Audiovisual Digital Preservation and DNA Storage Requirements

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

JPEG DNA Workshop
2021-04-09

Hard numbers

By 2025:

The amount of data being **stored** globally from all sectors will reach 175 zettabytes.*

Half of that data will be stored in the public cloud.

Mass data migrations from obsolete generation data tape and older drives will contribute to the data storage tsunami.

* *Data Age 2025: The Digitization of the World – From Edge to Core.* IDC, 2020.

Audiovisual data in this tsunami

Audiovisual data is created and/or held by:

- Media & entertainment companies
- Cultural heritage organizations (archives, libraries)
- Museums
- Creatives (artists, photographers, musicians)
- Corporations (marketing materials, event/meeting documentation)
- Governments
- Law enforcement (police body camera videos)
- Surveillance video

Audiovisual data in this tsunami

Feature film masters: 26 – 40 TB per film

graded and ungraded picture: DPX, EXR, TIFF (DCDM), DCP, IMP

visual effects

audio printmasters, stems

project files (Avid and ProTools)

Television (10-episode season): 86 TB per season

graded and ungraded picture: DPX

visual effects

audio printmasters, stems

Audiovisual data in this tsunami

Museums/Artists Time-Based Media Art (TBMA): storage varies

disk images

original files used in creating artwork

final works

software/apps

Libraries and Archives: storage varies (reformatted & born-digital)

newspapers, journals, photographs, video, film, audio

Audiovisual data in this tsunami

Corporations: storage varies

marketing videos, images

event documentation (meetings)

Governments (local, state, federal): storage varies

training videos

event documentation (meetings)

Law enforcement: storage varies (evidentiary data retained for more than 1 year)

Body-Worn Camera video

911 calls

DNA storage as archival storage

Not all 175 zettabytes need to be in active storage.

The majority of that data can be stored “offline” in archival storage.

DNA storage could be part of an HSM strategy as the “Tier 3” storage medium.

Core digital preservation actions beyond storage

But digital preservation is more than storage. Storage must <u>support</u> and <u>integrate</u> with core digital preservation actions.

Digital preservation is an ongoing, managed process.

Core digital preservation actions beyond storage

1. Obsolescence. The storage media might be viable in 500 years, but the data stored on it might not be.

Digital object obsolescence vulnerabilities and dependencies must be monitored:

1. format
2. codec
3. software
4. operating system
5. hardware

Core digital preservation actions beyond storage

1. **Obsolescence**. The storage media might be viable in 500 years, but the data stored on it might not be.

"At Los Alamos, we have some of the oldest digital-only data and largest stores of data, starting from the 1940s," [Bradley] Settlemyer said. "It still has tremendous value. Because we keep data forever, we've been at the tip of the spear for a long time when it comes to finding a cold-storage solution."

https://www.storagenewsletter.com/2021/04/08/from-los-alamos-national-laboratory-translation-software-enables-efficient-dna-storage/

Core digital preservation actions beyond storage

Because of obsolescence factors, digital objects/files could be transcoded over time into more viable formats.

Archivists need to:

1. retrieve specific obsolete files stored on DNA (random access)
2. transcode them
3. copy new versions to DNA

…. tracking all actions and relationships in a digital preservation system

Core digital preservation actions beyond storage

2. Monitoring bit health through scheduled **fixity checks**.

How will DNA storage impact this core principle?

(digital preservationists are a suspicious species….)

Error detection and validation while encoding needs to be flawless.

Core digital preservation actions beyond storage

3. Manage geographically dispersed **redundant copies**

Usually, data is migrated to new media (refreshed) over time as storage infrastructure is deprecated.

How will this be performed with DNA storage?

What Digital Bedrock does and where DNA storage could fit in our architecture

- Managed OAIS-compliant digital preservation **service**, built **specifically** for active preservation actions to support any type organization or individual.
- Files sent to storage **after** we perform intensive preservation processing. Our primary work is in the processing before storage, and maintenance after storage.
- The target storage can be agnostic (LTO, cloud, DNA?), but our current SOP is to 3 copies LTO7 (LTO9 in 2022), securely geographically dispersed. Some clients request a copy also pushed to cloud after processing.
- No license subscription, no hardware – we do the work as our clients' extended staff and infrastructure.
- Any format and content type, although specialize in audiovisual formats.

What Digital Bedrock does and where DNA storage could fit

- Verify clients' hashes (checksums) upon receipt, but also create and verify SHA-512 for ongoing annual scheduled SHA-512 fixity checks
- Extract extensive technical and embedded metadata (becomes indexed to access unstructured data in our digital preservation system and client's portal)
- Retain original directory structure for context; track file relationships
- Monitor obsolescence vulnerabilities (Digital Object Obsolescence Database; aka the "DOOD"; US patent 10592674)
- Perform annual fixity checks on data
- Retain audit trail of all actions
- Write three copies on LTO (LTFS), validate each copy, and store them offline in geographically dispersed secure locations.
- Migrate to future storage media
- Implement an open architecture (**no vendor lock-in and easy exit**)

DNA storage as archival storage

What are the requirements for DNA storage to be used as an archival storage medium in an overall digital preservation strategy?

Minimum Requirements for DNA Storage

1. **Interoperability** with other DNA write/read systems (LTFS model).

   DNA storage cannot be used for digital preservation until this is resolved.

2. **Portability** so data storage vendors can write/read in-house.

3. **Single file retrieval/indexing (random access)**. Need to be able to retrieve/restore one specific file, not a 2 PB string of data just to get to a 1 MB file. The DNA storage needs the equivalent of an indexing system in the storage so the files are indexed in the strand for retrieval.

Minimum Requirements for DNA Storage

4. **Cost** shouldn't be more than what consumers are accustomed to spending today for a physical storage medium. (Service is different)

   Consumer expectations: costs should be lower since the medium will be "write once forever" (no migration/refreshing costs), even if the technology write/read costs can be expensive.

5. **Write/read time** must at a minimum be 1 Gbps

6. **Form factor**. Form factor must be consistent across systems.

7. **Storage environment.** Must be reasonable and not require cold storage.

Minimum Requirements for DNA Storage

8. <u>**Lossless encoding.**</u> This is not optional for preservation applications.

9. <u>**Error detection and correction**</u>. If scheduled fixity checks are obviated, need proof data was encoded with no errors.

10. <u>**Green technology**</u>

    - Physical media should not be made from rare/heavy earth materials.

        Can it be made without silica sand?

    - Write/read processes should use minimal energy.

    - Physical storage should not require ultra-cold environments.

Contact

Linda Tadic

Founder/CEO

ltadic@digitalbedrock.com

www.digitalbedrock.com