OPEN ACCESS

Best Practices for using Cloud Services for Digital Data Archive and Disaster Recovery

Jeff Gerbracht[‡], Steve Kelling[‡]

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

Abstract

Managing digital data for long-term archival and disaster recovery is a key component of our collective responsibility in managing digital data and metadata. As more and more data are collected digitally and as the metadata for traditional museum collections becomes both digitized and more comprehensive, the need to ensure that these data are safe and accessible in the long term becomes essential. Unfortunately, disasters do occur and many irreplaceable datasets on biodiversity have been permanently lost. Maintaining a long-term archive and putting in place reliable disaster recovery processes can be prohibitively expensive, both in the cost of hardware and software as well as the costs of personnel to manage and maintain an archival system. Traditionally, storing digital data for the long term and ensuring the data are loss-less, safe and completely recoverable when a disaster occurs has been managed on-premises with a combination of on-site and off-site storage. This requires complex data workflows to ensure that all data are securely and redundantly stored in multiple highly dispersed locations to minimize the threat of data loss due to local or regional disasters. Files are often moved multiple times across operating systems and media types on their way to and from a deep archive, increasing the risk of file integrity issues. With the recent advent of an array of Cloud Services from organizations such as Amazon, Microsoft and Google to more focused offerings from Iron Mountain, Atempo and others, we have a number of options for long term archival of digital data. Deep archive solutions, storage where retrieval expected only in the case of a disaster, are

offered by many of these organizations at a rate substantially less than their normal data storage fees.

The most basic requirement for an archival system is storing multiple replicates of the data in geographically isolated locations with a mechanism for guaranteeing file integrity, usually using a checksum algorithm. Additional components that are integral to a robust archive include a simple metadata search and reliable retrieval.

In this presentation, we'll discuss the need for long term archive and disaster recovery capabilities, detail the current best practices of data archival systems and review a variety of archival options that have become available with Cloud Services.

Keywords

deep archive disaster recovery cloud-services digital data

Presenting author

Jeff Gerbracht