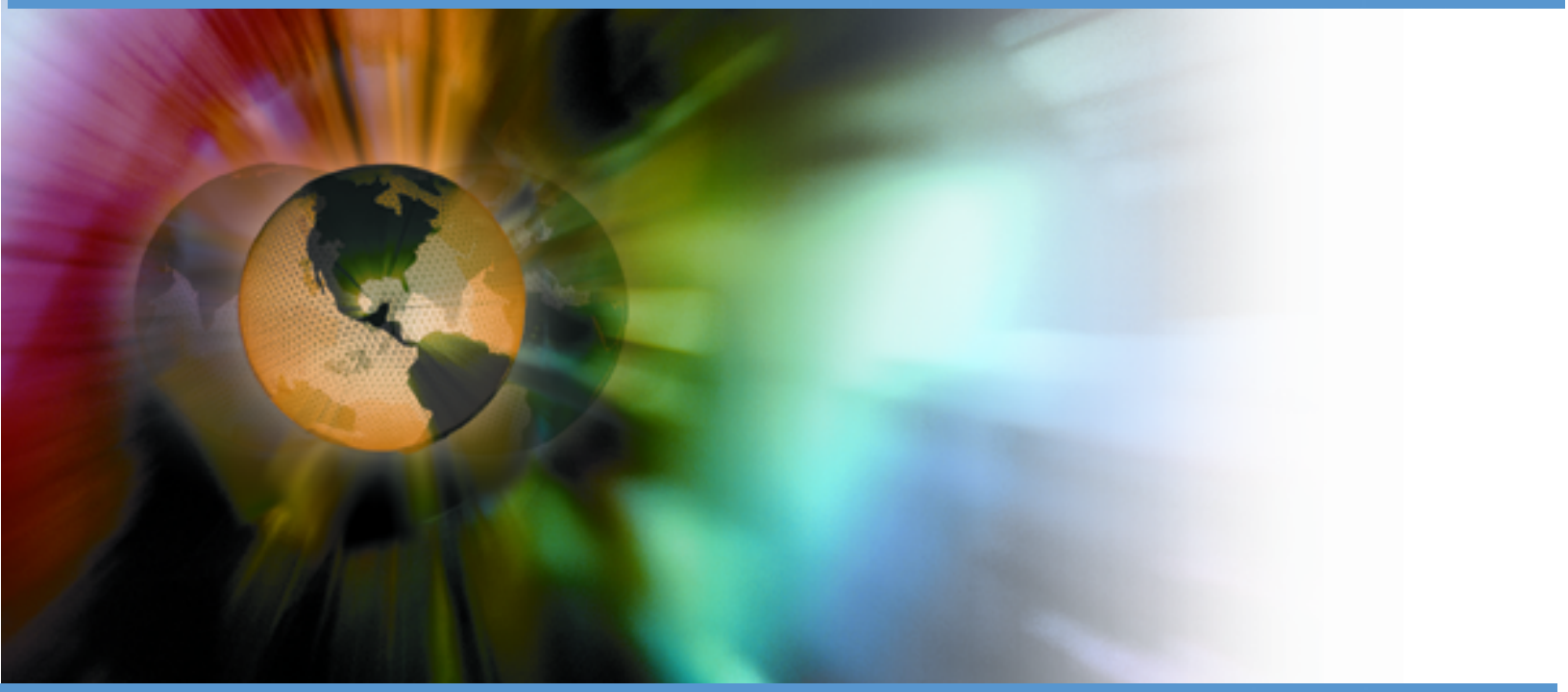
SNIA - Data Management Forum



The Coming Archive Crisis

November 16, 2006

SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

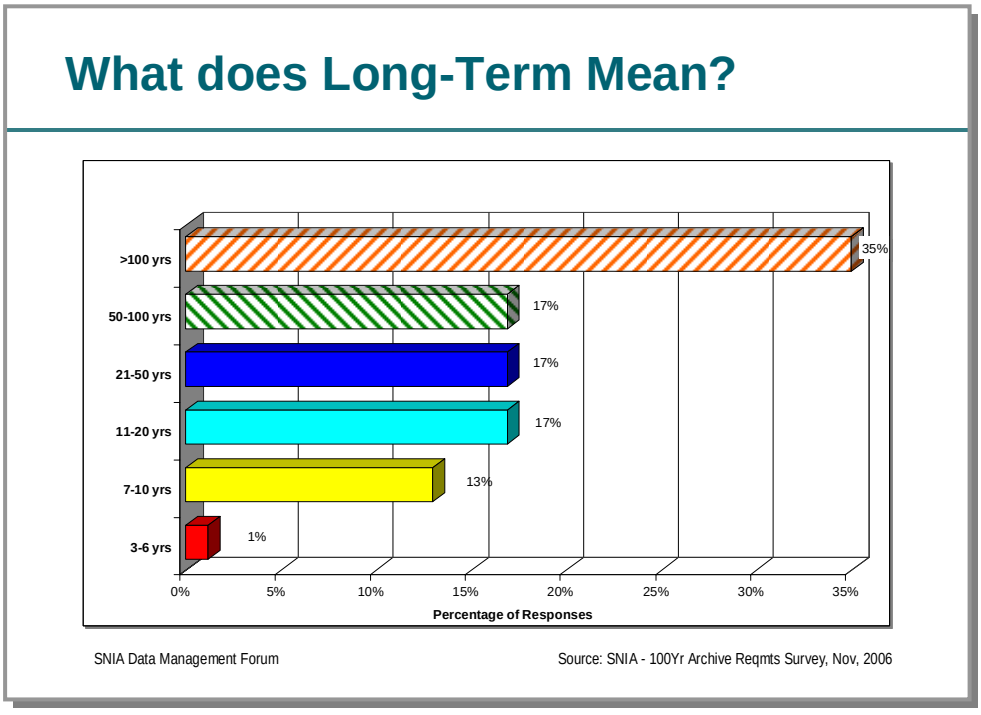
Author: Michael Peterson, President, Strategic Research Corp.
and Chief Strategy Advocate, SNIA Data Management Forum
www.sresearch.com and *www.snia-dmf.org*

The Coming Archive Crisis!
Requirements for Long-Term Digital Information Retention

Just like in the old TV game "Wordplay", all I have to do is say the word '*archive*' and I get dozens of different interpretations. Archivists, records and information managers (RIM), Information technologists (IT), legal, the business group, and the vendors all have different ideas of what 'archive' means. And, in the context of their practices, their viewpoints are valid. Here's the paradox. The archivist and records manager's practices are well-developed for preserving specific digital information collections, digital records, and their provenance. A number of International Standards Organization (ISO) standards and best practices exist on which they base their methods. In contrast, IT has at least four barriers that make this job particularly hard.

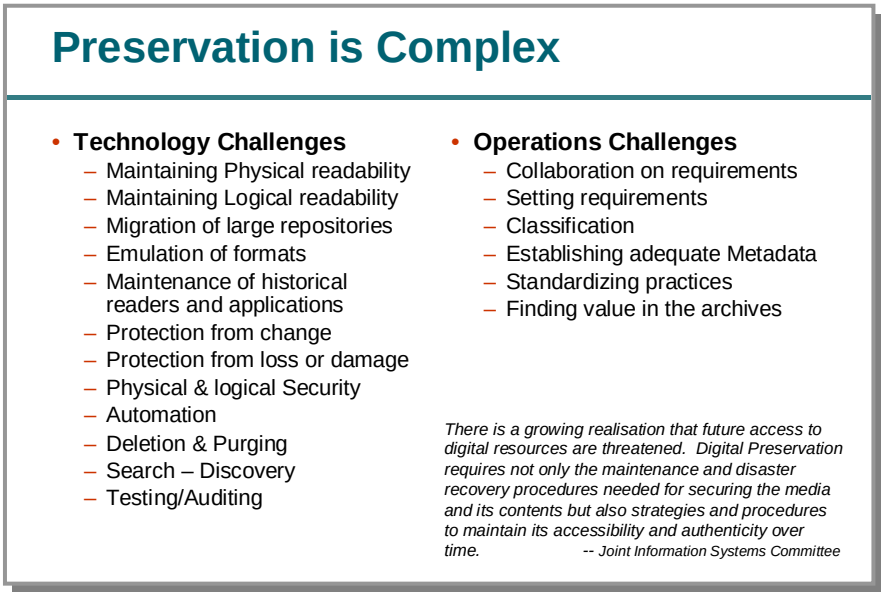
- Working in an organizational vacuum – RIM and IT don't communicate clearly. IT rarely knows the requirements even though there may be compliance policies in place. Add to this the obstacle that few IT shops see long-term "archiving" as important or even an honorable occupation and you have a better understanding of the dilemma.
- The notion of records does not necessarily exist to IT. Instead, IT is usually dealing with tens-to-hundreds of terabytes of data that it has to retain for varying periods of time.
- IT has little to no context about what these data objects are as part of a business record, where they came from (their provenance) and which object is the original or final version (as if that concept even makes sense in an IT practice) as many duplicates and many versions are spread all around to assure recovery of one of them.
- There are no storage practice standards or industry best practices for long term digital information retention in common use in the IT domain.

Let's further establish the problems that IT has to overcome. First, the notion of what "long-term" means varies widely across organizations based on their business, operating, legal, and



regulatory compliance needs. Most IT professionals assume that long-term means more than 7 years. It is not difficult to preserve information and be able to read it for 10 years, so why is this an issue? What about in 20 years, or 50 years, or more than a hundred years? Based on the SNIA-DMF's "100 Year Archive Task Force Requirements Survey"[1], most organizations have a long-term retention problem that exceeds 50 years and the respondents are far from confident that they can meet these requirements. Not only is there a disconnect between awareness and requirements, but as said before IT lacks the methods (and often the interest) for long term preservation.

The next point to understand about the nature of the long term preservation problem in an IT context is that it is very complex. The following chart illustrates many of the challenging factors such as maintaining physical and logical readability. Point one - long-term digital information retention is not a media problem. Even if storage media survived for fifty to a hundred years, the systems on which to read and interpret the information would also have to be archived along with spare components, software, and the knowledge on how to operate them. It is one thing to be able to physically read media and another to logically interpret it in the context of the application. Here is a simple, but pertinent example. Files written in early PDF[2] formats are not always readable by the current version PDF Reader. This is why in 2005 the ISO standard PDF/A format was adopted. Yet, even that standard warns us with these caveats[3]:



- *PDF/A alone does not guarantee preservation. PDF/A alone does not guarantee exact replication of source material*
- *The intent of PDF/A is not to claim that PDF-based solutions are the best way to preserve electronic documents*
- *But once you have decided to use a PDF-based approach, PDF/A defines an archival profile of PDF that is more amenable to long-term preservation*

And, most of us thought "PDF" was a 'standard' format and that was all we had to do? Guess again. It is not that easy.

[1] A summary of the SNIA's Data Management Forum 100 Year Archive Task Force Requirements Survey is available online at www.snia-dmf.org .
[2] PDF – refers to the Adobe Portable Document Format (PDF), PDF/A refers to the "Archive" version
[3] PDF/A, The Development of a Digital Preservation Standard, Abrams, Fanning, Helander, Sullivan – Aug 2005

The three preservation mantras of the archivist community[4] migration, encapsulation, and emulation also apply to the IT domain. Migrating has two dimensions, physical and logical. The United State's National Archives and Records Administration (NARA) offers these rules for physical migration: First, use the most current storage technology, then

- …if on disk, MIGRATE every 3 years
- …if on tape, MIGRATE every 5 years

So, what do you do if you have a very large repository like NARA's? How painful and costly is it to migrate a Petabyte per year? And, that is just the physical side. What about the logical? How are you going to guarantee that you can interpret the information in the long term? This is where methods such as emulation, encapsulation, rehosting, translating to 'more standard' formats such as images, PDF/A, or XML are in use, but not the end-all. The only thing you can guarantee is that the original application will not be around. Logical and physical migration are only two dimensions of the problem. The list of challenges is long and includes many technology and operations factors. Perhaps foremost on the operations side is the subtle but critical need to find value in the archives. After all, if there is no value, there is no budget, and who can afford to pay attention. In the end, it may be like compliance or legal risk; it is the price of failure that will be the motivator?

These dilemmas and challenges are well recognized by the archive communities. A classic statement exists in the ISO standard "Open Archival Information Systems"[5] (OAIS) which says:

> *"The fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media ARE AT ODDS with the key purpose of an Open Archival Information System: to preserve information over a long period of time. No matter how well an Open Archival Information System maintains its current holdings, it will eventually NEED TO MIGRATE much of its holdings to different media and/or to a different hardware or software environment to keep them accessible."*

Based on the work done by the 100 Year Archive Task Force, we need to make another very important point about the operational challenges of long-term digital information retention. We are experiencing extraordinary changes in the industry driven by factors such as regulatory compliance and legal and security risks. The survey confirms that retention periods are increasing. What is the impact? Here is an example. In most organizations e-mail has turned into a vital business record and individual e-mail messages are now the target for legal discovery. Consider the storage consequence and other IT issues this overwhelming volume creates. Without new automation methods, how can larger organizations manually process, classify, retain, dispose (securely delete), and manage millions of individual records a day? This is just a single manifestation of the crisis. It really is a different world now. It will take a new approach to solve these large-scale problems.

Recommendations

Instead of continuing to operate with two different domains (specifically RIM vs. IT), each with different operating requirements and methods, it is time the interests and needs of both

[4] See the Reference Model for an Open Archival Information System (OAIS) and other archiving standards.
[5] ISO 14721:2003, Reference Model for an Open Archival Information System (OAIS)

communities come together and move ahead with a single purpose and single approach. How do we do that? It begins with collaboration[6] and then setting in place a process called information classification[7] followed by implementation of Information Lifecycle Management (ILM) practices aimed at automating the IT infrastructure so that it will meet the business requirements for the information were trying to preserve. The RIM community has a key part of the expertise IT needs. Let's use it. To quote one respondent to the recent survey:

> *"Remember that IT doesn't own the information. RIM, Legal, Business units and IT all have a part to play in the decisions applied to business records and should be sitting down at the table together."*

These challenges must be addressed and standards and best practices developed consistent with the operating practice we call ILM. This is the charter of the Storage Networking Industry Association's (SNIA's) 100 Year Archive Task Force. The Task Force is an open multi-agency organization actively soliciting participation. You can participate and access it through www.snia-dmf.org/100year.

In summary, long term digital information retention is approaching a crisis. We are talking about the amalgamation and collection of information for which businesses and organizations are being held legally accountable. Is it not correct that the "archival" problem has expanded beyond the notion of "retaining records," to retaining "all relevant digital information" that is important to the business for periods of time specific to that information's lifecycle?

The problem of long-term digital information retention is soon going to be a multi-Petabyte problem in many organizations. It is not a problem that will be solved by casting it into the basement of IT practices. Rather, it requires a new and comprehensive systems engineering and automation approach otherwise it will overwhelm us. Let's face it, storage systems were not designed for this purpose. If we can solve the system problem then everyone benefits. The paradox as we see it today is that this isn't just a technology problem, it is also an organizational problem, requiring new information-based management practices which begin with collaboration and information classification.

*** end ***

Michael Peterson, President Strategic Research Corp. and Chief Strategy Advocate, for the SNIA Data Management Forum
www.sresearch.com and www.snia-dmf.org

[6] "Collaboration: The New Standard of Excellence", a SNIA-ARMA joint publication, October, 2006 and "Collaborate or Die!" by Michael Peterson, Nov. 2006. Both available at www.snia-dmf.org.

[7] "Managing Data and Storage Resources in Support of Information Lifecycle Management", Gelb, St.Pierre, Yoder, SNIA ILM-TWG, July 2006 and other relevant documents available at www.snia-dmf.org.

About the Author(s)

Michael Peterson is President of Strategic Research Corporation based in Santa Barbara California and Chief Strategy Advocate for SNIA's Data Management Forum. For the past 20 years he has been an energetic leader and catalyst for the storage industry, publishing insightful books and industry reports, consulting with the entire industry in business and market development, pioneering IT research on storage and management practices, creating innovative conferences, speaking internationally as an industry visionary, forming industry trade groups, and even developing new solutions and companies. Michael is the founder of the SNIA and was the past president from 1998 to 1999. He is currently the Chief Strategy Advocate for SNIA's Data Management Forum with responsibility for guiding and promoting the many programs of the DMF. SRC's website is www.sresearch.com.

About the Data Management Forum

The Storage Networking Industry Association's Data Management Forum (DMF) is a cooperative initiative of IT professionals, integrators and vendors working to define, implement, qualify and teach improved and reliable methods for the protection, retention and lifecycle management of electronic data and information. The DMF operates three initiatives: the Information Lifecycle Management Initiative, the Data Protection Initiative and the Long Term Archive and Compliant Storage Initiative. DMF also sponsors many industry wide task forces coordinating with a broad range of trade associations and agencies. To participate with the DMF or find out more about our programs, go to www.snia-dmf.org.

About the SNIA

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization, made up of more than 450 member companies and close to 8,000 active individuals spanning virtually the entire storage industry. SNIA members share the common goal of advancing the adoption of storage networks as complete and trusted solutions. To this end, the SNIA is uniquely committed to delivering standards, education and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at www.snia.org.