Physical and Logical Migration

Michael Peterson
September 2010
Putting Migration in Context

- SNIA’s role in Preservation
- Complexity Theory
- Information vs. Data (Preservation vs. IT)
- “You will lose information – how much?”
SNIA in Digital Preservation

● **Storage Networking Industry Association**
  - 400 member companies, ~10,000 individual members
  - The current mission statement positions the association to lead the (storage) industry in developing and promoting standards, technologies, and educational services to empower organizations in the management of information.

● **History in digital preservation**
  - Beginning 2003 - implementation focus
    - Workgroups on information lifecycle management, long-term retention, preservation, storage security, data protection
    - Focus – the business IT datacenter
The Storage Networking Industry Association is developing standards and practices which:

- Are helping move information retention and preservation more mainstream and easier to procure
- Minimize proprietary, custom implementations
- Better support industry expectations for retention, preservation, access and disposition of objects.

Examples of Standards in process:

- **SIRF** - Self-Contained Information Retention Format – in development
- **XAM** - eXtensible Access Method – ANSI std, going to ISO
- **CDMI** - Cloud Data Management Interface – in development
1st Principle:
- If you want to solve a complexity problem, “Stop doing it!”

1st Corollary:
- Automating a bad process still produces a bad process
- Moving a bad process (i.e. to the Cloud) is still a bad process

Message:
“We need new system approaches that include the business requirements, not just more technology!”
Is it Information or is it Data?

- **Block Object:**
  - Aggregation of bits

- **Data Object:**
  - Aggregation of blocks

- **Information Object:**
  - Data + Metadata
  - Preservation Object

The question is how much and is it a problem?

- Deletion during litigation hold or lack of clear retention policies
- Corruption or damage and the inability to recover or decrypt
- Can’t find it, read it, or interpret it
- Security theft or changes
- No longer have the ‘original’ records
- Inability to access 3rd party sites/systems
- Failure to control and prove the integrity and authenticity of the information and its metadata
- Migration/transformation to other formats (e.g. a tiff image, or a ‘pdf’ of spreadsheets that don’t provide formulas, macros, links...)

Physical Migration
Physical Migration

- **Goal** – keep the bits useable over time
  - Migrate physical media before data loss
- **No standards for when, only recommendations based on time**
  - 3 yr disk, 5 yr tape
- **Issues:**
  - Cost and complexity
  - Time and bandwidth
    - Very Large file systems overwhelm both
Migration Complexity

- **Physical Migration:**
  - Media, media handlers
  - Player/recorders
  - Computer systems
  - Operating systems
  - Library management
  - drivers, formats, protocols
  - bandwidth

- cost, complexity
- support infrastructure including people
- emulation software
- migration practices
Physical Migration Solutions

- **Self-healing, federated storage systems**
  - Disk and tape
  - Use hashing & error rates to identify thresholds for replacement
  - Continuous Auditing & policy-based correction
- **Self-healing file systems**
  - Global recovery
- **Apply SIRF**
  - Portable information objects

Federated Storage Resources:
- Tiered and Virtualized
- Distributed
- Scalable
- Multiple vendors
- Multiple architectures
- Interoperable
Example: audited replicated archive

Reliability vs. Auditing

- No auditing
- No latent errors
- With auditing
- With disk exercise penalty

Mean Time to Data Loss (yrs)

Auditing interval (years)


© 2009 Storage Networking Industry Association. All Rights Reserved.
What is SIRF?

- Logical data format of a mountable unit
  - File system, block device, stream device, object store, tape, etc.

- Includes a cluster of “interpretable” preservation objects
  - Self-describing – can be interpreted by different systems
  - Self-contained – all interpretation data contained in object cluster
SIRF Addresses Migration

**Without SIRF**

Sets of linked objects moved individually; referential integrity and context may be lost

Only original application that created the objects can read and interpret them

Export and import needed to migrate objects

Preservation Objects cannot be sustained long-term

**With SIRF**

Sets of linked objects moved between systems maintaining referential integrity and full context

Any SIRF compliant application can read and interpret the objects

Objects migrated without export and import

Preservation Objects can survive longer

---

**Diagram Description**

- **Application A** and **Application B** are shown with interfaces, preservation, retention, and storage subsystems.
- **Without SIRF**, data types are shown as dashed lines, indicating lost context.
- **With SIRF**, data types are solid lines, indicating maintained context.
- The use of SIRF ensures preservation and retention, allowing for long-term sustainability.

---

**Key Points**

- **Sets of linked objects** moved individually.
- **Referential integrity and context** may be lost.
- **Only original application** can read and interpret objects.
- **Export and import** needed for migration.
- **Preservation Objects cannot be sustained long-term**.

- **Sets of linked objects** moved between systems maintaining referential integrity and full context.
- **Any SIRF compliant application** can read and interpret objects.
- **Objects migrated without export and import**.
- **Preservation Objects can survive longer**.
Self-Contained Information Retention Format

- **Why SIRF?**
  - Reduce Migration cost, improve efficiency
  - Improves potential for interoperability
  - Creates a portable information object
    - Can embed security, audit, and readers
  - Support future storage systems
    - Needed for the “Cloud”

- **Unique Attributes**
  - Preserves collections of objects and their relationships
  - Includes generic metadata that can be extended with domain specific information for fast access
  - Can be mapped to and physically migrated between a wide variety of underlying storage systems
Preservation Object Evolution

- **Requirements**
  - Self-Contained – complete
  - Self-Describing - self interpretable
  - Extensible – meet future needs
  - Portable – addressable & independent of storage & location
  - Resilient – to failure and change

OAIS as foundation

Architecture
- Archival Information Package

Portable Information Objects
- SIRF

Metadata & Packaging Models
- PREMIS
- METS
- XFDU
- VEO

Containerization
- Bagit
- JHOVE
- TAR
- PREMIS
Logical Migration
Logical Migration

- **Goal**: keep the bits interpretable and reusable over time

- **Methods (migrate logical format over time)**
  - Encapsulation – embed readers in the objects
  - Emulation – recreate original operating platform via software emulation
  - Transformation – encode the data into new formats

- **Issues**
  - Retain context (interpretation) not just readability
  - Cost and complexity
  - Track and audit change over time
Migration Complexity

- **Logical Migration**
  - Application support
  - Encapsulation
  - Standard formats,
  - Container Interop
  - Transformations (i.e. video, images)
  - Emulation software
  - Readers

- Computers, OS, support, interfaces, etc
- Cost & complexity
- Business value
- People with knowledge
- Migration practices
Logical Migration Solutions

- Will always have to migrate – trying to slow down the frequency
  - All three techniques in use in various communities
- SIRF
  - Encapsulate readers so that each object is portable
    - create open source community to produce
  - Audit logs and clusters of objects accommodate transformations and emulation
How you can Help SIRF

- Talk it up, promote it
- Encourage vendors to support it
- Encourage SNIA to drive it
- Help with funding to move faster
  - A 100% Volunteer effort today
Resources

- **SNIA** - [http://www.snia.org](http://www.snia.org)
  - SNIA Technical Working Groups (including the LTR TWG) –
  - “100 Year Archive Requirements Survey'

- **IMERGE**
  - [www.ilm20.org](http://www.ilm20.org)