

Solving Digital Preservation in the Business Data Centre

In the last five years, a radical shift has taken place in the data centre. The requirements of the business now dictate that digital information be retained long-term. But what is long-term? And what new standards are required to meet this radical shift?

BY PETER MOJICA, MICHAEL PETERSON, SIMONA RABINOVICI-COHEN, AND GARY ZASMAN

Based on the research¹ of the SNIA's Data Management Forum (DMF), long-term today is "Any period over 10-15 years is long-term." It is the period of time beyond which a data centre storage practices start losing information because their retention and preservation technologies are inadequate.

The DMF's research shows that regulatory compliance risk, legal risk (discovery), security risk (privacy), and digital asset preservation requirements have changed the game. The DMF's "100 Year Archive Requirements Survey" (Jan 2007) provides confirmation. According to respondents, the top five external factors driving current retention requirements are all driven by the "business."

Top five Retention Drivers

- 1 Protection and **preservation** of the organisation's history
- 2 Meeting regulatory requirements
- 3 Concern with litigation protection
- 4 Protection of business or intellectual property assets
- 5 Protection of customer privacy

MOVING FORWARD – NEW STANDARDS REQUIRED

The time dimension of the business problem is that very few organisations believe that they can actually meet these long-term retention and preservation requirements. The industry lacks a cohesive set of standards and best practices that will

Story Snapshot

- New data storage challenges require new strategies and initiatives worldwide
- New and emerging standards developed by SNIA are addressing the urgent issues

allow retention and preservation practices to start small and yet scale to the requirements of the data centre.

It is one thing to keep several hundred GBs around for 50 years, where you can archive or try to virtualise the storage environment itself, periodically migrating the information to new and improved media solutions in order to preserve it. It is quite another thing to deal with the needs of a typical data centre, especially a petabyte size information repository that is growing at 30% to 50% per year. Just the migration costs and time to migrate alone are prohibitive, not to mention the issues of preservation.

In addition, all the cost and work associated with retaining information for the long-term may be useless unless you can prove its authenticity and integrity, verify its provenance, audit its accesses, and understand its context.

For example, what's the benefit for your company in presenting information in court and not being able to 'prove' its integrity and authenticity? Actually, the inverse might be worse – a fine for presenting unverifiable or incomplete information. Successful preservation requires maintaining a comprehensive set of information attributes such as those outlined in Figure 1 on page 22.

About the Authors

This article was jointly written for *IQ* by the leadership of the SNIA Data Management Forum, Long Term Archive and Compliance Storage Initiative (LTACSI):



PETER MOJICA, Co-Chair, SNIA DMF LTACSI Reference Guide Committee, AXS-One.



MICHAEL PETERSON, Chief Strategy Advocate, SNIA-DMF.



SIMONA RABINOVICI-COHEN, Chair, SNIA DMF LTACSI Standards and Practices Committee, IBM.



GARY ZASMAN, Chair, SNIA DMF Long-Term Archive and Compliant Storage Initiative, NetApp.

A WORLDWIDE EFFORT

The challenge of digital information preservation is complex. New practices and standards are needed. Many organisations are working on addressing this problem around the world as Figure 2, on page 23, illustrates.

SPECIFIC SNIA INITIATIVES – XAM AND SIRF

The SNIA eXtensible Access Method (XAM) and Self-contained Information Retention Format (SIRF) standards activities are designed to benefit all of the respective members of the information and application ecosystem – the storage vendors, application developers, and the information using communities.

The XAM interface specification “defines a standard access method (API) between *Consumers and Providers* (such as applications and storage systems) giving each the intrinsic knowledge needed to effectively participate in the long-term access and preservation of digital content.

One of the benefits of XAM is *interoperability* where applications can support any XAM-conformant storage system, allowing greater end-user flexibility with migration processes, across the ecosystem, and ensuring long-term readability. With XAM, technology-obsolescence is avoided through the standard application-independent structure.

The newly initiated SIRF standardisation effort, proposes a logical container format appropriate for the long-term preservation of digital information. The result will be a Self-Describing Self-contained Information Retention Format (SIRF) that will work in conjunction with other interfaces to ensure that users, applications and storage platforms can share information saved in a logical container, encapsulating both the content and associated *preservation* metadata, across various applications and storage systems.

This data format will provide data portability and accessibility even when the originating application is no longer available, providing content independence from proprietary applications.

WHERE IS THE WORLD GOING?

There is no turning back. New and emerging standards such as **XAM** and the recently initiated **SIRF** are underway and suited specifically to deal with the urgent issues of application independence, content mobility and accessibility, which are just a few of the many issues affecting retention and preservation.

These standards will eventually ensure that each member participating in the management and retention of digital information can effectively communicate and interoperate across physical and logical storage layers, and across a span

continued on page 49

Figure 1: Digital Preservation Requirements

PRESERVATION ACTIONS AND REQUIREMENTS		THINGS TO CONSIDER...
Future Readability	Ability to interpret the data in the future even when technologies for computer hardware, operating systems, data management products and applications are replaced with newer ones – and even as the data consumers (designated user-communities) change frequently.	Generic “viewing” technology and formats such as TIF and PDF are not adequate preservation methods for long-term retention. Properly preserving information includes protecting and providing “content and its context,” “meta-data plus accrued meta-data,” and “provenance and fidelity.” All of these are needed for complete and accurate [authentic] preservation.
Integrity	Data, including meta-data and log files, must be “correct,” “complete”, and unchanged.	Integrity extends to newly created data such as indices, and <i>correctness</i> and <i>completeness</i> extends to search and discovery. Both retention and preservation are rendered ineffective if search results yield incomplete or partial result sets or different copies of the same version.
Authenticity	Original information must be both maintained and verifiably proven to be unaltered. This is a paramount requirement for preservation.	Consider write-once, read-many (WORM) media as well accepted hashing methods for proving that your stored file is the original. Other proofs for authenticity include the access logs, provenance, chain of custody, and security records.
Meta-data	Both original meta-data and “accruing meta-data” must be preserved as “authentic” and maintained with original data (not separate and apart).	Meta-data considerations and planning should include re-indexing operations which can be performance and time intensive.
Audit	Un-audited data puts businesses at risk. Complete records of access and “use” must be maintained as “authentic.”	Consider “ <i>transparency</i> ” of electronic audit requirements through regularly scheduled quality control checks.
Electronic and Process Chain of Custody	Helps to establish the authenticity of evidence, proving it has been preserved from point of capture through its life cycle with complete audit of both manual and electronic reviews, and transportation.	Consider using binary signatures meta-data to log <i>How</i> and <i>Who</i> handled data – <i>chain of custody</i> includes electronic system processes including automated logical migrations, file integrity checks, compression and other system processes which interact directly against original data.

Figure 2: Example World-Wide Efforts on Long-Term Digital Information Retention and Preservation

ORGANISATION	STANDARD/PROJECT	WEBSITE	KEY ROLE(S)
CASPAR – Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval	Integrated Project co-financed by the European Union within the Sixth Framework Programme	www.casparpreserves.eu	CASPAR will research, implement, and disseminate innovative solutions for digital preservation based on the OAIS reference model (ISO:14721:2002)
InterPARES	InterPARES 1, 2, and 3	www.interpares.org	Dealing with issues of authenticity, reliability, and accuracy during the entire lifecycle of records
ISO	OAIS, Open Archival Information Systems, ISO 14721-2003	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683	The purpose of this standard is to establish a reference model and a system for archiving information, both digitalized and physical, with an organizational scheme composed of people who accept the responsibility to preserve information and make it available to a designated community.
NARA – National Archives and Records Administration	Electronic Records Archive, ERA	www.archives.gov/era	The “ERA” project will establish the “Archives of the Future,” which will preserve, manage and provide sustained access to all types of e-records, independent of any specific type of software or hardware.
NDIIPP – National Digital Information Infrastructure and Preservation Program	The Library of Congress – Digital Preservation	http://www.digitalpreservation.gov	The Library of Congress is the nation’s oldest federal cultural institution and serves as the research arm of Congress. It is also the largest library in the world, with millions of books, recordings, photographs, maps and manuscripts in its collections.
National Archives of Australia	Australian Government Recordkeeping Metadata Standard	http://www.naa.gov.au/records-management/publications/AGRkMS.aspx	Definition of recordkeeping metadata defined as structured or semi-structured information that enables the creation, management, and use of records through time and across domains. Recordkeeping metadata can be used to identify, authenticate and contextualise records and the people, processes and systems that create, manage, maintain and use them.
SNIA – Storage Networking Industry Association	XAM – (eXtensible Access Method)	www.snia.org/xam	XAM is a new standard in development by SNIA. It specifies an application-to-storage interface that allows storage independence and allows applications to add metadata to the data containers for management purposes.
SNIA – Storage Networking Industry Association	SIRF, Self-contained Information Retention Format	www.snia.org/forums/dmf	A new standard in development by SNIA to address logical migration and the long-term preservation and retention of digital information.