Preserving the Last Copy

Andrés Rodriguez
CTO and Founder
Archivas, Inc
andres@archivas.com

Abstract

The problem of preserving digital records for long-term access requires careful consideration about process and technology. Preserving digital information is more difficult than preserving records on materials such as paper or film. The sheer volume and the volatility introduced by digital demand a new software architecture capable of scaling and of preventing accidental changes to the records. Procedures need to be put in place to identify, classify, move, evolve, access and occasionally dispose of digital records. Library science and traditional archival practice provide an extensive body of knowledge that can be leveraged with technology to create a true modern archive.

1. Introduction

Archives are repositories for organizational records that are no longer in use but that may need to be accessed in the future. These are not working documents that can be modified. They are fixed content, files that have become records and that should not be modified. The primary goal of a digital archive is to preserve these records from change. But since saving information and later being unable to access it renders the archive meaningless, an archive also needs to provide the necessary means to find and retrieve the records it is preserving.

2. Lessons from Traditional Archives

Traditional archives have, over time, developed a number of general theories about the characteristics of archival records and best practices for managing them. An archivist's governing principles are provenance and original order. Provenance establishes that records generated by a person or group be kept together so that their context will be preserved. Original order requires that the sequence in which these records were found be maintained as well.

2.1. Provenance

Records in archives are arranged according to provenance, a principle that states records of different people or groups should never be mixed. Along with preserving the context of a file, provenance serves two other functions. It maintains the chain of custody. The chain of custody names the previous curators for a particular record. When a record's chain of custody is unclear, that record's value as evidence is lessened considerably. Retaining chain of custody is also an important intellectual property consideration. Without contextual information, much of the record often cannot be understood, or might be misunderstood, particularly after a long time has passed.

2.2. Original Order

The second principle of traditional archival organization is original order. This principle dictates that records should be kept arranged in the order in which they were found. Unlike the rule of provenance, the rule of original order can sometimes be broken - for example, if records were not kept in good order, or were kept in no discernible order. In addition to providing contextual information, original order is an existing system of organization, and saves spending time and resources to create a new system that would probably be less helpful. Original order is a universal organizational principle; hence it is the most likely arrangement to be understood in the future.

3. The Modern Archive

Digital records present extraordinary opportunities and challenges. Digital allows for perfect copies of the records to be made and, because digital records require no physical space, allows enormous amounts of information to be preserved. The records can also be indexed in

various ways simultaneously to ensure instant retrieval of files. But the challenges are very real. The ones and zeroes representing digital records are inherently unstable. There is no direct access to a digital record; computers must be used to convert the raw data into information. Data formats evolve over time, encumbering our ability to decode old records. These challenges can be overcome. Digital signatures can prevent records from being changed. Clusters of networked computers can replicate and distribute the records to ensure that there will always be enough copies and computers to access every record in the repository. Software running inside the cluster can evolve data from legacy formats to new standards while preserving the chain of custody.

3.1. Ingestion

Ingestion offers an archivist the opportunity to control how records are organized, and to impose provenance and original order. Ingestion is divided into appraisal and accession.

3.1.1. Appraisal

During appraisal the modern archivist determines what information sources in the organization should be publishing to the archive. The archivist is aided by applications responsible for extracting data from production systems. Original sources of records include email servers, file systems and data bases. The archive connects to the applications and ingests the stream of information that comes from that source. Each record has metadata and on entering the archive generally acquires more life-cycle metadata, such as when the record was archived. In ingestion, records are automatically checked to ensure that they are not corrupt or infected. Instead of examining individual records, the modern archivist manages information streams. Metadata added at this stage will become crucial to enabling later access to the records.

3.1.2. Accession

Accession is the actual movement of the records into the archive. Traditional archivists often do not accession all the records that are given them; archivists may determine that some of the records are not of permanent value or do not fit the archive's collection policy. Modern archivists have different challenges and more choices. In a digital archive the incremental cost of holding another record is negligible; digital records take no space. It is important that modern archivists do not save redundant or inaccurate information; however, from a cost perspective it is almost always more expensive to determine what to store than simply to store everything. During accession the application publishing to the archive needs to encapsulate the record into an object that contains all the information that will be necessary to interpret the record.

3.2. Disposition

The last step is often disposition, or the disposal of records. Keeping digital records online allows the archival software to retain control over disposition. At the end of their life cycle a compliance officer can decide if the records need to be kept onsite, or if can they be stored at a remote facility or even destroyed.

3.3. Preservation

The primary responsibility of the preservation layer of the modern archive is to keep the digital record intact. Because digital information is intangible, this is an enormous concern. Each record must be periodically refreshed, and when the hardware that created a record becomes obsolete, the record must be automatically moved to a new device. As data formats become outdated, records must be evolved to support the new standards. Keeping the records safe is an important aspect of preservation - organizational archives must be secure from either malicious or accidental intrusion.

3.4. Access

A modern archive is useful because it enables users and programs to access digital records. A modern archive should be thought of as a component in systems that enforce compliance, improve knowledge transfer and enable better corporate governance. One of the least intuitive aspects of access is that the needs of the users cannot be anticipated when the archive is created. Users will want to query the records in an archive in ways that were not anticipated by the curators of the archive; hence, the importance of providing access as a separate tier from preservation. Like a traditional archive, the preservation layer should make no assumptions as to how the records will be accessed, and the access layer should accommodate multiple access methods and be able to adapt to the needs of its users.

4. Conclusions

No single vendor can deliver the full modern archive. The problem involves many vendors and varies significantly depending on the organization. But the technology is developing quickly, and soon the modern archive will have become a familiar extension of the workplace, a tool and a resource that achieves compliance but then goes far beyond. Archivas provides an open software platform that delivers core ingestion,

preservation and access services to 3rd party applications so that organizations can build a true modern archive.