Advancing Data Integrity in a Digital Preservation Archive Ex Libris and the Church of Jesus Christ of Latter-day Saints

Nir Sherwinter (nir.sherwinter@exlibrisgroup.com) and Gary T. Wright (wrightgt@ldschurch.org)

1. 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, 138 temples, and thousands of seminaries and institutes of religion around the world that enroll more than 700,000 students in religious training.

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 later it was renamed the Church History Department.

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

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

Over the last two decades, the Church has developed state-of-the-art digital audiovisual capabilities to support its vast, worldwide communications needs. One such need is broadcasting semiannual sessions of General Conference, which are broadcast in high definition video via satellite to more than 7,400 Church buildings in 102 countries and are simultaneously translated into 32 languages. Ultimately, surround sound digital audio tracks for more than 90 languages are created to augment the digital video taping of each meeting—making the Church the world's largest broadcaster of languages.

Another communications need is producing weekly broadcasts of *Music and the Spoken Word*—the world's longest continuous network broadcast (now in its 84th year). Each broadcast features an inspirational message and music performed by the Mormon Tabernacle Choir. The broadcast is aired live by certain radio and television stations and is distributed to approximately 2000 other stations for delayed broadcast.

The Church's Publishing Services Department, which supports all these broadcasts, generates multiple petabytes of production audiovisual data annually. 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.

2. INTRODUCTION TO THE EX LIBRIS GROUP

Ex Libris is a leading provider of library automation solutions, offering the only comprehensive product suite for the discovery, management, distribution, and preservation of digital materials. Dedicated to developing the market's most inventive and creative solutions, Ex Libris products serve the needs of academic, research, national, and other libraries, such as the Church History Library. With more than 460 employees worldwide, Ex Libris operates an extensive network of eleven wholly-owned subsidiaries and twelve distributors, many of which are exclusive. Ex Libris corporate headquarters are located in Jerusalem, Israel.

3. BUILDING THE CHURCH'S DIGITAL RECORDS PRESERVATION SYSTEM

In order to build and maintain a large digital archive, the Church History digital preservation team realized that it would be critical to minimize the total cost of ownership of archival storage.

An internal study was performed to compare the costs of acquisition, power, data center floor space, maintenance, and administration 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 uses IBM 3500 Tape Libraries with LTO-5 and TS1140 tape drives for its digital preservation archive today.

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 needs.

Next, the digital preservation team implemented the Church History Interim Preservation System (or CHIPS) using Rosetta for a more comprehensive test. CHIPS used only disk for storage. 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—see Figure 1).

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 preservation team 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 reddish boxes in Figure 1). These software extensions will be discussed later.

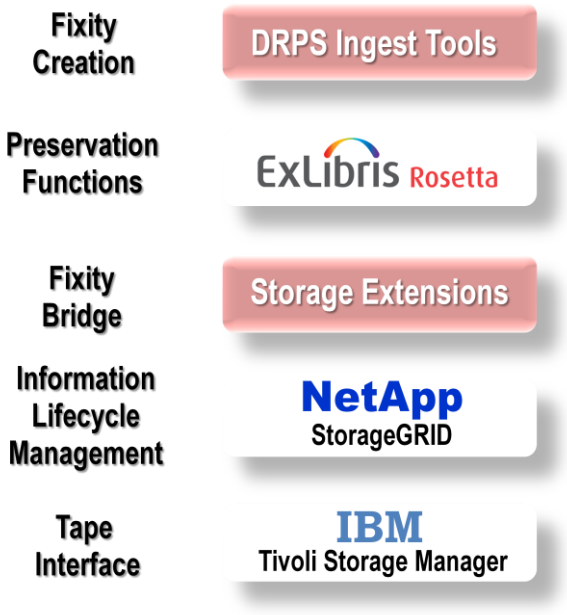

Figure 1
Components of the Church History Department's Digital Records Preservation System (DRPS)

4. DATA CORRUPTION IN A DIGITAL PRESERVATION TAPE 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 data errors.

5. 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 Figure 1).

Within Rosetta, SHA-1 fixity checks are performed three times—(*i*) when the deposit server receives a Submission Information Package (SIP) [1], (*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 corrupted 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 hash-based message authentication code (HMAC) message authentication digest is appended. This message digest can be verified using the HMAC key that is stored as part of the metadata managed by StorageGRID. Although the HMAC message digest takes more resources to implement than the CRC checksum described above, it is secure against all forms of tampering as long as the HMAC 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 HMAC message digest 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 the file 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 HMAC message digest checking—until the data is transferred to Rosetta for delivery to its requestor, thus completing the DRPS data integrity validation cycle.

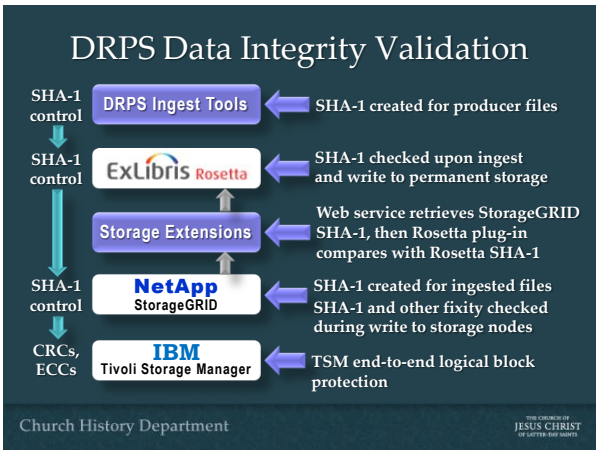

Figure 2
Summary of the DRPS data integrity validation cycle

6. 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 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 and other data errors.

7. LOOKING TO THE FUTURE

StorageGRID provides an HTTP API that automatically returns its SHA-1 hash values when called, but this API is not used at the present time because Rosetta currently only writes to permanent Network File System (NFS) storage using POSIX commands.

As a result of collaboration between Ex Libris and the Church History digital preservation team, the next version of Rosetta (3.1) will expose the Rosetta storage handler component as a Rosetta plugin. This will enable Rosetta to integrate with storage systems other than NFS, such as Amazon S3, storage systems which support CDMI (Cloud Data Management Interface), and others. The enhancement significantly expands Rosetta's reach into modern distributed file systems.

Ex Libris has committed to the Church a Rosetta plugin that will utilize the StorageGRID HTTP API and thus eliminate the need for the ICS-developed web service mentioned previously. This will provide a more elegant DRPS solution to fixity information hand off between Rosetta and StorageGRID.

As the size of the DRPS digital archive continues to grow, the need for increased Rosetta scalability is ever present. Fortunately for the Church, Ex Libris has been proactive in meeting its needs.

Subsequent to the original Rosetta scalability work mentioned earlier that was performed by the Church and Ex Libris together, significant improvements have been integrated to enhance Rosetta robustness and scalability.

For example, to fully leverage modern multicore processor technologies, a series of concurrent processing techniques have been implemented in Rosetta. Multi-threading is a programming and execution model that provides developers with a useful abstraction of concurrent execution. When applied to a single process, multi-threading permits parallel execution on a multiprocessor system, and also increases fault tolerance of the system.

Managing a concurrent flow has its challenges, however, since operations exclusivity and timing need to be continuously considered. To preclude errors, Java Messaging Service (JMS) was employed in Rosetta, allowing communications between different components of the distributed application to be loosely coupled, asynchronous, and reliable [3]. This enhancement provides robustness and fault tolerance, and guarantees that no work is lost.

Additional processing enhancements were implemented by using symbolic links for files during ingest and operational processes. These enhancements remove the need for copying files from one temporary location to another, thereby reducing I/O and improving network utilization as well as data integrity.

Large file ingest processing was also improved by incorporating DROID 6 file identification during the SIP validation stage. DROID 6 is substantially more efficient at identifying file formats of large files since it uses offsets to locate the file signature, and thus avoids a full scan of the entire file.

8. CONCLUSION

By working collaboratively with Ex Libris and utilizing advanced tape drives plus the sophisticated data integrity features of StorageGRID, the Church of Jesus Christ of Latter-day Saints has been able to advance the state of the art of data integrity and long term preservation in a rapidly growing digital preservation archive.

9. 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
[3] http://www.oracle.com/technetwork/java/jms/ index.html