



Building a Terminology Bridge

**Guidelines for Digital Information Retention and
Preservation Practices in the Datacenter**

September 2009

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

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 VP Product Strategy and Management, Unify

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

Disclaimer Note: This report is a product of the SNIA's Data Management Forum and does not represent all viewpoints within SNIA's overall charter. Its content is designed to address only the context of retention and preservation of digital information in the datacenter based on ILM-practices. It is hoped that this report will stimulate debate and collaborative agreement on how to define key practices in the datacenter as organizations go about the process of implementing information governance and service management methods to deal with the broad spectrum of challenges around enterprise information management.

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 as well. <http://community.snia-dmf.org>

Copyright © 2009 SNIA

All rights reserved. All material in this publication is, unless otherwise noted, the property of the SNIA. Reproduction of the content, in whole or in part, by any means, without proper attribution given to the publisher, is a violation of copyright law. This publication may contain opinions of the SNIA, which are subject to change over time. The SNIA logo and the Data Management Forum (DMF) logo are trademarks of the SNIA. All other trademarks are the property of their respective organizations.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Table of Contents

- Executive Summary** 1

- Retention and Preservation Terminology**

 - Active Information or Data 6
 - Archive 6
 - Audit Log (Audit Trail) 7
 - Authenticity 9
 - Classification 11
 - Data (Digital) 13
 - Data Deduplication 14
 - Deletion 16
 - Digital Fingerprinting 17
 - Disposition Policy 19
 - Electronically Stored Information 20
 - Emulation (System or Software Emulation) 21
 - Encapsulation (Information Encapsulation) 22
 - Expired Information or Data 23
 - Fixity 23
 - Inactive Information or Data 24
 - Information (Digital) 24
 - Information Object 25
 - Information (or Data) State 27
 - Ingestion 29
 - Integrity 30
 - Logical Format 31
 - Long-term 32
 - Long-term Digital Information Preservation 33
 - Metadata 33
 - Migration 34
 - Permanent Deletion 37
 - Preservation 38
 - Preservation Repository (Preservation Store) 39
 - Provenance 40
 - Record (Digital) 41
 - Reference Information or Data 41
 - Retention 42
 - Versions and Copies 42

- Appendix:**

 - Terminology Reference Sources & Bibliography 44
 - Index 45

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

EXECUTIVE SUMMARY

In 2004, the SNIA's Data Management Forum, DMF, began work on a project to define best practices for long-term preservation in the datacenter, under the name of "the 100 Year Archive Task Force." The first task undertaken by the Task Force was to reach out to the many preservation and archival communities and to investigate 'prior-art' such as existing standards and preservation "best-practices." What rapidly became clear was that no definitive standards existed clarifying how to operate these practices in the datacenter. Yes, there are standards such as those published by the International Standards Organization, ISO, for libraries, data acquisition, and historical, artistic, or cultural records preservation practices, but not for the litigation-constrained enterprise datacenter with its billions of unstructured, semi-structured, and structured files and data-objects growing at 50% per year. In 2004, only a few governmental agencies¹ had defined some storage-specific requirements that were applicable. In the literature of the period, is discussion of an interim requirement for physical migration, "If it is stored on disk, migrate the information every 3 years and if it is stored on tape, migrate it every 5 years." All other practices merely said in effect, "use whatever current storage practices exist." Nothing has changed in the ensuing years. These results leave two huge and urgent gaps that need to be solved². First, it is clear that digital information is at risk of being lost as current practices cannot preserve it reliably for the long-term, especially in the datacenter. Second, the explosion of the amount of information and data being kept long-term make the cost and complexity of keeping digital information and periodically migrating it prohibitive.

These two forces of cost and complexity are poised to overwhelm the datacenter as well as the preservation repositories of the world³. Take the physical migration requirement as a simple example. If a datacenter has 1.0 Petabyte of information today that it is keeping long-term, then it will have 2.25 PB after 3 years at a 50% annual growth rate. In year 3, it should migrate all of year #1's information and data that is on disk to new disk systems. Let's assume that they have 25% of the total storage on disk and 75% on tape and analyze the cost. The cost of 250TB of installing new enterprise-class disk arrays is ~\$400,000 at today's prices, the labor to perform the entire migration process is another ~\$100,000⁺, and the time to plan, install, configure, and migrate at 24x7 operations plus error correction ~3-4 months. By year 5, the datacenter now has 5.1PB under management and the migration load is 750TB of tape plus 560TB of disk. In total, they now have 1.3PB to migrate and the costs to do so on the order of \$2Million. Just the time to migrate will take ½ year of 24/7 operations with no errors. Many people would say it would take over a year to physically migrate this much content. When you add all the planning, configuration, implementation, and operations time, the time required for physical migration of a PB is

¹ Source: "Data Management Services for Global Change Research," Distributed Active Archive Centers (NASA), 2004

² Source: "The 100 Year Archive Requirements Survey", SNIA-DMF, January 2007

³ Source: Library of Congress, Blue Ribbon Task Force Report: "Sustaining the Digital Investment" December 2008

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

potentially full-time employment for a bunch of people. Now add the parallel demands of logical migration and by year 6 or year 7 in this scenario, we run out of time and money for migration.

Take note. The requirement is to save and preserve digital information for potentially hundreds of years, not 6 or 7. Who is going to do it? Who can afford it? The current process is broken, does not scale, is pure overhead, and paradoxically not where any business organization wants to be spending its time and money.

Two interesting things happened along the path of developing best practices and terminology for long-term retention and preservation of digital information in the datacenter. First, In December of 2006, the revised amendments to the Federal Rules of Civil Procedures, FRCP, were published. In 2007, the amendments began to be broadly interpreted and put to use strengthening the focus on discovery of “electronically stored information,” ESI. In 2008, recognition of the role and importance of metadata was elevated⁴. Now, not only is all ESI discoverable, but its metadata may also be required. IT systems and business applications were not designed to fully utilize and preserve metadata. Changes in IT and storage practices are needed. From an information management perspective this is good news. Expanded and updatable metadata is the key to implementing ILM-based practices⁵. Emphasis placed on metadata by the legal community only helps catalyze this change; consequently the requirements for retention and preservation are shifting as well. What does this statement mean? An operative example is found today in the litigation discovery

process. When information is discovered that meets the profile that is being searched for, a copy is usually captured and ingested into a separate, controlled repository. Via this ingestion process, its history and chain of custody are logged and controlled from that point on so that the digital evidence can be verified as


Figure 1

The New Reality

- ▶ You never know when you will have to produce ESI as evidence
 - ◆ You can still ‘delete’ based on implementation and control of proper policies and practices as defined with your legal department
- ▶ Retention practices are key
 - ◆ Consistent adherence to policies more important than the actual practice

Preservation Starts Day #1

Retention:
To keep and control information or data for specific periods of time
(Source: DMF and SAA)



⁴ Source: Sedona” Commentary on ESI Evidence,” March 2008 and the case: “Aguilar v. Immigration” November 2008

⁵ The SNIA’s eXtensible Access Method, XAM, standard provides just such a metadata container

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

genuine and free of tampering by the discovery and evidence review process in court. In this example, preservation begins upon ingestion into the controlled discovery repository. What happened before this capture event may or may not be logged in the information's metadata or may not help identify things like what is an original, a version, or a copy. Evidentiary rules take over to deal with these issues. In the near future, expect that retention and preservation practices will have to be initiated upon creation of information to accommodate legal, security, risk management, and compliance requirements. An important impact of this anticipated change is that the old approach to preservation, as an event that starts when records are no longer 'active' and are ingested into a long-term preservation repository (an electronic archive), is too late and no longer appropriate for the datacenter.

The second evolutionary change to the plan for this terminology 'bridge' came about while making revisions to accommodate the first one. As illustrated in the discussion of the growth and cost problem on page 1, it is paramount that the operating processes used as best practices in the datacenter be instrumented and automated as much as possible otherwise they won't scale cost-effectively. The framework for datacenter infrastructure management recommended is the practice that the SNIA's Data Management Forum has been developing since late 2003 called Information-Lifecycle Management⁶.

Information-Lifecycle Management, ILM, is a management framework that supports a service management-style approach to defining IT practices and it can effectively be applied to governance, compliance, risk, and information management. ILM-based practices uniquely use the value of and requirements for information to define the policies that the datacenter infrastructure and IT have to support.

What began as a paper focused at developing a terminology set to improve communication around the long-term preservation of digital information in the datacenter based on ILM-practices, has now evolved more broadly into explaining terminology and supporting practices aimed at stimulating all information owning and managing departments in the enterprise to communicate with each other about these terms as they begin the process of implementing any governance or service management practices or projects related to retention and preservation. Agreeing on terminology, just like managing information successfully in the face of legal, business, and security risk is difficult. Both require active collaboration by all information owning and administrating parties⁷ regardless of the governance or management practices in which they are engaged.

⁶ To IT and the storage industry, ILM was defined in 2004 by the SNIA as "The policies, processes, services, and tools used to align the business value of information with the most appropriate and cost-effective infrastructure from the time information is created through its final disposition." To the records management industry, ILM includes every phase in the lifecycle of a 'record' from its beginning to its end. ILM is also a part of the overall management approach of enterprise content management and records management processes. ILM is a classic term describing an important and broad set of practices with many different interpretations and similar need for alignment.

⁷ See the SNIA-ARMA paper: "*Collaboration: The New Standard of Excellence*", 2006

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Usually, this group of stakeholders includes at least legal, RIM, IT, security, and the business group often under the guise of some multi-departmental group such as the information governance committee⁸. Each department comes to this meeting with a different understanding of terminology and their role in the management framework, consequently a common perspective and a common language is needed to empower communication.

This document is intended to contribute to structuring that bridge. It does not have to be all things to everyone to contribute to helping the process. Its role is to identify terminology in a context that looks forward to how practices should operate in a risk and compliance driven datacenter, not necessarily how it may be done today. It supports and articulates best practices for retention and preservation of digital information in large and scalable datacenter environments using ILM-based practices to implement supporting services.

While developing these definitions and their explanations, valuable insight was received from some records managers that participate in the Data Management Forum. They suggested that a way to bridge the definitional differences between the goals of this work by the SNIA and the many organizations world-wide working on this problem would be to write it as a white paper and to provide everyone's definitions for comparison. Examples of these other organizations range from trade associations such as⁹ ARMA, AIIM, OSTA, SAA, the Sedona Conference, or the Research Libraries Group, to projects such as CASPAR, InterPARES, or Midess, to governmental agencies such as NARA, DOD, NIST, to legal documents such as the Federal Rules of Civil Procedures or the Federal Rules of Evidence, and to existing standards such as the ISO's Open Archival Information System standard, OAIS. In developing this terminology, glossaries from 20 groups were referenced. With many groups, many departments, and many companies in many different industries, terminology is a mess. Clearly, a terminology bridge is needed to enable communication so that collaborative Information-Lifecycle Management methods can begin.

Again, this paper is not a dictionary or encyclopedia encompassing all points of view. Instead, it provides an explanation of what the terms mean in the context of retention and preservation of digital information and in light of what the DMF believes to be best practices based on ILM-based practices for the datacenter. Clearly, there are other contexts to be considered and two stand out, legal and security. Legal interpretations in an evidentiary context overlap due to eDiscovery activity. Security and risk management groups have another viewpoint. You will need to work

⁸ See "The Future of Enterprise Information Governance", Published by the Economist Intelligence Unit, 2008

⁹ Association & agency acronyms: ARMA International– the records and information management association; AIIM – Association for Information and Image Management; OSTA – Optical Storage Trade Association, SAA – Society of American Archivists; CASPAR – Cultural, Artistic, and Scientific knowledge for Preservation, Access, and Retrieval; InterPARES – International Research on Permanent Authentic Records in Electronic Systems; Midess – Management of Images in a Distributed Environment with Shared Services; NARA – National Archives and Records Administration; DOD – Department of Defense; NIST – National Institute of Standards and Technology

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

with your internal legal counsel, security, and risk management departments to resolve any differences appropriate for your organization and come to consensus.

The approach taken in this document, hopefully, allows the reader to understand each term and its associated practices better and improves communication between disciplines that may use the same term differently. Use the definitions from the other reference sources for comparison, but always consider their context to understand the differences. The bottom line is that this document is designed to help organizations get started implementing Information-Lifecycle Management practices today. To that end, comments are encouraged and welcomed. This is a working document and will be periodically updated. Below is a summary of the “Terminology Bridge’s” objectives:

TERMINOLOGY BRIDGE OBJECTIVES:

- **Aid in and stimulate adoption of ILM-based practices:** Agreeing on terminology and practice objectives is one of the key starting points in implementing an ILM-based service management style practice methodology. 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 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

As you read this document, keep in mind the objectives and how the practices these terms support affect your ability to succeed in a risk-filled environment. Again, we encourage you to start implementing Information-Lifecycle Management methods and invite your review, analysis, comments, and feedback. Please give us that feedback through the public DMF community site (<http://community.snia-dmf.org>) or through the Storage Technology Online Community at (<http://community.stortoc.org>).

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

RETENTION AND PRESERVATION TERMINOLOGY

(Context: ILM-based practices)

Active Information or Data:

(See “Information-State”)

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 (ESI),” “Preservation,” “Preservation Repository,” “Information Object,” and “Retention”]

Examples of use:

One example of an archive is today’s “email-archive” repositories. The original purpose for email-archive systems was to capture copies of emails and associated attachments (upon transmission and after delivery), instant messaging, plus other data types and store them in a dedicated repository that has preservation, discovery, and protection attributes. These repositories are now commonly being used for retaining other information types for compliance, eDiscovery, and long-term preservation because they meet the criteria for a preservation repository and are more than just an ‘email archive.’

Typical misuses of the term archiving often include examples such as moving data or information from a Tier-1 or Tier-2 array into a Tier-3 tape system just to reduce cost or

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

power consumption (called “tiering”) or to compress and/or move information just to save storage space and improve disk performance (an activity called capacity optimization). These services and the associated stores do not meet the requirements of preservation. They do not define or utilize an archive.

The roots for the broad use of ‘archive’ are quite visible in the definitions used by other associations. However, several clear exceptions have emerged. PREMIS, OAIS, and its successor CASPAR take an approach more compatible with the datacenter by recognizing that an archive is a preservation store and by focusing on the services required to support preservation not just the repository.

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)
- Data stored off-line - An attribute in some file systems, typically used to indicate that a file has changed since it was backed up. (Source: Society of American Archivists)
- Archival Processing --The activities of accessing, arranging, describing, conducting access review, and properly storing documentary material. (Source: NARA-ERA Glossary)
- To conduct all activities related to caring for records of continuing value. (Source: ARMA)
- (Noun) 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)
- (Verb) The process of ingesting data into an Archive. (Source: SNIA Dictionary)
- Archival Storage: The OAIS entity that contains the services and functions used for the storage and retrieval of Archival Information Packages. (Source: OAIS)
- Open Archival Information System (OAIS). An OAIS is an archive, consisting of an organization of people and systems that have accepted the responsibility to preserve information and make it available for a Designated Community. (Source: OAIS)
- Preservation Repository: Repository that, either as its sole responsibility or as one of multiple responsibilities, undertakes the long-term preservation of the Digital Objects in its custody. (Source: PREMIS)

Audit Log (Audit Trail):

These two terms are often confused with each other. A ‘log’ is a record; a collection of events recorded in some index whether digital or analog and does not have to be all-inclusive. Multiple logs exist and are used to establish a historical record. An audit trail is a “*chronological record of system activities to enable the reconstruction and examination of the sequence of events and/or changes in an event.*”¹⁰

An audit log provides a location to record a history of accesses, changes, events, custodianship, and system activity. Logs take on many forms such as a database, index, metadata, or even hand

¹⁰ Source: National Information Assurance Glossary, 2006

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

written records and log-books. Their purpose centers principally on helping to ensure authenticity of digital objects over time whether in a preservation or evidentiary context.

The challenge in the datacenter is that creating and maintaining an audit log is not a native capability of most applications, file systems, and operating systems. Normal file system practices or external information, data, or storage services have little to no provision to maintain an audit log as part of the information object's metadata, because they can't write directly to the metadata. Consequently these events have to be tracked externally in a database, index, or even manually by process administrators and that association maintained over the lifecycle of the associated digital objects if they are to be effective.

Examples of use:

Audit logs are found in the datacenter in several locations and for multiple activities ranging from security logs of accesses, encryption key management, and data in transit, change, event, and activity logs, and even deletion logs. For example, a business policy that requires all email to be permanently deleted using a specific process after 180 days is meaningless as a defensible 'safe harbor' in a litigation event unless there is also a consistent audit log of that policy in operation. Logs that demonstrate that business policies are correctly and consistently adhered to are of great value in validating conformance to those policies and mitigating risk.

Audit logs, in a retention and preservation context, are usually applied and most useful at a point when the protected information objects are 'captive' to a specific repository and under the control of a specific entity where access can be controlled and logged and chain of custody established and tracked in an external database, such as in a preservation store or an evidence repository. An evidence repository is a special case of a repository where information identified by the eDiscovery process is copied and securely held as potential evidence for litigation. Repositories such as those developed for email archives and preservation stores are often designed to provide these services.

In the eDiscovery process, the audit log records a history of how the discovery process was done such as where information or data was found including location, content owner(s), and what media it was resident on, who the custodians are, and how the information was processed, including records like a digital timestamp and digital fingerprint. In a preservation repository, the audit log provides a chronological access and transaction history, including a history of integrity audits, migrations, transformations, control entities, and other activity records about the protected digital objects. The purpose in all cases is to help provide evidence of authenticity and a history of events.

Other associations have consistently defined an audit log or audit trail in one of these two capacities.

Reference definitions:

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- **Audit Trail:** A chronological record of activities that is sufficient to enable the reconstruction, review, and examination of the sequence of environments and activities. (Source: ARMA International)
- A record showing who has accessed an Information Technology (IT) system and what operations the user has performed during a given period. (Source: NIST)
- Who, what, when, where a record was accessed. (Source: NARA ERA)
- Information about transactions or other activities which have affected or changed entities (e.g. metadata elements), held in sufficient detail to allow the reconstruction of a previous activity. Note: an audit trail generally consists of one or more lists or a database which can be viewed in that form. The lists can be generated by a computer system (for computer system transactions) or manually (usually for manual activities); but the former are the focus of this specification. (Source: MOREQ)
- In computer security systems, a chronological record of when users logged in, how long they were engaged in various activities, what they were doing, and whether any actual or attempted security violations occurred. An audit trail is an automated or manual set of chronological records of system activities that may enable the reconstruction and examination of a sequence of events and/or changes in an event. (Source: Sedona Conference Glossary 12/07)
- Documentation of all the interactions with records within an electronic system in which any access to the system is recorded as it occurs. (Source: INTERPARES)
- **Audit Trail:** A chronological record of system activities to enable the reconstruction and examination of the sequence of events and/or changes in an event. (Source: National Information Assurance Glossary, 2006)

Authenticity:

Digital authenticity has become an important issue driven by the need for authentic, trustworthy evidence in litigation and the preservation of authentic digital records. InterPARES defines authenticity as being genuine and bona fide. The Federal Rules of Evidence defines it generally as “evidence that the matter in question is what its proponent claims.” The Sedona “Commentary on ESI” helpfully qualifies authenticity further by stating “First, and most important, the act of storing the information does not establish authenticity, the validity of the information depends on the process that placed it there.” Digital retention and preservation practices are in alignment.

In a digital retention and preservation context, authenticity describes 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 practices by themselves say nothing about what “original” means only that the object being tracked hasn’t changed without authorization since the point in time it came under control. Typically this occurs when a digital object is ingested into a preservation-class repository. Authenticity for preservation practices requires maintenance of digital ‘integrity’ (the consistency, accuracy, and correctness of stored or transmitted data or information) through preventing byte-level change or corruption, verification that it is the same object that was ingested, and an auditable history contained in its associated metadata and logs. Metadata is a key to verifying authenticity, as authenticity testing in a digital preservation context requires processes for periodic auditing for change, maintaining security,

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

maintaining a record of access and change through an audit log including a history of migrations and ownership/custodial tracking, and providing a means to detect change (not just prevent it) typically accomplished through reliable digital fingerprinting (hashing) and digital time-stamping methods. This means that the ability to write and update metadata and reference information and keep it associated with the original data is important, even though these records may not be contained within the same digital object.

Examples of use:

Controlling authenticity in the datacenter with today's IT capabilities is a practice requiring an ingestion process into a preservation-class repository. Ingestion into a controlled repository is typically a software-based process with capabilities and features that establish a suite of digital mechanisms to define the preservation object such as a digital fingerprint, digital timestamp, metadata records, unique digital object naming, and an audit log. What transpires before ingestion is generally uncontrolled from a digital preservation perspective.

Authenticity over the long-term gets complicated when migration begins to occur. For example, logical migration can include transformations to a different format. The problem is that hash codes are broken when the bits change because of a transformation. Digital objects are not necessarily inauthentic if only the physical integrity has changed. For example, converting a document from one coding format (EBCDIC or ASCII) to PDF-A, XML, or a tiff-image changes the hash value of the data but not necessarily the logical contents (or presentation) of the document.

To deal with this issue, in 2008 the preservation community introduced a new concept for tracking long-term authenticity called "Significant Properties." Significant properties are defined as "essential characteristics of a digital object (defined in advance) which must be preserved over time for the digital object to remain accessible and meaningful."¹¹ This creates a provision that allows authenticity to be validated even if data bits change during a logical migration (such as a transformation) event. Note, how the process of controlling and verifying authenticity requires the ability to write and update metadata and keep it associated with the original data, even though these records may not be contained within the same digital object.

Compliance requirements and the need for legally admissible digital evidence are changing the need for authenticity controls in the datacenter from just that being held in the litigation evidence process, to include all information that may ever have to be produced in a lawsuit or to prove compliance. Examples of good datacenter practices that support improved authenticity can be found in email archives, compliance stores, preservation repositories, and eDiscovery evidence repositories because of their comprehensive preservation services.

Reference definitions:

¹¹ Source: Digital Preservation Coalition, JISC/BL/DPC workshop, April 2008

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- **Authenticity:** Property that a Digital Object is what it purports to be. (Source: PREMIS)
- The quality of being genuine, not a counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context. (Source: Society of American Archivists)
- The property of a documentary material that it is what it purports to be and has not been corrupted. (Source: NARA-ERA)
- The trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that is free from tampering or corruption. [Context: Archives] (Source: InterPARES2)
- The quality of being genuine. (Source: MoReq Glossary)
- The quality of being authentic, or entitled to acceptance. The term authentic means ‘worthy of acceptance or belief as conforming to or based on fact’ and is synonymous with the terms genuine and bona fide. Genuine implies actual character not counterfeited, imitated, or adulterated and connotes definite origin from a source. Bona fide implies good faith and sincerity of intention. From these definitions, it follows that an authentic record is a record that is what it purports to be and is free from tampering or corruption. (Source: InterPARES – Authenticity Task Force Report 2001)
- **Authenticate:** To verify the identity of a user, user device, or other entity, or the integrity of data stored, transmitted, or otherwise exposed to unauthorized modification in an information system, or to establish the validity of a transmission. (Source National Information Assurance Glossary)
- **Fixity Information:** The information which documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is a Cyclical Redundancy Check (CRC) code for a file. (Source: OAIS)
- The process of verifying that a record is what it purports to be. (Source: ARMA)
- Synonym for data integrity. (Source: SNIA Dictionary)
- **Significant Properties:** essential characteristics of a digital object which must be preserved over time for the digital object to remain accessible and meaningful. (Source: CASPAR)

Classification (Classify or Categorize):

The objective of classification is to collate an organization’s information and data assets into categories, collections, or classes so that the process of managing these digital objects can be simplified by operating on groups rather than on individual objects. Classification methods may be based on various taxonomies and ontologies, but the methods chosen need to be appropriate for the organization’s requirements.

While similar in practice, information classification is not to be confused with security classification. The two have identical purposes, but a security classification is a subset of an information class. Security classes allow the application of business requirements for security to collections of information just like, and in concert with, other business requirements for information. However, since a full set of requirements are needed to properly manage information, security requirements are just another business requirement.

With regards to retention and preservation, all classes of information or data should be associated with a retention policy and a disposition policy. Consequently, classification is essential to implementing a comprehensive and successful retention and preservation program.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of use:

Within information-lifecycle management, ILM, methods the chief purpose of a classification scheme is to associate service level objectives with groups of information and data based on the requirements of the business. Many companies try to reduce the number of classes they have to manage to no more than five-to-ten. The use of more than this introduces complexities that existing systems do not support well. It is expected that as improvements in ILM tools occur plus the adoption of more standard metadata definitions and logical containers at an application level the ability to implement more classification granularity will happen. SNIA is working on two projects that are focused on supporting metadata for ILM and preservation purposes through an enhanced logical container, eXtensible Access Method, XAM and the Self-Contained Information Retention Format, SIRF.

Reference definitions:

- Classify (verb) – The systematic identification and arrangement of business activities and/or records into categories according to logically structured conventions, methods, and procedural rules represented in a classification scheme. (Source: ISO 15489)
- Classification scheme: often represented as a hierarchy. (Source: MOREQ)
- Data classification [context: Data Management] - An organization of data into groups for management purposes. A frequent purpose of a classification scheme is to associate service level objectives with groups of data based on their value to the business. (Source: SNIA)
- Classification - organizing information and data into groups for management purposes based on some taxonomy or ontology (Source: SNIA)
- Taxonomy: The science of categorization, or classification, of things based on a predetermined system. In reference to Web sites and portals, a site's taxonomy is the way it organizes its ESI into categories and subcategories, sometimes displayed in a site map. Used in information retrieval to find documents that are related to a query by identifying other documents in the same category. (Source: Sedona Conference)
- 1. The organization of materials into categories according to a scheme that identifies, distinguishes, and relates the categories. – 2. The process of assigning materials a code or heading indicating a category to which it belongs. – 3. The process of assigning restrictions to materials, limiting access to specific individuals, especially for purposes of national security; security classification.
 - Notes: Classification may involve physically arranging the materials or use of a class code to index and retrieve documents stored in a different order. For electronic documents, classification may involve assigning a class code used to index and retrieve the document. In some schemes, documents may be assigned to more than one class. (Source: SAA)
- Classification Scheme: (also classification plan), n. ~ A diagram or chart that describes standard categories used to organize materials with similar characteristics.
 - Notes: Classification schemes are often hierarchical in nature and frequently associating codes with each class. Typically used in an office of origin to file active records or in archives as a finding aid. Libraries commonly use either the Library of Congress Classification System or the Dewey Decimal Classification to organize their books. These bibliographic standards have only limited use in archives, which maintain the records in their original order. (Source: SAA)
- Classifying: The act of analyzing and determining the subject content of a document and then selecting the subject category under which it will be filed. (Source: ARMA)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Data (Digital):

Digital data is on one hand “everything digital” and on the other just the content, the payload, in an information object. How you use the term depends on the frame of reference. From a retention and preservation perspective, it is important to recognize and distinguish between data and metadata. The objective in doing this is to elevate the role and visibility of metadata in the datacenter. The ‘old school’ view that IT and IT practices only operate on data in ignorance of the metadata has to change to meet today’s litigation, compliance, and business requirements. So, in this context, what is data? Years ago the archival and preservation communities (such as OAIS and PREMIS) took a systems approach to the question and defined information as an object made up of data plus its metadata where data is the content and metadata is the system and descriptive representation information about the data that provides context and relevance.

What this means, from a storage system’s perspective, is that digital data can be described as a logical collection of associated digital blocks forming readable content representative of specific values or conditions such as bit-mapped images, text, cells in a database, or sound. By recognizing data as a block object (made up of associated blocks) and information as a data object (made up of associated data and metadata) the supporting data and information services can discriminate between the two and recognize, protect, and secure the appropriate metadata when it exists. [See also: “Information,” “Information Objects,” and “Metadata”]

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of use:

The objective of looking at data and information objects in this fashion is to assure that metadata is discriminated and preserved. IT and IT infrastructure practices are not use to thinking about data and supporting services this way. Consequently, all that is intended is to be clear about the requirements and propose an approach to solve the problem. The good news is that the archival community figured this out years ago, and successfully designed and implemented preservation systems around this definition¹².

Reference definitions:

- Data: Examples of data include a sequence of bits, a table of numbers, the characters on a page, or the recording of sounds made by a person speaking. (Source: OAIS)
- Digital (Data) Object: An object composed of a set of bit sequences. (Source: OAIS)
- Facts, ideas, or discrete pieces of information, especially when in the form originally collected and unanalyzed. Traditionally a plural noun, data – rather than datum – is now commonly used with a singular verb. Data often is used to refer to information in its most atomized form, as numbers or facts that have not been synthesized or interpreted, such as the initial readings from a gauge or obtained from a survey. In this sense, data is used as the basis of information, the latter distinguished by recognized patterns or meaning in the data. The phrase 'raw data' may be used to distinguish the original data from subsequently 'refined data.' Data is independent of any medium in which it is captured. Data is intangible until it has been recorded in some medium. Even when captured in a document or other form, the content is distinct from the carrier. (Source: SAA)
- Groups of characters that represent a specific value or condition. Data provide the building blocks of information. (Source: ARMA)
- The digital representation of anything in any form. (Source: SNIA Dictionary)
- Reference Information: Information that identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the content information. It also provides identifiers that allow outside systems to refer to particular content information. An example is an ISBN. (Source: Caspar Glossary of Terms, 2008)
- Representation Information: The information that maps a data object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e. a data object) is mapped into a symbol. (Source: Caspar Glossary of Terms, 2008)

Data Deduplication:

Capacity optimization is the umbrella-term for methods used to increase the efficiency of storing or transmitting digital objects. Typical methods include data deduplication, sub-file data deduplication, single instance storage, and compression.

- Data deduplication is the replacement of multiple copies of data with references to a shared copy in order to save storage space and/or bandwidth. The granularity of data deduplication varies based on the specifics of the technologies and processes used.

¹² For examples of this, refer to the OAIS standard ISO 14721 and the work of the CASPAR Preserves on preservation stores.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- Sub-file data deduplication is a form of data deduplication that operates at finer granularity than an entire file. Examples of subfile objects include data sets, data objects, or even the I/O stream.
- Single-Instance-Storage operates at a file or information object level and replaces duplicate objects with references to a shared copy.
- Compression is complementary to deduplication because deduplicated objects can still be compressed. Rather than make comparisons across objects, compression is a computational method of encoding duplicated bits within digital object such as a file. Compression reduces the number of binary digits required to represent the object.

In all data deduplication processes, the definition of "what is a duplicate" is predicated upon the method used to evaluate, identify, track and avoid duplication. After a computational analysis (usually some form of hashing) and comparison, the data deduplication process includes updating tracking information, storing and/or sending data or metadata that is new and unique, and disregarding any data or metadata that is a duplicate. The principle is that data deduplication methods do not eliminate information or data that needs to be stored nor do they change the ability to subsequently access a copy or a version. Data deduplication eliminates the redundancy in the data at various levels of granularity depending on the method chosen.

Examples of use:

Data deduplication methods may operate at many places such as in the I/O stream, in the application, or within the storage repository and each approach has specific benefits and tradeoffs. An example is data deduplication of the backup stream. Traditional backup processes have a lot of redundancy in the backup data because they use repetitive full, differential and incremental backup cycles. For example, each full backup is a complete copy of everything and is redundant to the previous full backup with the exception of the small percentage of information that has changed between backup periods. Backup data deduplication processes are commonly achieving greater than 20 times data reduction over time.

Capacity optimization methods can meaningfully be compared only under the same set of working conditions. The amount of duplicate data or information objects in a specific environment is determined by the characteristics and access patterns of the data and by the operational policies and practices in use. The space savings actually achieved depends on which data is deduplicated and on the effectiveness and efficiency of the specific technologies used to perform deduplication. As in any storage process, capacity optimization methods are expected to maintain the integrity of the data and metadata.

Reference definitions:

- Deduplication ("De-Duping") is the process of comparing electronic records based on their characteristics and removing or marking duplicate records within the data set. The definition of "duplicate records" should be agreed upon, i.e., whether an exact copy from a different location

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

(such as a different mailbox, server tapes, etc.) is considered to be a duplicate. De-duplication can be selective, depending on the agreed-upon criteria. (Source: Sedona)

- Data deduplication - [context: Capacity Optimization] The replacement of multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth. (Source: SNIA Dictionary)
- Single instance storage - [context: Capacity Optimization] A form of data deduplication that operates at a granularity of an entire file or data object. See data deduplication, subfile data deduplication. (Source: SNIA Dictionary)
- Subfile data deduplication - [context: Capacity Optimization] A form of data deduplication that operates at a finer granularity than an entire file or data object. See data deduplication, single instance storage. (Source: SNIA Dictionary)

Deletion:

Once information has reached the end of its retention period, it is a candidate to be deleted in accordance with the organization's disposition policies. These policies generally state that deletion may only occur if there are no litigation holds or deletion suspension holds in place and if the retention period has been reached. One more important constraint has to be considered. Deletion must be a standard process performed in accordance with the organization's documented policies and all the files/records/data being deleted must be logged, so that adherence to the organization's operating policies can be demonstrated. Failure to follow the policies or introduce new practices for a specific set of information will potentially cause suspicion that something is being hidden. The point is that while retaining information past its expiration date puts organizations at risk from litigation, deleting it inappropriately or differently than normal can be viewed as spoliation.

An important potential challenge to some of these retention-oriented deletion rules exists in dealing with legacy data and information that was never incorporated into a retention control process. This pool of information needs to be reviewed and dealt with rather than just letting it sit.

Deletion practices have two major forms, file system deletion and permanent deletion. When only the term deletion is used, assume it means file system-level deletion. This type of deletion uses normal operating system or file-system methods to remove data, indexes, pointers, and metadata associated with information or data objects. The actual data and metadata are not permanently removed, but are simply lost to the file system. Good forensic analysis techniques can easily recover information deleted by a file system. File-system deletion (sometimes called 'soft-deletion') is convenient, but risky as the information and associated metadata is easily recoverable. If information presents risk to the organization if it is recoverable or if it leaks out of the organization's control, then permanent deletion methods should be used. [See also: "Permanent Deletion" and "Disposition Instructions"]

Examples of use:

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

When information or data has 'expired' based on its retention requirements a typical disposition policy may be to delete all copies of the object(s) from all locations within the storage systems using a method appropriate for that class of information. (The challenge of finding and identifying all copies, versions, backups, and snapshots in all possible locations is not a trivial matter.) These policies must be authorized by legal and the line of business.

Deletion must be done appropriately including fulfilling the last step which is to create a record that the deletion event occurred. This record is a log including what was deleted and how it was deleted. The deletion log can be important in the case of proving conformance to company policies in the case of litigation.

Reference definitions:

- Deletion does not necessarily make data unreadable. The original information may remain intact and, if not overwritten, could be recovered using special software tools. (Source: Society of American Archivists)
- Deleted data may remain on storage media in whole or in part until they are overwritten or "wiped." Even after the data have been wiped, directory entries, pointers or other information relating to the deleted data may remain on the computer. "Soft deletions" are data marked as deleted (and not generally available to the end-user after such marking), but not yet physically removed or overwritten. Soft-deleted data can be restored with complete fidelity. (Source: Sedona)
- Destruction - The definitive obliteration of a record beyond any possible reconstitution. [See also disposition.] (Source ARMA)

Digital Fingerprinting

The concept of a digital fingerprint is analogous to a digitally computed numerical identifier based on the 'hash' of the information or data object (or a subset of the object, file, or data.) The principle is that by computing a hash-value of the contents of an object, using a rigorous mathematical algorithm such as MD5 or SHA, the generated numerical value (the hash) is unique. What makes a good hash algorithm is that there is a very low probability that two different objects will produce the same hash-value. (How low this probability needs to be is up to the organization to decide.) This 'fingerprint' is a mathematically unique identifier. Fingerprinting is basically the same mathematical process used to compute encryption 'hash-values', but in this case, decryption is not of concern so there are no keys to manage or lose.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of use:

Digital fingerprints (hash-values) have three principal uses in retention and preservation.

- They are used to deduplicate information and data to reduce storage requirements since two objects that have the same fingerprint can be concluded to be the same.
- Another use for fingerprints and hash-values is to aid in assuring integrity and authenticity by verifying that the content of the object has not changed over time. Not prevent, but verify. If the original hash-value is stored and protected (in the metadata or an index – Note, this metadata is called ‘fixity’ by the preservation community.) it can be tested over time by periodically re-hashing the object and comparing the numerical values, thereby verifying that no change has occurred, provided the hash-value is held securely. This auditable record provides evidence of authenticity over time and when changes are detected, can be used to flag those changes and initiate a recovery. Fingerprinting is strengthened as a tool for authenticity testing when combined with a digital timestamp.
- Digital fingerprints also provide a verifiably unique name for digital objects useful for addressing (or locating) objects in preservation systems and providing location independence.

Reference definitions:

- Hash value: A unique numerical identifier that can be assigned to a file, a group of files, or a portion of a file, based on a standard mathematical algorithm applied to the characteristics of the data set. The most commonly used algorithms, known as MD5 and SHA, will generate numerical values so distinctive that the chance that any two data sets will have the same hash value, no matter how similar they appear, is less than one in one billion. “Hashing” is used to guarantee the authenticity of an original data set and can be used as a digital equivalent of the Bates stamp used in paper document production... (Source: Federal Rules of Civil Procedure)
- The process of using a mathematical algorithm against data to produce a numeric value that is representative of that data. .. (Source: NIST)
- Hash: A mathematical algorithm that represents a unique value for a given set of data, similar to a digital fingerprint. Common hash algorithms include MD5 and SHA. Hash Coding: To create a digital fingerprint that represents the binary content of a file unique to every electronically-generated document; assists in subsequently ensuring that data has not been modified. See also Data Verification, Digital Fingerprint and File Level Binary Comparison. Data Verification: Assessment of data to ensure it has not been modified. The most common method of verification is hash coding by some method such as MD5. See also Digital Fingerprint and File Level Binary Comparison and Hash Coding. (Source: Sedona Conference 12/07)
- Fixity Information: The information which documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is a Cyclical Redundancy Check (CRC) code for a file. (Source: OAIS)
- Hash-value: A value deterministically derived from data and assumed to be unique enough within the domain of that data for the purposes of its application. (Source: SNIA Dictionary, 2008)

Disposition Policy:

Disposition is a records management term defining a business policy that specifies the process

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

that happens when data or information's retention requirement is met. Ideally, this policy would be defined upon the creation of any information based on its classification as part of the overall business requirement set. That way, when the retention period is over, the disposition can be automatically acted upon. Typical examples of the options called for in a disposition policy are to delete the information using a pre-described method, find all copies and versions and delete all of them (providing no litigation or deletion holds are in effect), or call for a review. Disposition policies that call for a record to be placed in a long-term preservation store upon expiration are too late. That instruction should have been initiated upon creation, not expiration. [See also "Information State" and "Retention"]

Examples of use:

In an ILM context, disposition instructions (policies such as 'permanently delete by method 'XYZ', 10 days after expiration') need to be established concurrent with retention policies so that the system can manage information throughout its lifecycle. A barrier to this is that legal departments generally remain reluctant to implement automated information deletion processes. Business, IT, legal, finance, and records managers must collaborate to establish these policies in advance, so that automated deletion can become a new standard in the datacenter. At hand is the core issue "Is it safe to delete information that has expired?" The irony is that the inverse policy, the 'do nothing, save everything strategy,' is far more costly and risky than setting in place proper practices. In the end, this is a question for an organization's general counsel, but more and more legal precedent is being set that says if you have litigation hold, retention, deletion, and 'deletion suspension' practices in place, proven, and they are your normal (and reasonable) practices, your organization is protected from adverse judgements.

Continuing to retain expired information or data presents a legal liability to the organization as well and contributes to additional operating overhead costs. However, keep in mind that deletion policies and practices are dependent upon freedom from any relevant litigation holds.

Reference definitions:

- Those actions taken regarding records no longer needed for the conduct the regular current business of the creator. The instructions contained in a disposition agreement that mandate what is to be done with documentary material at certain points in their lifecycle. Disposition Instructions may consist of: specification of the length of time material should be retained by their creator or custodian (a retention period), conditions under which the creator or custodian should terminate retention, physical or legal transfer of material to another custodian, destruction of records, or stipulation that the material is not to be destroyed. (Source: NARA-ERA)
- A final administrative action taken with regard to records, including destruction, transfer to another entity, or permanent preservation. (Source: ARMA)
- The final business action carried out on a record. This action generally is to destroy or archive the record. Electronic record disposition can include "soft deletions" (see Deletion), "hard deletions," "hard deletions with overwrites," "archive to long-term store," "forward to organization," and "copy to another media or format and delete (hard or soft)." (Source: Sedona Principles)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- The authoritative instructions governing the process of determining the transfer and destruction of records. (Source: InterPARES2 Glossary)

Electronically Stored Information, ESI:

ESI is a term adopted by the legal community used to distinguish digital and electronically stored information and data from 'paper' documents and physical objects. ESI includes analog and digital data or information captured, stored, or retained in or on an 'electronic' based medium.

Electronic media¹³ is defined "as relating to technology having electrical, digital, magnetic, wireless, optical, electromagnetic, or similar capabilities." The Federal Rules of Civil Procedures Rule 34 gives some examples of ESI when it states, (ESI) "...includes writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations stored in any medium from which information can be obtained."

In the datacenter, ESI can include all information and data stored or in transit in any form on any media, network, or device. It can include metadata, all states of data or information, and all types of information regardless where it resides. Discovery of ESI potentially includes transitory or mobile data that may be in instant messaging, voice mail, mobile devices, or in transit. It is also expected that the definition of what is ESI will evolve with technology and the global distribution of information. [See also: "Information" and "Information Objects"]

Examples of use:

Most litigation events now involve the discovery of ESI and in parallel the preserving of ESI so that it can be evaluated and potentially used as evidence. Becoming discovery ready is complex. It involves putting in place litigation hold processes, establishing retention and deletion policies and implementing those practices carefully. It also requires the preservation of information, in many cases from creation, appropriately. Even deletion suspension processes are necessary steps in protecting evidence from spoliation (damage or change) and avoiding adverse consequences in court such as fines or other adverse findings. Email is the most common ESI requested in litigation and consequently the top focus of discovery efforts. Metadata attributes can also become discoverable and should be managed and protected just like data, in accordance with the organization's retention and disposition policies.

¹³ Source: The National Conference of Commissioners on Uniform State Laws, August, 2007

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Reference definitions:

- Electronic Record -- Information recorded in a form that requires a computer or other machine to process it and that otherwise satisfies the definition of a record. (Source: DoD 5015)
- ...including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations stored in any medium from which information can be obtained. (Source: Federal Rules of Civil Procedure, Rule 34)

Emulation (System and Software Emulation):

One of the core problems in long-term preservation is maintaining the ability to read and interpret the digital objects being preserved because the original computer hardware, operating system, application and even the expertise to operate the system may be no longer available. A technique for addressing part of this problem is called emulation. Emulation is a method used to maintain access to and readability of digital information objects by mimicking or recreating the original hardware and software environment by using contemporary computer and software technologies. Its purpose is to provide the ability to access and/or view information in its original context long after the original hardware and software environments are no longer available. The hope is that emulation will reduce the amount or number of logical migrations required to sustain the ability to read and interpret information that is being preserved. [See also “Encapsulation” and “Migration”]

Examples of use:

“Emulation uses software to recreate the original digital operating environment to enable the original performance of the software to be recreated on current computer systems. The result is that the original data format is preserved and may be accessed in an environment that allows for the recreation of the original look and feel of the information.” (From the AU National Archives “Approach to Digital Preservation”) This approach avoids the need for continual conversion, but emulators must be developed for every software/hardware environment and configuration, which makes it far more expensive over time than migration.

The tradeoffs in using emulation include:

- It reduces the preservation problem to one of preserving the emulation platform
- The cost is proportional to the number of formats (not information objects) that need to be preserved
- Emulation has an up-front investment and then periodic costs to verify the operation of the emulation systems.
- People skills to operate obsolete systems have to be maintained. Unfortunately, emulation does not remove this requirement.
- It is effective only for information that is coupled to specific applications.
- It does not allow for new interpretations of the information that may be introduced in the future.

Reference definitions:

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- The use of one system to reproduce the functions and results of another system. (Source: Society of American Archivists)
- A proposed approach to the problem of software and hardware dependence is emulation, which aims to preserve the original software environment in which records were created. Emulation mimics the functionality of older software (generally operating systems) and hardware. This technique seeks to recreate a digital document's original functionality, look, and feel by reproducing, on current computer systems, the behavior of the older system on which the document was created. (Source: GAO 2002)
- Preservation strategy for overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers. (Source: PREMIS Working Group)
- A strategy for continuing access to digital materials that mimics or re-creates the digital object's original technical environment using current technology. Access to the object relies on a copy of the original byte stream (as deposited) and an emulation of its original operating environment. Emulation can take place at either the hardware or software level. (Source: RLG-OCLC)
- Emulation can be considered to be the development of software/hardware combinations to replicate the behavior of obsolete processes. These hardware processes may include systems such as interfaces, operating systems or hardware configurations. By replicating the behavior of obsolete processes this enables digital material stored using these obsolete processes to be accessed by modern systems. (Source: Midess Project)
- Emulate: in hardware terms, the creation of software for a computer that reproduces in all essential characteristics (as defined by the problem to be solved) the performance of another computer of a development environment while the newer computer is still being fabricated. In software preservation terms, the creation of software that analyzes the software environment of a document such that it can provide a user interface to the document that substantially reproduces the essential characteristics of the document as it was created by its originating software. (Source: Universal Preservation Format Glossary)

Encapsulation (Information Encapsulation):

Another method for solving the long-term preservation issue of reading and interpreting digital objects long after the original operating systems and applications are no longer available is called encapsulation. It is the process of grouping information into an information object (data and metadata) along with instructions on how its contents should be accessed and interpreted to enable future interpretation. Encapsulation may be used to combine data (content), associated metadata, and a viewer to render the combination into a single object. [See also: "Information Object," "Emulation," and "Migration"]

Examples of use:

As an alternative to migrating information from obsolete formats to a new logical format, encapsulation retains information in its original form by putting it into a logical container with a set of instructions on how the original should be interpreted. This approach needs a detailed formal description of the encapsulation file format and what the information means using a method such as XML. The idea, in general, is that using simpler, standard formats

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

makes encapsulation easier. Other approaches to encapsulation take the approach of including software readers in with the object or placing links to external software readers.

The positive benefits of encapsulation include:

- It is the most flexible approach and can be combined with emulation and migration.
- It allows new interpretations to be made.
- It is OAIS compliant.

The shortcomings of encapsulation include these:

- The interpretation information in encapsulation is very high level and may not be adequate in the future.
- Collecting all this 'representation information' and placing it in each object is potentially very hard and costly.

Reference definitions:

- A technique of grouping together a digital object and anything else necessary to provide access to that object. (Source: NIST)
- The combination of several digital objects into a single unit that preserves the association of those objects. Encapsulation may be used, for example, to combine an electronic record, metadata, and a viewer to render the record. (Source: Society of American Archivists)
- A way to group together all the relevant material for the digital object and to manage the resulting digital object as one. (Source: Midess)

Expired Information or Data:

(See [Information-State](#))

Fixity:

A term used by the preservation community (as in the ISO standard, OAIS) for 'hashing' with an algorithm to compute a hash value to be used with authenticity verification testing. It is one use for a digital fingerprint. [See also: [Digital Fingerprinting](#)]

Reference definitions:

- The process of using a mathematical algorithm against data to produce a numeric value that is representative of that data. .. (Source: NIST)
- Hash: A mathematical algorithm that represents a unique value for a given set of data, similar to a digital fingerprint. Common hash algorithms include MD5 and SHA. Hash Coding: To create a digital fingerprint that represents the binary content of a file unique to every electronically-generated document; assists in subsequently ensuring that data has not been modified. See also Data Verification, Digital Fingerprint and File Level Binary Comparison. Data Verification: Assessment of data to ensure it has not been modified. The most common method of verification is hash coding by some method such as MD5. See also Digital Fingerprint and File Level Binary Comparison and Hash Coding. (Source: Sedona Conference 12/07)
- Fixity Information: The information which documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

undocumented manner. An example is a Cyclical Redundancy Check (CRC) code for a file. (Source: OAIS)

Inactive Information or Data:

(See [Information-State](#))

Information (Digital):

As discussed under the term ‘data’, the objective in revising the definitions of digital information and data in the datacenter is to assure that information and data services properly and efficiently support the organization’s requirements for information and empower the use of information lifecycle management, ILM, based practices to cope with the volume and scale of the information-management problem. An important key for legal, retention, and preservation purposes is to recognize when metadata is present and if it is, handling it with the same care given to the 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 community¹⁴ 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. [See also “Data,” “Information Object,” and “Metadata”]

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 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

¹⁴ The preservation community consists of “archivists” and those engaged in the provision of long-term preservation services as typically found in the digital library, historical, cultural, and governmental preservation arenas.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

object-oriented concepts which try to hide these implementation issues. This presents a significant challenge to the preservation of digital information.”

Reference definitions:

- **Information:** Any type of knowledge that can be exchanged. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees Celsius (the representation information). (Source: OAIS)
- Data that has been given value through analysis, interpretation, or compilation in a meaningful form. (Source ARMA)
- 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 is data that is interpreted within a context such as an application or a process. (Source: SNIA Dictionary, 2008)
- **Reference Information:** Information that identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the content information. It also provides identifiers that allow outside systems to refer to particular content information. An example is an ISBN. (Source: Caspar Glossary of Terms, 2008)
- **Representation Information:** The information that maps a data object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e. a data object) is mapped into a symbol. (Source: Caspar Glossary of Terms, 2008)

Information Object:

An object is a logical concept used to connote an associated grouping of content, properties, and methods of interpretation or operations. A digital information object is any discrete collection of information and data, along with all the metadata and reference information required to represent the collection as a single conceptual entity including a unique identifier. The content of the object may include groupings of digital information, files, content (data), metadata, properties, methods, and reference information. Digital information is an object by definition as it is a logical container such as a file made up of discrete components; data, metadata, and reference information. Information objects may also contain links to other pieces of information or data. It may be that combinations of information hold particular relevance and those relationships need to be maintained in a single object. It is even possible for several information objects to share common content. [See “Information” and “Encapsulation”]

Examples of use:

The concept of information objects is useful in both conceptualizing the contents of digital information programmatically and in establishing a model for the contents of the container that constructs or bounds the object. For example, when multiple files are hyper-linked, file systems have no mechanism for maintaining those links external to each file. Move one file and the links are broken. But, when those files are combined into a single information object, the

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

links are maintained because they are bound together. Think of an information object as a container to hold related or associated content and metadata. Some other examples include a set of productivity applications that cross-link files and metadata or a set of linked html files as in a web site. All produce information objects that if separated, are no longer useful. Keeping them together or at least properly associated is important.

Reference definitions:

- 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.
 - Data Object: Either a Physical Object or a Digital Object.
 - Digital Object: An object composed of a set of bit sequences. (Source: OAIS)
- Digital Object: A unit of information that includes properties (attributes or characteristics of the object) and may also include methods (means of performing operations on the object). The concept of digital object comes from object-oriented programming. Objects typically include properties and methods. Objects may belong to classes and inherit properties and methods from a parent class. Similarly, an object may have child objects that inherit its properties and methods. Digital objects are an abstraction that can refer to any type of information. The object may be simple or complex, ranging from values used in databases to graphics and sounds. An object called name may include properties such as title, first name, and last name, as well as methods for returning the value of the name in natural language or inverted order. An object called graphic may include properties that define an image, such as dimensions, color scale, and encoding scheme and might include methods that make that image data available at different resolutions. Objects are not necessarily self-contained. For example, a graphics object may require an external piece of software to render the image. In addition to the data that makes up the fundamental content, the object often includes metadata that describes the resource in a manner that supports administration, access, or preservation. (Source: Society of American Archivists)
- Digital Object: Discrete unit of information in digital form. A Digital Object can be a Representation, File, Bitstream, or Filestream. Note that the PREMIS definition of Digital Object differs from the definition commonly used in the digital library community, which holds a digital object to be a combination of identifier, metadata, and data. (Source: PREMIS)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Information (or Data) State:

State is an abstraction useful for categorizing how digital objects align with business processes. Information or data 'state' refers to a way to categorize and describe how information or data is being used and accessed such as, is it changing, does it need to be 'readily available', does it exist online merely for reference and periodic access, or has it reached the end of its lifecycle and is now 'expired?' As a descriptor for the access and usage mode of digital objects, information state corresponds to lifecycle independent of other requirements such as retention period. A digital object's state can be described by one of four usage modes: active, inactive, reference, or expired and these states may change back and forth over time as information objects can be reused or reactivated.

Business process state and operational state describe two additional state models useful to describe how information objects and data or groups of information objects align with supporting services and business requirements: examples are "pending (some action)," "in transit," "litigation hold," or "deletion hold."

Examples of use:

ACTIVE: when information or data is created it generally begins its lifecycle in an 'active state', meaning it is "currently in use and subject to change." Information remains "active" as long as there is regular reference being made to it, it is available for change, and it has a storage policy that includes a requirement for it to be "immediately" available.

INACTIVE: Once "active" information or data falls out of current, in-process, production, or regular use, it tends to no longer be modified and is accessed infrequently. The change in usage state typically moves from 'active' to 'inactive' or 'reference' as the information no longer needs to be updated. Retention periods for inactive information vary depending on the classification of the information or data not based on their inactivity as these are independent variables. Inactivity can also be used as a qualifier for ILM-based practices such as tiering since inactive information no longer needs to be "readily accessible."

REFERENCE: Reference information or data is defined as information or data that will not be modified for the remainder of its lifetime, but needs to be readily accessible. (In storage terms this is called "read-only.") An example is a document or presentation turned into a 'pdf' or image format, secured so that it cannot be changed, and placed online or in a reference library. Other examples may include information such as training materials, marketing materials, image libraries, and similar content. Reference information is often kept online for ease of access and 'reference' by users. The term "fixed content" is analogous to "reference information" and is sometimes used interchangeably.

EXPIRED: This state describes information or data that are no longer required to be retained for any reason and are candidates for deletion. Examples of when information may become expired are when it has reached its defined retention period or when an event makes it

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

obsolete and it has no further value to the organization. How much expired information remains in the pool of digital assets in a datacenter is a measure of how well its retention and deletion policies are operating.

The value of information state is that it provides a classification schema corresponding with information lifecycle to which global business or governance policies can be applied. Classifying information or data into categories based on their usage state is very helpful in efficiently providing the right level of services, at the right cost, for each class of information. For example, in tiering practices, policies may exist that ‘active information’ or ‘active data’ only reside on Tier-1 or Tier-2 systems in order to meet application performance requirements or ‘reference information’ may exist on Tier-2 or “inactive information” only on Tier-3 storage depending on the application or business requirements. Another policy may be to permanently delete ‘expired information’ 10 days after expiration using established protocols.

The legal, records and information management (RIM), and archivist communities use similar concepts with the term ‘active records.’ The key differences in the DMF’s use of these terms are twofold. First, all data and information in the datacenter must be included because all are subject to discovery in litigation and all must be managed, not just specific “records” of business or organizational value. Second, DMF recognizes that state and retention period are independent variables, not to be confused. The reason retention period has nothing to do with state can be illustrated by several examples. If a digital object has a long-term retention requirement or a short-term requirement, those requirements have nothing to do with whether or not the object is active or reference. The retention period is chosen based on the business requirements for the object, not its state. Similarly, if an object is classified as ‘disposable information’, then the point in time at which its retention period expires is irrespective of its state.

Reference definitions:

- State: A record’s usage or access modes, described as transitory, active, semi-active, inactive, unscheduled (in a Federal environment), archival or permanent. (Source ARMA)
- Active Records: Records that continue to be used with sufficient frequency to justify keeping them in the office of creation; current records. [Context Computing] Information stored on computer systems that can be readily accessed by the operating system or software without a need to reload media, undelete the information, or reconstruct it from other sources. (Source: Society of American Archivists)
- Active Data is information residing on the direct access storage media (disk drives or servers) of computer systems, which is readily visible to the operating system and/or application software with which it was created and immediately accessible to users without restoration or reconstruction. (Source: Sedona Principles)
- Active Records: Active Records are those records related to current, ongoing or in-process activities and are referred to on a regular basis to respond to day-to-day operational requirements. An active record resides in native application format and is accessible for purposes

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

of business processing with no restrictions on alteration beyond normal business rules. See Inactive Records. (Source: Sedona Principles)

- Active Record - A record needed to perform current operations, subject to frequent use, and usually located near the user. See also “semi-active record” and “inactive record.” (Source: ARMA)
- Active Record: Information stored on computer systems that can be readily accessed by the operating system or software without a need to reload media, undelete the information, or reconstruct the information from other sources. (Source: InterPARES2)
- Active Data: Data that is immediately accessible to an application without the need to stage it in from a lower tier of storage. (Source: SNIA Dictionary)
- Inactive records are no longer used in the day-to-day course of business, but which may be preserved and occasionally used for legal, historical, or operational purposes. Inactive records are often stored out of the office of creation in a records center or on offline media. They may either be destroyed when their frequency of use falls so low that they have lost all value or they may be transferred to an archival repository for permanent retention. (Source: Society of American Archivists)
- Inactive records are those records related to closed, completed, or concluded activities. Inactive records are no longer routinely referenced, but must be retained in order to fulfill reporting requirements or for purposes of audit or analysis. Inactive records generally reside in a long-term. (Source: Sedona)
- Inactive record: A record that is no longer needed to conduct current business but preserved until it meets the end of its retention period. (Source: ARMA)
- Reference: A copy of a record kept for easy access to the information it contains, as opposed to its intrinsic or evidential value. (Source: Society of American Archivists)
- Reference File/Copy: A copy of a record used primarily for consultative purposes. (Source: ARMA)
- Reference: Content that does not change, fixed content. (Source: SNIA Dictionary)
- Temporary Record: A record approved by the appropriate authority for disposal, either immediately or after a specified retention period. (Source: NARA-ERA)
- Disposal: The transfer of records, especially noncurrent records, to their final state, either destruction or transfer to an archives. (Source: Society of American Archivists)
- Non-current records: Non-current records are often stored out of the office of creation in a records center or on offline media. Either they may be destroyed when their frequency of use falls so low that they have lost all value or they may be transferred to an archival repository for permanent retention. (Source: Society of American Archivists)

Ingestion:

Ingestion was originally an archivist’s term that is relevant in the datacenter as preservation, discovery, and authenticity practices become more important. It is a process of preparing, indexing, and adding information or data into a managed domain within an information repository or preservation repository. Ingestion may consist of processes such as hashing and time-stamping to get baselines for authenticity verification established and defining a unique naming convention, updating metadata fields, deduplication, indexing content, retention, and classification information, transcoding into standard formats, creating containers for the information and its metadata, encryption, etc.

Examples of use:

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Ingestion is another set of services operating to assure preservation, discoverability, efficiency, authenticity, integrity, availability, protection, and confidentiality in a preservation repository. When information or data is discovered in an eDiscovery process, one method is for a copy to be made that is ingested into a separate preservation repository where it can be preserved and verified authentic.

Reference definitions:

- The process of moving records into the Electronic Records Archive, ERA, system. (Source NARA-ERA)
- Process of adding objects to a Preservation Repository's storage system. In the context of OAIS, Ingest includes services and functions that accept Submission Information Packages (SIP) from Producers, and transforms them into one or more Archival Information Packages (AIP) for long-term retention. (Source: PREMIS & OAIS)
- In the Open Archival Information System (OAIS) model, processes related to receiving information from an external source and preparing it for storage. (Source: Society of American Archivists)
- The OAIS entity that contains the services and functions that accept Submission Information Packages from producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the OAIS compliant repository. (Source: RLG-OCLC)
- Data Ingestion: A process for depositing data onto a storage system. (Source: SNIA Dictionary)

Integrity:

Services that maintain integrity are focused on assuring the consistency, accuracy, and correctness of stored or transmitted data or information. Integrity is often provided through the use of 'hashing' methods such as digital fingerprinting or through locking the media from change by techniques, such as write-once-read-many, WORM. The objectives of integrity processes are to be able to detect and prevent change or corruption.

Examples of use:

Integrity is often confused with authenticity. The key difference between the two is that integrity alone does not always prevent change and may not always detect it. Computer systems can be spoofed and data forged in a manner such as replacing the original and re-hashing it or replacing a WORM image with another. Authenticity brings to bear the second dimension of preventing and detecting change, by including elements such as digital time stamping and an audit log with a documented and controlled chain of custody. Integrity merely tries to prevent and detect change. In a preservation context, integrity is a component of authenticity and a necessary feature of any storage process or repository.

Reference Definitions

- [Context Data Security] Data Integrity: The property that data has not been altered or destroyed in an unauthorized manner [ISO 7498-2:1988] (Source: SNIA Dictionary).
- Condition existing when data is unchanged from its source and has not been accidentally or maliciously modified, altered, or destroyed. (Source National Information Assurance Glossary)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- File Integrity: The ability to retrieve and use a document without the chance of it being lost or misfiled. Also refers to the thoroughness of a file. (Source: ARMA)
- File Integrity - 1. Being accurate, complete, and in original order. 2. being free of corruption. (Source: ARMA)
- Ensure the integrity of data by providing the capability to do system evaluation, data validation, and data integrity checks. (Source: DoD 5015)
- Data integrity service ensures that data is not altered or destroyed in an unauthorized manner. (Source: OAIS)
- The quality of being whole and unaltered from loss, tampering, or corruption. (Source: Society of American Archivists)
- Safeguarding the accuracy and completeness of information and processing methods. (Source ISO 17799:2000)

Logical Format:

In a file-system context, a logical format defines the internal recorded structure of a stored object. Each application, utility, file system, or operating system writes its information and data to storage using a particular logical format that defines the way digital objects such as a files, data, or metadata are stored, accessed, interpreted, and shared. The problem is that all these formats change over time making long-term retention and preservation a challenge, as migration between logical formats has to be accommodated; otherwise the ability to read and interpret information across time is lost.

Examples of use:

Migration of logical formats is achieved by several means today. As upgrades to new versions of applications, utilities, and operating systems are integrated, testing backward read compatibility is an essential practice. It is not uncommon for products to lose backward compatibility after 3 or 4 version changes, so periodic migration to newer versions may need to be a standard practice. It is a common “best-practice” that migration of logical formats must be accommodated before any applications are obsoleted or changed. Another good practice is the transformation to simple and standard formats (such as XML) where possible to reduce the future need for logical migration. [See “Migration” and “Encapsulation”]

Reference definitions:

- The internal data structure of a stored object. (Source: SNIA Dictionary)
- The internal structure of a file, which defines the way it is stored and used. Specific applications may define unique formats for their data (e.g., “MS Word document file format”). (Source: Sedona Principles)
- Specific, pre-established structure for the organization of a File, Bitstream, or Filestream. (Source: PREMIS)
- A collection of related data elements treated as a conceptual unit, independent of how or where the information is stored. A logical record is defined by a particular data structure in an application, independent of the physical characteristics and constraints of the storage medium. (Source: Society of American Archivists)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- A standard for the representation and exchange of data in machine-readable form. (Source: RLG-OCLC)
- The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, aggregations of these data types, and mapping rules which map from the underlying data types to the higher level concepts needed to understand the Digital Object are referred to as the Structure Information of the Representation Information object. (Source: OAIS)

Long-term:

“Long-term” is talked of as a period in excess of 10-to-15 years and extending to ‘forever.’ When asked what is ‘long-term’, archive practitioners consistently express that the point where they begin losing information is after several migrations, in the 10 to 15 year range. “Long-term” in this context is the threshold where migration must deal with technology and logical format obsolescence or even with changes in the managing community or face losing the ability to access, read and interpret the information or to compromise its authenticity due to other factors. [See also: “Long-Term Digital Information Preservation”]

Examples of use:

This is the threshold beyond which retention and preservation processes are at risk of beginning to lose digital information. Information can be effectively lost several ways in a preservation context. It may no longer be able to be verified authentic and consequently its value is compromised. It or the media it is on may become irreversibly damaged or corrupted and not recoverable or readable. Or, it may simply not be interpretable, as there is no way to read or interpret the format as the application or computer systems that generated it are long gone. Even the people who once understood the information may no longer be available.

Reference definitions:

- Long term: A period long enough to raise concern about the effect of changing technologies, including support for new media and data formats, and of a changing user community. (RLG-OCLC)
- Long Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future. (Source: OAIS)
- Records that have enduring value to the organization. (Source: ARMA)

Long-Term Digital Information Preservation:

Long-term digital information preservation can be thought of as the practices of preserving digital information and data for extended periods of time including ‘forever.’ Preservation requires that many information and data services operate in unison to control and maintain the information across extended time-periods. Examples of these services that make up the core of preservation

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

practices are logical and physical migration, authenticity, integrity, confidentiality, availability, and protection. To provide these services in a controlled environment, the use of a specialized preservation repository is often chosen. In addition to a preservation repository, there are currently three main methods of preserving information and data when the hardware or software used needs to be managed for obsolescence: logical and physical migration, emulation, and encapsulation. [See also: “Preservation,” “Migration,” “Emulation,” and “Encapsulation”]

Examples of use:

Digital objects being held for regulatory compliance and for business history are the leading forms of digital information being preserved. Not all information or data in a datacenter has long-term retention requirements. But, enough does that it is a major expense and challenge. Without good preservation practices digital information and data are at great risk of being lost after several migrations.

Reference definitions:

- Long-term data retention: The practice of archiving data for extended periods of time, including ‘forever.’ (Source: SNIA Dictionary)
- Long Term Preservation: The act of maintaining information, in a correct and independently understandable form, over the Long-Term. (Source: OAIS)
- Long-term records - Records that have enduring value to the organization. (Source: ARMA)

Metadata:

Metadata is commonly described as ‘data about other data’, but that is too simple a definition to understand the role and importance of metadata in retention and preservation. Many different forms and levels of metadata exist; some have to do with the file system, application, interpretation, access methods, indexing, or storage extents, some deal with the integrity of the data, and some with reference, representation, or historical information. Note, not all metadata is relevant to the maintenance of authenticity or preservation, nor is all metadata stored within an information object.

It is the specific types of metadata that include the contextual, processing, custodial, reference, representation, and use information needed to identify and certify the scope, authenticity, and integrity of digital information objects, that are of concern for retention and preservation. In the legal world, metadata has two principle functions, it is essential to authenticate digital information and metadata itself may be evidence that must be authenticated.

Examples of use:

Besides supporting the basic functions of providing context and relevance to information, metadata is a necessary attribute for all information objects to help verify authenticity and to support legal evidence and preservation. From an information-services perspective, metadata is the distinguishing component between information and data and must be protected and preserved with the same vigor as applied to data. [See also: “Information” and “Data”]

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Reference definitions:

- A characterization or description documenting the identification, management, nature, use, or location of information resources (data). (Source: Society of American Archivists)
- Metadata is information about a particular data set which describes how, when and by whom it was collected, created, accessed or modified and how it is formatted (including data demographics such as size, location, storage requirements and media information). (Source: Sedona)
- Data about other data. (Source: OAIS, SNIA Dictionary)
- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. (Source: ARMA)
- Metadata (data about data) includes all the contextual, processing, and use information needed to identify and certify the scope, authenticity, and integrity of active or archival electronic information or records. Metadata can come from a variety of sources. It can be created automatically by a computer, supplied by a user, or inferred through a relationship to another document. Metadata is created, modified and disposed of at many points during the life of electronic information or records. Some metadata, such as file dates and sizes, can easily be seen by users; other metadata may be hidden or embedded and unavailable to computer users who are not technically adept. Metadata is generally not reproduced in full form when a document is printed. (Source: Sedona Principles)
- The importance of metadata in electronic archives: The key to maximizing the utility of an electronic archive is the availability of record metadata--especially metadata that cannot be easily derived from the record content--and record management data (such as the business owner, the planned disposition date, various retention factors, etc.) along with the native record. This additional data may add value for searching, reporting and analysis purposes. By adding value for business or user processes, electronic archive systems can present a positive situation for all parties within an organization. (Source: Sedona Principles)
- Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage and information resource. There are three main types of metadata: descriptive, structural, and administrative. Preservation metadata is a form of administrative metadata. Metadata can also describe resources at any level of aggregation such as a collection, a single resource, or a component part of a larger resource. Metadata can be embedded in a digital object or it can be stored separately and linked. (Source: NISO, "Understanding Metadata")
- Usually the metadata describes the contents, physical description, location, type and form of the information, and information necessary for management including migration history, expiry dates, security, authentication, and file formats. (Source: Universal Preservation Format Glossary)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Migration:

Migration is a practice in the datacenter described as the movement or copying of information or data between information systems, storage media, or logical-formats to ensure continued access, readability, and proper interpretation. Migration is required when a datacenter upgrades or changes systems or software, when it has to integrate disparate software and hardware systems such as from an acquisition, and it is a process of transparently moving digital objects between tiers of storage to improve performance, cost, or storage efficiencies. The underlying requirement of migration practices, whether physical or logical, is that the process must not compromise the integrity and authenticity of the original. Physical and logical migration are the two principle technical challenges of long-term preservation. Any migration event needs to be documented and logged as part of the object's history.

Examples of use:

Migration has two principle independent forms, physical and logical.

- **Physical Migration:** refers to moving information from one physical system or location to another or from one physical media-format to another, such as from an older generation tape drive technology to a new higher density tape drive technology, or between tiers of storage to maintain physical readability, accessibility, and integrity, or to achieve other storage and efficiency benefits.
- **Logical migration:** refers to moving information from one logical-format to another, such as from an old application version to a new version, to preserve readability, interpretability, and integrity. Logical migration converts the content and representation information of an information object into a new information object (a transformation) and maintains an audit trail of the change, documenting the conversion event.

The tradeoffs with migration include:

- It allows media, systems, and formats to be retired, obsolete, or replaced before they fail.
- Migration can allow new uses and interpretations
- Migration may 'introduce noise' into the data or information causing errors or lost data
- The cost of migration is proportional to the amount of data to migrate and is continuous over time.

Reference definitions:

- Moving files to another computer application or platform which may require changing their formats. (Source Sedona Principles)
- Migration of digital material is the potential ability to transfer the material into a new digital format without losing any of the digital content within the original migrated file. Thus there is a transfer of the file from one hardware/software configuration to another (updated) hardware/software configuration. A major issue with migration is that there is the potential for information to be lost in the migration. Also since the resultant migrated file is not an exact copy of the original file, then

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

to what extent this technique of migration can be said to meet the requirements of digital preservation is open to debate. Systems are being developed which periodically migrate files from their original formats to more readable formats automatically. These systems store the new migrated file and either retain the original file or discard it. (Source: Midess Project)

- The process of moving data from one information system or storage medium to another to ensure continued access to the information as the system or medium becomes obsolete or degrades over time. (Source: SAA)
- Preservation strategy in which a Transformation creates a version of a Digital Object in a different Format, where the new Format is compatible with contemporary software and hardware environments. Ideally, Migration is accomplished with as little loss of content, formatting and functionality as possible, but the amount of information loss will vary depending on the Formats and content types involved. Also called “format migration” and “forward migration.” Note: Migration and Media migration are used in preference to the definition of “digital migration” in the OAIS Reference Model. OAIS defines digital migration as the “transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes: 1) a focus on the preservation of the full information content; 2) a perspective that the new archival implementation of the information is a replacement for the old; and 3) an understanding that full control and responsibility over all aspects of the transfer resides with the OAIS.” (Source: PREMIS)
- The process of moving data from one information system or storage medium to another. Note: Migration is done to ensure continued access to the information as the system or medium is replaced, becomes obsolete, or degrades over time. (Source: ARMA)
- The primary types (of migration), ordered by increasing risk of information loss, are: (Source: OAIS)
 - Refreshment: A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by copying the bits on the medium used to hold AIPs and to manage and access the medium. As a result, the existing Archival Storage mapping infrastructure, without alteration, is able to continue to locate and access the AIP.
 - Replication: A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication may require changes to the Archival Storage mapping infrastructure.
 - Repackaging: A Digital Migration where there is some change in the bits of the Packaging Information.
 - Transformation: A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content.
- Refreshing: To copy digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit-stream. Migration: The periodic transfer of digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. (Source: NIST)
- Significant Properties: essential characteristics of a digital object which must be preserved over time for the digital object to remain accessible and meaningful. (Source: CASPAR)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- **Data migration:** One current strategy for providing continuing access to archived digital materials over time. It is perceived to be the most reliable strategy for providing continuing access to many types of digital materials because it has been used for years for routine migration of homogeneous digital materials. However, the library community lacks documented practical experience that shows it is a reliable approach for heterogeneous digital collections such as multimedia CD-ROMs or electronic journals, so it has yet to see widespread adoption for these materials. For the purposes of this report, the definition of “data migration” is based on that provided in the CPA/RLG report: “a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. (Source: RLG-OCLC)

Permanent Deletion:

In contrast to file-system deletion, which is easily recoverable, permanent deletion is the process of reliably and provably eliminating the ability to discover, recover, and read specific digital objects from digital media. The degree to which specific classes of information need to be permanently deleted is a consideration that must be made as part of the disposition policy. Other terms in use analogous to permanent deletion include purging, expunging, destroying, or data shredding. [See also: “[Deletion](#)” and “Disposition Policy”]

Examples of use:

The deletion process has two phases, locate and then delete appropriately. The first phase is identifying all of the instances (copies, versions, replicas, backups, etc. including their physical locations) of the information to be deleted regardless where located, and the second phase is permanently destroying all traces of the information including the ability to recover by analysis of the media. Depending on the level of security required, complete physical destruction and obliteration of the media might be necessary. The permission to delete is also predicated upon freedom from any litigation holds or deletion suspension directives and the process(es) used must be the organization’s standard process(es) approved by the appropriate legal and business entities. Permanent deletion should span all media types, devices, and all repositories. To do so in any environment requires good indexing and discovery tools. Permanently deleting has to, as the Department of Defense manual states, “preclude recognition or reconstruction.” This approach is very different than file system-level deletion, which only masks the objects from the file or operating system. File system-level deletion and many other methods can be recovered by various forensic techniques.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Reference definitions:

- The process of permanently removing, erasing or obliterating recorded information from a medium, especially a reusable magnetic disk or tape. (DoD 5015)
- Deletion does not necessarily make data unreadable. The original information may remain intact and, if not overwritten, could be recovered using special software tools. (Source Society of American Archivists)
- Deleted data may remain on storage media in whole or in part until they are overwritten or “wiped.” Even after the data have been wiped, directory entries, pointers or other information relating to the deleted data may remain on the computer. “Soft deletions” are data marked as deleted (and not generally available to the end-user after such marking), but not yet physically removed or overwritten. Soft-deleted data can be restored with complete fidelity. (Source: Sedona Principles)
- Destruction - The definitive obliteration of a record beyond any possible reconstitution. [See also disposition.] (Source ARMA)
- Expunge: To completely remove documentary material from NARA's physical custody and all related information about the material such as that no trace of the material's existence or its audit trail remains. (Source: NARA-ERA)
- The process of permanently removing, overwriting, or obliterating information from an erasable storage medium. Destruction - The definitive obliteration of a record beyond any possible reconstitution. (Source: ARMA)
- The method of destruction must preclude recognition or reconstruction of the classified information or material. (Source: DoD 5220.22-M National Industrial Security Program Operating Manual (NISPOM) January 1995)

Preservation:

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 ‘archive’ 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”]

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of use:

Examples of preservation repositories in common use in the datacenter are compliance stores, eMail archives, eDiscovery repositories, and database archives. Practices such as eMail archive capture and ingest incoming messages and attachments upon creation and preserve them in a secure repository based on retention policies.

Litigation and compliance requirements now require information be retained for a specific period of time and preserved from creation. Consequently, preservation practices need to be tightly coupled to retention requirements and policies across the entire retention period for information and data. Preservation is no longer just a practice relegated to the domain of the long-term “archive.”

Reference definitions:

- The process of ensuring retention and protection from destruction or deletion all potentially relevant evidence, including electronic metadata. (Source: Sedona Guidelines)
- Preservation Metadata: Information a Preservation Repository uses to support the digital preservation process. Preservation Repository: Repository that, either as its sole responsibility or as one of multiple responsibilities, undertakes the long-term preservation of the Digital Objects in its custody. (Source: PREMIS)
- The act of maintaining correct and independently understandable information over the long term. (Source: PREMIS)
- Preservation Description Information (PDI): The information that is necessary for adequate preservation of the Content Information; it can be categorized as Provenance, Reference, Fixity, and Context Information. (Source: RLG-OCLC, OAIS, CASPAR)
- The professional discipline of protecting materials by minimizing chemical and physical deterioration and damage to minimize the loss of information and to extend the life of cultural property. The act of keeping from harm, injury, decay, or destruction, especially through noninvasive treatment. Law - · The obligation to protect records and other materials potentially relevant to litigation and subject to discovery. (Source: Society of American Archivists)
- Preserve, v. To keep for some period of time; to set aside for future use. Conservation · To take action to prevent deterioration or loss. Law - · To protect from spoliation. (Source: Society of American Archivists)
- Processes and operations involved in ensuring the technical and intellectual survival of authentic records through time. (Source: NARA-ERA)
- Process and operation involved in ensuring the technical and intellectual survival of authentic records through time. (Source: ARMA, duplicate of)
- Long Term Preservation: The act of maintaining information, in a correct and Independently Understandable form, over the Long Term. (Source: OAIS)

Preservation Repository (Preservation Store):

A storage domain designed and operated to provide access to, as well as retain, protect, and preserve authentic digital objects within the control of the repository over their retention period. This system may range in scale from a single array to a large distributed and federated set of repositories using many types of storage systems including disk, tape, and optical disk.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of Use:

What is unique about a preservation repository is that it operates a set of preservation services in a controlled storage domain and is designed to meet the organization's business requirements for preservation of digital information over their retention period. Preservation is not just for the long-term. For example, information captured in a discovery process has similar preservation requirements for the life of the litigation as information being retained forever. In the case of litigation, a properly operating preservation repository may also be designed to implement policy-based litigation and deletion holds.

A preservation repository can be designed to deal with mitigating the cost and complexity of physical migration due to storage system failure or drive or media technology obsolescence by implementing self-healing systems. This capability reduces one of the difficult barriers of long-term preservation, physical migration, to only the case where a migration to a completely different repository is to be conducted.

Reference definitions:

- A repository for electronic records is a direct access device on which the electronic records and associated metadata are stored. Sometimes called a "records store" or "records archive." (Source: Sedona Guidelines)
- Preservation Repository: Repository that, either as its sole responsibility or as one of multiple responsibilities, undertakes the long-term preservation of the Digital Objects in its custody. (Source PREMIS)
- An Open Archival Information System (OAIS) or archive, consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities that allows an OAIS archive to be distinguished from other uses of the term "archive." (Source: OAIS)
- A place where things can be stored and maintained, a storehouse. (Source Society of American Archivists)

Provenance:

Provenance is an 'archival' term used to express the documentation of the history (including things like origin, source, and changes) and chain of custody of a digital record that is being ingested into a preservation repository, such as those found in libraries, museums, or historical archives. It has not found distinct use in the datacenter with the breadth and complexity of preserving ESI but could become useful. Authentic metadata and a chain of custody are the closest analog to provenance for ESI in the datacenter. [See also: "Authenticity", "Chain of Custody," and "Audit Log"]

Reference definitions:

- Provenance Information: The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration. (Source: RLG and OAIS)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

- Digital Provenance: Documentation of processes in a Digital Object's life cycle. Digital Provenance typically describes Agents responsible for the custody and stewardship of Digital Objects, key Events that occur over the course of the Digital Object's life cycle, and other information associated with the Digital Object's creation, management, and preservation. (Source: PREMIS)
- The origin or source of something. – 2. Information regarding the origins, custody, and ownership of an item or collection. Provenance is a fundamental principle of archives, referring to the individual, family, or organization that created or received the items in a collection. The principle of provenance or the 'respect des fonds' dictates that records of different origins (provenance) be kept separate to preserve their context. (Source: SAA)
- Provenance - The organization or individual that created, accumulated, and/or maintained the documentary material in the conduct of business prior to their legal transfer to NARA. Note: The archival principle of provenance states that documentary material of the same provenance must not be intermingled with those of any other provenance. (Source: NARA-ERA)

Record (Digital):

A record is electronically stored information, ESI, that has value to an organization and is retained and preserved for its historical or business value, as possible (future) digital evidence or for compliance. The term is a derivative of the concept of a 'business record'.

Examples of use:

The concept of a 'record' is common to the information management community but not generally in IT or storage. It is useful to distinguish between information that is important and valuable to an organization and must be retained and preserved from information that is not, and is disposable. When records are classified and defined, such as by a document or content management system, it makes deciphering the differences between versions, copies, and replicas much easier.

Reference definitions:

- Recorded information, regardless of medium or characteristics, made or received by an organization that is evidence of its operations, and has value requiring its retention for a specific period of time. (Source: ARMA)
- Document(s) produced or received by a person or organization in the course of business, and retained by that person or organization. (Source: MOREQ)
- A unit of recorded information created, received, and maintained as evidence or information by an organization or person, in pursuance of legal obligations or in the transaction of business. (Source: NARA)
- Information, regardless of medium or format, that has value to an organization. Collectively the term is used to describe both documents and electronically stored information. (Source: Sedona Principles)

Reference Information or Data:

(See "Information State")

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Retention:

Retention is described as the process of keeping and controlling digital objects for specific periods of time. Retention is implemented as a policy defining a retention period. Retention policies ideally define the period of time the digital objects are to be retained along with their administrative, legal, fiscal, historical, business, security, or other disposition requirements.

Examples of use:

The retention period in effect specifies an information or data object's expiration date. Some classes of objects have short retention periods and others have long periods depending on factors such as value to the organization, compliance requirements, legal or statutory requirements, or business practices. Events or new usage can trigger retention periods to change so they are not static and need periodic review.

Reference definitions:

- The prevention of data from deletion. The length of time a compliance volume or file must be maintained undeleted and unchanged. (Source: SNIA Dictionary)
- The length of time records should be kept in a certain location or form for administrative, legal, fiscal, historical, or other purposes. (Source: Society of American Archivists)
- Retention periods are determined by balancing the potential value of the information to the agency against the costs of storing the records containing that information. Retention periods are set for record series, but specific records within that series may need to be retained longer because they are required for litigation or because circumstances give those records unexpected archival value. (Source: Society of American Archivists)
- Retention period: The length of time a given records series must be kept, expressed as either a time period (e.g., four years), an event or action (e.g., audit), or a combination (e.g., six months after audit). (Source Sedona Principles)
- Retention period - The length of time a record must be kept to meet administrative, fiscal, legal, or historical requirements. (Source: ARMA)
- The act of maintaining correct and independently understandable information over the long term. (Source: RLG-OCLC)

Versions and Copies:

As files or records are created, edited, polished, and finally approved, they usually produce many versions and copies. A version is a generational variant based on changes made to the content (data) or the metadata over time, as the content evolves and changes over its active state. Each successive change is a 'version' of the original object. A copy occurs for many reasons ranging from distribution for accessibility to data protection and business continuity. In most environments multiple versions and copies of versions exist and are distributed across the enterprise.

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Examples of use:

Versions, copies, backups, test copies, and replicas are legitimate and valuable ‘duplicates’ of information or data and present challenges in managing and discovering information. Those challenges include maintaining clear visibility of which ones are the official record, their location, ownership, chain of custody, authenticity, etc. Versions and copies of versions are often distributed across an organization residing in backup systems, email archives, user’s disks, various servers, and at disaster recovery sites. Versions and copies make authenticity, discovery, and deletion practices particularly difficult as prior versions of a final record may actually have information relevant to a particular litigation or other business process, and it may be difficult to locate all copies.

Reference definitions:

- Version: a particular form of or variation from an earlier or original record. For electronic records, the variations may include changes to file format, metadata or content. (Source: Sedona Principles)
- Version: an attribute of an Archival Information Package, AIP, whose information content has undergone a transformation on a source AIP and is a candidate to replace the source AIP (Source: OAIS)
- Authentic copy or version: A copy of documentary material for which the official custodian attests the authenticity. (Source: NARA-ERA)
- Version: a variant with differences from an earlier form. Often a particular instance of something made during development. A revision is a different version. A version may indicate a different form, translation, or adaptation. (Source: Society of American Archivists)

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

APPENDIX

TERMINOLOGY REFERENCE SOURCES:

- ARMA International: Glossary of Records and Information Management Terms, [ARMA_RIM-Glossary_4-05.pdf]
- Caspar: Glossary, <http://www.casparpreserves.eu/caspar-glossary-of-terms.pdf/download>
- Committee on National Security Systems, "National Information Assurance (IA) Glossary, 2006" http://www.cnss.gov/Assets/pdf/cnssi_4009.pdf
- Dept of Defense: DOD 5015 and 5022, <http://jrtc.fhu.disa.mil/recmgt/p50152s2.doc>
- Federal Rules of Civil Procedures, [Federal-Rules-of-Civil-Procedure.pdf]
- Federal Rules of Evidence, www.uscourts.gov/rules/Evidence_Rules_2007.pdf
- International Standards Organization, ISO Archiving Standards - Reference Model Papers http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
- InterPARES(2): The International Research on Permanent Authentic Records in Electronic Systems <http://www.interpares.org/>
- InterPARES Project: "Authenticity Task Force Final Report," Oct. 2001
- Midess Project: http://www.leeds.ac.uk/library/midess/MIDESS_Workpackage_5_Digital_Preservation.pdf
- MoReq2: Model Requirements for the Mgmt of Electronic Records, 2008 <http://www.moreq2.eu/downloads.htm>
- National Institute of Standards and Technology, NIST: "The State of the Art and Practice in Digital Preservation", Journal of Research of the National Institute of Standards and Technology, January 2002, and Appendices to Guide for Mapping Types of Information and Information Systems to Security Categories, Pub 800-60, August 2008 http://csrc.nist.gov/publications/nistpubs/800-60-rev1/sp800-60_Vol2-Rev1.pdf
- National Information Assurance Glossary, 2006 http://www.cnss.gov/Assets/pdf/cnssi_4009.pdf
- National Information Standards Organization, NISO, document "Understanding Metadata" <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>
- National Archives and Records Administration, NARA – Electronic Records Archive Glossary: <http://www.archives.gov/era/about/glossary.doc>
- Open Archival Information Systems: the glossary is internal to the ISO standard: <http://public.ccsds.org/publications/archive/650x0b1.pdf> (also available as ISO 14721 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683)
- PREMIS: Data Dictionary for Preservation Metadata, May 2005 <http://oclc.org/research/projects/pmwg/premis-final.pdf>
- Research Libraries Group – RLG and OCLC – Online Computer Library Center <http://www.oclc.org/support/documentation/glossary/oclc/glossary.htm>
- Sedona Conference Glossary: http://www.thesedonaconference.org/dltForm?did=TSCGlossary_12_07.pdf
- SNIA Dictionary: <http://www.snia.org/education/dictionary/>
- Society of American Archivists, SAA, Glossary of Archival & Records Terminology: <http://www.archivists.org/glossary>
- Universal Preservation Format Glossary: <http://info.wgbh.org/upf/glossary.html>

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

Index

Active Information.....	26	ILM.....	3, 12, 23, 26
Archival storage.....	7	Inactive Information.....	26
Archive	6, 37	Information.....	13, 21, 23
Audit log.....	7	Information assets.....	11
Audit Log.....	10, 15, 16, 29, 34	Information lifecycle management.....	3, 12, 23, 26
Authenticate.....	11	Information object.....	13, 21, 23, 24, 32, 34
Authenticity.....	9, 11, 17, 22, 29, 32, 41	Information state	25
Backup.....	15	Ingestion	28
Business process state.....	26	Integrity.....	6, 9, 29
Capacity optimization.....	14	Legacy data.....	16
Categorize.....	11, 12, 25	Litigation hold.....	15, 18, 36
Chain of custody.....	10, 29, 39, 41	Logical format.....	21, 30, 33
Classification	11, 26	Logical migration.....	10, 34
Classification scheme.....	12, 26	Long-term	31
Collaborate.....	3, 18	Long-term preservation.....	20, 21, 30, 31, 33
Compliance.....	10	Media migration.....	35
Compression.....	14	Metadata.....	9, 13, 19, 21, 23, 24, 32
Cost of migration.....	34	Migration.....	6, 10, 20, 21, 30, 31, 33, 35, 38
CRC.....	11	Ontology.....	11
Data	13, 23	Open Archival Information System (OAIS).....	7
Data deduplication	14, 17	Operational state.....	26
Data loss.....	34	Overwrite.....	16
Data object.....	23	Permanent deletion.....	16, 36
Data shredding.....	36	Physical Migration	34
Deletion	15, 18, 36	Preservation.....	6, 31, 32, 37
Deletion log.....	8	Preservation community.....	23
Deletion suspension.....	18, 36	Preservation repository.....	6, 7, 8, 28, 38
Destruction.....	17, 36	Provenance	39
Digital evidence.....	10	Purging.....	36
Digital fingerprinting.....	10, 17, 22, 29	Record (digital)	40
Digital name.....	17	Reference information.....	26
Digital object.....	8, 13, 14, 20, 21, 25	Refresh.....	35
Disposition policy.....	15, 18, 40	Representation information.....	22, 23, 24
eDiscovery.....	8, 19, 28	Retention.....	6, 40
Electronically Stored Information, ESI	19	Retention period.....	6, 15, 18, 27, 37, 38
Email archive.....	6	<i>Security classification</i>	11
Emulation	20, 31	Significant properties.....	10, 11, 35
Encapsulation	21, 31	Single-instance-storage.....	14
ESI.....	6, 37, 40	Spoilation.....	37
Evidence repository.....	8	Storage media.....	33, 34
Expired information.....	16, 18, 26	Taxonomy.....	11, 12
Expunge.....	36	Tiering.....	6, 26, 34
File system deletion.....	16, 36	Version and copies	41
Fixity.....	11, 17, 22	WORM.....	29
Forensic analysis.....	16, 36	XAM.....	12
Hashing.....	10, 14, 22	XML.....	21, 30
Hash-value.....	17		

Building a Terminology Bridge: Guidelines for Retention and Preservation Practices

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, long-term information retention, preservation, database archiving, and discovery. For more information, visit www.snia.org/forums/dmf .

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.