Assessing and Mitigating Bit-Level Preservation Risks, and NDSA Infrastructure Working Group Panel Discussion
DigitalPreservation 2012 Session
July 25, 2012
4:15 – 5:15 pm

Presenters: John Spencer, BMS/Chace LLC; Priscilla Caplan, Florida Virtual Campus; Andrea Goethals, Harvard University; Micah Altman, MIT Libraries

Attendees: 38

Highlights
The information presented here was informed by a survey sent out to NDSA members on bit-level preservation and storage.

Bit Threat

- Physical and hardware
- Insider and external attacks
- Media
- Organizational failure
- Software
- Curatorial error

Encoding

- Compression
    - Format-based file compression (JPEG2000)
    - Tape hardware compression at the drive
    - NAS compression via appliance or storage device
    - Data de-duplication
- Compression tradeoffs
    - Space savings allows more copies at same cost
    - But makes files more sensitive to data corruption

Encryption

- Two contexts
    - Archiving encrypted content
    - Archive encrypting content
- Reasons to encrypt
    - Prevent unauthorized access
    - To enforce DRM
    - Legal requirements (HIPAA, state law)
- Concerns
    - Increased file size
    - Performance penalty
    - Additional expense
    - Makes files more sensitive to data corruption
    - Complicates format migration
    - Complicates legitimate access

    - Risk of loss of encryption keys
    - Difficulty of enterprise level key management
    - Obsolescence of encryption formats
    - Obsolescence of PKI infrastructure

Mitigating risks

- Redundancy (multiple duplicates)
- Diversity (variations)
- Likely candidates for failure
    - Storage component faults
    - Organizational disruptions
- Bit-Level Fixity
    - Process: implementing fixity checks into ingest and migration workflows
    - Product: creating a duplicate copy and making sure the bits are the same in each
- Auditing and repair
    - Fixity mitigates risk only if you use it to audit
    - Functions of auditing
        - Detect
        - Verify
        - Repair
    - Audit design choices
        - Audit regularity and coverage
        - Fixity check
        - Auditing scope
    - Auditing mitigates risk only if you use it to repair
    - Reviewed auditing systems for DuraCloud, iRODS, and SafeArchive

Discussion

- The size of the collection and the size of the file and the complexity of the file can factor into determining how often one should run fixity checks.
- Typically people don't measure the fixity of data on discs, so there's not a lot of literature on that.
- The safer you want to be, the more it's going to cost you. Random checking is better than having a regular schedule. Systematic random sampling, as in every quarter you'll do a check on a third of the collection, works well.

Action Items

The discussion focused on ways to make the matrix better:

- Technology obsolescence needs most work. It is currently two things together—media and software. What are the things you depend on to store the stuff? Perhaps change it to infrastructure obsolescence?
- Storage—different administration and technology stacks is missing?

- Online vs offline is not specified. If one of the two copies is offline, does that count as Level 1? Some become dependent on other levels; can't do file fixity without being online.
- Glossary would be helpful, with links to tools and other resources to get more information.
- Add rights protection into data integrity?
- Move transactions in fixity to Level Three? How do we define transactions? Migration action, changing content, or metadata.
- For information services, levels one and two could be policies, levels three and four could be enforcement of policies.
- Audit of logs in level 4 for information security. Combine levels 1 and 2.