Using Rosetta, StorageGRID, and New IBM Tape Solutions to Implement State-of-the-Art Digital Preservation

ELUNA Conference 2012 (May 2012)

Gary T. Wright

Introduction to the Church of Jesus Christ of Latter-day Saints

The Church of Jesus Christ of Latter-day Saints is a worldwide Christian church with more than 14.4 million members and 28,784 congregations. With headquarters in Salt Lake City, Utah (USA), the Church operates three universities, a business college, 136 temples, and thousands of seminaries and institutes of religion around the world that enroll more than 700,000 students in religious training.

Salt Lake Temple
photographed by Henok Montoya

The Church has a scriptural mandate to keep records of its proceedings and preserve them for future generations. Accordingly, the Church has been creating and keeping records since 1830, when it was organized. A Church Historian's Office was formed in the 1840s, and in 1972 it was renamed the Church History Department.

The Church History Department

Today, the Church History Department has ultimate responsibility for preserving records of enduring value that originate from its ecclesiastical leaders and within the various Church departments, the Church's educational institutions, and its affiliations.

Church History Library
at Temple Square in Salt Lake City

With such a broad range of record sources, the array of digital record types requiring preservation is also extensive. However, as explained below, the vast majority of storage capacity in the Church History Department's digital preservation archive is allocated to audiovisual records.

Church Audiovisual Capabilities

Over the last two decades, the Church has developed state-of-the-art digital audiovisual capabilities to support its vast, worldwide communications needs. To illustrate one such need, twice a year the Church holds General Conference in its remarkable Conference Center, which seats 21,000.

Conference Center
at Temple Square in Salt Lake City

Church members from all over the world travel to attend these conferences in order to hear timely, relevant counsel from inspired ecclesiastical leaders.

Thomas S. Monson
President

The Church of
Jesus Christ
of Latter-day Saints

However, since the majority of the 14.4 million members are unable to attend in person, the meetings are broadcast in high definition video via satellite to more than 7,400 Church buildings in 102 countries. The broadcasts are simultaneously translated into 32 languages. In addition, the meetings are streamed live on the Church's website (lds.org) and on the Mormon Channel (radio.lds.org).

Ultimately, surround sound digital audio tracks for 96 languages are created (comprising 1.75 TB of raw audio content) to augment the digital video taping of each meeting, making the Church of Jesus Christ of Latter-day Saints the world's largest language broadcaster.

Because of their exalting and enduring value, all General Conference meeting videos, along with associated audio tracks, are preserved digitally.

The Church's advanced audiovisual capabilities are also used to support weekly broadcasts of *Music and the Spoken Word*—the world's longest continuous network broadcast (now in its 83rd year).

Each broadcast features an inspirational message and music performed by the Mormon Tabernacle Choir and the Orchestra at Temple Square. The broadcast is aired live by certain radio and television stations and is distributed to approximately 2000 other stations for delayed broadcast.

World-famous Mormon Tabernacle Choir
and Orchestra at Temple Square

The Mormon Tabernacle Choir was named "America's Choir" by President Ronald Reagan and was awarded the National Medal of Arts (the United States' highest honor for artistic excellence) by President George W. Bush. And in 2010, the *Music and the Spoken Word* broadcast was inducted into the National Radio Hall of Fame. It is no wonder that The Church of Jesus Christ of Latter-day Saints considers broadcasts of *Music and the Spoken Word* worthy of digital preservation.

As a gift to the world, the Church of Jesus Christ of Latter-day Saints launched a new website (biblevideos.lds.org) on December 4, 2011 that provides ***free*** Bible videos of the birth, life, death, and resurrection of the Lord Jesus Christ.

Nativity scene from biblevideos.lds.org

Viewable with a free mobile app, these videos are faithful to the biblical account. The text also accompanies each video scene. This remarkable gift provides a new and meaningful way to learn about Jesus Christ. Of course, all these digital videos will be preserved by the Church.

The Church's Publishing Services Department, which created the Bible videos, generates multiple petabytes of production audiovisual data annually. Typically, about 40% is targeted for preservation. In addition, a multi-petabyte backlog is currently being ingested into the digital preservation system.

In just ten years, Publishing Services anticipates that it will have generated a cumulative archival capacity of more than 100 petabytes for a single copy.

This means that the Church History Department's digital preservation archive for audiovisual data will become one of the largest in the world within a decade. And all the other Church records of enduring value will add to the total capacity of this world-class digital archive.

Architecting the Church History Department's Preservation System

Late in 2008, the Church History preservation team discovered that the National Library of New Zealand had developed a thorough set of business requirements for digital preservation. Steve Knight, Program Director for Preservation Research and Consultancy at the Library, was kind enough to share these requirements with the Church—which are probably the most comprehensive available anywhere, even today. The requirements provided an excellent basis from which the preservation team developed business requirements for the Church.

One such requirement was minimizing the total cost of ownership of archival storage. An internal study was performed to compare the costs of acquisition, maintenance, administration, data center floor space, and power to archive hundreds of petabytes of digital records using disk arrays, optical disks, virtual tape libraries, and automated tape cartridges. The model also incorporated assumptions about increasing storage densities of these different storage technologies over time.

Calculating all costs over a ten year period, the study concluded that the total cost of ownership of automated tape cartridges would be 33.7% of the next closest storage technology (which was disk arrays).

Consequently, the Church History Department today uses IBM 3500 Tape Libraries with LTO-5 and TS1140 tape drives.

Another requirement was scalability. Clearly, a multi-petabyte archive requires a system architecture that enables rapid scaling of automated ingest, archive storage capacity, access, and periodic validation of archive data integrity.

After several discussions with qualified, relevant people, concerns over the ability of open source repositories to adequately scale eliminated these potential solutions from consideration.

Ex Libris Rosetta was evaluated next. In order to determine if it would be able to scale to meet Church needs, a scalability proof of concept test was conducted.

The Rosetta evaluation involved joint scalability testing between Ex Libris and the Church History Department. Results of this testing have been published on the Ex Libris website (exlibrisgroup.com). The white paper is titled "The Ability to Preserve a Large Volume of Digital Assets—A Scaling Proof of Concept."

Results of the scalability test indicated that Rosetta would be able to meet Church History Department needs.

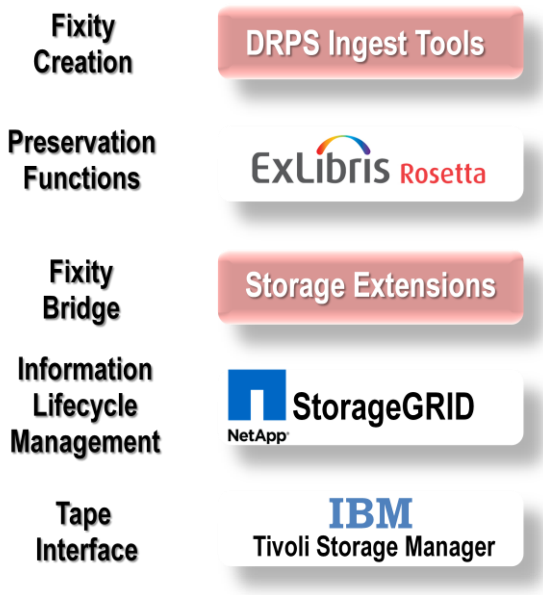
Next, the preservation team implemented CHIPS (Church History Interim Preservation System) using Rosetta for a more comprehensive test. When the CHIPS proof of concept test was completed with successful results, the Church History Department decided to move forward with Rosetta as the foundation for its Digital Records Preservation System (DRPS).

Rosetta provides configurable preservation workflows and advanced preservation planning functions, but only writes a single copy of an Archival Information Package[1] (AIP—the basic archival unit) to a storage device for permanent storage. An appropriate storage layer must be integrated with Rosetta in order to provide the full capabilities of a digital preservation archive, including AIP replication.

After investigating a host of potential storage layer solutions, the Church History Department chose NetApp StorageGRID to provide the Information Lifecycle Management (ILM) capabilities that were desired. In particular, StorageGRID's data integrity, data resilience, and data replication capabilities were attractive.

In order to support ILM migration of AIPs from disk to tape, StorageGRID utilizes IBM Tivoli Storage Manager (TSM) as an interface to tape libraries.

DRPS also employs software extensions developed by preservation team members from Church Information and Communications Services (shown in the red boxes below). These software extensions will be discussed later.



Architecture of the Church History Department's Digital Records Preservation System (DRPS)

DRPS is a dark archive—meaning that delivery of requested records is only provided to the department or institution that produced the records. Likewise, records in the archive may only be discovered by the organization that produced the records. Furthermore, authorized access requests are serviced by DRPS staff; thus producers of records do not have direct access to the DRPS archive. This arrangement enhances security of the archive.

Data Corruption in a Digital Preservation Archive

A critical requirement of a digital preservation system is the ability to continuously ensure data integrity of its archive. This requirement differentiates a tape archive from other tape farms.

Modern IT equipment—including servers, storage, network switches and routers—incorporate advanced features to minimize data corruption. Nevertheless, undetected errors still occur for a variety of reasons.

Whenever data files are written, read, stored, transmitted over a network, or processed, there is a small but real possibility that corruption will occur. Causes range from hardware and software failures to network transmission failures and interruptions. Bit flips (also called bit rot) within data stored on tape also cause data corruption.

Recently, data integrity of the entire DRPS tape archive was validated. This validation run encountered a 3.3×10^{-14} bit error rate.

Likewise, the USC Shoah Foundation Institute for Visual History and Education has observed a 2.3×10^{-14} bit error rate within its tape archive, which required the preservation team to flip back 1500 bits per 8 petabytes of archive capacity.[2]

These real life measurements—one taken from a large archive and the other from a relatively small archive—provide a credible estimation of the amount of data corruption that will occur in a digital preservation tape archive. Therefore, working solutions must be implemented to detect and correct these bit errors.

DRPS Solutions to Data Corruption

In order to continuously ensure data integrity of its tape archive, DRPS employs fixity information.

Fixity information is a checksum (i.e., an integrity value) calculated by a secure hash algorithm to ensure data integrity of an AIP file throughout preservation workflows and after the file has been written to the archive.

By comparing fixity values before and after files are written, transferred across a network, moved or copied, DRPS can determine if data corruption has taken place during the workflow or while the AIP is stored in the archive. DRPS uses a variety of hash values, cyclic redundancy check values, and error-correcting codes for such fixity information.

In order to implement fixity information as early as possible in the preservation process, and thus minimize data errors, DRPS provides ingest tools developed by Church Information and Communications Services (ICS) that create SHA-1 fixity information for producer files *before* they are transferred to DRPS for ingest (see the DRPS architecture shown previously).

Within Rosetta, SHA-1 fixity checks are performed three times—(*i*) when the deposit server receives a Submission Information Package[1] (SIP), (*ii*) during the SIP validation process, and (*iii*) when an AIP file is moved to permanent storage.

Rosetta also provides the capability to perform fixity checks on files after they have been written to permanent storage, but the ILM features of StorageGRID do not utilize this capability. Therefore, StorageGRID must take over control of the fixity information once files have been ingested into the grid.

By collaborating with Ex Libris on this process, ICS and Ex Libris have been successful in making the fixity information hand off from Rosetta to StorageGRID.

This is accomplished with a web service developed by ICS that retrieves SHA-1 hash values generated independently by StorageGRID when the files are written to the StorageGRID gateway node. Ex Libris developed a Rosetta plug-in that calls this web service and compares the StorageGRID SHA-1 hash values with those in the Rosetta database, which are known to be correct.

Turning now to the storage layer of DRPS, StorageGRID is constructed around the concept of object storage. To ensure object data integrity, StorageGRID provides a layered and overlapping set of protection domains that guard against data corruption and alteration of files that are written to the grid.

The highest level domain utilizes the SHA-1 fixity information discussed above. A SHA-1 hash value is generated for each AIP (or object) that Rosetta writes to permanent storage (i.e., to StorageGRID). Also called the Object Hash, the SHA-1 hash value is self-contained and requires no external information for verification.

Each object contains a SHA-1 object hash of the StorageGRID formatted data that comprise the object. The object hash is generated when the object is created (i.e., when the gateway node writes it to the first storage node).

To assure data integrity, the object hash is verified every time the object is stored and accessed. Furthermore, a background verification process uses the SHA-1 object hash to verify that the object, while stored on disk, has neither become corrupt nor has been altered by tampering.

Underneath the SHA-1 object hash domain, StorageGRID also generates a Content Hash when the object is created. Since objects consist of AIP data plus StorageGRID metadata, the content hash provides additional protection for AIP files.

Because the content hash is not self-contained, it requires external information for verification, and therefore is checked only when the object is accessed.

Each StorageGRID object has a third and fourth domain of data protection applied, and two different types of protection are utilized.

First, a cyclic redundancy check (CRC) checksum is added that can be quickly computed to verify that the object has not been corrupted or accidentally altered. This CRC enables a verification process that minimizes resource use, but is not secure against deliberate alteration.

Second, a key-based hash value is appended. This value can be verified using the key that is stored as part of the metadata managed by StorageGRID.

Although this hash value takes more resources to implement than the CRC checksum described above, it is secure against all forms of tampering as long as the key is protected.

The CRC checksum is verified during every StorageGRID object operation—i.e., store, retrieve, transmit, receive, access, and background verification. But, as with the content hash, the key-based hash value is only verified when the object is accessed.

Once a file has been correctly written to a StorageGRID storage node (i.e., its data integrity has been ensured through both SHA-1 object hash and CRC fixity checks), StorageGRID invokes the TSM Client running on the archive node server in order to write the file to tape.

As this happens, the SHA-1 (object hash) fixity information is not handed off to TSM. Rather, it is superseded with new fixity information composed of various cyclic redundancy check values and error-correcting codes that provide *TSM end-to-end logical block protection* when writing the file to tape.

Thus the DRPS fixity information chain of control is altered when StorageGRID invokes TSM. Nevertheless, validation of the file's data integrity continues seamlessly until it is written to tape.

The process begins when the TSM client appends a CRC value to file data that is to be sent to the TSM server during a client session. As part of this session, the TSM server performs a CRC operation on the data and compares its value with the value calculated by the client.

Such CRC value checking continues until the file has been successfully sent over the network to the TSM server—with its data integrity validated.

Next, the TSM server calculates and appends a CRC value to each logical block of the file before transferring it to a tape drive for writing. Each appended CRC is called the "original data CRC" for that logical block.

When the tape drive receives a logical block, it computes its own CRC for the data and compares it to the original data CRC. If an error is detected, a check condition is generated, forcing a re-drive or a permanent error—effectively guaranteeing protection of the logical block during transfer.

In addition, as the logical block is loaded into the tape drive's main data buffer, two other processes occur—

(1) Data received at the buffer is cycled back through an on-the-fly verifier that once again validates the original data CRC. Any introduced error will again force a re-drive or a permanent error.

(2) In parallel, a Reed-Solomon error-correcting code (ECC) is computed and appended to the data. Referred to as the "C1 code," this ECC protects data integrity of the logical block as it goes through additional formatting steps—including the addition of an additional ECC, referred to as the "C2 code."

As part of these formatting steps, the C1 code is checked every time data is read from the data buffer. Thus, protection of the original data CRC is essentially transformed to protection from the more powerful C1 code.

Finally, the data is read from the main buffer and is written to tape using a read-while-write process. During this process, the just written data is read back from tape and loaded into the main data buffer so the C1 code can be checked once again to verify the written data.

A successful read-while-write operation assures that no data corruption has occurred from the time the file's logical block was transferred from the TSM client until it is written to tape. And using these ECCs and CRCs, the tape drive can validate logical blocks at full line speed as they are being written!

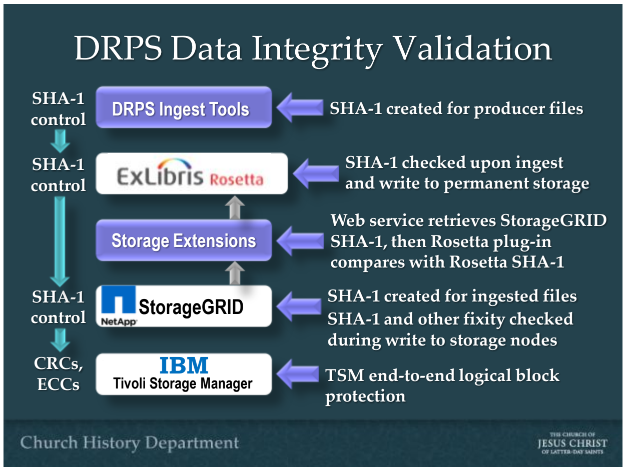
During a read operation (i.e., when Rosetta accesses an AIP), data is read from the tape and all three codes (C1, C2, and the original data CRC) are decoded and checked, and a read error is generated if any process indicates an error.

The original data CRC is then appended to the logical block when it is transferred to the TSM server so it can be independently verified by that server, thus completing the TSM end-to-end logical block protection cycle.

This advanced and highly efficient TSM end-to-end logical block protection is enabled with state-of-the-art functions available with IBM LTO-5 and TS1140 tape drives.

When the TSM server sends the data over the network to a TSM client, CRC checking is done once again to ensure integrity of the data as it is written to the StorageGRID storage node.

From there, StorageGRID fixity checking occurs, as explained previously for object access—including content hash and key-based hash value checking—until the data is transferred to Rosetta for delivery to its requestor, thus completing the DRPS data integrity validation cycle.



DRPS Data Integrity Validation

SHA-1 control — DRPS Ingest Tools — SHA-1 created for producer files

SHA-1 control — ExLibris Rosetta — SHA-1 checked upon ingest and write to permanent storage

Storage Extensions — Web service retrieves StorageGRID SHA-1, then Rosetta plug-in compares with Rosetta SHA-1

SHA-1 control — NetApp StorageGRID — SHA-1 created for ingested files SHA-1 and other fixity checked during write to storage nodes

CRCs, ECCs — IBM Tivoli Storage Manager — TSM end-to-end logical block protection

Church History Department

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS

Summary of the DRPS data integrity validation cycle

Ensuring Ongoing Data Integrity

Unfortunately, continuously ensuring data integrity of a DRPS AIP does not end once the AIP has been written correctly to tape. Periodically, the tape(s) containing the AIP needs to be checked to uncover errors (i.e., bit flips) that may have occurred since the AIP was correctly written.

Fortunately, IBM LTO-5 and TS1140 tape drives can perform this check without having to stage the AIP to disk, which is clearly a resource intensive task—especially for an archive with a capacity measured in hundreds of petabytes!

IBM LTO-5 and TS1140 drives can perform data integrity validation *in-drive*, which means a drive can read a tape and concurrently check the AIP logical block CRC and ECCs discussed above (C1, C2, and the original data CRC). Status is reported as soon as these internal checks are completed. And this is done without requiring any other resources!

Clearly, this advanced capability enhances the ability of DRPS to perform periodic data integrity validations of the entire archive more frequently, which will facilitate the correction of bit flips.

Other DRPS Solutions

A digital preservation best practice is to preserve records at the highest resolution affordable and practical.

Most of the still images produced by the Church are created as TIFF (Tagged Image File Format) files. While TIFF provides very high resolution, it also consumes considerable archive storage capacity.

To provide image resolution equivalent to TIFF while significantly reducing archive capacity, the Church History Department preserves still images in the lossless JPEG 2000 file format.

Tests conducted by Church software engineers have shown that the original bit stream of a TIFF image can be recreated from its lossless JPEG 2000 archival version. Yet reduced storage capacity benefits ranging from 50% to 60% or more are consistently realized with lossless JPEG 2000.

Since most DRPS archive content will come from audiovisual records, the Church History Department is focused on preserving very large audiovisual files (hundreds of gigabytes in size) produced by Publishing Services.

These audiovisual files are packaged with the MXF (Material Exchange Format) container format. As explained previously, a single General Conference session MXF SIP consists of the Conference video plus 96 audio tracks. Also, an American Sign Language video is included.

Ideally, the Church History Department would like DRPS to do file validation of the components within the MXF wrapper and extract technical metadata from them. Unfortunately, no tools are known to exist at the present time that will do all this.

The open source tool MediaInfo can extract technical metadata, but only from the MXF wrapper itself. Therefore, ICS engineers developed an MXF Extraction Tool that is a Rosetta plug-in which allows extraction of technical metadata from the MXF wrapper. This tool is currently being used to ingest MXF SIPs.

Rosetta does not presently support ingest of repeating tracks (such as multiple audio tracks). In order to ingest MXF SIPs such as General Conference sessions that have such repeating tracks, the MXF Extraction Tool described above concatenates data from each track in the metadata it extracts. This method of ingest is acceptable to the Church History Department until Ex Libris provides Rosetta ingest support for repeating tracks.

In the meantime, to help other institutions deal with this challenge, the Church History Department has made a modified version of the MXF Extraction Tool available to Ex Libris for distribution to Rosetta customers. The modified version limits extraction to one video and one audio track, which should meet the needs of such customers.

Dual, State-of-the-Art Digital Archive Facilities

Many physical Church records and artifacts of priceless value are currently stored in the Granite Mountain Records Vault. This unique facility features six tunnels that are bored into the side of a solid granite mountain (one of several prominent mountains that surround and protect the Salt Lake Valley). The tunnels have ambient conditions that are naturally conducive to records preservation.

Granite Mountain Records Vault

In order to efficiently preserve the burgeoning capacity of digital records the Church is generating, plans have been completed to renovate the Granite Mountain Records Vault in order to equip two of the vaults exclusively for digital preservation storage media.

These special vaults, which will become the Church's "deep" digital archive, will be physically isolated from the other four vaults in the facility. Such isolation will allow archival environmental conditions for magnetic tape to be maintained consistently for the automated tape cartridges that will be stored in the vaults. Isolation will also provide a high degree of physical security.

An "active" preservation facility, to be located in a different disaster zone, is also planned.

Active Preservation Facility design concept courtesy of Integrated Design Group

This archive will provide environmental, fire protection, and security facilities similar to the deep archive in the Granite Mountain Records Vault, and will allow more copies of the Church's rapidly growing collection of priceless digital records to be preserved.

The active archive will be the primary repository used to send copies of preserved records to authorized requestors. Both facilities will archive the same records, but the deep archive will be used for access only if the active archive is unable to service an authorized access request.

These redundant, state-of-the-art archive facilities will help the Church safely and securely preserve digital records of exalting and priceless value so they can be shared with the world—now and into the future.

Conclusion

The Church of Jesus Christ of Latter-day Saints is making a substantial investment to preserve records of its proceedings for use by future generations. The benefits of preserving these exalting and inspiring records cannot be measured in financial terms, however. Those benefits include building character and strengthening families—both of which are designed to foster both personal and family happiness.

References

[1] CCSDS 650.0-B-1BLUE BOOK, "Reference Model for an Open Archival Information System (OAIS)," Consultative Committee for Space Data Systems (2002)

[2] Private conversation with Sam Gustman (CTO) at the USC Shoah Foundation Institute August 19, 2009