Data Management for DAS experiments

Javier Quinteros with material from DAS group at GFZ, GEOFON

Geo-INQUIRE is funded by the European Commission under project number 101058518 within the HORIZON-INFRA-2021-SERV-01 call.

Some challenges related to big datasets

- Formats and services specifications designed in a different landscape.
  - How to provide data from large experiments to users with our current standards?
- If data increases 1, 2 orders of magnitude, technical problems are expected.
- Data formats:
  - How to archive data in the data centre?
  - How to provide data to users?
- Storage space needed.
- Definition of metadata that allows proper processing and interpretation.

Focus Section: European Seismic Networks

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

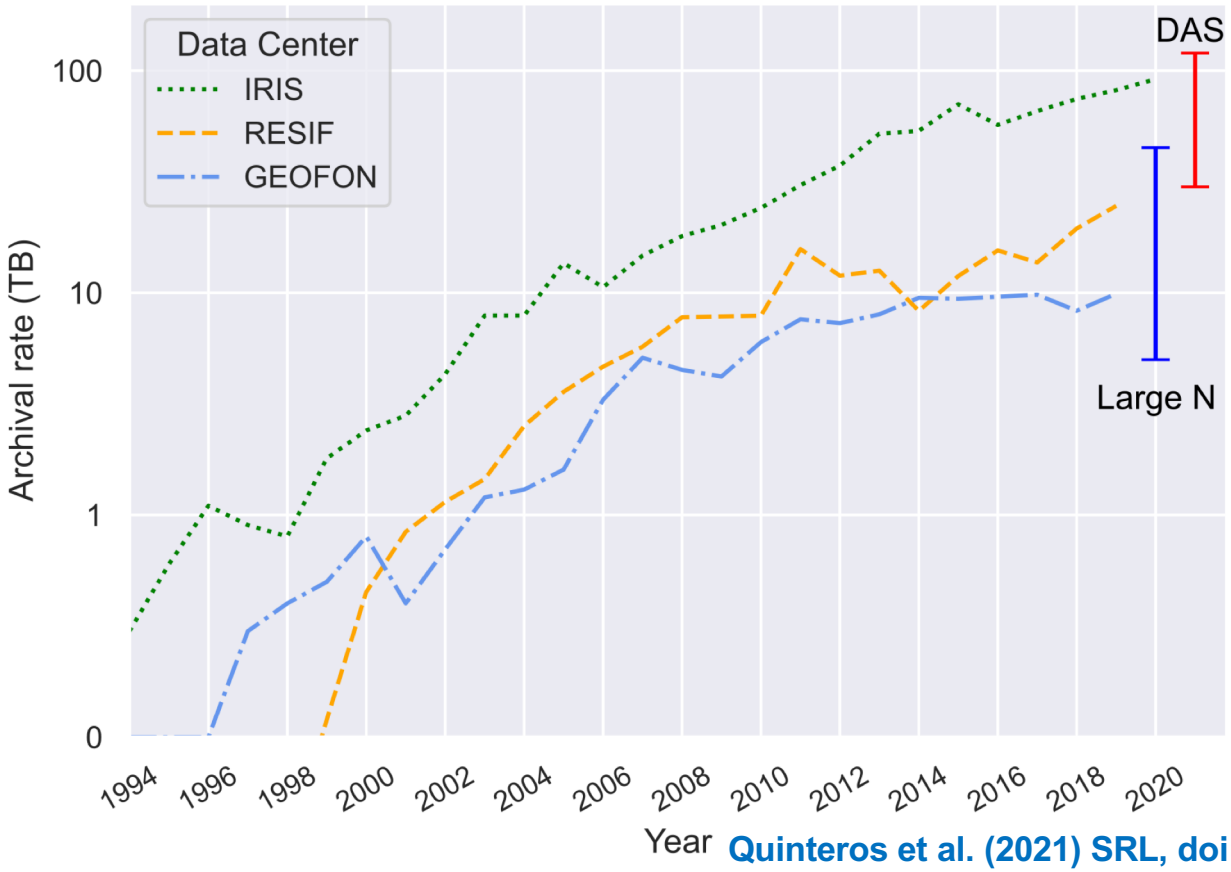
Javier Quinteros[*1], Jerry A. Carter[2], Jonathan Schaeffer[3], Chad Trabant[2], and Helle A. Pedersen[3,4]

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.

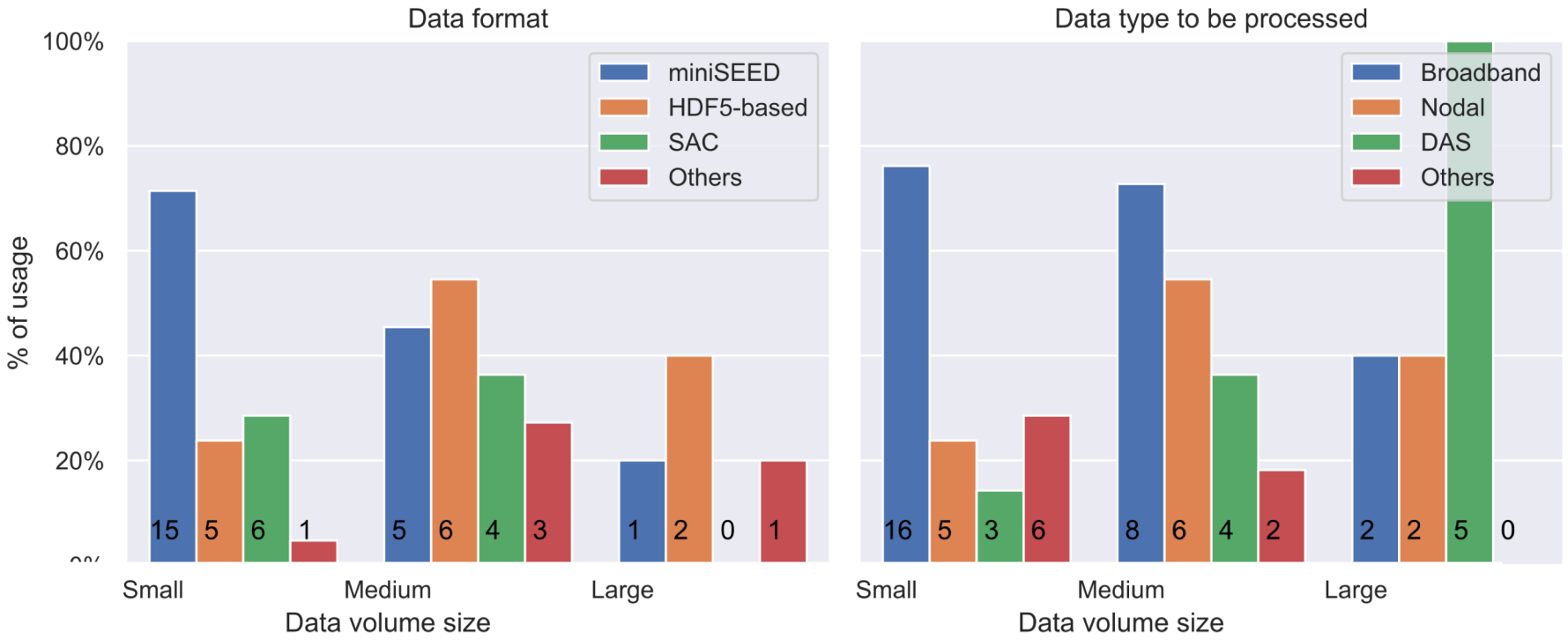
The Incorporated Research Institutions for Seismology—Réseau sismologique et géodésique français and GEOForschungsNetz data centers—carried out a survey and

Archival Rate per Year



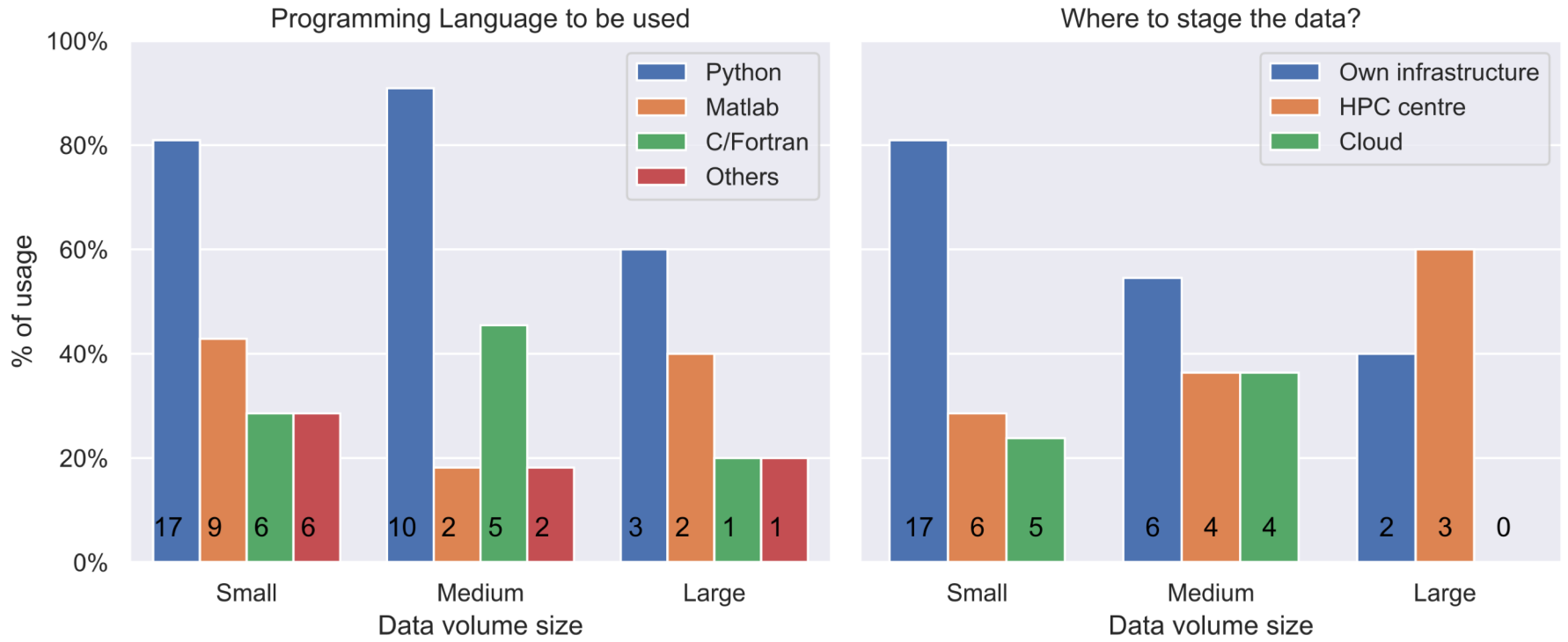
Quinteros et al. (2021) SRL, doi:10.1785/0220200390

User Survey – Big Datasets



Quinteros et al. (2021) SRL, doi:10.1785/0220200390

User Survey – Big Datasets



Quinteros et al. (2021) SRL, doi:10.1785/0220200390

Data formats

Proprietary formats

- TDMS (Silixa)
- HDF5 (OptoDAS)

Community:

- Seg-Y (some manufacturers)

Other solutions:

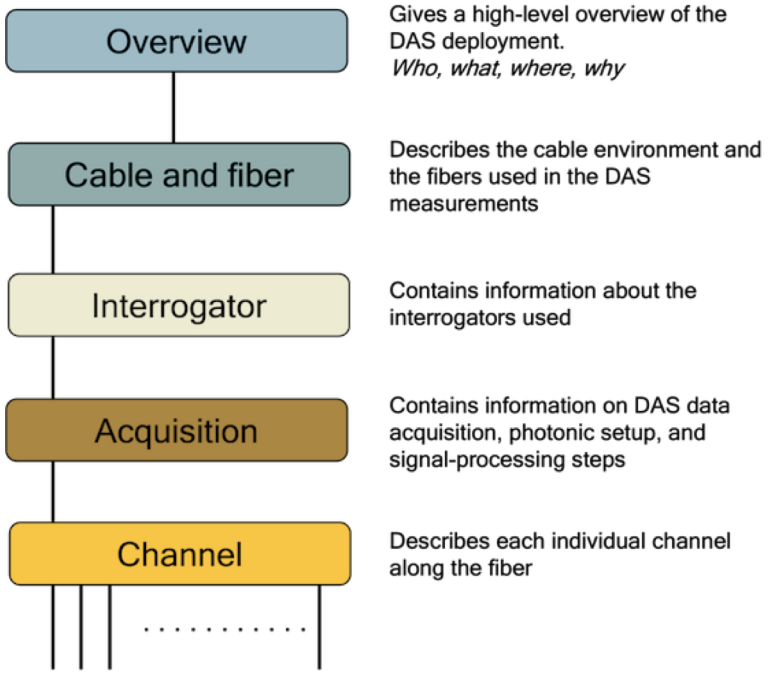
- Ad-hoc user-tailored formats (usually HDF5-based)
- miniSEED

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
  - IRIS candidate

Metadata definition

- DAS-RCN Data Management Working Group proposes a starting point for a common DAS metadata standard for archival purposes and to guide data collection at experiments.
- The intent is that this metadata standard should be independent of the specific implementation and the emphasis is on content.



https://das-metadata.gitbook.io/das-metadata-standard-by-das-rcn/

dastools

Project description

pypi v0.7.3 | python 3.6 | 3.7 | 3.8 | 3.9 | format source | status stable

Tools to work with data generated by DAS systems.

Overview

This package provides a set of tools to read, manipulate and convert seismic waveforms generated by DAS systems. In particular, the ones saved in TDMs format.

dasconv

This utility lets you convert and manipulate seismic waveforms in TDMs format and export them into MiniSEED.

Data acquired from experiments with DAS systems are usually stored in one folder. Files within this folder have names indicating the experiment and the start time of the waveforms saved. An example of the files generated in a test experiment is shown below.

```
$ ls -l
total 1577352
-rwxrwxrwx  1 user  staff   49965056 May  8 09:38 default_UTC_20190508_093735.409.tdms
-rwxrwxrwx  1 user  staff   49965056 May  8 09:38 default_UTC_20190508_093805.409.tdms
```

Quinteros (2021), doi:10.5880/GFZ.2.4.2021.001

What do we offer in dastools?

- Python library to natively read TDMS/OptoDAS(alpha) data (obspy compatible).
- Import and use it in your own python code.
- Dasconv: converts from DAS to miniSEED.
- FDSN web services (station and dataselect) on top of the TDMS files.
- Quality control and reports of the DAS data (WIP)
- Continuous development and improvement of the package. Feedback is more than welcome!

Quinteros (2021), doi:10.5880/GFZ.2.4.2021.001

Check also the work done in Pyrocko!

Our preferred data workflow today

- Stage data in our server.
- Copy raw data to cold storage (e.g. tapes).
- Convert data to standard miniSEED (spatial/temporal downsampling).
- Reserve network code at FDSN (including DOI).
- Archive in your preferred seismological data centre (e.g. GEOFON).
- Discuss internally at each institution how to make raw data available to users.

First steps in the community - PubDAS

Name	IU	T. span (d)	Format	Sps (hz)	Vol. (Gb)	GL (m)	CL (m)	CS (m)	units
Fairbanks	iDAS	59⋆	TDSM	1,000	10,441	10	4,000	1	$\dot{\epsilon}$
FORESEE	iDAS-v2	365	HDF5	$125^{\div}$	29,338	10	4,900	2	$\dot{\epsilon}$
FOSSA	iDAS-v2	7	TDSM	500	11,680	10	23,300	2	$\dot{\epsilon}$
LaFarge	iDAS	2⋆	SEG-Y	1,000	45	10	1,120	1	$\dot{\epsilon}$
Stanford-1	ODH3	940	SEG-Y	50	18,908	7.14	2,500	8.16	ϵ
Stanford-2	ODH3	14	SEG-Y	250	2,887	20	10,200	8.16	ϵ
Stanford-3	ODH4	6	SEG-Y	$\sim$	92	$\sim$	2,500	8.16	ϵ
Valencia	A1-R	7	HDF5	$250^{\div}$	3,213	30.4	50,000	16.8	$\dot{\epsilon}$

Table 1. List of the data sets currently available on PubDAS and their main characteristics. IU: Interrogator Unit; T. Span: Time span in days; Sps: Samples Per Second in Hertz; Vol.: Volume in Gigabytes; GL: Gauge Length in meters; CL: Cable Length in meters; CS: Channel Spacing in meters; $\dot{\epsilon}$: strain rate; ϵ: strain; A $\star$ means data contain active sources. A $\sim$ means that this value may vary; $^{\div}$: means the dataset is downsampled. Name abbreviations are the same as in Fig. 1.

Spica et al. (2023), SRL, doi:10.1785/0220220279

First steps in the community - PubDAS

Origin	Destination	Files transferred	Bytes transferred	Effective Speed (MB/s)	Time
UM	CSM	1585	509.02 GB	491.76	17 m 15 s
UM	UNAM	2744	5.94 TB	102.73	16 h 5 m 5 s
IGN	UM	1	240.08 GB	29.05	2 h 15 m 55 s
ERI	UM	7389	15.64 TB	90.81	1 d 23 h 50 m 40 s
CTC	UM	8868	2.91 TB	21.74	1 d 13 h 17 m 27 s
UM	Caltech	3241	1.08 TB	163.22	1 h 51 m 8 s

Table 2. Examples of data upload and download using Globus and using different network speeds. UM: University of Michigan; CSM: Colorado School of Mines, USA; UNAM: Universidad Nacional Autónoma de México, Mexico; IGN: Instituto Geográfico Nacional, Spain; ERI: Earthquake Research Institute, Japan; CTC: Cordova Telephone Cooperative, Alaska, USA; Caltech: California Institute of Technology, USA.

Spica et al. (2023), SRL, doi:10.1785/0220220279

Thank you for your attention!

Geo-INQUIRE is a joint effort of 51 institutions

Geo-INQUIRE is funded by the European Commission under project number 101058518 within the HORIZON-INFRA-2021-SERV-01 call.