The Need for a Terminology Bridge

May 2009

Principal Author:

Michael Peterson Chief Strategy Advocate for the SNIA's Data Management Forum, CEO, Strategic Research Corporation and TechNexxus

Supporting Authors:

Bob Rogers Chair of the Data Management Forum's ILM Initiative, and CTO, Application Matrix

And DMF member companies:

Produced and supported by the SNIA Data Management Forum. The Data Management Forum, DMF, operates a public website at www.snia.org/forums/dmf Comments to this document can be discussed online at the DMF's community site -- http://community.snia-dmf.org.

When data is moved off of a disk array is it called migration, tiering, archiving, disk grooming, hierarchical storage management, deletion, move, disposition, or something else? What kind of migration? What kind of archiving? Is that file-system deletion, permanent deletion, shredding, purging, wiping, destroying, overwriting, soft deletion or secure deletion? Is tiering the same as migration or different? The answers you'll find in a typical organization depend as much on what department you ask as which dictionary you choose to reference. Terminology in the datacenter can be very confusing. At least 20 different dictionaries exist and are available on the internet. In addition, IT, records management, security, legal, compliance, and the business groups all have their own vernacular.

The issue is not about having one standard terminology set, one glossary, or one master universal dictionary. Rather, it is about what is right for your organization. How do you make those decisions? How do you even get people in a room to discuss it and get them to care? There has to be a larger, over-riding reason and purpose and the good news is that there is.

Whether setting business requirements for service management or information management purposes in an information governance[1] context or trying to establish better practices in support of cross-departmental operational needs, having a consistent terminology set will accelerate the ability of any project to succeed. Building any of these collaborative practices begins with pulling together a committee from legal, IT, records management, security, and the business group. Among the first tasks is to agree on terminology. People need to be able to communicate with each other and understand how practices and services work in the datacenter and how they can be applied to better address the business requirements.

In 2007, the Data Management Forum's, Long-Term Archive and Compliant Storage Initiative set out to develop a terminology set that crossed departmental boundaries and addressed the management of digital information for retention and preservation in the datacenter. After over a year of correlating that terminology internally and externally, the report from this effort is now available. "*Building a*

[1] See the SNIA-ARMA paper: *"Collaboration: The New Standard of Excellence"*, 2006

Terminology Bridge: Guidelines for Digital Information Retention and Preservation Practices in the Datacenter" is designed to aid in agreeing on a common language and practices between disparate departments. It will help organization's get started in implementing their collaborative governance committees, service management methods, and Information-Lifecycle Management, ILM, based practices in a datacenter environment.

"*TERMINOLOGY BRIDGE*" OBJECTIVES

- **Stimulate adoption of ILM:** by reducing communication barriers, thereby building bridges between departments and encouraging implementation of ILM-based practices
- **Improve communications:** by creating a comparative terminology between an ILM-based context and other key information management, archival, 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

The following are three examples of the types of terminology discussion found in the "*Building a Terminology Bridge"* report. The format for the discussion around each term in the report is an explanation of how the term is used, some examples, and a set of relevant reference definitions taken from among 20 industry glossaries. These reference definitions give a broader understanding of how the terms are used in other contexts and help to understand why the differences exist when using an ILM-based methodology.

EXAMPLES FROM THE "*TERMINOLOGY BRIDGE*"

- **Preservation**: is a collection of services that maintain and ensure the readability, accessibility, usability, security, and the genuine character of information over the long-term. Examples of these services include: the ability to read and interpret information in its original context over time and across hardware and software obsolescence, to protect it from loss or change, to verify and protect its authenticity, availability, and security for the entire information object, including its data and metadata. The unique change from the old way of thinking about 'archival' practices is that preservation services are required to deal with legal, security, and compliance risk as well as long-term retention requirements from creation to expiration. Placing a copy of business or compliance records in a preservation store after they have become 'inactive' or as a final disposition event is too late and too costly.

  The big technical problem and operational expense in long-term retention is dealing with hardware and software obsolescence. Over long periods of time not only do applications and hardware change and become obsolete, but people lose expertise to operate and interpret them. These problems are not solved and may never be fully. Today, there are three main methods used to maintain access to and readability of digital objects in the face of

obsolescence: encapsulation, emulation, and migration. [See also: "Long-Term Digital Information Preservation," "Encapsulation," "Emulation," "Migration", "Authenticity," and "Archive"]

- Sample reference definitions:
  - Process and operation involved in ensuring the technical and intellectual survival of authentic records through time. (Source: ARMA and National Archives and Records Admin)
  - The act of maintaining correct and independently understandable information over the long term. (Source: PREMIS)

- **Archive**:
  An archive is a specialized repository (including the supporting processes, policies, hardware, and software) used to preserve information and data for the long-term. This repository is more currently being called a 'preservation store' or a 'preservation repository. The capabilities of an archive or a preservation repository are the same. They include the ability to preserve, protect, control, maintain authenticity and integrity, accommodate physical and logical migration, and guarantee access to information and data objects over their required retention period.

  Archive and archiving (the verb form of archive) are inconsistently-used historical terms whose use needs to be updated because of the current risk-driven compliance and litigious business environment. This transformation is being driven by a changing use model that requires all electronically stored information, ESI, including that being preserved long-term, be subject to legal discovery. ESI must be able to be located, indexed, and controlled. The practices of "retention and preservation" more precisely define the requirements now than does "archive." But, if you chose to use the term "archive", make sure you define it and its context to reduce confusion. Retention is applicable because all information and data should have a defined retention policy that may range from short-term to long-term or even to forever. Consequently, the concept of "archiving" as a separate practice often performed when information is no longer considered "active" or an "archive" as a unique collection of assets can be replaced by retention and preservation for all the information and data. The legal, security, and business risk requirements mandate that to cope with the scale and complexity of information in the datacenter, information assets must be preserved over their retention period. [See also "Electronically Stored Information," "Preservation," "Preservation Repository," "Information Object," and "Retention"]

  - Sample reference definitions:
    - Noun: Archives are long term repositories for the storage of records. Electronic archives preserve the content, prevent or track alterations and control access to electronic records. (Source: Sedona Principles)
    - A collection of data objects, perhaps with associated metadata, in a storage system whose primary purpose is the long-term preservation and retention of that data. (Source: SNIA Dictionary)

- **Information vs. Data**: The debate of what is digital information and what is data can be confusing and highly opinionated. Most definitions apply human, application, or process interpretation as the limiter or definer to distinguish information from data. For example, the SNIA's current definition for information is "*Information is data that is interpreted within a context such as an application or a process."* This definition is valid but not useful for specifying retention or preservation practices. In the IT realm, more precision is needed so that information and data services can take responsibility to recognize and preserve the data and its associated metadata. The preservation community[2] addressed this dilemma when they defined information as a digital object made up of metadata plus data. Consequently, the solution to the debate is to use the term "information object" instead of 'information' to describe information in this context.

  Examples of use:

  Examples of information objects include a report or results from a database query (not the 'data' in each database cell), or a file, or a postscript image. A file is an 'information object' by this approach because it is a digital object that encapsulates the data with system and user metadata and other reference and representation information, that give it context and relevance. It is in the context of preserving metadata that maintaining a distinction between information and data is especially important. As the Open Archival Information System, OAIS, reference model says, "In general, it can be said that data interpreted using its representation information yields information. In order for this information object to be successfully preserved, it is critical for an OAIS to clearly identify and understand the data object and its associated representation information. For digital information, this means the (archival system itself) must clearly identify the bits and the Representation Information that applies to those bits. This required transparency to the bit level is a distinguishing feature of digital information preservation, and it runs counter to object-oriented concepts which try to hide these implementation issues. This (difference in approach) presents a significant challenge to the preservation of digital information." [See also "Data," "Information Object," and "Metadata"]

  - Sample reference definitions:
    - In the digital library community, the definition commonly used for a digital object is a combination of identifier, metadata, and data. And, a digital object is defined as a discrete unit of information in digital form. (Source PREMIS)
    - Information Object: A Data Object together with its Representation Information. An Information Object is composed of a Data Object that is either physical or digital, and the Representation Information that allows for the full interpretation of the data into meaningful information. (Source: OAIS)

[2] The preservation and digital library community consists of "archivists" librarians, IT experts, and those engaged in the provision of long-term preservation services as typically found in the digital library, historical, cultural, and governmental preservation arenas.

Use the *Terminology Bridge*

If the objective of your organization is to empower a collaborative governance-style committee to work together and to have good communications across departmental boundaries, then a common terminology is essential. One of the first steps for this committee is to agree on the terminology used for each practice or methodology being employed.

The SNIA-DMF has just published an important report for this purpose called *Building a Terminology Bridge: Guidelines for Digital Information Retention and Preservation Practices in the Datacenter*. It will give you an understanding of the terminology and a framework consistent with any information management or governance practice. You can obtain a copy of *Building a Terminology Bridge* from the DMF's website at http://www.snia.org/forums/dmf/knowledge/white_papers_and_reports/ and you can participate in active discussion around it and information-lifecycle management at the DMF Community site, http://community.snia-dmf.org. Your feedback is welcomed and encouraged.

About the Data Management Forum:

The SNIA Data Management Forum is a cooperative initiative of IT professionals, vendors, integrators, and service providers working together to conduct market education, develop best practices and promote standardization activities that help organizations become Information-Centric Enterprises. Areas of focus include the technologies and services that support information lifecycle management, data protection, and information retention, and preservation. For more information, visit www.snia.org/forums/dmf or participate in our open online community http://community.snia-dmf.org

About the Storage Networking Industry Association:

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization, made up of some 400 member companies spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies, and educational services to empower organizations in the management of information. 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.