



Education

Cloud Archive and Long Term Preservation Challenges and Best Practices

Chad Thibodeau, Cleversafe Inc.
Sebastian Zangaro, HP

Author: Chad Thibodeau, Cleversafe Inc.

- ◆ The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- ◆ Member companies and individual members may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced in their entirety without modification
 - ◆ The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- ◆ This presentation is a project of the SNIA Education Committee.
- ◆ Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- ◆ The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

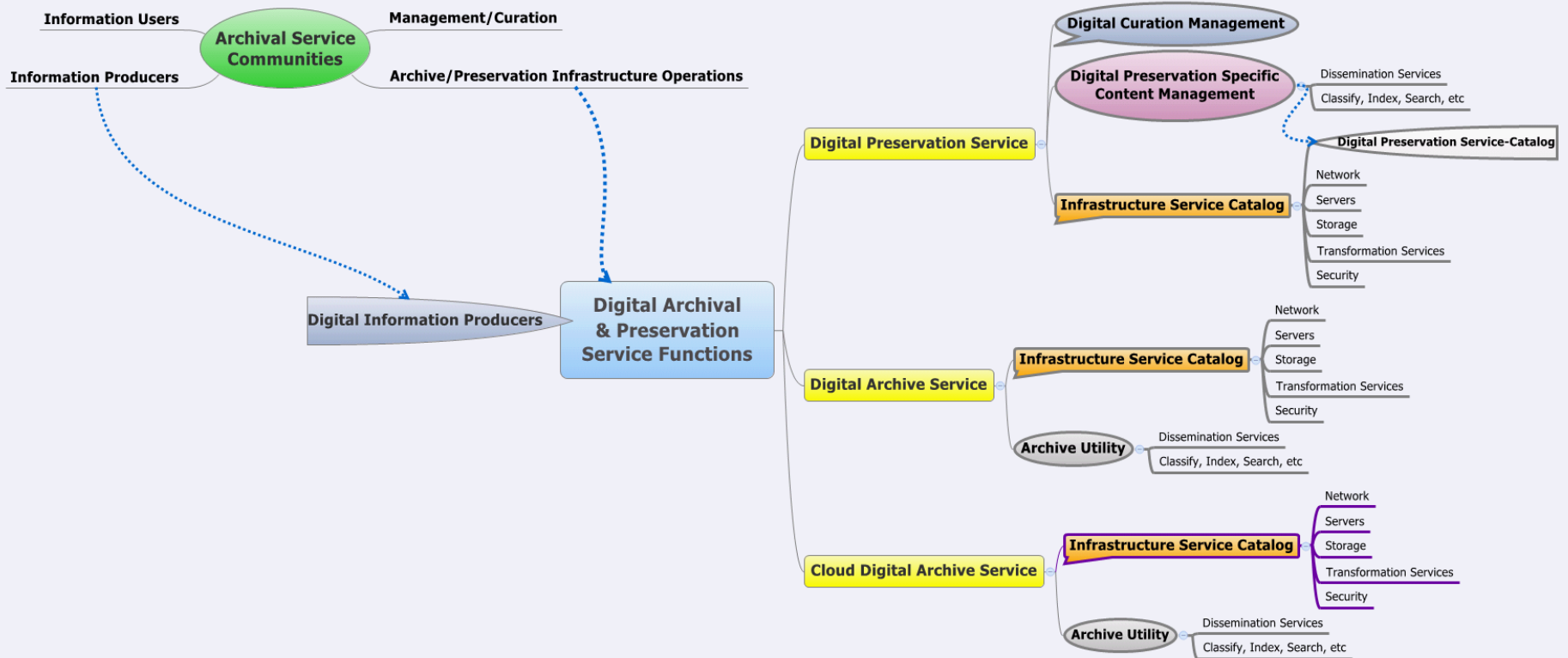
NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

➤ Cloud Archive Challenges and Best Practices

- ◆ This session will appeal to Storage Vendors, Datacenter Managers, Developers, and those seeking a basic understanding of SNIA's Cloud Archive and Preservation Special Interest Group and its work towards promoting the use of the cloud for archival purposes.
- ◆ This session will examine current challenges within the Public Cloud Storage Industry, delve into some specific services profiles, and address some best practices for utilizing cloud storage for archive and preservation needs.

- Cloud Archive and Preservation Definitions
- Introduce SNIA Cloud Archive and Preservation SIG
- Challenges within Public Cloud Storage
- Solutions – from data to information

Digital Archive and Digital Preservation Taxonomy



Source: MPeterson, www.ltdprm.org

➤ Cloud Digital Archive Service:

- ◆ A cloud-base service providing a specialized online storage repository for the purposes of compliance, litigation support, and/or retention for extended periods of time, not including “long-term.”
 - › Can be utilized as a component of a complete digital preservation service.
 - › Does not necessarily provide adequate services to accomplish digital preservation.

➤ Cloud Digital Preservation Service

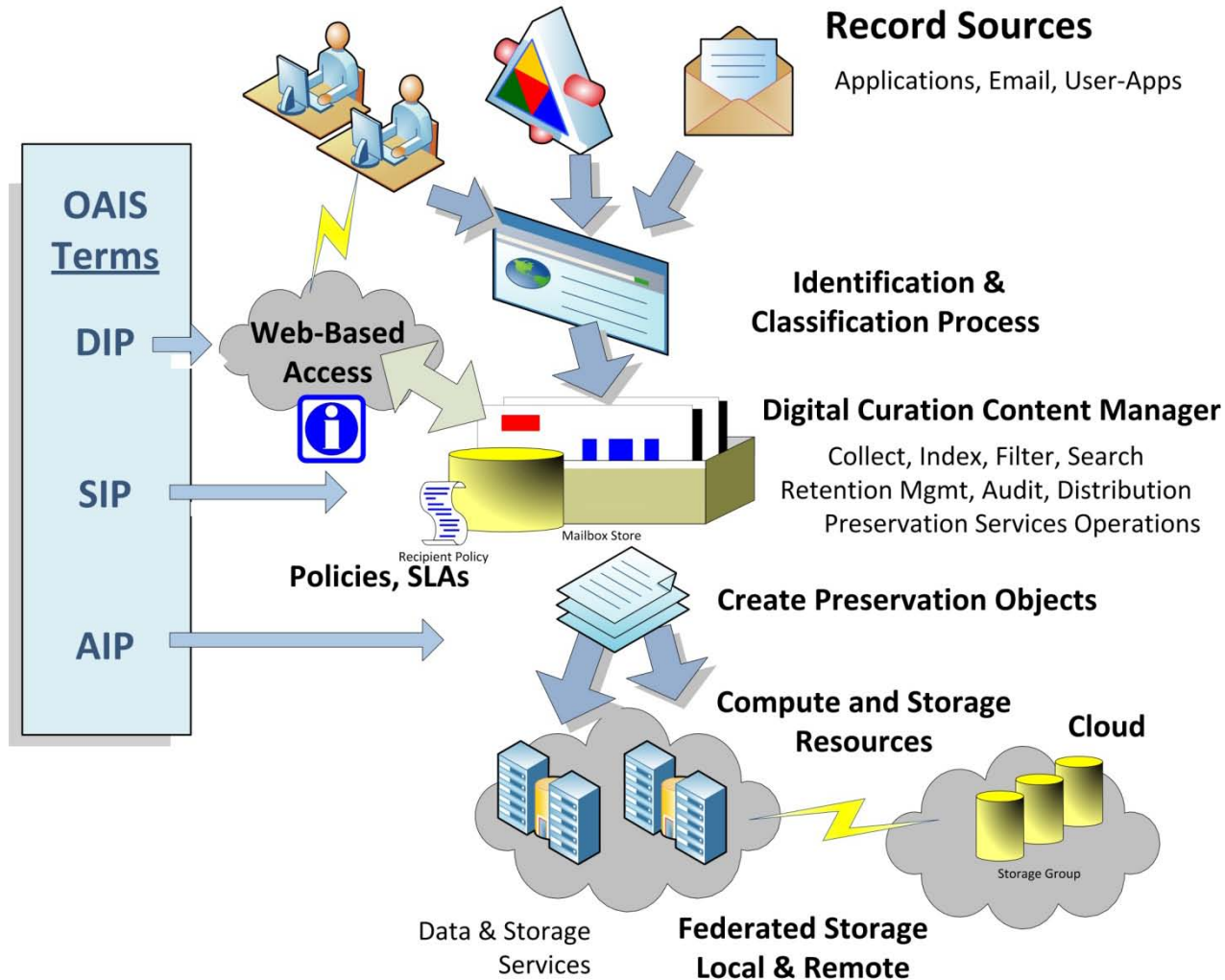
- ◆ A cloud service providing digital preservation of information and data.
- ◆ A digital preservation service includes a comprehensive management and curation function that controls:
 - › Supporting Infrastructure
 - › Information
 - › Data
 - › Storage Services

➤ Digital Preservation Object

- ◆ A special type of a digital information object consisting of indexes, fixity, audit logs, data files, reference information, and metadata wrapped into a single/compound digital container.
- ◆ A preservation object provides the functionality required to use, secure, interpret and verify authenticity of the metadata, information and data in the *container* and is the foundational element for digital preservation of information.

CDMI is the foundation for a Cloud-intelligent **Preservation Object**

Digital Preservation Framework



Source: www.ltdprm.org

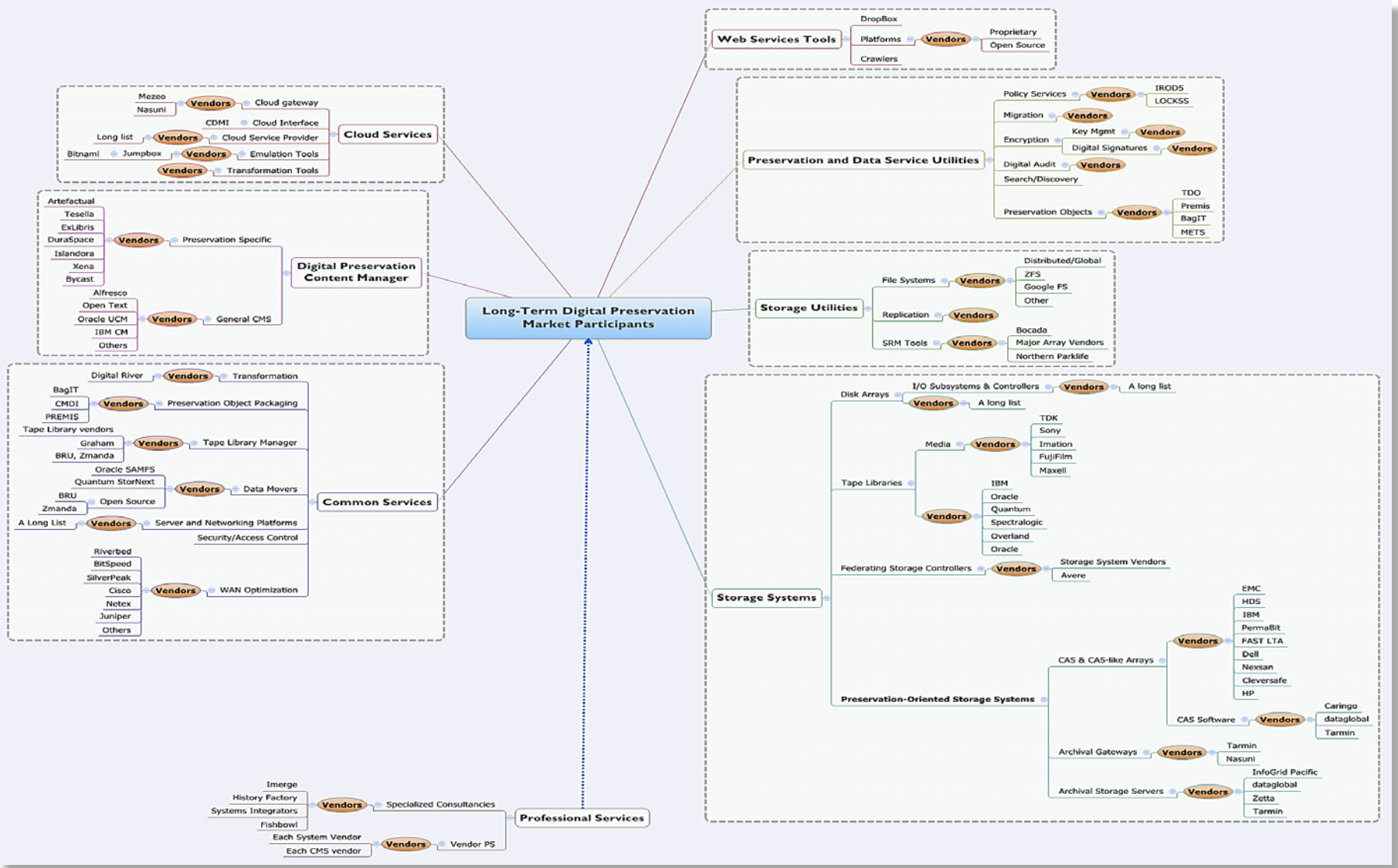


Education

Introduction to the SNIA Cloud Archive and Preservation (CA&P) Special Interest Group

- Advance the use of public, private and hybrid clouds for archival services and long term retention
 - ◆ CDMI
 - ◆ Market Education
 - ◆ Best Practices
 - ◆ Services Profiles
 - ◆ Standards Promotion
 - ◆ Industry Liaison
 - ◆ Interoperability Demonstrations/Certifications and Plugfests
 - ◆ Implementation Reference Model

Digital Preservation Market



➤ Participating companies:

- ◆ BlueArc, Cleversafe, Computer Associates, EMC, HP, Hitachi Data Systems, IMERGE Consulting, Iron Mountain, NetApp, Novell, Oracle, SNIA, Spectra Logic

➤ **New Potential members:**

- ◆ Hardware vendors
- ◆ Content management software developers
- ◆ Professional services
- ◆ Cloud services providers
- ◆ Archival infrastructure providers
- ◆ End-User organizations



Education

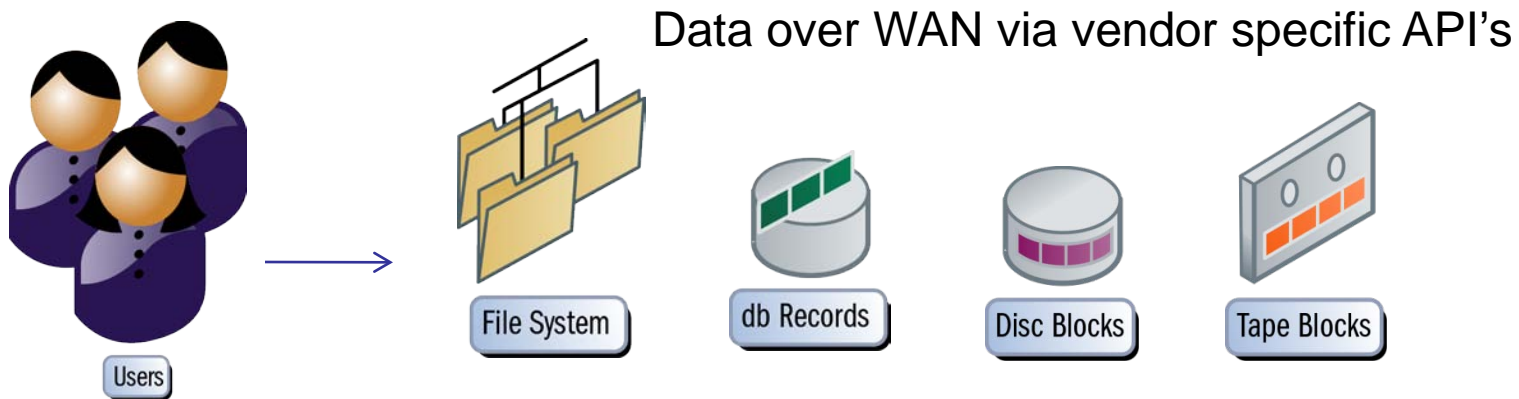
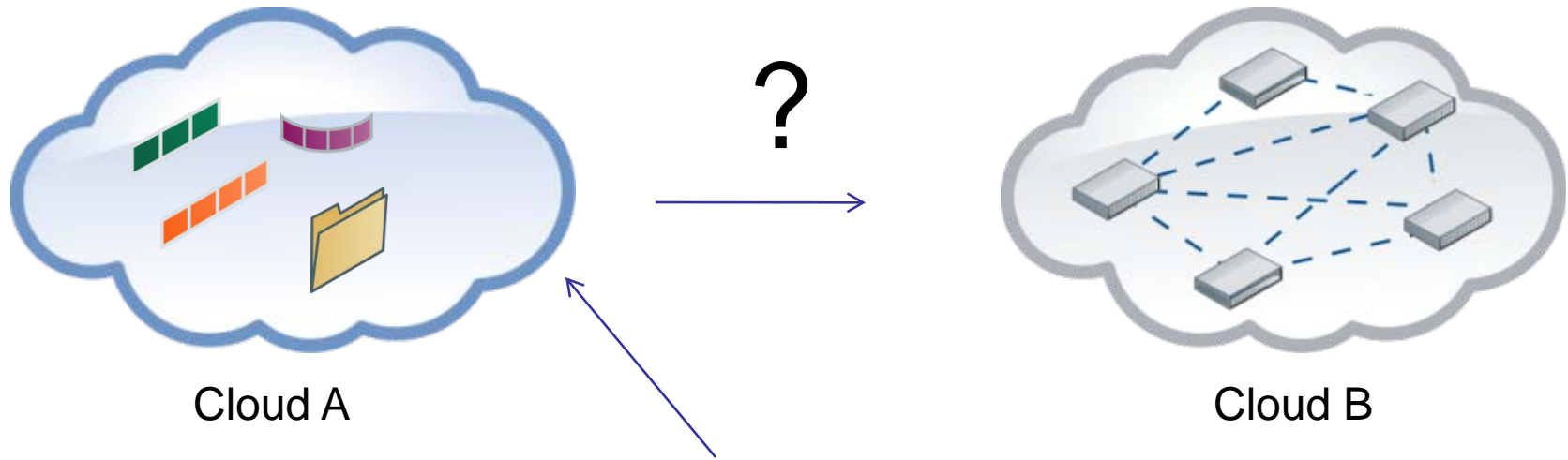
Current challenges with the Public Cloud Storage Industry

- Lack of uniform semantics and standard interfaces
- Interoperability between public cloud providers
- Managing data format changes over time
- Authenticity verification
- Compliance and Governance
- Risk Management & Litigation

THE UNITED STATES
LIBRARY OF CONGRESS IS STORING
235 TERABYTES OF DATA.
88% OF US INDUSTRIES
HAVE MORE DATA THAN THAT!



A new class of challenges

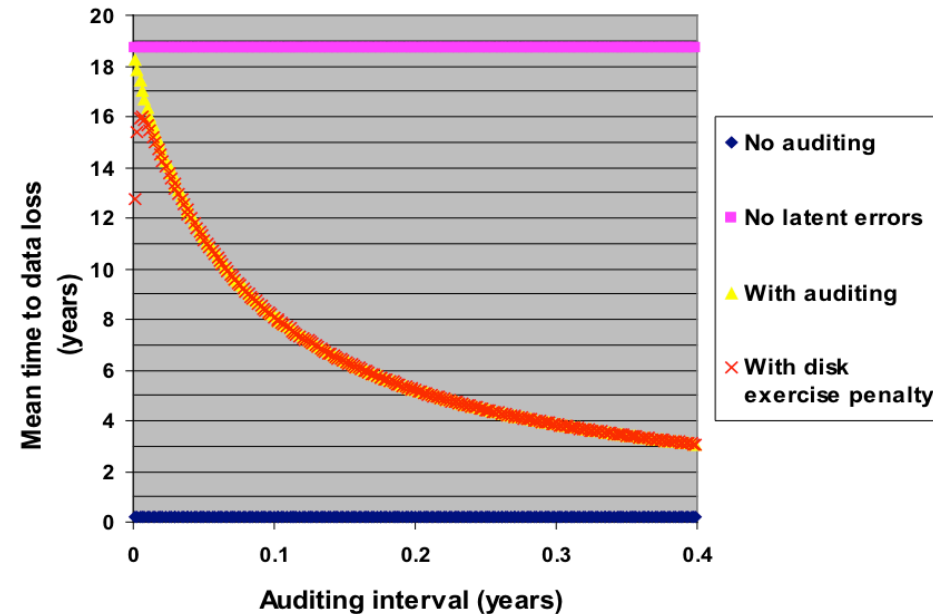


➤ Authenticity Verification

- ◆ Talk about digital auditing
- ◆ Migration
- ◆ Compliance & Governance
- ◆ Risk and Litigation

Baker et al., "A Fresh Look at the Reliability of Long-term Digital Storage." *EuroSys 2006*.

Reliability vs. Auditing



Retaining Information for 100 Years

© 2010 Storage Networking Industry Association. All Rights Reserved.

➤ Data Migration

- ◆ Every time a data object is moved it runs the risk of corruption due to error rates migration technologies.
- ◆ Even data at rest can be corrupted via bit-rot, malicious attack, or user error

➤ Bit-integrity preservation

- ◆ Data management and curation can monitor the integrity of data stored in a Digital Archive.
 - › Often checksums are used to verify that there is no variance between two copies of the same piece of data over time.
 - › A Cloud Archive must ensure that data migration does not affect the integrity of the data object stored.

➤ Governance and Compliance



Education

Solutions – from Data to Information

IT'S THE INFORMATION STUPID!



What is already standardized?

- **SNIA's Cloud Data Management Standard (CDMI)**
 - ◆ Standardized Data Path (Access) to the Cloud
 - ◆ Standardized metadata to express the Archive requirement for the Data put in the cloud
- **Requirements for Archiving:**
 - ◆ Multiple primary copies – distributed geographically
 - ◆ Secondary copies (backup) if the data changes
 - ◆ Hash of the data to check for changes and corruption
 - ◆ Immutability in some cases

- Cloud Client needs to discover what archiving capabilities are provided by the cloud
 - ◆ CDMI does this through *Capabilities* – a type of resource that acts like a service catalog for the functions that the cloud offers customers
 - ◆ If the cloud offers the capability, the customer marks the data objects and containers with metadata (Data System Metadata) that specifies the requirements
 - ◆ Lastly the Cloud provider has a way of expressing what is actually being provided through metadata also
- What does this metadata look like?

- Section 16.4 of CDMI addresses Data System Metadata
 - ◆ A standardized means to express the archive requirements for a single data object or an entire container
- Multiple Primary Copies....
 - ◆ ***cdmi_data_redundancy*** - Contains the desired number of complete copies of the data item to be maintained. This number determines the minimum number of primary copies of the data that the cloud shall maintain. Additional primary copies may be made to satisfy demand for the value.

- ... distributed geographically
 - ◆ ***cdmi_data_dispersion*** - Contains the desired distance (km) between the infrastructures supporting the multiple copies of data. This metadata is used to separate the (cdmi_infrastructure_redundancy number of) infrastructures by a minimum geographic distance to prevent data loss due to site disasters.
 - ◆ ***cdmi_infrastructure_redundancy*** - Contains the number of desired independent storage infrastructures supporting the data. This metadata is used to convey that, of the primary copies specified in cdmi_data_redundancy, these copies shall be stored on this many separate infrastructures. Any two infrastructures may not share common elements, such as a network or power source.

➤ Secondary copies

- ◆ ***cdmi_RPO*** - Contains the largest acceptable duration in time between an update and when the update may be recovered, specified in seconds. This metadata is used to indicate the desired backup frequency from the primary copy(s) of the data to the secondary copy(s).
- ◆ ***cdmi_RTO*** - Contains the largest acceptable duration in time to restore data, specified in seconds. This metadata is used to indicate the desired maximum acceptable duration to restore the primary copy(s) of the data from a secondary backup copy(s).

➤ Hash

- ◆ ***cdmi_value_hash*** - If present, this metadata lists the hash algorithm/ lengths supported. If absent, objects shall not present a hash value as system metadata. Values are in the form of "Algorithm Length", for example, "SHA256". When a CDMI implementation supports hashing, the system-wide capability of "cdmi_security_data_integrity" shall be set to "true". Otherwise, it shall not be present, indicating that there is no hashing support.

➤ Immutability

- ◆ ***cdmi_retention_period*** - Contains an ISO-8601 time interval during which the object is under retention. Only the end-date may be altered when updated. If an object is under retention, the object may not be deleted and its value may not be altered. After the retention date has passed, the object may be deleted.

Other Requirements...

- Archived data may have other requirements besides those related to archiving
- i.e. access latency, throughput, encryption, geographic placement, etc.
 - ◆ All standardized by CDMI!
- But, need implementation by Cloud Archiving Providers
 - ◆ Ask for these from your favorite storage cloud...

➤ Benefits of Industry standards:

- ◆ Allows storage vendors and developers to easily integrate with any cloud infrastructure.
- ◆ Allows Data Object Migration between heterogeneous systems:
 - › End User site to Public Cloud
 - › Public Cloud A to Public Cloud B
 - › From Public Cloud back to the End User
- ◆ Standards already exist such as CDMI (The Cloud Data Management Interface)

Example Services Profiles

Digital Cloud Archive	Digital Preservation Cloud	Backup Cloud
Retention for extended periods of time	Retention for extended periods of time (indefinitely)	Variable retention periods (may not be long-term)
Permanent Deletion	Permanent Deletion	Incremental Deletion
Long-term information	Long-term information	Data recovery
Availability	Interpretation	Device specific
Interpretation	Authentication	Policies
Authentication		Data protection
Security	Security	Security
Data Integrity	Information Integrity	Data Integrity

Define Profiles using CDMI JSON attributes
Certify compliance at Plugfests

- Please send any questions or comments on this presentation to SNIA: tracktutorials@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.
SNIA Cloud Archive and Preservation SIG**

**Chris Marsh
Michael Peterson
Don Post
Thomas Rivera**

**Mark Carlson
Ray Clark
Bob Rogers**