

Building a Terminology Bridge:

Guidelines for Digital Information Retention and Preservation Practices in the Datacenter

Bob “Mister” Rogers

Chair, Information Lifecycle Management Initiative
and CTO, Application Matrix



Data Management Forum

◆ Principal Author:

- ◆ **Michael Peterson** Chief Strategy Advocate for the **SNIA's Data Management Forum**, and CEO of Strategic Research and of TechNexus

◆ Supporting Authors:

- ◆ **Gary Zasman** Chair **Long-Term Archive and Compliant Storage Initiative** and **WW Practice Director**, NetApp
- ◆ **Jeff Porter** **2008 Chair-Emeritus**, **SNIA Data Management Forum** and **Senior Technologist**, SSG Office of the CTO, EMC
- ◆ **Peter Mojica** Chair **DMF-LTACSI Reference Guide Committee** and **Consultant** in Risk Management and Compliance Practices
- ◆ **Edgar St. Pierre** **Co-Chair SNIA Data Management Forum's ILM Initiative** and **Senior Technologist**, Office of the CTO, EMC
- ◆ **Bob Rogers** **Co-Chair SNIA Data Management Forum's ILM Initiative** and **CTO** Application Matrix

Why the Terminology Bridge?

- If your organization is operating an information governance-style committee or developing service management practices, and needs to develop business requirements for information assets, then this report provides the common terminology and understanding you need to communicate retention and preservation practices among all stakeholders
- If your organization needs to better understand retention and preservation principles and have a common terminology that spans internal departments and business units, external partners, customers, and vendors
- If your organization is dealing with eDiscovery, litigation holds, reducing risk and exposure, regulatory compliance, and/or long-term preservation of digital information, you need a tool to guide the development of key information management practices. This report will help develop a common understanding of practices and requirements within the datacenter.

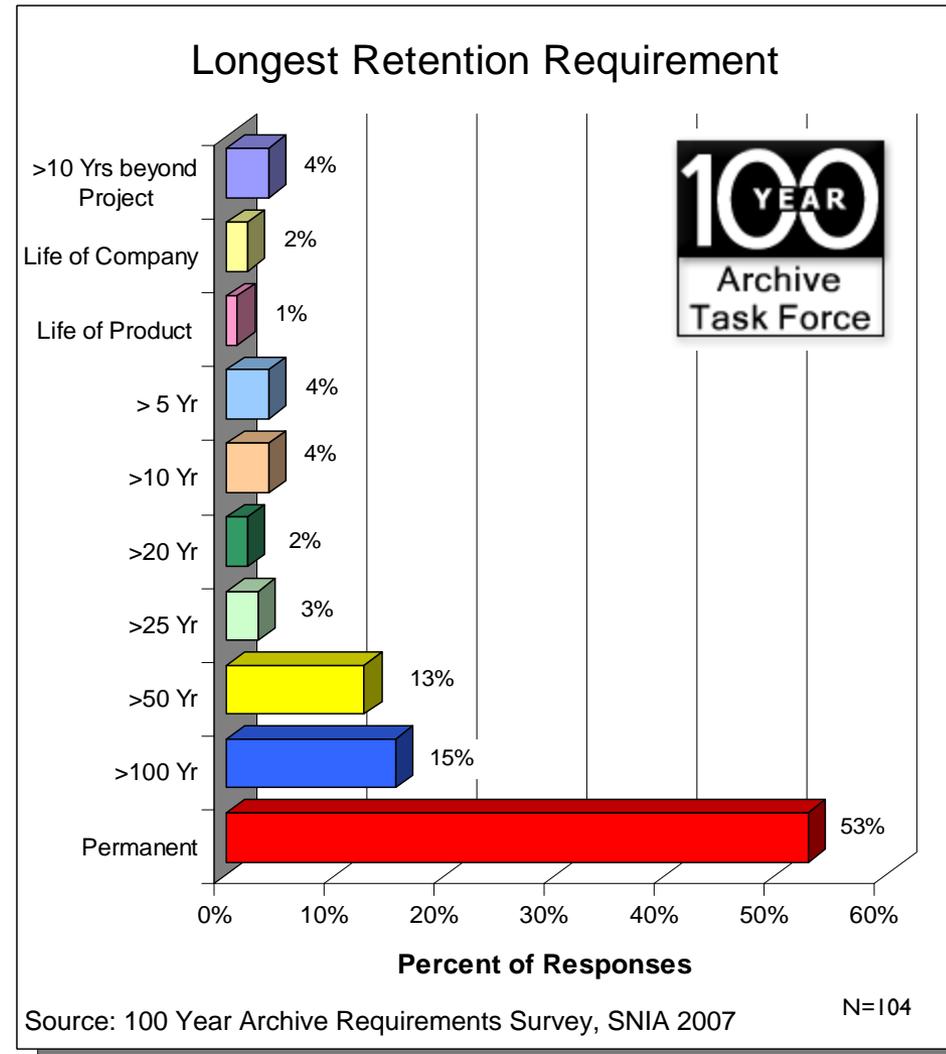
Why the Terminology Bridge? (cont'd)

“Achieving business alignment has been the holy grail of IT for 20 to 30 years, but the dialogue is broken between business and IT. The reason? They don’t have a language that works. The business wants to talk in business terms, and IT talks in technical terms” says Rudy Puryear, global IT practice head at Bain & Co.

Source: CIOInsight: “Why IT and Business Can’t Get In Sync”
Tony Kontzer, 2009-06-04

Longest Retention Requirement

- Long-term needs are real:
 - ◆ 68 % over 100 Years
 - ◆ 83% over 50 Years
- Requirements vary by organization type, information type, and compliance rules/risk
Leading Organizations: Education, gov., IT services, manufacturing, finance, insurance, pharma, & health-care
 - ◆ Leading Info-Types: Source-files, government, history, customer, & database records



SNIA 100 Year Archive Task Force Highlights

- Long-term retention needs are real and that many organizations have long-term requirements.
 - ◆ 80% of respondents declared they have information they must keep over 50 years
 - ◆ 68% of respondents said they must keep it over 100 years.
- Long-term generally means greater than 10 to 15 years – the period beyond which multiple migrations take place and information is at risk.
- Database information (structured data) was considered to be most at risk of loss.
- Over 40% of respondents are keeping e-Mail records over 10 years. E-Mail is not just a short-term problem.
- Physical migration is a big problem. Only 30% declared they were doing it correctly at 3-5 year intervals. The rest of the sample group is placing their digital information at risk.
- 60% of respondents say they are 'highly dissatisfied' that they will be able to read their retained information in 50 years.
- Help is needed – current practices are too manual, too prone to error, too costly and lack adequate coordination across the organization.
- Collaboration and classification were recognized as very important practices to get the organization working together setting requirements for the management of their information.

Report available at: http://www.snia.org/forums/dmf/programs/ltacsi/100_year/

Objective of the Terminology Bridge?

- **Aid in and stimulate adoption of ILM-based practices:** This report is designed to **'build a bridge' between disparate departments** and to guide organizations in developing terminology and practices suitable for their needs.
- **Improve communications:** by creating **a comparative terminology** between an ILM-based context and other key legal, information management, archival, security, and preservation oriented industry glossaries to act as a bridge to better communications within the datacenter
- **Explain terminology and practices:** by **improving the understanding** of what each retention and preservation oriented service attempts to achieve as a datacenter practice in the context of ILM-based practices

What it is Not

- It is not a dictionary
- It is not an architectural model
- It is not a legal (or security) position
- It is not a universal glossary
 - ◆ That is the wrong approach and not what the report says
 - ◆ Rather it is to stimulate each organization to refine the terminology and practices to meet their needs

- End-users are asking for help and guidance
 - ◆ Terminology and better understanding of practices are key to moving ILM as a service management practice ahead
 - ◆ Many end-users care and are asking for this tool
- Alliance Partners
 - ◆ ARMA, AIIM, SIM, SAA, CASPAR, StorTOC
- We are being copied by others now

Retention and Preservation Terminology

- ▶ Active Information or Data
- ▶ Archive
- ▶ Audit Log (Audit Trail)
- ▶ Authenticity
- ▶ Classification
- ▶ Data (Digital)
- ▶ Data Deduplication
- ▶ Deletion
- ▶ Digital Fingerprinting
- ▶ Disposition Policy
- ▶ Electronically Stored Information
- ▶ Emulation (System or Software Emulation)
- ▶ Encapsulation (Information Encapsulation)
- ▶ Expired Information or Data
- ▶ Fixity
- ▶ Inactive Information or Data
- ▶ Information (Digital)
- ▶ Information Object
- ▶ Information (or Data) State

Retention and Preservation Terminology (cont'd)

- Expired Information or Data
- Fixity
- Inactive Information or Data
- Information (Digital)
- Information Object
- Information (or Data) State
- Ingestion
- Integrity
- Logical Format
- Long-term
- Long-term Digital Information Preservation
- Metadata
- Migration
- Permanent Deletion
- Preservation
- Preservation Repository (Preservation Store)
- Provenance
- Record (Digital)
- Reference Information or Data
- Retention
- Versions and Copies

- The report advocates that IT practices adopt a more consistent usage of the term 'archive' to facilitate interaction with other departments within the organization. To the archival, preservation, and records communities, an archive is a specialized repository with preservation services and attributes.
- Typical IT use of the verb “archiving” actually refers to a practice based on ILM called “tiering,” the migration of inactive or expired information to a lower tier of storage to reduce cost and improve storage efficiencies.
- Another IT misuse happens when ‘archive’ is confused with backup. Backup does not create an archive (a preservation store) nor should backup media be used for such.

The Preservation Example

- Managing information in today's datacenter with requirements to safeguard information assets for eDiscovery, litigation evidence, security, and regulatory compliance requires that many classes of information be preserved from time of creation.
- Preservation is a set of services that protect, provide availability, integrity and authenticity controls, include security and confidentiality safeguards, and include an audit log, control of metadata, and other practices for each preservation object.
- The old IT practice of placing information into an archive when it becomes inactive or expired no longer works for compliance or litigation support, and only adds cost. Thus, we see products and practices like eMail Archive, Compliance Storage, Preservation Stores, and Database Archive being used to capture and preserve key classes of information and data upon creation.

- Authenticity is defined in a digital retention and preservation context as a practice of verifying a digital object has not changed or is not corrupted. Authenticity attempts to identify that an object is currently the same genuine object that it was “originally” and verify that it has not changed over time unless that change is known and authorized. Authenticity verification requires the use of metadata.
- The critical change for IT practices is that metadata is now very important and must be safeguarded with the same priorities the data is. IT practices that damage, merge, ignore, or scramble metadata are no longer appropriate.

How do I get a copy of the report?

- You can download your own copy from
 - ◆ www.snia.org/forums/dmf/knowledge/term_bridge
- Your comments and input are welcome at
 - ◆ <http://community.snia-dmf.org>

