Digital Preservation 2013

Bill Ying

CIO/VP of Technology

ARTstor

ARTstor Shared Shelf Preservation Plan Based on the NDSA Levels of Digital Preservation

- This is the story of ARTstor's attempt to implement a preservation plan for Shared Shelf (a media management software) using the NDSA Levels of Digital Preservation document as a guide line.

ARTstor

- The ARTstor Digital Library is a nonprofit resource that provides over 1.6 million digital images in the arts, architecture, humanities, and sciences with an accessible suite of software tools for teaching and research. Our community-built collections comprise contributions from outstanding international museums, photographers, libraries, scholars, photo archives, and artists and artists' estates.

ARTstor By the Numbers

- Subscribers
  - United States Institutions 1,222
  - International Institutions 284
  - All Institutions 1,506
  - Countries/Regions 45

- Registered Users
  - Student 419,322
  - Instructor 31,605

- Images Available by Region
  - United States 1,627,829
  - International 1,355,368

- Collections Online 216
- Collection Contributors 206

Shared Shelf Overview

- Shared Shelf media management software enables institutions to manage, store, use, and publish their institutional and faculty media collections within their institution or publicly on the Web.
  - **Cataloging tools**
  - **Vocabulary warehouse**
  - **Digital asset storage**
  - **Publishing and export tools**

Shared Shelf By the Number

- Subscribers
  - United States Institutions 73
  - International Institutions 6
  - All Institutions 79
  - Countries/Regions 3

- Contents
  - Collections 334
  - Objects in Projects 1,164,502
  - Image 1,163,529
  - QTVR Video 85
  - Audio 39
  - document 360
  - Video 489

Preservation

- We want to provide “Preservation” service to our Shared Shelf users
- So we decide to “try to” implement the NDSA Preservation guide line.
- Here is our story.

NDSA Levels of Digital Preservation

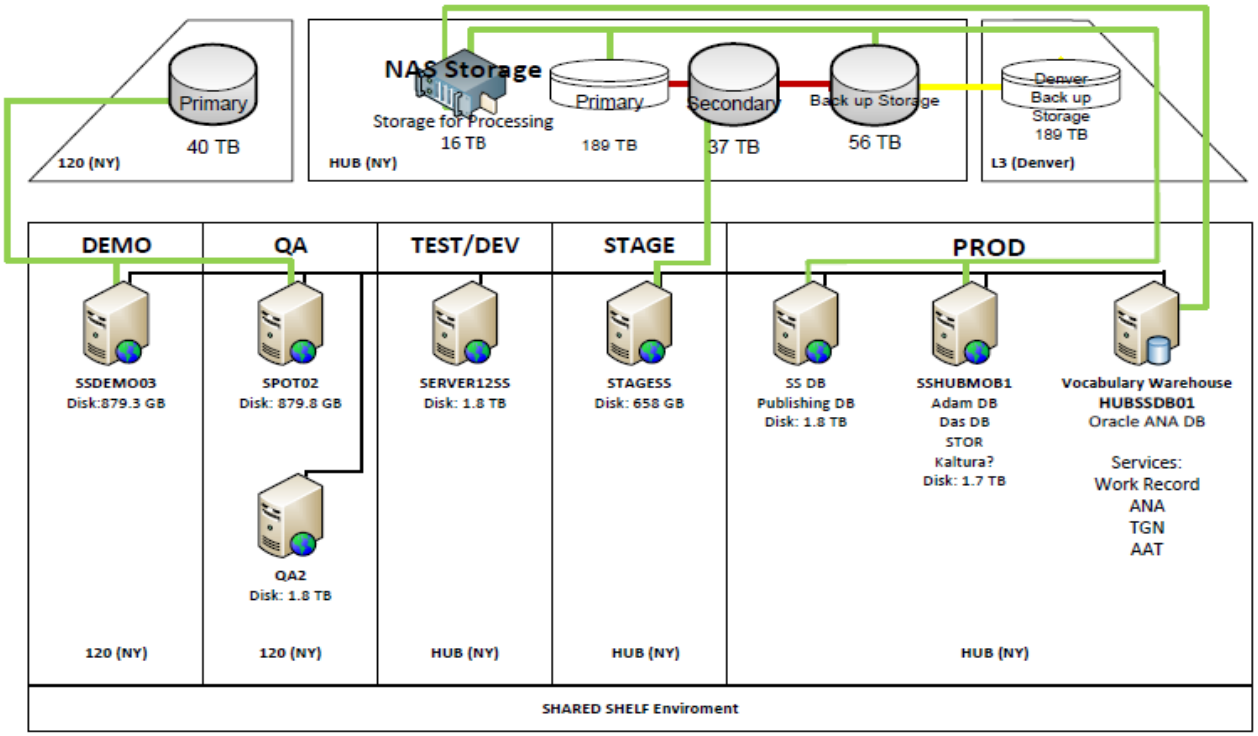
	Level One (Protect Your Data)	**Level Two (Know Your data)**	**Level Three (Monitor Your Data)**	**Level Four (Repair Your Data)**
Storage and Geographic Location	Two complete copies that are not collocated For data coming in on heterogeneous media (optical disks, hard drives, floppies) get the digital content off the medium and into your storage system	Three complete copies At least one copy in a different geographic location Document your storage system(s) and storage media and what you need to use them	At least one copy in a geographic location with a different disaster threat Start an obsolescence monitoring process for your storage system(s) and media	All copies in geographic locations with different disaster threats Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems.
File Fixity and Data Integrity	Check fixity on ingest if it has been provided with the content Create fixity info if it wasn't provided	Check fixity on all ingests Use write-blockers when working with original media Virus-check high risk content	Check fixity on all transformative acts Check fixity of sample files/media at fixed intervals Maintain logs of fixity info; supply audit on demand Ability to detect corrupt data Virus-check all content	Check fixity of all content in response to specific events or activities Ability to replace corrupted data
Information Security	Identify who has read, write, move, and delete authorization to individual files Restrict who has those authorizations to individual files		Maintain logs of who has accessed individual files	Maintain logs of who performed what actions on files, including deletions and preservation actions Perform audit of logs
Metadata	Inventory of content and its storage location Ensure backup and non-collocation of inventory	Store administrative metadata Store transformative metadata and log events	Store standard technical and descriptive metadata	Store standard preservation metadata
File Formats	Encourage use of limited set of known and open file formats and codecs	Inventory of file formats in use	Validate files against their file formats Monitor file format obsolescence threats	Perform format migrations, emulation and similar activities

1. Storage and Geographic Location: Level One (Protect Your Data)

- Two complete copies that are not collocated
  - Yes, see diagram next slide
- For data on heterogeneous media (optical disks, hard drives, etc.) get the content off the medium and into your storage system
  - Yes, as soon as possible

Shared Shelf Infrastructure

6-4-13



LEGEND
- DB and image file storage
- Automated backup
- Remote replication

All SS servers are running on Linux
DB and Image files are stored with the same methods

2. File Fixity and Data Integrity: Level One (Protect Your Data)

- Check file fixity on ingest if it has been provided with the content
  - So far, no fixity file was provided for all content ingestion
  - Finally, we have a “partner” that store their digital asset in Amazon S3. So when we request to download the image, we can also request a MD5sum. We can finally do fixity during ingestion!

3. Information Security: Level One (Protect Your Data)

- Identify who has read, write, move, and delete authorization to individual files:
  - Yes, we provide all these on Role based by collection level.
- Restrict who has those authorizations to individual files:
  - We cannot restrict by individual file level

4. Metadata: Level One (Protect Your Data)

- Inventory of content and its storage location:
  - We have a database that store all these information
- Ensure backup and non-collocation of inventory:
  - These data are all backup remotely just like the actual digital asset

5. File Formats: Level One (Protect Your Data)

- When you can give input into the creation of digital files encourage use of a limited set of known open file formats and codecs:
  - We accept almost anything!
  - Shared Shelf currently supports:
    - Image types: .png, .jpg, .jpeg, .tif, .tiff, .mov(qtvr), .jp2
    - Video types: .asf, .qt, .mov, .mpg, .mpeg, .avi, .wmv, .mp3, .mp4, .m4v, .3gp
    - Document types: .doc, .ppt, .xls, .pdf, .docx, .pptx, .xlsx

6. Storage and Geographic Location: Level Two (Know Your data)

- At least three complete copies:
  - Yes, we actually have four copies as I explained earlier in the diagram
- Document your storage system(s) and storage media and what you need to use them:
  - I am sure we do not do enough “documentation”

7. File Fixity and Data Integrity: Level Two (Know Your Data)

- Check fixity on all ingests
  - See answer before
- Use write-blockers when working with original media:
  - We try, but not all the time! But we handle all disk drives very carefully.
- Virus-check high risk content
  - Not yet! But we are looking at three options:
    - Client digital asset upload tool (check pre-upload)
    - After upload; server check in real time (ClamAV)
    - Asynchronous background virus check later

8. Information Security: Level Two (Know Your Data)

- Document access restrictions for content:
  - Obviously we need to provide a lot more documentation which we are not good at!

9. Metadata: Level Two (Know Your Data)

- Store administrative metadata:
  - We have some admin metadata and also looking at
    - DC admin metadata: http://dublincore.org/groups/admin/
    - METS admin metadata: http://www.loc.gov/standards/mets/METSOverview.v2.html
  - We want to hear from other what they are storing?
- Store transformative metadata and log events:
  - We are doing all kinds of transformation; but is not doing a good job keeping track of them! We do have logs!

10. File Formats: Level Two (Know Your Data)

- Inventory of file formats in use:
  - Yes, we do that.

11. Storage and Geographic Location: Level Three (Monitor Your data)

- At least one copy in a geographic location with a different disaster threat:
  - We are ready for another “Super Storm: Sandy”
- Obsolescence monitoring process for your storage system(s) and media:
  - We are saying “YES”! But are we really ready to migrate all the digital media to a new format? How long would it takes?

12. File Fixity and Data Integrity: Level Three (Monitor Your Data)

- Check fixity of content at fixed intervals:
  - We are now doing a complete fixity check on all the digital media using the staging server with the “secondary” copy of the media files every three months. So far we have not discovered any “corrupted” files.
- Maintain logs of fixity info; supply audit on demand:
  - Yes, we have logs in our databases
- Ability to detect corrupt data:
  - Does it mean “metadata”?
- Virus-check all content:
  - We are still in step 7. We are not yet there!

13. Information Security: Level Three (Monitor Your Data)

- Maintain logs of who performed what actions on files, including deletions and preservation actions:
  - We have some!

14. Metadata: Level Three (Monitor Your Data)

- Store standard technical and descriptive metadata:
  - Descriptive is easy, we do it all the time.
  - What Technical metadata?
    - NISO/CLIR/RLG: Technical Metadata Elements for Images Workshop
    - http://www.niso.org/news/events/niso/past/image/
    - NISO Metadata for Images in XML (NISO MIX)
    - http://www.loc.gov/standards/mix/

15. File Formats: Level Three (Monitor Your Data)

- Monitor file format obsolescence issues:
  - Of course I am going to say « YES »!
  - Do you think I am going to say « NO »?

16. Level 4 (Repair your data)

- Run out of time to even think about these!