BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25811
doi: 10.3897/biss.2.25811

OPEN ACCESS

Conference Abstract

COSA: Cloud Object Storage Archive for deep archival of digital data

Jeff Gerbracht ‡

‡ Cornell Lab of Ornithology, Ithaca, NY, United States of America

Corresponding author: Jeff Gerbracht (jag73@cornell.edu)

Received: 13 Apr 2018 | Published: 22 May 2018

Citation: Gerbracht J (2018) COSA: Cloud Object Storage Archive for deep archival of digital data. Biodiversity Information Science and Standards 2: e25811. https://doi.org/10.3897/biss.2.25811

Abstract

The Cornell Lab of Ornithology gathers, utilizes and archives a wide variety of digital assets ranging from details of a bird observation to photos, video and sound recordings. Some of these datasets are fairly small, while others are hundreds of terabytes. In this presentation we will describe how the Lab archives these datasets to ensure the data are both loss-less and recoverable in the case of a widespread disaster, how the archival strategy has evolved over the years and explore in detail the current hybrid cloud storage management system.

The Lab runs eBird and several other citizen science programs focused on birds where individuals from around the globe enter their sightings into a centralized database. The eBird project alone stores over 500,000,000 observations and the underlying database is over a terabyte in size. Birds of North America, Neotropical Birds and All About Birds are online species accounts comprising a wide range of authoritative live history articles maintained in a relatively small database. Macaulay Library is the world's largest image, sound and video archive with over 6,000,000 cuts totaling nearly 100 TB of data. The Bioacoustics Research Program utilizes automated recording units (SWIFTs) in the forests of the US, jungles of Africa and in all seven oceans to record the environment. These units record 24 hours a day and gather a tremendous about of raw data, over 200 TB to date with an expected rate of an additional 100TB per year. Lastly, BirdCams run by the lab add a steady stream of media detailing the reproductive cycles of a number of species. The lab is committed to making these archives of the natural world available for research and

conservation today. More importantly, ensuring these data exist and are accessible in 100 years is a critical component of the Lab data strategy.

The data management system for these digital assets has been completely overhauled to handle the rapidly increasing volume and to utilize on-premises systems and cloud services in a hybrid cloud storage system to ensure data are archived in a manner that is redundant, loss-less and insulated from disasters yet still accessible for research. With multimedia being the largest and most rapidly growing block of data, cost rapidly becomes a constraining factor of archiving these data in redundant, geographically isolated facilities. Datasets with a smaller footprint, eBIrd and species accounts allow for a wider variety of solutions as cost is less of a factor. Using different methods to take advantage of differing technologies and balancing cost vs recovery speed, the Lab has implemented several strategies based on data stability (eBird data are constantly changing), retrieval frequency required for research and overall size of the dataset. We utilize Amazon S3 and Glacier as our media archive, we tag each media in Glacier with a set of basic DarwinCore metatdata fields that key back to a master metadata database and numerous project specific databases. Because these metadata databases are much smaller in size, yet critical in searching and retrieval of a required media file, they are archived differently with up to the minute replication to prevent any data loss due to an unexpected disaster. The media files are tagged with a standard set of basic metadata and in the case where the metadata databases were unavailable, retrieval of specific media and basic metadata can still occur.

This system has allowed the lab to place into long term archive hundreds of terabytes of data, store them in redundant, geographically isolated locations and provide for complete disaster recovery of the data and metadata.

Keywords

cloud object store deep archive cornell multimedia

Presenting author

Jeff Gerbracht