# Information Migration

**Raymond A. Clarke**
Sr. Enterprise Storage Solutions Specialist, Sun Microsystems - Archive & Backup Solutions
SNIA Data Management Forum, Board of Directors

# SNIA DMFs 100 Year Task Force Study Objectives

- Study the business and operational requirements for long-term digital information retention in the data center
  - **Goal:** Determine the requirements for long-term digital information retention in the data center. These requirements are needed to frame the definition of best practices and solutions to the retention and preservation problems unique to large, scalable data centers
  - **Research Hypothesis:** Practitioner's experiences with terabyte-size archival systems are adequate to define the business and operating requirements for petabyte-size information repositories in the data center

# Key Concern

- Logical and physical migration do not scale cost-effectively
  - Only operating standard today is to migrate information physically (to new media) every three to five years and logically (to new formats) before the applications and readers die and become obsolete (every 5-10 years)
    - **A never ending, costly cycle of migration**
  - Practitioners are struggling to keep up with migration requirements. Only 30% claimed to be doing physical migration correctly on disk & none on tape or optical. Only 20% claimed they were confident in their ability to logically migrate some of the data.
    - **Information is at risk long-term!**

# Key Findings

- The problems of logical and physical retention
  - > Practitioners are struggling – information is at risk long-term
  - > Problems are real and generally understood
- Long-term generally means over 10-15 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 organizations reporting have a need to retain information over 50 years and 68% report a need of over 100 years.

*"This is the problem with 'Digital Archive', you are not thinking long enough into the future."* *(Source: Respondent)*

# Requirements for Long-Term Retention

- Accommodate Drivers
  - > Legal risk
  - > Compliance requirements
  - > Business risk
  - > Security risk
  - > Preserving history
    - > Organizations
    - > Institutions
      - – Universities
      - – Libraries
      - – Archives
    - > Nations
- Overcome Inhibitors/Barriers
  - > Executive mgmt commitment
  - > Maintaining readability
    - > *Logical & physical migration*
  - > Collaboration between information owners and administrators
  - > Cost and complexity
  - > Professional status

100 YEAR Archive Task Force

*It's very scary to me that the administration is so cavalier about business records.*

*(Source: Respondent)*

# Requirements for Long-Term Retention

- Practitioner's Needs
  - > Solve logical and physical migration
  - > Solve technology obsolescence
  - > Improve business commitment
  - > Reduce operating cost
  - > Better management tools
  - > Better collaboration

*Distribution of state on disk must match the ongoing business value of the data – automatically. If not, it's an unsolvable problem, since humans cannot keep up with the data onslaught. (Source: Respondent)*

- Technology Issues
  - > Solve logical and physical migration challenges
  - > Solve the scaling problem to keep up with the growing volume
  - > Information classification
  - > Include provenance & metadata
  - > Dealing with growth & technology change/obsolescence
  - > Include databases and email
  - > Include legacy Information
  - > Better discovery and deletion

# Task Force's Recommendations

- Response to business drivers
  - > Solutions & best practices that are developed have to be compatible with requirements for compliance, discovery, integrity, privacy, protection, etc.
- Inhibitors to overcome
  - > The three technical inhibitors (migration, cost, & complexity) are essential elements of proper solutions
- Target solving the top storage-related problems
  - > Physical & logical Migration
  - > Integrating meta-data
  - > Reducing management cost and reduce operating costs through automation
  - > Keep information available, discoverable, protected, private & secure
  - > Integrate with XAM, ILM , and other existing standards & practices

# Long-Term Retention Reference Model

- Glossary (in review)
- Physical Migration
  - A virtualized, federated information repository in which self-healing eliminates need for physical migration
  - Add all required services (de-duplication, hash-based unique naming, location independence, encryption …
  - Meta-data: thru XAM
- Logical Migration
  - **SIRF: Self-contained Information Retention Format**
  - a container based on OAIS' Archival Information Package integrated with XAM
    - Through XAM applications can write archival formats containing metadata, information, and a reader.
    - XAM encourages support
- Best Practices

# Self-Contained Information Retention Format (SIRF)

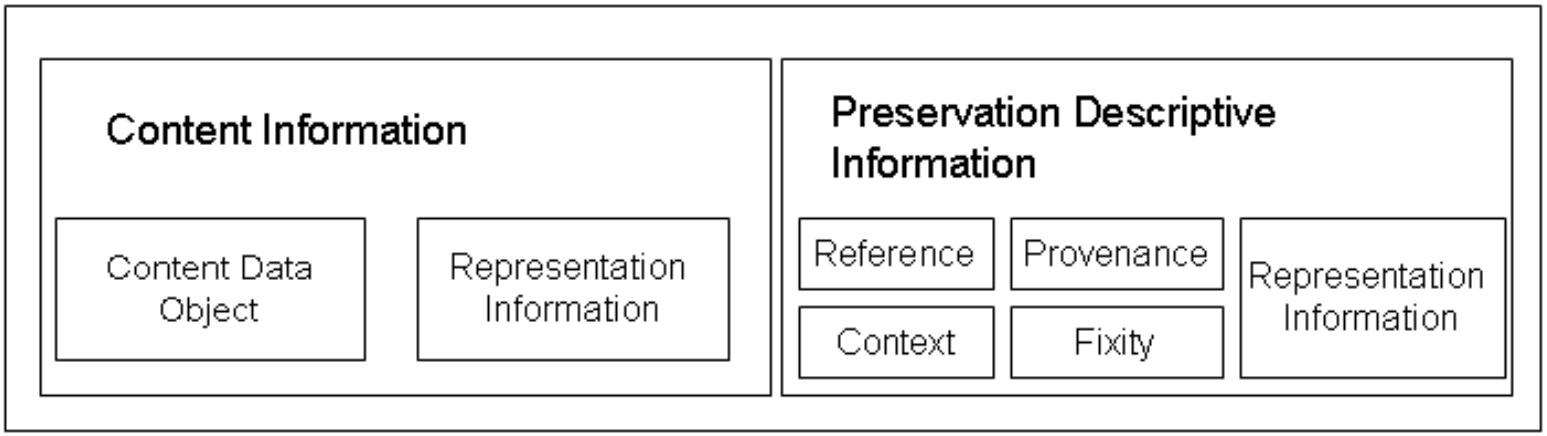
# Requirements for the Data Layer Format (1/2)

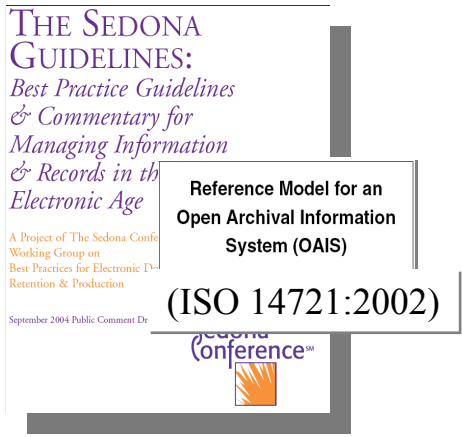
- Media agnostic
  - > Tape, disk, future media
- Vendor and Platform agnostic
- Self-describing
- Support self-contained data
  - > Include means to represent internal links and cross references
- Performance
  - > Needs to have good performance even for large datasets, including text and binaries
  - > Enable parallel reads and writes

# Requirements for the Data Layer Format (2/2)

- Interoperability
  - Need to be able to migrate data between different systems without loss of data
  - Can be interpreted in the future
- Extensible
  - Additional information which may be added in the future
  - Vendor specific extensions
- Cost
  - Free parsers
- Readable by both humans and machines
  - Ability to do offline inspection
- Support additional functions on the data
  - compression, encryption, cryptography

# OAIS AIP Logical Structure




- 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.
- Reference – globally unique and persistent identifiers for the content information.
- 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.
- 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.

# Proposal

- SIRF is a proposal for an open logical format standard based on marrying the OAIS AIP with XAM.
- The data format will include OAIS concepts
  - RepInfo, reference, provenance, fixity, context, etc.
- Add inter-link and external-link mechanism
- Include TOC that points to the various AIPs on the media

- VTL could be a natural translation point and search staging area for off line data

# Example Helpful Practices

- Classify your information (into a few common buckets)
- Set retention periods and delete 'expired' information
  - Free up space, only store what is required
  - Include your databases
- Control the number of copies for protection and operational recovery and their locations
- Set policies for audits and perform them
  - Measure and improve


Sun
microsystems
Thank You
for
Your Time and
Attention
Raymond.Clarke@Sun.com
(212) 558-9321