

ANSS ARCHIVE REQUIREMENTS

1. Introduction

The foundational document of ANSS, Circular 1188, states: “Long-term investigations of earthquake and volcanic processes and effects require investment in data management facilities to organize and distribute raw seismic data for research purposes... specialized facilities are needed to archive and distribute raw seismic data.”

This document describes the general requirements for such an archive. In broad terms the purpose of the archive is to:

- Assemble raw ANSS digital seismic data into a unified, well-managed collection
- Insure the safe, secure, and permanent preservation of those data
- Provide tools and interfaces for addition of new data
- Insure access to the data by providing tools and interfaces for searching, marshalling, and retrieving data sets

For the purposes of this document the term “archive” refers to a collection of digital seismological ground motion records of varied types which have been selected by ANSS for permanent preservation and that must be organized and managed to allow fast and easy access by users. It is intended to encompass only “raw” digital waveform data and, therefore, must include the metadata required to correctly interpret those waveforms, and an event catalog sufficient to function as an index into the data. It is not intended to be a comprehensive repository of all seismological products of every kind.

Because of its intended long duration the archive is not defined by any particular format, media type, location, or storage strategy. Any specific formats, media, protocols, etc. specified in this document are descriptive and reflect the state of current technology, but it is anticipated that the concrete implementation of the archive will change over its lifetime. Thus, the archive’s digital holdings are inherently “virtual” (the data), not physical (the media), and will migrate to new archive technologies as they become available.

2. Archive Holdings

The archive “holdings” are the digital contents of the archive and are independent of the actual design or implementation of the archive. The holdings will include the following continuous time-series data as described below. Other earthquake-related products derived from waveform and other geophysical data are not within the scope of the archive and will be developed, produced, and delivered by other means.

2.1. Waveforms

Curation of digital continuous seismic waveform data is the primary task of the archive. Waveforms are also called time-series, time-histories, ground motion data, records, etc. A waveform is a digital representation of ground motion over time. The type of ground motion (displacement, velocity or acceleration) depends on the sensor and how the data are processed. All waveform data can be assumed to be sampled at a regular, known sample rate. As a minimum, the archive must accommodate sample rates in the range 0.001 to 2000 samples per second. Other sample rates may be archived in the future.

2.1.1. Continuous Waveforms

The archive must accept, store, and index continuous waveform data streams. A continuous data stream is one that grows forward in time in near real-time. Data providers and ANSS will determine which channels are appropriate for archiving in continuous mode.

All waveform data, even “continuous”, is discrete or packetized with some granularity. Therefore, gaps, overlaps, out-of-order, and late packets of continuous data must be handled correctly.

2.1.2. Triggered Waveforms

The archive must accept, store, and index non-continuous waveform data. These waveform “snippets” are frequently produced by systems that save only time windows likely to contain ground motions of interest in order to economize some resource like bandwidth or storage. For the sake of data modeling within the archive, triggered waveforms may simply be treated as very “gappy” continuous data, thus blurring the distinction between continuous and triggered waveforms.

2.2. Non-Seismic Time Series

ANSS may choose to archive time series related to seismic monitoring other than ground-motions; e.g. mass positions, state-of-health, barometric pressure, temperature, time code, etc. The archive must handle these correctly.

2.3. Metadata

In this context “metadata” means the information required to identify where, when, and how each waveform was recorded. The primary requirement is that actual ground motion can be recovered for every waveform. At a minimum the metadata must include:

- 2.3.1. The location of the sensor, within 10m in both the horizontal and vertical direction within a standard, well documented reference system (e.g. WGS84).
- 2.3.2. The orientation of the sensor within 5° with respect to true north and true vertical.
- 2.3.3. Sufficient information to recover the UTC time of each sample in the waveform to the accuracy achievable by the original data collection method. In other words, the archiving operation must cause no loss of timing accuracy or precision. Leap seconds must be properly accounted for.
- 2.3.4. The instrument response described by a standard, documented seismological technique so as to allow recovery of the true ground motion within the limits of the sensor.

2.4. Event Catalog

Data users commonly want waveforms representing ground motions caused by particular events. Therefore, an event catalog indexed to associated waveform snippets or windows in continuous data streams must be supported by the archive. This catalog will be provided to the archive by ANSS and ANSS will be free to modify and update the catalog.

2.5. Total Data Volume

The total volume of data to be archived by a fully implemented ANSS of 1,000 stations is estimated to be less than 29 Tb/year (Appendix A). The actual total may be larger if ANSS elects to include data from cooperating and contributing stations.

3. Archive Attributes

The archive must have the following attributes.

3.1. Data Input from Providers

The archive will provide a well defined IP service or services, approved by ANSS, by which it will receive continuous data streams or packages of new and updated data from data producers via the internet. All data, including current and historic data, will be submitted to the archive via these services by ANSS data producers.

The archive will provide a well defined IP service or services, approved by ANSS, by which it will receive historic, new and updated metadata from data producers via the internet.

These data transfer protocols must guarantee delivery and correctly handle late, out-of-order, duplicate, and missing data. The protocols must recognize and flag

malformed data. Additional data integrity checks may be applied.

3.1.1. Data Storage Formats

Data must be archived in a format that guarantees the maximum integrity of the data. Information must not be lost or corrupted in format conversion. Copying to additional formats to facilitate the efficient operation of the archive system is acceptable. It is anticipated that most data will be delivered in SEED or mini-SEED format. Data compression schemes must be lossless.

3.2. Data Organization

All archived data must be organized and indexed in a manner that will allow efficient inventory, searching, and retrieval of the holdings.

The organization of the archive must correctly account for duplicate, overlapping, out-of-order, and missing data. Auditing and versioning must document any processing that changes the data from its raw form as received from the data providers (e.g. time corrections, reformatting, compression, etc.)

Source attribution for each data stream must be maintained.

3.3. Quality Assurance

The quality of the data archived is primarily the responsibility of the data providers and not the archive. However, the archive should check for obvious blunders like malformed data as it enters the archive. Updates and revisions submitted by data providers should be tracked by some form of version control. Only the most current and, presumably, best version will normally be presented to users unless they specifically request alternative data sets.

3.4. User Data Access

Archive data will be made available to scientific and engineering users via web-based interfaces. Requested data will be delivered via the internet. Other delivery mechanisms may be added to keep pace with evolving technology and user demand.

3.4.1. Anticipated Users

The primary users are expected to be seismologists, engineers, and other professionals. While archive users will have a relatively high level of seismological and computer sophistication the archive interfaces and tools should be user-friendly, intuitive, and as simple as the task allows. The archive must document its holdings, practices, and interfaces to allow users

to correctly interpret and access the data.

3.4.2. User Interface

The web-based user interface will support at least two modes of accessing the archive holdings. The first will allow the user to specify sets of channels and time windows. The second will allow users to select seismic events and will return sets of time series that have significant signals associated with those events. The user tools should conform to best practices and standards of the seismological community.

The public web interface of the archive must be presented to users in the *usgs.gov* domain. It must conform to USGS web guidelines and will be subject to review and approval by USGS.

The archive must keep information that tracks usage like who accesses the archive, what data types and formats are used, user feedback, etc.

3.5. Archive Reliability and Performance

The archive must be highly available, maintaining an up-time of 95% or better for both input of new data and access by users. No single period of unavailability should exceed 24 hours.

3.5.1. Input

The archive input services will be available 95% of the time or better. No single period of unavailability should exceed 24 hours. In the event of a period of unavailability the data providers can be expected to buffer the data for up to 24 hours. Upon recovery from a period of unavailability the archive will “catch up” by accepting and archiving the buffered data.

3.5.2. Data Access

The archive’s public data access interface will be available 95% of the time or better. In most cases routine user requests should be delivered within minutes or less. Unusually complex or large but reasonable requests should take no more than 24 hours to deliver. Actual data download/upload rates should be limited only by the network on the user end of the connection. The data access interfaced must be capable of meeting the data service demands that are expected after a noteworthy event.

3.6. Archive Duration

The archive holdings are to be kept in perpetuity. The archive must have a written plan to insure the long term preservation of the holdings. The design of

the archive must anticipate the future need to migrate to new storage technology in the future. This may include changing the organization, format, media, or design of the archive.

3.7. Backup and Recovery

The archive must devise and implement a Continuity of Operations Plan to insure operation in the event of a disaster at the archive facility. This plan must be reviewed at least biannually. The archive must maintain at least two full backups. Incremental backups of new and modified data must be performed daily. Copies of these backups must be stored off-site on media with a long shelf life.

3.8. Security

The archive infrastructure must comply with U.S. Government computer security requirements.

4. Ownership of Holdings

Most of the archived data interpreted to be Federal Records and, as such, U.S. federal law places all archive data in the public domain. The USGS will retain unlimited rights to all data stored in the archive and any software produced to meet the archive requirements (FAR 27.4). The archive operator may not sell archive-related data or services or otherwise profit from its role (apart from any fee or profit paid by the USGS under the contract).

Appendix A: Estimate of Data Volume

The total volume of data to be archived per year by a fully implemented ANSS is estimated to be less than 33 Tb/year. The actual total may be different. Any data volume estimate depends on the following factors:

1. The total number of stations to be archived
2. Whether those stations/channels are archived in continuous or triggered mode
3. The number of channels archived for each station
 - a. The digitization rate of each channel
 - b. The sample size (bits/sample)
4. The data format
 - a. The archive packet size and data header “overhead”
 - b. The efficiency of data compression, if used

The following is an estimate based on current common practices:

1. Total target station count for ANSS = 1,000
2. Continuous recording of all 1,000 stations
3. Three broad-band and three strong-motion channels per station (six total)
 - a. 100 samples-per-second sampling rate on all six channels
 - b. Minimum 24-bits per sample represented as a 32-bit number
4. Data format (miniSEED):
 - a. Header is 64 bytes of a 512 byte packet or 12.5% of the data volume. That means 87.5% of the data volume is waveform.
 - b. A compression ratio of 3.0 will be achieved, on average, on the waveform portion of the SEED record (Steim-2).

Data Volume Estimate :

One channel (raw)

$$= 100 \text{ samples/sec} * 4 \text{ byte/sample} * 3.15 * 10^7 \text{ sec/year} = \mathbf{12.6 \text{ Gb/year}}$$

One channel (SEED format, account for header overhead and compression)

$$= (\mathbf{12.6 \text{ Gb/year}} / 3.0) / 87.5\% = \mathbf{4.8 \text{ Gb/year}}$$

One station (6 channels)

$$= \mathbf{4.8 \text{ Gb/year}} * 6 \text{ channels} = \mathbf{28.8 \text{ Gb/year}}$$

1,000 stations

$$= \mathbf{28.8 \text{ Gb/year}} * 1,000 = \mathbf{28.8 \text{ Tb/year}}$$

In addition data streams digitized at lower sample rates or non-seismic channels may also be archived. Inclusion of such channels would increase the volume of the archive no more than 15% to a total of about 33 Tb/year.

There is a trend toward higher sample rates of 200sps or even 500sps in engineering applications. If continuous streams at these higher rates are archived, the total volume of the archive will increase. However, this type of data is more likely to be saved in as triggered segments.