:01	2 MB	Do	8			
	GE.DSBBHE.D.1993.354	Ŷ	19. Aug 2024 at 12:44	2,1 MB	Do				
	GE.DSBBHE.D.1993.355	Ą	19. Aug 2024 at 09:32	1,9 MB	Do				
MacBook Pro	GE.DSBBHE.D.1993.356	Ą	16. Aug 2024 at 08:59	2 MB	Do				
🗁 Macintosh HD	GE.DSBBHE.D.1993.357	Ŷ	18. Aug 2024 at 19:04	1,9 MB	Do				
CloudMounter - GE	GE.DSBBHE.D.1993.358	ф	19. Aug 2024 at 22:19	1,9 MB	Do				
2 CloudMounter - GE	GE.DSBBHE.D.1993.359	ф	20. Aug 2024 at 08:13	1,9 MB	Do				
A CloudMounter - 3U2023	GE.DSBBHE.D.1993.360	Ą	19. Aug 2024 at 04:39	1,9 MB	Do	GE.DSB	BHE.D.1	993.352	ረታን
Network	GE.DSBBHE.D.1993.361	Ą	18. Aug 2024 at 19:04	1,9 MB	Do	Document	- 1,9 MB		
	GE.DSBBHE.D.1993.362	Ą	20. Aug 2024 at 15:58	1,9 MB	Do	Informatio	n		
	GE.DSBBHE.D.1993.363	Ą	17. Aug 2024 at 14:45	1,9 MB	Do	Created	20. Aug	ust 2024 a	t 15:5
Red	GE.DSBBHE.D.1993.364	Ð	18. Aug 2024 at 00:16	1,9 MB	Do	Modified	20. Aug	ust 2024 a	t 15:58
	GE.DSBBHE.D.1993.365	Ð	19. Aug 2024 at 15:43	1,9 MB	Do		~		
Gray	> 🚞 BHN.D		Today at 08:44		Fc		\odot		
😑 Blue	> 🚞 BHZ.D	Ð	Today at 08:44		Fc		More.		
Purple	s 🖿 i he d		NN'RN te vehoT		Fr				
			1 of 59 selected, 519 GB avail	able					

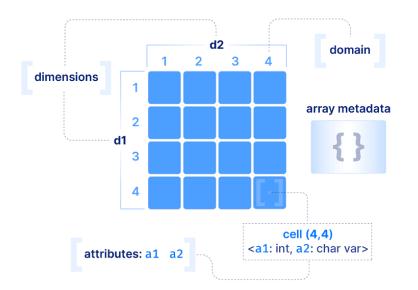
Endpoint: https://s3.gfz-potsdam.de/ Bucket: gc.ge Size: 18 TB. Objects: 17.355.337



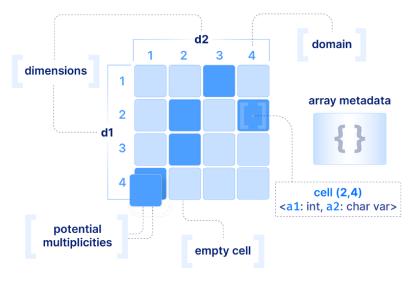
- Support for both dense and sparse arrays
- Support for dataframes and key-value stores
- Optimized for object stores (AWS S3, Google Cloud Storage, Azure Blob Storage)
- Chunked (tiled) arrays
- Tiling and compression
- Parallel IO
- Data versioning (rapid updates, time traveling)
- Groups
- Arbitrary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



Dense array



Sparse array

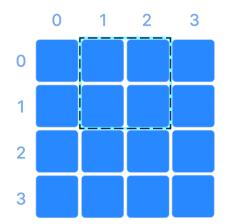




Slicing

A [0:2, 1:3]

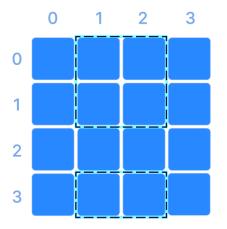
SELECT attr FROM A WHERE d1>=0 AND d1<=1 AND d2>=1 AND d2<=2



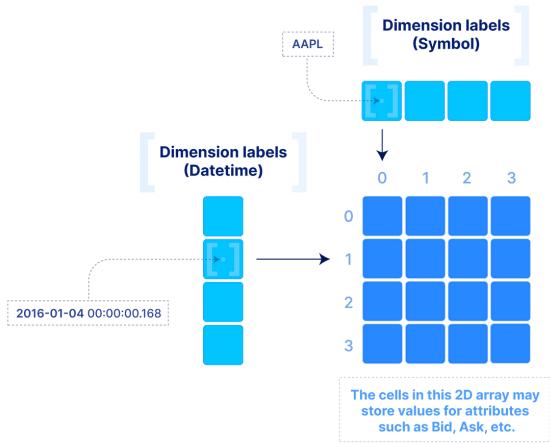
Multi-range Slicing

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



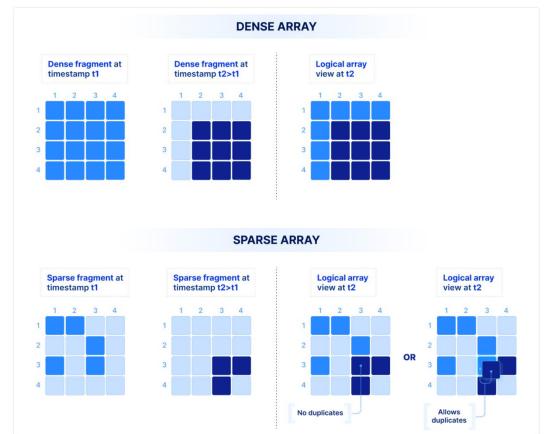






my_array	#	array directory Copy
	 fragments	
	<pre>- <timestamped_name></timestamped_name></pre>	<pre># fragment directory</pre>
		# fragment metadata
	a0.tdb	<pre># fixed-sized attribute</pre>
	a1.tdb	<pre># var-sized attribute (offsets)</pre>
	a1_var.tdb	<pre># var-sized attribute (values)</pre>
	a2_validity.tdb	<pre># validity of fixed- or var-sized attribute</pre>
	d0.tdb	# fixed-sized dimension
	d1.tdb	<pre># var-sized dimension (offsets)</pre>
	d1_var.tdb	<pre># var-sized dimension (values)</pre>
	L	







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!

