Solving the Coming Archive Crisis - the 100 Year Dilemma

Michael Peterson,
President,
Strategic Research Corp.
Chief Strategy Advocate
SNIA Data Management Forum
SNIA Legal Notice

• The material contained in this tutorial is copyrighted by the SNIA.

• Member companies and individuals may use this material in presentations and literature under the following conditions:
  – Any slide or slides used must be reproduced without modification
  – The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.

• This presentation is a project of the SNIA Education Committee.
Abstract

• Solving the Coming Archive Crisis
  – The volume of disparate digital information sources being kept online for long term preservation is overwhelming and leading to a crisis of cost, business risk, and complexity!
  – Long-term accessibility, accurate full-text search, and data extraction in “original form” with all original properties in-tact is just one of many requirements exasperated by the volume of disparate information.
  – New methods and practices are required to deal with the avalanche.

• This presentation will discuss the following:
  – Best practices for long term digital information retention based on the work of the 100 Year Archive Task Force
  – New technologies you can deploy that will help such as de-duplication and CAS
  – Future technologies and solutions that we can expect
About SNIA and the DMF

About the Storage Networking Industry Association (SNIA)
- SNIA’s primary goal is to ensure that storage networks become complete and trusted solutions across the IT community
- For additional information about SNIA see [www.snia.org](http://www.snia.org)
- SNIA’s “Dictionary of Storage Networking Terminology” is online at [www.snia.org/dictionary](http://www.snia.org/dictionary)

About the SNIA Data Management Forum (DMF) [www.snia-dmf.org](http://www.snia-dmf.org)
- The DMF is a sub-group of SNIA acting as the worldwide authority on Data Management, Data Protection and ILM
- The DMF is a collaborative storage industry resource available to anyone responsible for the accessibility and integrity of their organization’s information.

<table>
<thead>
<tr>
<th>DMF</th>
<th>Data Protection Initiative (DPI)</th>
<th>Information Lifecycle Management Initiative (ILMI)</th>
<th>Long term Archive and Compliance Storage Initiative (LTACSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining new approaches and best practices for data protection and recovery</td>
<td>Developing, teaching and promoting ILM practices, implementation methods, and benefits</td>
<td>Addressing challenges in developing, securing, and retaining long-term digital archives</td>
<td></td>
</tr>
</tbody>
</table>
“As an archivist, I oversee a collection that we've maintained for 40 years which includes published materials from as early as 1460 and manuscript materials from as early as 1390. In terms of data and automation system, the collections include Hollerith cards, Dictaphone tapes, 5 ¼ disks, 3.5 disks, CDs, DVDs, etc. We also maintain a small collection of 386 and 486 machines, drives, as well as various software packages (Windows 3, Wordperfect 4.2, 5.0, 5.1, 6, etc.) in order to read and migrate and convert.”

Source: 100 Yr Archive Requirements Study, An archivist at a major University
Or Here? (Clue - TIME!)

- 81% of respondents have a retention problem over 50 years
  - How will they do that?
- What is long term?
  - Beyond which logical & physical migration is required
  - >15 years

Source: 100 Yr Archive Requirements Survey, 12/06, n=276
Or Here?  VOLUME & $$

- How do you migrate 10 PB per year?
- What does it cost?

Source: Ken Thibodeau, Director ERA
Does Your Businesses Know the PAIN of Preserving Information?

Have you ever “tried” to retrieve ‘old’ information?…
Preservation is Complex

• Technology Challenges
  – Harvesting data sources
  – Maintaining Physical readability
  – Maintaining Logical readability
  – Migration of large repositories
  – Emulation of formats
  – Maintenance of historical readers and applications
  – Protection from change
  – Protection from loss or damage
  – Physical & logical Security
  – Automation
  – Deletion & Purging
  – Search – Discovery
  – Testing/Auditing

• Operations Challenges
  – Collaboration on requirements
  – Setting requirements
  – Classification
  – Establishing adequate Metadata
  – Standardizing practices
  – Finding value in the archives

The rate of change in computing technologies is such that information can be rendered inaccessible within a decade. Preservation is therefore a more immediate issue for digital than for traditional resources. Digital information will not survive or remain accessible by accident: pro-active preservation is needed.

-- JISC, 2002

Solving the Coming Archive Crisis © 2007 Storage Networking Industry Association. All Rights Reserved.
What types of Gaps Exist?

Sedona Conference, 2004
- Digital preservation requires routine efforts to MIGRATE records to overcome software and technological obsolescence and from deteriorating media.
  - The results of NIST’s (Nat’l Institute of Standards & Techn) evaluations are controversial and do not agree with manufacturer and independent testing

NARA (National Archives & Records Admin)
- Use current storage technology
  - …if on disk, MIGRATE every 3 years
  - …if on tape, MIGRATE every 5 years

OAIS (Open Archival Information System)
- 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 OAIS: to preserve information over a long period of time.
- No matter how well an OAIS 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.
100 Yr Archive Reqmts Survey: Executive Summary

• Key Observations
  – The problems of logical and physical issues affecting our ability to achieve long term retention are real and generally, but not well, understood
  – Long term generally means over 15 years. IT can manage to migrate and retain readability for about this long. For longer periods, processes begin failing, become too costly, and the volume of information becomes overwhelming.
  – Long-term retention requirements are real. Over 80% of respondents report a need to retain information over 50 years and 68% report a need of over 100 years.
  – Help is desperately needed. Current practices are too manual, too dependent on human intervention, too costly, too prone to error, and not adequately coordinated across internal information owning organizations.
What is Needed to Succeed

- **Good Practices**
  - Modeled after OAIS, Sedona, and others...

- **Commitment**
  - Sr. Management Support
  - Budget
  - Authority

- **Requirements**
  - Collaboration
  - Information Classification

- **Experience**

- **A storage repository**
  - OAIS-like
    - Virtualization, CAS, integrated data-services
    - Automated management
      - ILM-based practices

- **Scalable migration capabilities**
  - Physical
  - Logical
    - Encapsulation, Emulation, Standard archive formats
Getting Attention and Commitment

- 4 factors drive need for long-term retention:
  - Legal, Compliance, Business, and Security Risk
  - ‘Other’ is principally the needs to preserve the business or organization’s history which is at risk

Source: 100Yr Archive Reqmts Survey, Dec, 2006, n=276
Setting Requirements

- **Information Convergence**
  Compliance, Legal Discovery, Security Risk

- **Collaborate**
  Stakeholders: Business Group, Legal, Finance, I.T., RIM, Security

- **Information-Centric Enterprise**

- **Information Classification**
  - Set requirements based on value to organization

- **ILM Practices**
  - Data Classification
  - Automate Services:
    - Information, Data, Security, Storage
What are the Storage Problems

- How do you know Digital Records being preserved for the required retention period
  - Are the requirements clear?
  - How do you ensure records are accessible, readable, and usable throughout their scheduled retention?
  - Do you have the right selection criteria?
  - What are you doing about duplicates?
  - How do you purge?
- Identify and preserve the originals
  - Including provenance and finding-aids
- Are there appropriate controls & plans in place to address the:
  - Obsolescence of software (Application and OS)
  - Obsolescence of compute hardware & drivers
  - Obsolescence of the storage media & hardware/software
  - Ability to read or interpret the data or information over their lifecycle
- How do you cost-effectively MIGRATE large volumes?
- How do you maintain logical readability?
- How do you protect and secure?
Create an OAIS

Electronic Archive:
- (Electronic) Archives are repositories for the long term storage (*meaning logical and physical preservation) of information. Electronic archives preserve the content, prevent or track alterations and control access to electronic records.

Sedona Conference, 2004
Federated means “many-to-many with central mgmt.”
Requires all information mgmt functions & services, not just a virtualized storage pool – must deal with issues ranging from privacy to logical & physical migration
Leverages XAM for application interfaces and metadata
Solving Storage Problems

- **A Federated Information Repository** (an OAIS)
  - **Storage Systems** – Federate & virtualize a set of tiered storage resources with centralized ILM-based management, ILM-based placement, retention, deletion, protection, and business continuity
    - Eliminate physical migration with self healing systems
  - **Integrity** – CAS or WORM practices
  - **Scalable:** – Transparent growth with de-duplication
  - **Access Model:** – Open standards-based with persistent names and meta-data management
    - Identifying and making searchable all information, provenance, meta-data, & content
  - **Security** – Maintain ACLs and authorization over time
Building Preservation Environments with Data Grid Technology
(NARA Research Prototype Persistent Archive)

Reagan Moore
Richard Marciano
Arcot Rajasekar
Michael Wan
Wayne Schroeder
Antoine de Torcy
Sheau-Yen Chen

http://www.sdsc.edu/NARA/
http://www.sdsc.edu/srb/

San Diego Supercomputer Center    NARA Research Prototype Persistent Archive
What about Logical Migration?

- Encapsulation (XML)
- Emulation
- Exporting & Translation into a ‘standard’ archive format
- Problem areas:
  - Persistent Naming
  - Provenance
  - Integrity

- ‘Self-Describing Data Formats’
  - OAIS – AIP (Archival Information Package)
  - SNIA, 100 Yr Archive TF
    - XAM: eXtensible Access Method
- New archive media formats
  - ECMA TC-15
- Meta-data standards
  - ISO, NISO, AIIM, ARMA
Move to “Information Objects”

**OAIS ARCHIVAL INFORMATION PACKAGE (AIP) Logical Structure with XML**

- **Content Data Object** - the raw data that is the focus of the preservation.
- **Representation Information** – the information required to interpret the raw data to its designated community.
- **Provenance** – the history and the origin of the content information and any changes that may have taken place since it was originated, and who has had custody of it since it was originated.
- **Reference** – globally unique and persistent identifiers for the content information.
- **Context** – documents reason for creation of the content information and relationship to its environment.
- **Fixity** – a demonstration that the particular content information has not been altered in an undocumented manner.
Advice from the Sedona Guidelines

• An information and records management policy should identify and prescribe time periods for the retention of information and records that are **APPROPRIATE TO AN ORGANIZATION’S NEEDS AND LEGAL RESPONSIBILITIES.**

• There is general consensus that regardless of the various capabilities of different **BACKUP SYSTEMS**, those systems are designed for the purpose of business continuity and **SHOULD NOT BE USED** as a substitute for records management.

• What must be stored in order to achieve this goal and the manner and length of storage time will generally be decided by an organization’s information technology professionals (**WITH SUBSTANTIVE INPUT FROM THE OTHER DISCIPLINES** — operational, records management and legal) as the individuals who will be relied on to manage the recovery.

• The responsible handling of electronic information and records should be considered **A CORE VALUE** of an organization.
Peer Recommendations

ORGANIZATIONAL

• “Get RIM and IT at the same table. Create a relationship. Both need to be included up front to develop solutions that will work on both sides.”

• “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.”

TECHNICAL

• “Metadata is important, and it's better to implement the metadata at the front end rather than populating the metadata after the record has been saved into the repository.”

• “Stay with industry standards”

• “…use mass storage devices“

OTHER GUIDANCE

• “Prayer“

• “Stay with paper? (just kidding)“
100 Year Archive Project Map

• Gather Requirements

• Analyze & Identify Gaps
  • Develop, Report, Publish

• Target Deliverables
  • Report/Publish

Business Requirements

Gaps

Develop & Validate

Best Practices Document

Logical Standard

Archivist Practices

RIM Practices

IT Practices

Standards

Solving the Coming Archive Crisis © 2007 Storage Networking Industry Association. All Rights Reserved.
Resources

- www.snia-dmf.org/100year
  - 100 Yr Archive Task Force

- NARA:
  www.archives.gov/era/presentations/index.html

- OAIS
  - ISO Reference model: Open Archival Information System

- Sedona Conference
  - http://www.thesedonaconfere nce.org/
Q&A / Feedback

• Please send any questions or comments on this presentation to SNIA: (track-datamgmt@snia.org)

Many thanks to the following individuals for their contributions to this tutorial.

SNIA Education Committee

100 Year Archive Task Force
Gary Zasman   Peter Mojica   Simona Cohen

www.snia-dmf.org/100year

Get Involved!
- Find a passion
- Join a committee
- Gain knowledge & influence
- Make a difference

www.snia-dmf.org