Digital Bedrock and Cloud Storage

Linda Tadic, Founder/CEO
Digital Bedrock
ltadic@digitalbedrock.com

NDSA Cloud Working Group
January 25, 2019

Today's call

- Describe Digital Bedrock's services (high level)
- Discuss cloud storage and how it intersects with Digital Bedrock
- Digital Bedrock's approach to fixity checks

Digital Bedrock

- Managed OAIS-compliant digital preservation **service**, built **specifically** for active preservation actions to support any type organization or individual.
- Founded from an observed need – many organizations didn't have budgets or staff to build or buy infrastructure to perform digital preservation in-house.
- No license subscription, no hardware – we do the work as your extended staff and infrastructure.
- Any format and content type, although specialize in audiovisual formats.
- No required metadata. Flexible database supports all types of data and metadata.
- Not a digital asset management or content management system. It's pure preservation.

Digital Bedrock's entrepreneurial founding

- Founder Linda Tadic has decades of experience in archives, libraries, media & entertainment, and the arts. Sample positions: Director of Operations, ARTstor; Director of the Media Archive & Peabody Awards Collection, University of Georgia; Manager Digital Library, HBO; Digital Projects Coordinator, Getty Research Institute.
- Adjunct professor, UCLA GSEIS; formerly taught in NYU's MIAP.
- Founding member & past president, Association of Moving Image Archivists (AMIA).
- MLIS: UC Berkeley. MFA: UC San Diego. BFA: California Institute of the Arts

What we do

- Verify clients' hashes (checksums) upon receipt, but also create and verify SHA-512 for ongoing scheduled SHA-512 fixity checks
- Extract extensive technical and embedded metadata (becomes indexed)
- Retain original directory structure for context; track file relationships
- Monitor obsolescence vulnerabilities (Digital Object Obsolescence Database; aka the "DOOD")
- Retain events audit trail
- Write three copies on LTO7 (LTFS), storing offline in geographically dispersed secure locations. Operations in an ISO 27001 compliant data center
- Migrate to future storage media
- Implement an open architecture (no vendor lock-in and easy exit)

Who we help

Clients are diversified, so our system must support all types of organizations and individuals, as well as all file format types.

While we can help any type organization, our current clients come from these categories:

media and entertainment

creators (artists, filmmakers, composers, photographers)

public broadcasters

libraries, archives, museums

law firms

Preserved to date (since January 2017)

- 10 million files (x3 = 30 million)
- 550 TB (x3 = 1.65+ PB written to tape)
- 100% accuracy

Smallest client: 3.18 TB

Largest client: 400 TB

Additional services related to digital preservation

- **Media migration:**

  Migrate files from legacy storage media (HDD, LTO2-6) to LTO7, LTO8*, HDD, or cloud

- **Software development:**

  Example: open source PBCore Cataloging Tool for WGBH

- **Consulting:**

  Digital preservation and collection assessments, metadata audits and schema development

* When LTO8 stock becomes available

Deep archive with object storage benefits and cloud connectivity

Deep archive:

Preserved files stored offline (3 copies geographically dispersed to secure storage)

Object storage benefits:

All technical and embedded metadata extracted from files and indexed for unstructured data searching

Cloud connectivity (optional):

Location in direct connect facility with 10GB connections to major cloud provider servers

Digital Bedrock's cloud storage connectivity

Our operations are in a high security data center in Los Angeles. This enables us to offer clients hybrid off-cloud managed digital preservation services with cloud connectivity/storage.

- The data center is a direct connect facility for US-West regions for:

  AWS, Microsoft Azure, Google Cloud, IBM Cloud

Benefit: secure, faster (10GB) connection than going over public internet. Data can be pushed up from DB's servers to the cloud provider, or pulled down from cloud to DB's servers for digital preservation or migration actions.

Cloud storage pros and cons

PROs

- Scalable storage
- Minimizes local infrastructure (costs and staff)
- Easy access to data (although egress costs)
- Facilitates collaboration through easy access
- Store files that don’t require frequent access on lower-tier storage (egress costs) – or not in the cloud at all

Cloud storage pros and cons

CONs

- Scalable storage up to a point (always consider costs and effort to exit any provider; the more data held by a provider, the harder it is to leave)
- Should replicate data in multiple regions (extra storage + transfer costs)
- Must still perform and manage digital preservation actions (fixity) and connect to your local systems/databases; this is just storage
- Beware when uploading data that the file “modification date” isn’t changed
- Security concerns (but any online storage has security concerns)

Cloud storage and the environment

Environmental impact of cloud storage (data center energy use)

… but consider this with ALL storage/infrastructure options

Greenpeace "*Clicking Clean*" report

https://www.greenpeace.org/international/publication/6826/clicking-clean-2017/

The Environmental Impact of Digital Preservation (December 2018 update)

https://www.digitalbedrock.com/resources-2/

General cloud hidden costs

- Your infrastructure/bandwidth to connect to cloud
- Egress (download/data out)
- Running fixity checks: do those count as “transactions”?
- Multiple regions for geographic dispersal
- “Snowball” and other similar sneakernet file delivery (47 lb device; consider shipping costs)
- Development costs to connect your systems to cloud storage (especially if hybrid on-premise and cloud; Azure and .NET)
- Costs to exit the provider

Cloud storage transfer costs in the Digital Bedrock hybrid scenario

Transfer costs are above the data storage costs.

Using Amazon Web Services (AWS) S3 as an example:

Our data center is in the US-West (Northern California) region.

AWS costs

- If a client's data is already in this region, the AWS cost to copy that data in to DB's servers is $0.00/GB. [regular: moving from AWS out to internet is $0.09/GB up to 10TB]
- If a client's data is not already in this region, the client pays AWS $0.02/GB to move the data from one region to US-West (Northern California)

Cloud storage transfer costs in the Digital Bedrock hybrid scenario

AWS costs (continued)

- If a client wants their data copied out from DB's servers to AWS in this same US-West (NoCal) region, the AWS cost is $0.00/GB.
- If a client wants to push their data out from US-West (NoCal) to another region in the US, Canada, or Europe, the AWS cost is $0.02/GB.

https://aws.amazon.com/directconnect/

https://aws.amazon.com/directconnect/pricing/

Regular S3 transfer: https://aws.amazon.com/s3/pricing/

Cloud storage transfer costs in the Digital Bedrock hybrid scenario

Digital Bedrock's transfer costs:

Transfer costs are above costs associated with any DB preservation or migration work.

$ 0.10/GB for the first 16 TB (16,000 GB) that touches DB's direct connect servers (paid once, whether in or out)

(eg, data moved down to DB's servers from AWS is $ 0.10/GB. If the data is then moved out to another region, no DB charge – only AWS)

$ 0.05/GB for data above 16 TB

Example transfer costs (not including DB services)

1 TB (1,000 GB)

Example 1: Data is stored in US-West (NoCal).

Moved to DB servers for digital preservation actions.

AWS	$ 0.00
DB	$100.00

TOTAL transfer: $100.00

Example transfer costs (not including DB services)

1 TB (1,000 GB)

Example 2: Data is stored in US-East (NoVa).

Moved to US-West (NoCal)

Moved to DB servers for digital preservation actions.

AWS NoVa to NoCal	$ 20.00
DB	$100.00
TOTAL transfer:	$120.00

Example transfer costs (not including DB services)

1 TB (1,000 GB)

Example 3: Data is sent via media to DB for preservation
After preservation, DB moves out to US-West (NoCal)

AWS	$ 0.00
DB	$100.00
TOTAL transfer:	$100.00

Digital Bedrock's approach to fixity checks

Digital Bedrock's off-cloud storage and fixity checks

- Client's files received on HDDs (we can loan), LTO, Aspera, or now can be received via cloud.
- SIPs built using our package creator tool (de-dupes, verifies client checksums, creates SHA-512, retains directory structure and modification dates, creates manifest, copies files to delivery media)
- SHA-512 checked upon ingest into the digital preservation system.
- After digital preservation processing, simultaneously written to 3 copies LTO7 using LTFS. Only the client's data is written to their tapes.
- SHA-512 verified after files are written, before storage. Once verified, files deleted from servers (security and small carbon footprint)

Digital Bedrock's off-cloud storage and fixity checks

- **Copy A:** stored in safe in our office in the data center (requires passing through 6 levels of biometric security to even get to our office)
- **Copy B:** locked in Turtle case (only the client's LTO tapes) with no visible client identification. Stored in 2nd location 7 miles away in facility built as secure media vault storage (only we have access to our vault). Requires 4 levels of coded access to get to our vault.
- **Copy 3**: locked in Turtle case (same as Copy B). Shipped via FedEx to secure storage 3,200 miles away at a partner organization, and placed in our dedicated vault. Protocols in place to track delivery. Third party doesn't know who our clients are, or what is in the Turtles.

Digital Bedrock's off-cloud storage and fixity checks

Scheduled fixity checks: For our first two years, ran fixity every 6 months on our local LTO7.

After 8M files and no errors, we changed our policy to annual checks.

No need to run a tape through the "read" cycle so frequently, so long as the tapes were verified throughout the writing process, and are stored in a stable environment with no use. We find LTO7 (barium ferrite magnetic particles) robust.

Our software

- **Package creator tool** (client selects files; tool de-dupes & packages files with file-level SHA-512 checksums, retaining original context)
- **Digital Preservation Application (DPA)**
- **Client portal**
- **LTFS software** to manage the tape libraries
- **Digital Object Obsolescence Database (DOOD)**

Patent filed July 2017 for DIGITAL OBSOLESCENCE AVOIDANCE SYSTEMS AND METHODS, the key algorithm behind the DOOD. The DOOD tracks and monitors format vulnerabilities through research and proprietary algorithms.

Contact:

Linda Tadic
ltadic@digitalbedrock.com
(888) 938-7386 x700
www.digitalbedrock.com