Archives, Backups, and NAS
The Open Archive
Andres Rodriguez, Hitachi Data Systems
SNIA Legal Notice

• The material contained in this tutorial is copyrighted by the SNIA.
• Member companies and individuals may use this material in presentations and literature under the following conditions:
  – Any slide or slides used must be reproduced without modification
  – The SNIA must be acknowledged as 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.
Agenda

During this presentation we will cover:

- What are archives?
- Backups are not archives
- The Open Archive
- Approaching NAS

Check out SNIA Tutorial:
Archiving and Compliance Infrastructure

Check out SNIA Tutorial:
The Coming Archive Crisis
Fixed content data repositories storing data objects that have long-term value, that do not change over time, and are easily accessible.
Back up to Tape & Offsite Storage for Disaster Recovery

- Email app changing every 5 years
- ERP system changing every 10 years
- Document management changing every 3 years
- Hardware upgrades on a 3 year cycle
- Data stored off-site and off-line
- Data consciously growing in amount and complexity
Access from Tape and Offsite Storage

Data ingestion
Data preservation

**Data access**

- Data is off-line
- Too much data – cannot find files
- Data corruption and damaged tapes
- Cannot read files – unable to recreate the original environment
- Manual, error-prone and time-consuming data recovery process
- **Major risks during a legal discovery process**

![Image of data access issues]

Media sitting off-site
Birth of the Archival Tier

- Over 80% of the information in data centers today is fixed content
- 60% of all new information is fixed content
- Fixed content is growing 90% CAGR
- ILM has mostly failed to deliver on its promise
  - Customers didn’t know what they wanted
  - Products couldn’t deliver what the vendor claimed
The Open Archive Approach (OAIS)

- Driven by National Archive and Records Administration (NARA)

- Ingestion
  - Prepare ingestion package for each data type
  - Add provenance metadata (METS)
  - Convert original file to relevant Open standard

- Preservation
  - Individual object storage
  - Set policies (retention, authentication, de-dup, …)
  - Design for scale and reliability (parity protection, number of copies, replication)

- Access
  - Index all data and metadata
  - Design for high performance (sub-second queries)
  - Provision discovery to compliance group and end-users
Metadata Encoding and Transmission Standard (METS)

- Driven by the Library of Congress
- Descriptive Metadata
  - May be part of an object or come from another metadata object or both
- Administrative Metadata
  - May be created externally or internally
  - Examples could include DRM, object creation and storage
- File Groups
  - Groups files together into logical groups and may include nesting and versioning
- Structural Map
  - Establishes associations between files or metadata in an archive
- Behavior
  - Executable policies of how a file should be treated in an archive
Unlocking Application Gridlock

User

Production Server

Active Archive

online

earline

archive

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.
What is Long Term?

Beyond the applications

Email Server
- NFS
- Tape Library

Document Management
- CIFS
- Optical Jukebox

General Accounting
- NFS
- NAS

Web Applications
- HTTP
- RAID Array

Beyond the hardware
The Open Archive

- Decouple archive objects from the ingestion tools
  - Support Open data standards
  - Support Open network access protocols
- Decouple archive objects from hardware storage layer
  - Run portable archiving software
  - Assure interoperability at the network layer - not device specific
- Decouple archive objects from archiving software
  - Provide network export protocol (NDMP)
  - Support an Open transport format for archive objects
## Archive Requirements

Repositories that are collections of documents, records and other materials

### The Archiving Process

<table>
<thead>
<tr>
<th>Data Ingestion</th>
<th>Data Preservation</th>
<th>Data Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Appraisal:</strong> Is this the right archive for the records and do duplicates exist?</td>
<td><strong>Integrity:</strong> What condition are the records in? Should they be transcribed to a new format?</td>
<td><strong>Policies:</strong> What type of access policies should records have?</td>
</tr>
<tr>
<td><strong>Accession:</strong> Archives should record how and when records were included as well the author and owner of the records.</td>
<td><strong>Preservation:</strong> What are the environmental needs of the records? What type of enclosure is required?</td>
<td><strong>Arrangement:</strong> Group records according to provenance (where they came from)</td>
</tr>
<tr>
<td><strong>Disposition:</strong> Do particular records need to be on site or remote? Should the records ever expire?</td>
<td><strong>Security:</strong> What type of security controls are required?</td>
<td><strong>Search/Retrieval:</strong> A finding aid, and description of the record group. Can be online &amp; searchable.</td>
</tr>
</tbody>
</table>
Approaching NAS
NAS Challenges

- **Migration**
  - How do I allocate new storage to existing users and migrate from old hardware to new hardware

- **Single point scalability**
  - How large can my access point scale

- **Backup windows (and recovery time)**
  - How much of what is being backed up hasn’t changed since the last backup

- **Management**
  - Since NAS delivers functionality needed to support changing information shared among many users, it has more tasks to be preformed than a collection of static items in an archive
Global Name Space

Separation of the physical and the logical layers
Open Archive: NAS

- 80-90% Data in Archive
- Small NAS backup
- Scalability
- Data Integrity
- Single Control Point
- Search Files
- Less Management
References

• http://www.loc.gov/standards/mets/METS%20Documentation%20draft%20070310p.pdf
• http://www.loc.gov/standards/mets/METSObservable.html
Q&A / Feedback

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

Many thanks to the following individuals for their contributions to this tutorial.
SNIA Education Committee

Andres Rodriguez          Mike Fahey