

## Data management plan

Pop Up Archive collects audio files and associated metadata (text and images). All files and records that a user chooses to make public are available for public access at [popuparchive.org](http://popuparchive.org) immediately after they finish processing (within seconds or minutes).

Pop Up Archive is capable of processing audio files from any public location on the web and offers public and private preservation options. The application is designed with availability and preservation in mind. It is hosted on Amazon EC2 (<http://aws.amazon.com/ec2/>) via Heroku, a cloud application platform. Heroku uses web dynos to manage HTTP requests and worker dynos to run background jobs (transcoding audio, for instance). If a Heroku dyno fails, a new one will replace it automatically, and if the site experiences heavy load more dynos can be launched to easily scale the service. New Relic (<http://newrelic.com/>) and Logentries (<https://logentries.com/>) are used for monitoring and logging the application. These services provide alerts in the event of an error. However, Heroku routinely handles new instance creation in the event of a crash or exceeded memory limits.

Data is stored in Postgres database (<https://addons.heroku.com/pgbackups>). Heroku automatically generates a backup of the database daily; each backup is retained for one month. Pop Up Archive uses Elasticsearch to provide a search engine for the archive. All of the data from this hosted service is duplicative of what is in the Postgres database, and can be duplicated easily in the case of data loss by reindexing the database.

Audio files are stored at The Internet Archive or in the Pop Up Archive Amazon S3 bucket. As a free, scalable solution for preserving public material, the Internet Archive (<https://archive.org/>) has partnered with Pop Up Archive as a trusted backup digital repository. The mission of the Internet Archive is to offer "permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format." Pop Up Archive bundles each audio file with an XML file containing standardized metadata that is then stored as a record at the Internet Archive. Audio uploaded to the Internet Archive is transcoded into lower fidelity formats (.wav, .mp3, .aif, .ogg) to achieve compatibility with more systems and, thus, wider distribution. Users choose whether or not they would like to back up their material at the Internet Archive; however, in order to prevent accidental deletion, users must contact Pop Up Archive or the Internet Archive directly to remove individual files or collections.

Digital files are stored using the Internet Archive's hardware, which consists of PCs with clusters of IDE hard drives. Data is stored on DLT tape and hard drives in various appropriate formats, depending on the collection. Web data is received and stored in archive format of 100-megabyte ARC files made up of many individual files. Alexa Internet (currently the source of all crawls in Internet Archive collections) is proposing ARC as a standard for archiving Internet objects. The Internet Archive is also mindful of the obsolescence of digital formats. They note, "As advances are made in software applications, many data formats become obsolete. We will be collecting software and emulators that will aid future researchers, historians, and scholars in their research." (<http://archive.org/about/about.php>)

Pop Up Archive has also evaluated the LOCKKS model for preservation, and is currently acting under their advice that the Internet Archive provides adequate backup storage. However, our

team is open to reconsidering LOCKKS as a storage mechanism for the software artifacts created over the course of this project.

Pop Up Archive stores private material on Amazon S3 (<http://aws.amazon.com/s3/>), which is itself a distributed and highly resilient file store. According to the Amazon S3 documentation, the service:

*Is designed to provide 99.999999999% durability of objects over a given year. This durability level corresponds to an average annual expected loss of 0.000000001% of objects. For example, if you store 10,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000,000 years. In addition, Amazon S3 is designed to sustain the concurrent loss of data in two facilities.*

The service also uses checksums to routinely detect data corruption.

We are working with the Knight Foundation and PRX to safeguard and ensure the continued existence of the software we engineer and the audio we archive. We track our open-source software development progress on Github: <https://github.com/PRX/pop-up-archive>. GitHub uses the git revision control system to track iterations of the Pop Up Archive software and provide a backup of the codebase. It also provides a platform for collaboration and sharing.

The Pop Up Archive system outputs PBCore-compliant XML records and JSON via an Application Programming Interface (API) that can be used to distribute data to other institutional repositories such as the Library of Congress (LOC) and the Digital Public Library of America, and we are working to enable one-click data exchange with these organizations. Enabling data exchange among individual producer and oral history archive websites as well as through the repositories of these significant cultural institutions will expand the number of probable users of archival audio artifacts to the tens of thousands.