DESIGNING STORAGE ARCHITECTURES FOR PRESERVATION COLLECTIONS

Designing Storage Architecture for Digital Collections

LIBRARY OF CONGRESS

Survey Questionnaire Responses

Lunch time presentation: NDSA Infrastructure Working Group Storage Survey

- The NDSA Infrastructure Survey, conducted between August 2011 and November 2011, received responses from 58 members of the 74 NDSA member organizations who are preserving digital content. The goal of this survey was to get a snapshot of current storage practices within the organizations of the National Digital Stewardship Alliance. These organizations included consortia groups, professional organizations, university departments, funders and vendors.
- The following questions on and about the survey were provided to the Designing Storage Architecture for Preservation Collections participants
  - What are the biggest challenges you see for Storage Architectures for Digital Preservation over the next five years?
  - What do you see as under-tapped or untapped opportunities for meeting these challenges?
  - What question (s) would you like to see added to the Storage Survey conducted by the National Digital Stewardship Alliance for the next time it is conducted?
  - Do you have any suggestions for the Levels of Preservation document? Do you think the document accurately provides advice on mitigating risks of loss?
- The subsequent slides provide responses from the participants.
- (note: see http://blogs.loc.gov/digitalpreservation/2012/08/prognosticating-digital-preservation-infrastructure-finals-results-from-the-ndsa-storage-survey/ for results of survey)
- (note: see http://blogs.loc.gov/digitalpreservation/2012/09/help-define-levels-for-digital-preservation-request-for-public-comments/ for the levels of preservation blog post)

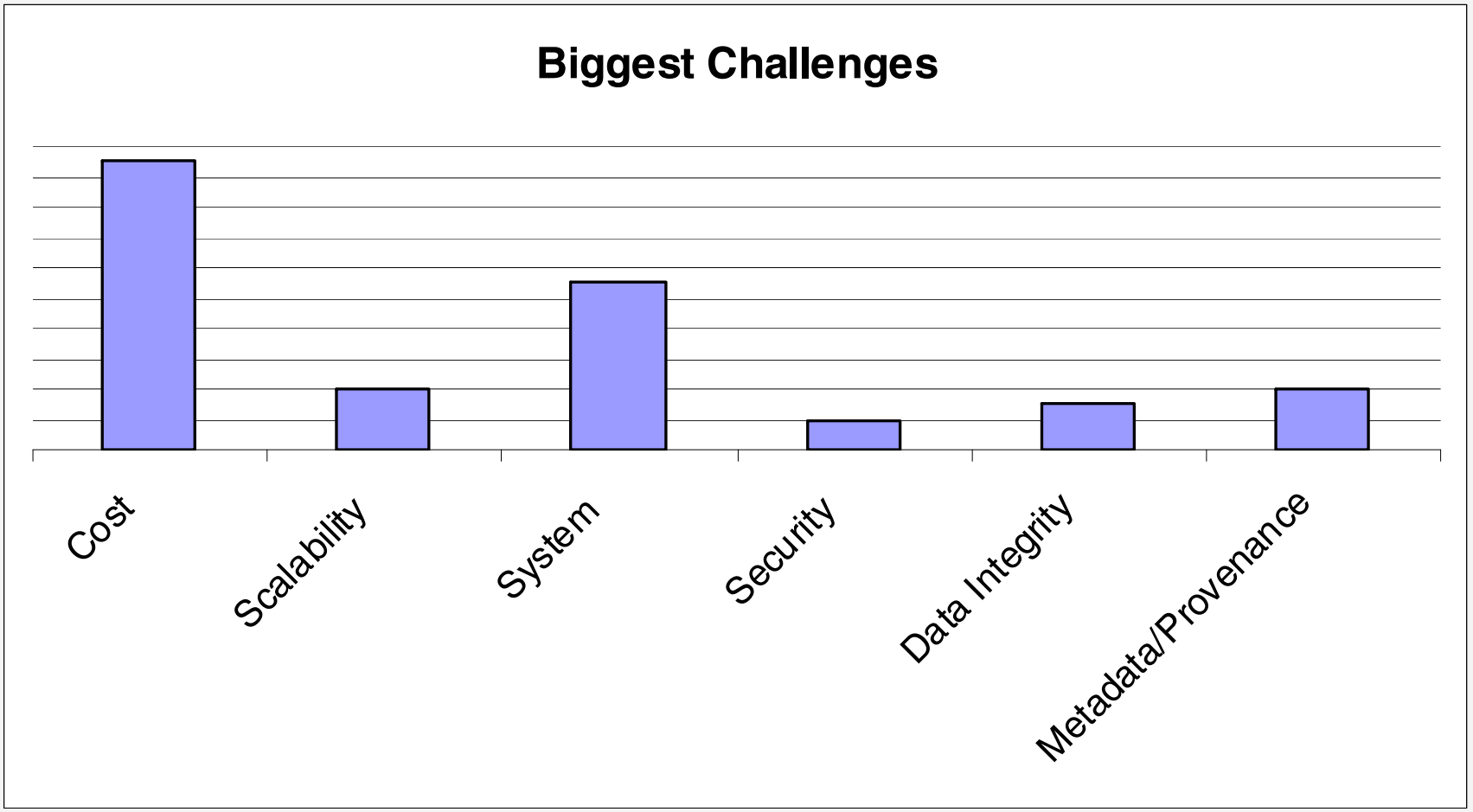
What are the biggest challenges you see for Storage Architectures for Digital Preservation over the next five years?

- Cost
  - Cost and space
  - Large data, video, audio.
  - Capacity needs vs cost. The tension of meeting capacity needs in a cost effective, sustainable way given that Moore's Law no longer applies to this area.
  - Cost-everyone wants everything stored forever.
  - 3-D scanning will also increase.
  - Increasing digitization for video and sound files will push demand on storage needs. All of which will need indefinite period of storage.
  - Understanding (as much as possible) how to pay upfront at content creation time for its long term storage.
  - Budget
  - Dealing with the competition for limited resources-space, processing power, personnel, equipment.
  - Tremendous growth in borne-digital content and the shrinking budget for digital preservation.
  - Human Capital
  - Metadata management-the ability to reference and cross-reference data to queries in mass storage preservation will drive massive amounts of extra storage.
  - Space Requirements
  - Budgetary- are research libraries and the like going to be able to support the business?
  - Scale/cost to institutions that "can't afford it"
  - Not storing "junk" or copies of copies of copies.
  - Costs and security, validity of content as archives expand and threats increase.
- Metadata/Provenance
  - Integrating provenance in a meaningful way.
  - Provenance
  - All have attached metadata so the preservation system will rely on inputted data.
  - Long term preservation and provenance – not yet built into commercial systems in any end-to-end fashion.

What are the biggest challenges you see for Storage Architectures for Digital Preservation over the next five years?

- Scalability
  - Scaling to meet needs.
  - Data growing faster than the archives' ability to hold and serve it.
  - Scale – size of data and number of objects affects management, migration, etc.
  - Retrieving data. We have gotten very good at storing data but it is getting harder to find what was stored and get it back.
- Data integrity
  - Data integrity.
  - How do we deal with data errors from an institutional policy standpoint.
  - Built in fixity checks.
- System
  - Costs and security, validity of content as archives expand and threats increase.
  - Software compatibility between different systems.
  - Connectivity limitations when leveraging cloud services.
  - Use of tapes without concern of reliability of media.
  - Understanding impact of storage technology changes.
  - Scalable, distributive search that isn't centralized (search media). Preservation of software used to create the data/interfaces/equipment of copy everything forward.
  - Ease of use for digital curators and non-technical users
  - Transparency/simplicity of internals.
  - Migration > ease of migrating to a digital format from existing music, art, literature, etc. Scale and retrieval speed.
  - Standardization
  - Storage formats
- Security
  - Key management if PKI technologies are leveraged.
  - Security that is "future-proof"

What are the biggest challenges you see for Storage Architectures for Digital Preservation over the next five years?



What do you see as under-tapped or untapped opportunities for meeting these challenges? (page 1 of 2)

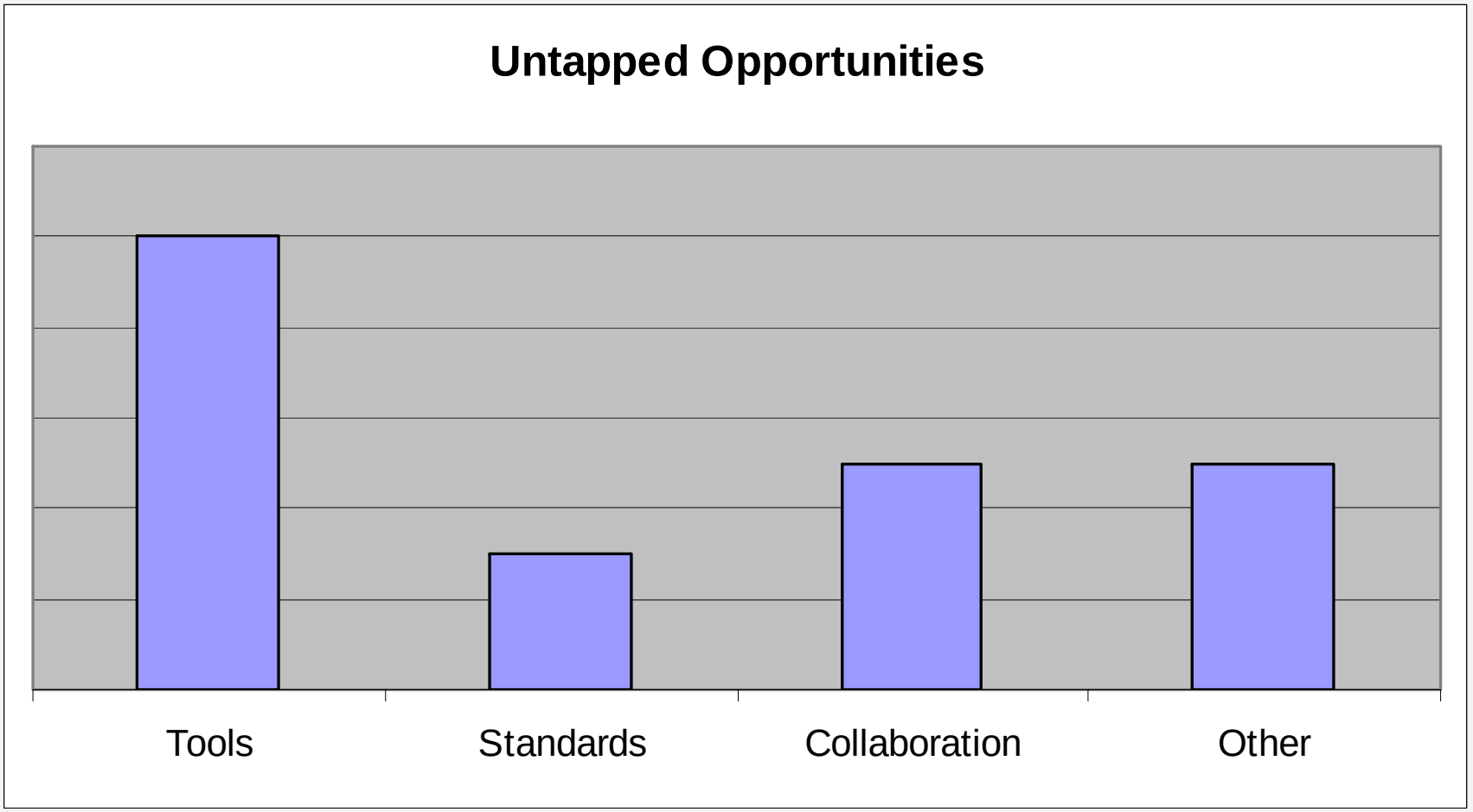
- Tools
  - Erasure coding technologies should remove the cost factor as we move beyond raid. This technology may also improve the durability of the archive. Unfortunately the threats will always exist as long as there are humans with differing thoughts.
  - Tools that allow less-skilled staff to do some of the work.
  - High performance file systems.
  - Cloud solutions (though security is always an issue).
  - Integration with mobile devices that are easily uploaded to cloud server.
  - Leverage erasure coding rather than replication.
  - Long-term preservation on tape using LTFS.
  - Need tools to generate statistically valid random samples of large datasets.
  - Specialized tools for specialized work.
  - Extensions to today's content management systems could begin to more fully meet these preservation challenges
- Collaboration
  - Collaboration with other repositories, like DPN.
  - Increased collaboration between CS researchers and preservationists.
  - More consortiums, group meetings.
  - Consortial and collaborative storage infrastructures
  - Engage the open source community

What do you see as under-tapped or untapped opportunities for meeting these challenges? (page 2 of 2)

- Standards
  - Open discussion of the allowable tolerance for errors in digital storage.
  - Willingness of top tier storage vendors to support standards established by the community to support digital preservation.
  - New preservation standards are in place to make preserved files metadata easy to add on to.

- Other
  - Need more data curation, curators with expert knowledge need to be able to sample and explore data.
  - For some metadata, crowd-sourcing is key.
  - Academia governed and sources cloud storage
  - Existing work in provenance in workflow systems and provenance in general purpose provenance libraries.
  - Leverage erasure coding rather than replication.

What do you see as under-tapped or untapped opportunities for meeting these challenges



What question (s) would you like to see added to the Storage Survey conducted by the National Digital Stewardship Alliance for the next time it is conducted?

- What sort of access do you provide for your archive:
  - open access over the internet w/o authentication?
  - access to registered users for free?
  - access for $?
  - closed access limited to my organization or users?
- I'd like to know whether organizations think that they need a full TRAC or ISO16363 repository, or whether they think something less would suffice,
- Also whether they think that their organization is capable of meeting the TRAC/ISO16363 requirements on what kind of timeline.
- What resources (Money) are available that are still obscure?
- Something about collection retention decision framework/collection overlap.
- Digital security requirements of archives and repositories.
- Do you allow users to query your archive?
- Do they have the ability to query all the metadata?
- Many questions discussed the presence of a plan, policy, etc. A good follow-up would be the level of confidence and satisfaction, particularly with real world tests of e.g. failure, recovery, etc.
- Distinction between active and passive fixity checks is important (i.e. can I get bad data?)
- Private/internal cloud as an option distinct from commercial.

Do you have any suggestions for the Levels of Preservation document? Do you think the document accurately provides advice on mitigating risks of loss? (page 1 of 2)

- "Repair your data" is confusing. A different phrase should be used. Makes me think that the server will spell check my documents and remove viruses.
- Fill in the blank spot in the grid with something.
- Figure out how curation and processing or transformation of data fits into this grid- maybe it's a new level between know/monitor or maybe it's a type of monitoring.
- Provenance
- End to end integrity
- Not clear how much of the levels and processes around them can be fully automated.
- More on data security could be added.
- Might want to consider technical heterogeneity (storage systems hardware/software) and administrative heterogeneity (no one person being able to admin all of the data) as storage librarian.
- What does it mean to check fixity on transformative acts?

Do you have any suggestions for the Levels of Preservation document? Do you think the document accurately provides advice on mitigating risks of loss? (slide 2 of 2)

- What does performing auditing on logs mean?
- What is the distinction between administrative, transformative, technical and preservation metadata?
- Staffing should be part of it, how many people are required to run an archive of specific magnitude.
- Re: # of copies, copies on the “best” storage you can afford that is fit for the purpose.
- Explicit mention of asset cataloging.
- How do we deal with the data loss/content loss continuum?
- What technologies are you using to reduce your storage costs, i.e. compression, data deduplication, etc.
- What percentage of your content has an associated SLA or some question to determine how many people have content that could expire and if there is a common acceptable internal review survey 3, 5, 7 or 10 years?
- What to do with data integrity loss?