

Developing workflows for the vectorization of legacy seismograms

Corbin Díaz, Elisa Duan, Brian Kim, Sze Lam, Felix Morales, Yoweri Nseko, Emile Okal,

Lucas Schirbel, & Suzan van der Lee

Northwestern University, Chicago, Illinois USA

Northwestern
University

UCDAVIS

Lorraine J. Hwang

University of California Davis, Davis, California USA

INTRODUCTION AND MAIN RESULTS

We embarked on a 5-year project to create online workflows to vectorize legacy data.

In Y1, we:

- reorganized *SKATE* to meet software best practices, upgraded to Python3, and began algorithmic improvements to meanline calculations.
- Piloted the FOLDS FDSN metadata standard in developing a database schema for legacy data with a few modifications.
- Applied ML to remove metadata from records

|

Introduction

The discovery and vectorization of legacy seismic data are two of the barriers to the (re)use of seismic data recorded on physical media. Challenges include imaging the large numbers of records, metadata discovery and curation, and the creation of time series accessible to modern digital data processing methods. Here, we describe and demonstrate progress on the development of an open-source web-enabled data pipeline that aligns with FAIR practices and incorporates emerging FOLDS FDSN data exchange standards for legacy seismic data. The target data sets are the WWSSN scans from the Historical Seismogram Filming Project, scans from the Northwestern University Seismogram Archive Facility and Southern California Seismic Network (SCSN).

Here we describe progress in Year 1.

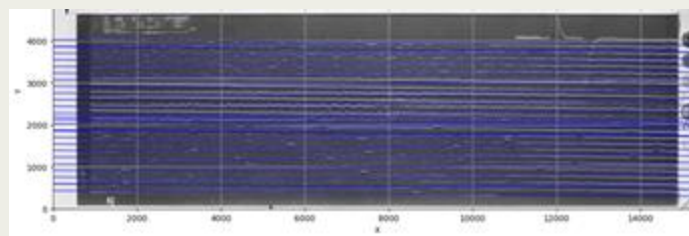
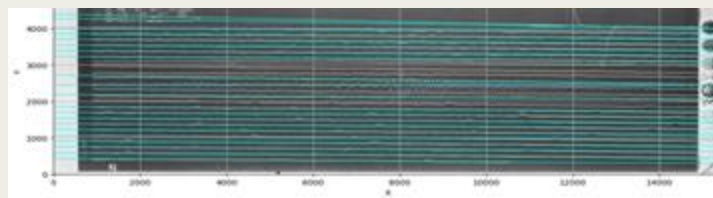
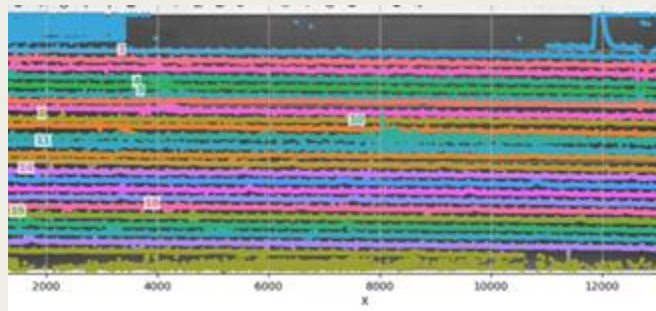


What are your workflow pain points?

OTHER Y1 IMPROVEMENTS include a data overlay tool and an option to choose an image type (negative film or positive). **FUTURE IMPROVEMENTS:** improving segment connectivity by using meanline extrapolation polynomial fitting for a smooth time series, and using a YOLO model (ML) to detect real-time objects requiring labelled data.

Meanlines

SKATE relies on calculating meanlines through pixel data for proper segment assignments and connection algorithms. Meanlines for well-behaved seismograms can be accurately detected (top).



Example of meanline assignment before algorithmic improvements (middle) and after (bottom).



WWSSN
film chip
scans



SKATE

Northwestern
University

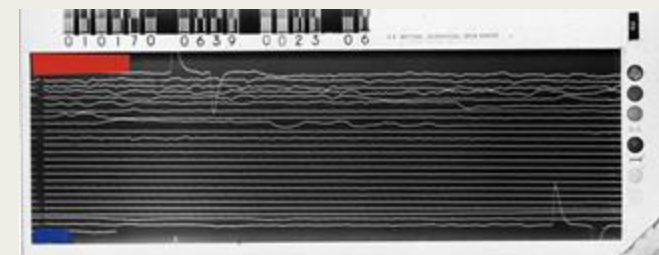
UCDAVIS

BP BORN
PHYSICAL

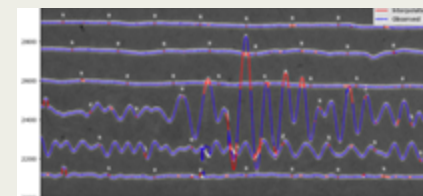
SD STUDIED
DIGITALLY

Metadata Extraction

The initial goal of this part of the project is to automate this process and interpret them as FOLDS metadata, thus saving time and improving accuracy. We are applying ML algorithms in recognizing the WWSSN film chip scan's dogtag and other metadata.



Examples of labelling efforts using WWSSN film chip scans. Here we identify 2 regions of interest: 1. The dogtag is the primary target (red), and 2. Other regions (blue). Both will be scraped and stored as separate layers before passing to the vectorization algorithms in SKATE.



ALGORITHMIC IMPROVEMENTS here are helping to interpret signal gaps and associate orphan segments.

Database Design

One seismogram can have multiple images.
One image only pertains to one seismogram.
Image ID is unique.

image	
image_id	autogenerated for now (PK)
PID	integer (FK)
date_scanned	date
DOI	text
resolution*	integer
x_pixel	integer
y_pixel	integer
format	char(4)
size	integer
length	float
width	float
phase_markings**	boolean
bulletin	text
exclusions**	boolean
recording_type*	text
signal	boolean
timestamp	text ("positive", "negative", "null", "unknown")
notes	text
seism_contact	text
location_record	text
vectorized	text ("true", "false", "null")
recording_gain	int4
image_path	text

One seismogram belongs to one station.
One station can have multiple seismograms.
PID should be unique.

data	
PID	autogenerated for now integer (PK)
network_code*	text (FK)
station_code	text (FK)
location_code	text (FK)
channel_name	text (FK)
physical_location	text
start_time	timestamp (placeholder for now)
end_time	timestamp (placeholder for now)
time_correction	float
polarity	text ("up", "down", "unknown")
data_notes	text
source_of_info	text
date_creation	date

Bold: Required Fields
Pink: To be Imported
White: To be scraped
* Drop down menu
** Drop down boolean

network	
network_code*	text (FK)
network_name*	text

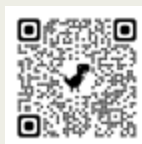
One network to multiple stations.
One station is linked one network.
However, station_code is only unique within the network.

station	
network_code*	text (FK, FK)
station_code	text (FK)
site_name*	text
longitude	float8
latitude	float8
elevation	int4
open_date	date
close_date	date

One station can have multiple location (sensors).
One location is linked to one station.
However, location is only unique within the network.

location	
network_code*	text (FK, FK)
station_code	text (FK, FK)
location_code	text (FK)
sensor*	text
sensor_name	text
install_date	date
depth	int4

channel	
network_code*	text (FK, FK)
station_code	text (FK, FK)
location_code	text (FK, FK)
channel_name	text (FK)
free_period	float8
damping	int4
sensor_serial_number	int4
dp	float8
azimuth	float8
recording_serial_number	float8
period_of_gain	float8
recording_name	text
paper_speed	float
r	float
i	float
b	float
a	float
FDSN_time_series	text



FOLDS
metadata
on Zenodo

Analog vs. Digital

Example 1: Instrumentation



WWSSN 3 component
SP + 3 component LP



STS-2 triaxial
seismometer

Analog - each component is a separate instrument package
vs.

Digital - 3 components can be bundled into the same instrument package.

Example 2: Data packages



Analog - 1 record each day, some metadata on record.
vs.

Digital - miniSEED continuous waveform 512 or 4096 byte.

Added:

location.install_date
channel.FDSN_time_series

Subtracted:

horizontal 2 dip/azimuth

Disagree:

"PID" needs further discussion



What other digital concepts do not apply to analog data?

Should time series be "repaired" or similar to digital data, do we allow for signal drop outs?

What about minute markings?

Legacy Data Resources



Looking for legacy data resources?
Start at the Legacy Seismic Data (LSD) website for links to data, software, references and the FOLDS specification.



Have a correction or addition?
Contribute directly to the GitHub repository.

lsd-sphinx.readthedocs.io/



Seismica is a community-driven, *Diamond Open Access* journal publishing peer-reviewed research in seismology and earthquake science.

<https://seismica.library.mcgill.ca/>

SCSN Data Availability



Late Spring we began reorganizing and inventorying the analog records from Southern California Seismic Network (**SCSN**) with records from 1925 -1990's. Researchers may request scanned images from this collection. This corpus will be used in the development of the BP/SD pipeline.



Geodynamica, the latest in the DOAJ family, is now open for submission. Geodynamica focuses on the **understanding of geodynamic processes that shape the Earth and (exo)planets** and welcomes studies based on observations, experiments, simulations, and models. geodynamica.org

Publish: Special issue



Publish your work using in this **special issue** Analog Seismograms in the Digital Era: Methods, Applications, and Perspectives. Deadline March 1, 2026.

<https://link.springer.com/collections/bgbgegjbgd>