SNIA 100-Year Archive Survey 2017

Thomas Rivera

Co-Chair, SNIA Data Protection & Capacity Optimization Committee

Participating SNIA Groups

- Long-term Retention Technical Working Group (LTR TWG)
  - SNIA's LTR TWG developed a SNIA standard for a logical container format called the Self-contained Information Retention Format (SIRF)
    - This new standard enables long-term hard disk, cloud, and tape-based containers a way to effectively & efficiently preserve and secure digital information for many decades, even with the ever-changing technology landscape
  - For more info on LTR TWG: http://www.snia.org/ltr
- Data Protection & Capacity Optimization (DPCO) Committee
  - The mission of the DPCO is to foster the growth and success of the market for data protection and capacity optimization technologies
  - For more info on the DPCO Committee: https://www.snia.org/dpco

The Need for Digital Preservation

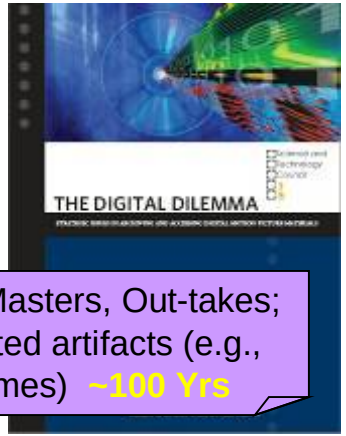
- Regulatory compliance and legal issues
  - Sarbanes-Oxley, HIPAA, FRCP, GDPR, intellectual property litigation
- Emerging web services and applications
  - Email, photo sharing, web site archives, social networks, blogs
- Many other fixed-content repositories
  - Scientific data, intelligence, libraries, movies, music
- Internet of things
  - Creating large sensory datasets

Scientific & Cultural

Satellite data is kept **for ever**

Digital art is kept **for ever**

Media & Entertainment



Film Masters, Out-takes; Related artifacts (e.g., games) **~100 Yrs**

Healthcare

X-rays often stored for periods of **75 yrs**

Records of minors are needed until age **20 to 43 yrs**

Extreme Examples

- Defense Innovation Unit Experimental (DIUX)
  - Orbital Insight - Synthetic Aperture Radar (SAR) micro-satellites have enabled daily imaging of the entire Earth.
- CERN
  - On 29 June 2017, the CERN DC passed the milestone of 200 petabytes of data permanently archived in its tape libraries.…Particles collide in the Large Hadron Collider (LHC) detectors approximately 1 billion times per second, generating about one petabyte of collision data per second.

Goals of Digital Preservation

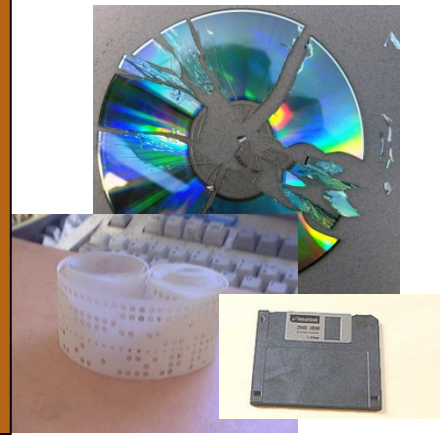
- Digital assets stored now should remain
  - Accessible
  - Undamaged
  - Usable
- For as long as desired – beyond the lifetime of
  - Any particular storage system
  - Any particular storage technology
- And at an ***affordable cost***

Threats to Long-term Assets

- Large-scale disaster
- Human error
- Media faults

- Component faults
- Economic faults
- Attack
- Organizational faults

Long-term content suffers from more threats than short-term content



- Media/hardware obsolescence
- Software/format obsolescence
- Lost context/metadata

So what has changed in a decade?

Emerging Changes

- Growth of the cloud
- Growth of SaaS preservation services
- Continued growth of data volume
- Migration from Content Addressable Storage Object & WORM NAS
- Assessing the continued viability of tape & storage mediums
- Emergence of preservation strategies and formats
- Growth in the need (and regulation) for security and privacy

Drivers for Cloud Growth

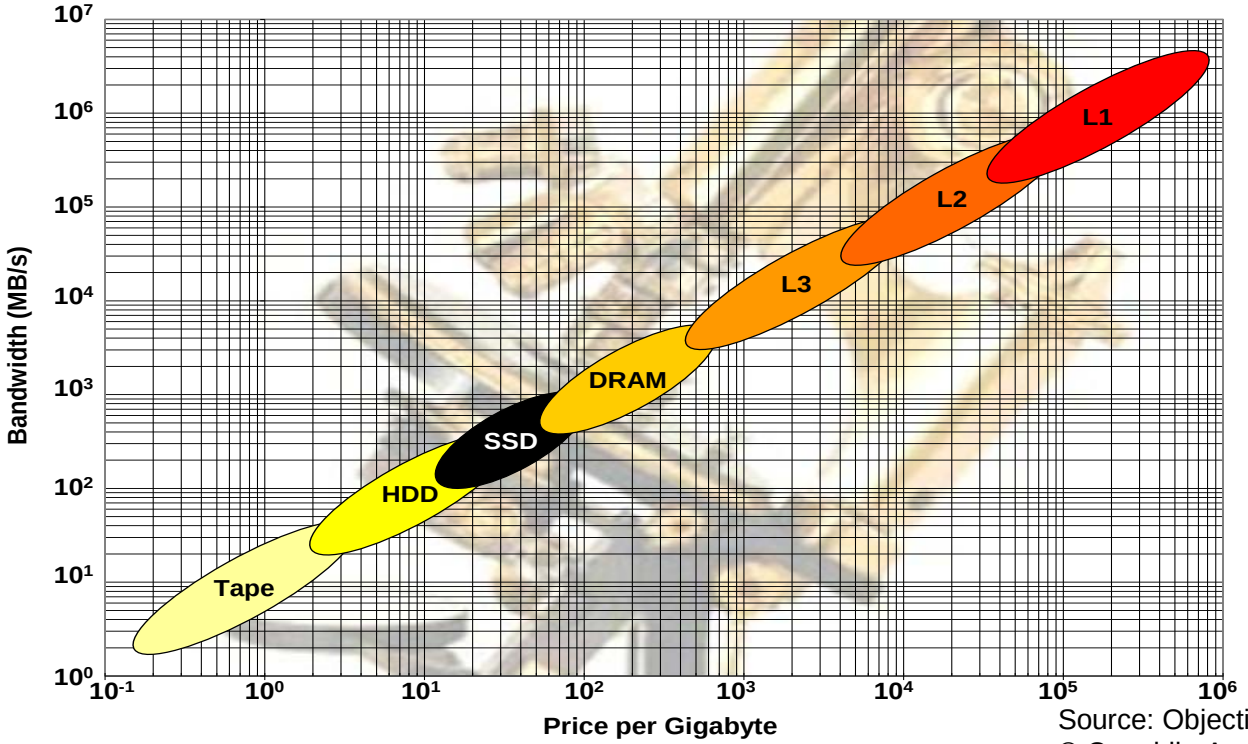
- “Cloud compliance requirements, enforcing mobility, and uninterrupted business continuity are forecast to steer the cloud storage market”
- “Data archiving solution type is anticipated to grow at the highest compound annual growth rate (CAGR) during the outlook period”

Source: Cloud Storage Market: Global Forecast until 2022; Feb 2018; https://www.reportlinker.com/p04141095/Cloud-Storage-Market-by-Solution-Service-Deployment-Model-Organization-Size-Vertical-Region-Global-Forecast-to.html

The Flash Revolution

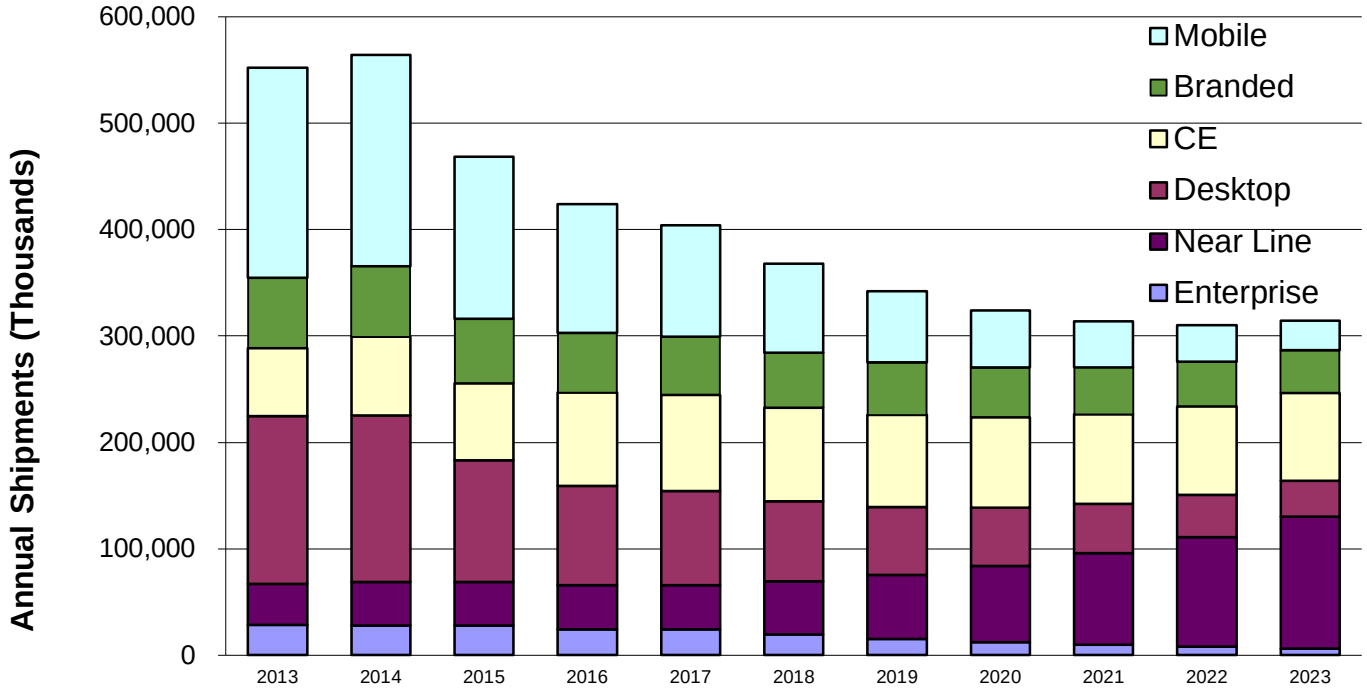
- Flash has changed the cost model of spinning disks
- Data processing speeds accelerating
  - Leads to more analytical models and datasets
  - Data becomes more transitory
- Archiving in an all-flash data centre?
- Is tape dead? Is disk dead?

Memory and Storage Technologies by Price per Gb and performance



Source: Objective Analysis, 2018

Shipped Disk Drive Units vs Time by Application



Source: Coughlin Associates

Data Preservation and cyber security

- Organizations now have the challenge of their data being hacked and how to secure long term data assets
- General Data Protection Regulation (GDPR)
  - Art. 17 - Right to erasure ('right to be forgotten'); how to remove long term data

The development of Cloud as an archive

- “Tape storage has been declared dead so many times that it has become a trope in technology journalism. The truth is more complicated and is ultimately less about tape’s demise than it is about the steady encroachment of cloud services.”[1]

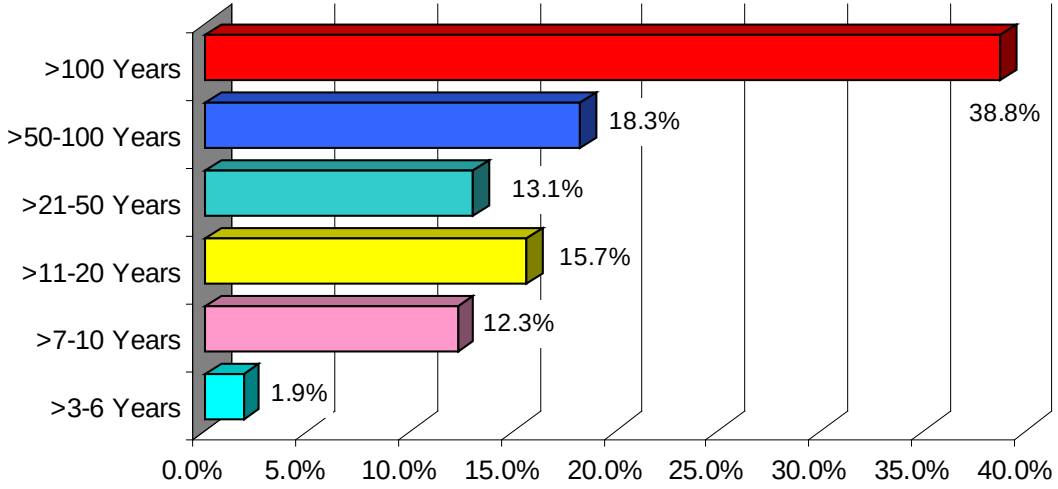
[1] https://www.cio.com/article/3129306/infrastructure-storage/is-tape-storage-dead-again.html

2007 SNIA Archive Survey

- Ten years ago, the SNIA 100 Year Archive Task Force developed a survey with the following goal and hypothesis:
  - Goal: Determine the requirements for long-term digital information retention in the data center. These requirements are needed to frame the definition of best practices and solutions to the retention and preservation problems unique to large, scalable data centers*
  - Research Hypothesis: Practitioner's experiences with terabyte-size archival systems are adequate to define the business and operating requirements for petabyte-size information repositories in the data center*

* Quoted from the SNIA 100 Year Archive Requirements Survey Report (2007)

2007 SNIA Archive Survey



What does Long-Term Mean?

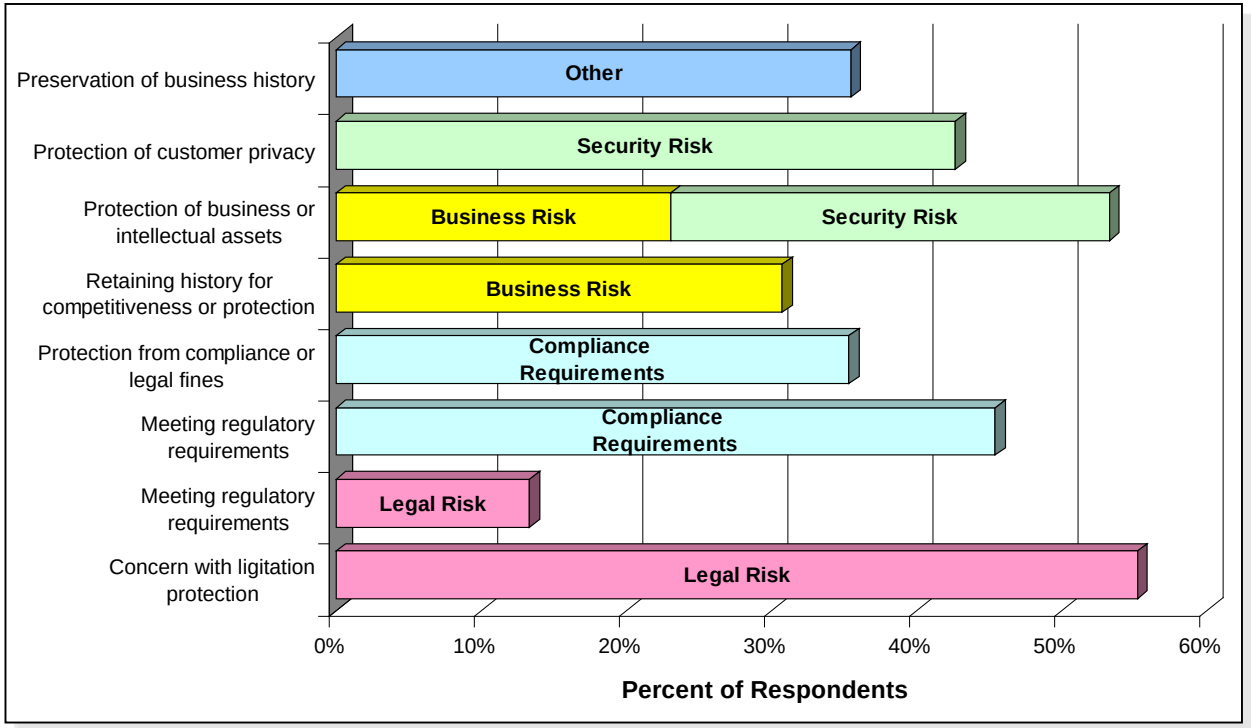
- **Retention of 20 years or more is required by 70% of responses**

Source: SNIA-100 Year Archive Requirements Survey, January 2007

2007 SNIA Archive Survey

***Top* External Factors Driving Long-Term Retention Requirements:**

- Legal Risk
- Compliance Regulations
- Business Risk
- Security Risk



Source: SNIA-100 Year Archive Requirements Survey, January 2007

2007 SNIA Survey

- Key Findings
  - The problems of logical and physical retention
    - Practitioners were struggling - information is at risk long-term
    - Problems were real and generally understood
  - Long term generally means over 10-15 years
    - IT can manage to migrate and retain readability for about this long
    - For longer periods, processes begin failing, become too costly, and the volume of information becomes overwhelming
  - Long term retention requirements are real
    - >80% of organizations reporting have a need to retain information over 50 yrs
    - 68% report a need of over 100 yrs

The SNIA 100-Year Archive Survey 2017

10 Years Later…

- Solutions are now becoming available
  - Standards – OAIS, VERS, MoReq, …
  - Storage formats - SIRF, OpenAXF, PREMIS, BagIt….
  - Software – Fedora, LOCKSS, DSPace, Arkivum, iRods, Rosetta, ….
  - Cloud Services – Preservica, Duracloud, Chronopolis, Dternity, Glacier, ….
- But, their usage is still limited
  - Primarily used in government agencies, libraries, & highly regulated industries
- Why aren't organizations using these solutions
  - Lack of education or understanding
  - Lack of: need, will, funding, penalties, etc.
  - Short term focus

100 Year Archive Survey 2017

- The LTR TWG and DPCO developed a new 100 Year Archive survey
  - Focused on IT best practices, not just business requirements
  - Survey the IT staff charged with long term retention requirements
  - Assess the impact of the cloud
- The goal of this new survey was:
  - Who needs to retain long term information
  - What information needs to be retained, with appropriate policies
  - Are organizations meeting their needs, have practices evolved in 10 years
  - How is long term information is
    - Stored – systems, services, storage technology, etc.
    - Secured – encryption, access control, etc.
    - Preserved – migration, preservation formats, etc.

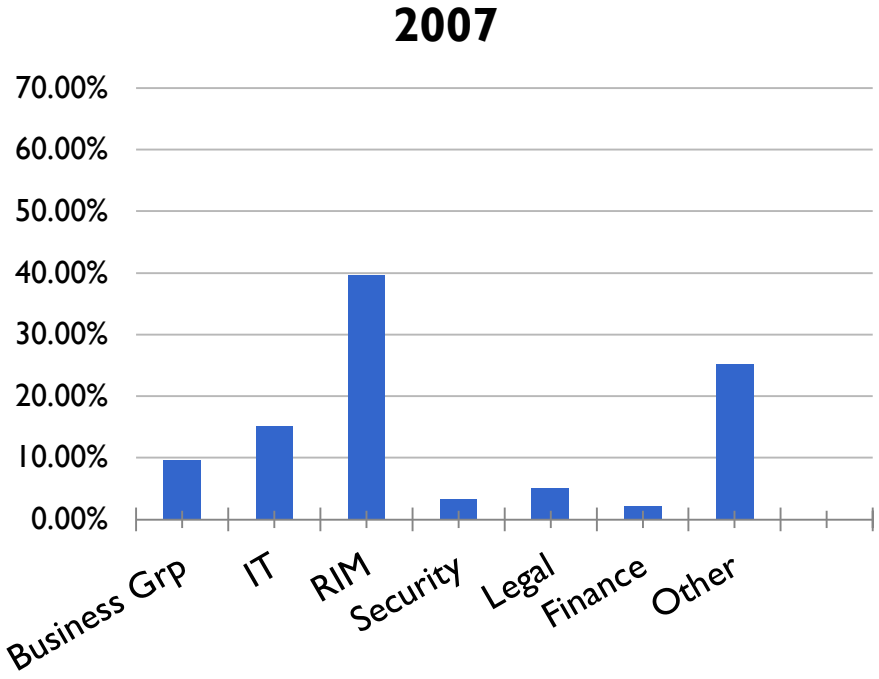
Target Respondents

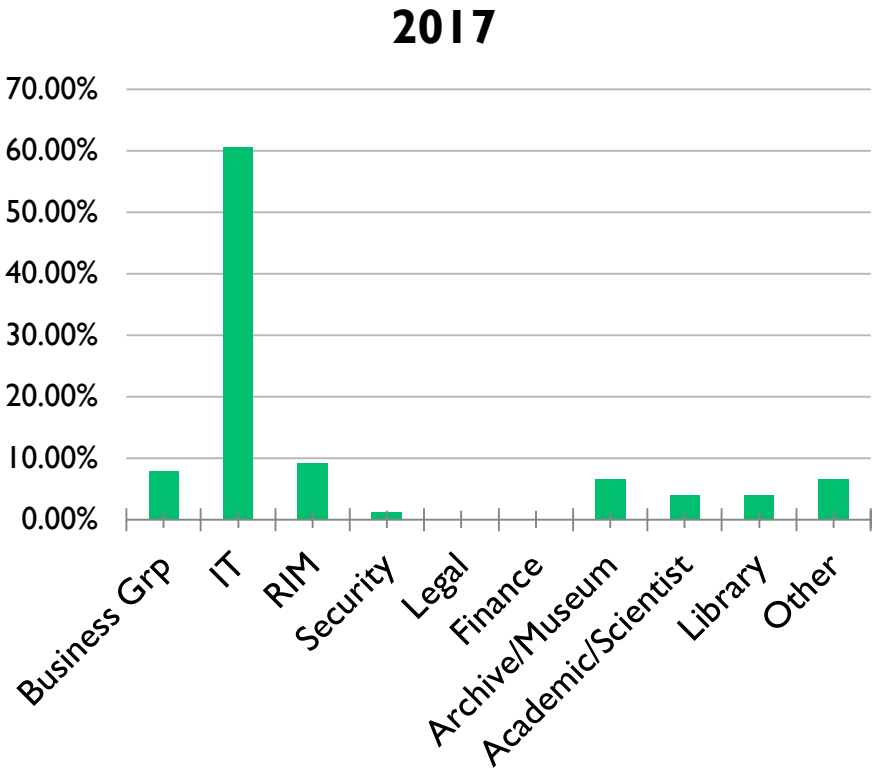
- Primary target is IT staff associated with archives
  - To get better information about systems and technologies
- Other respondents representing
  - A business group
  - Records and Information Management (RIM)
  - Archive/Museum
  - Academic/Scientist
  - Library
  - Security
  - Legal
  - Finance

Topics Covered

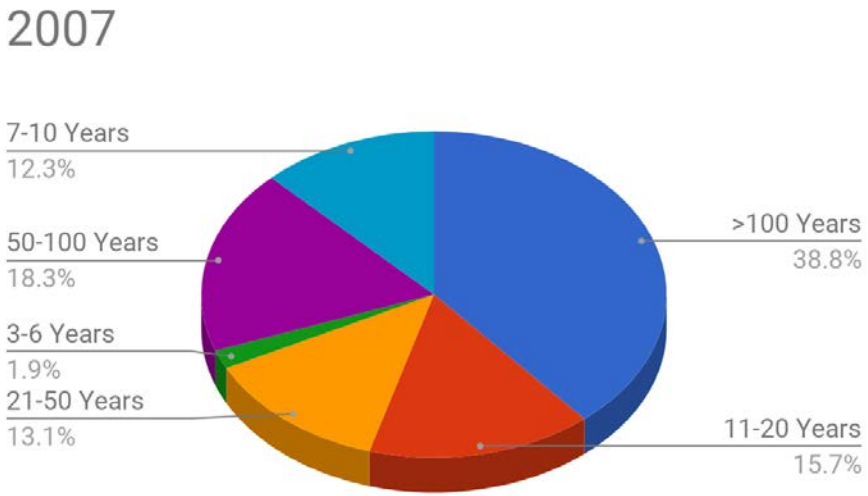
- Demographics
- Business Drivers
- Policies
- Sources
- Storage
- Practices/Experience
- Preservation
- Security/Privacy

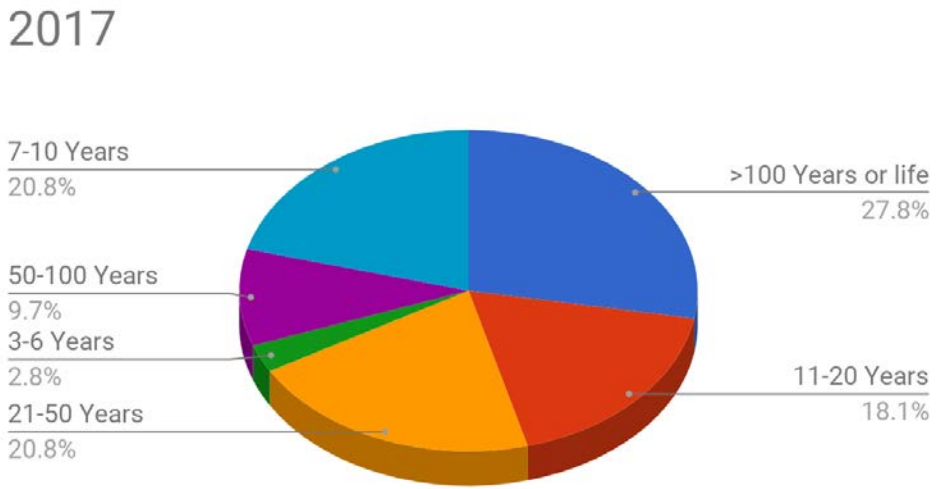
Which group do respondents represent?



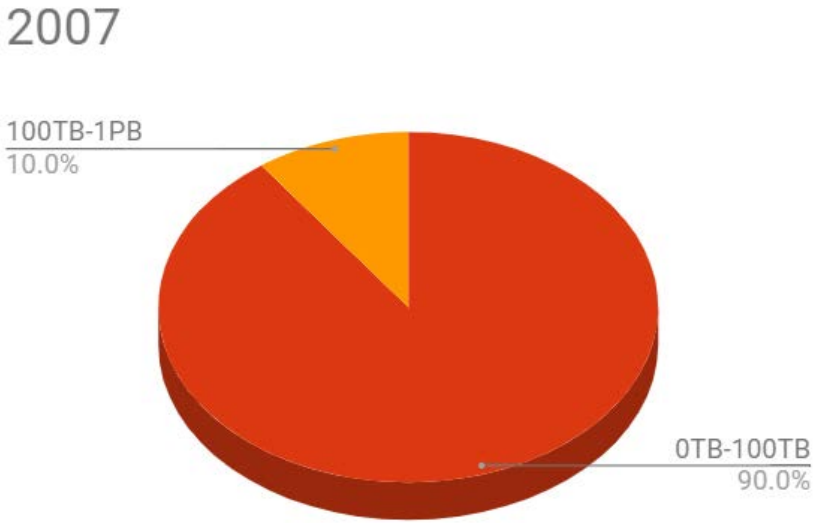


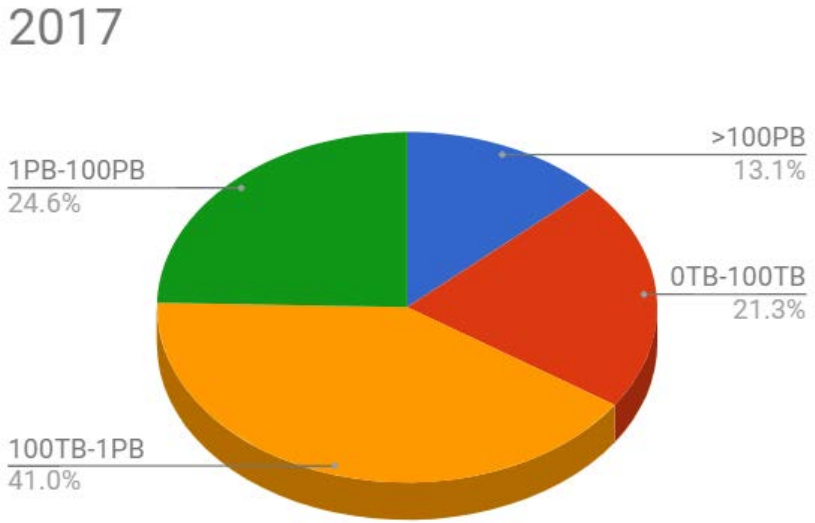
What does ‘Long Term’ mean to your organization?





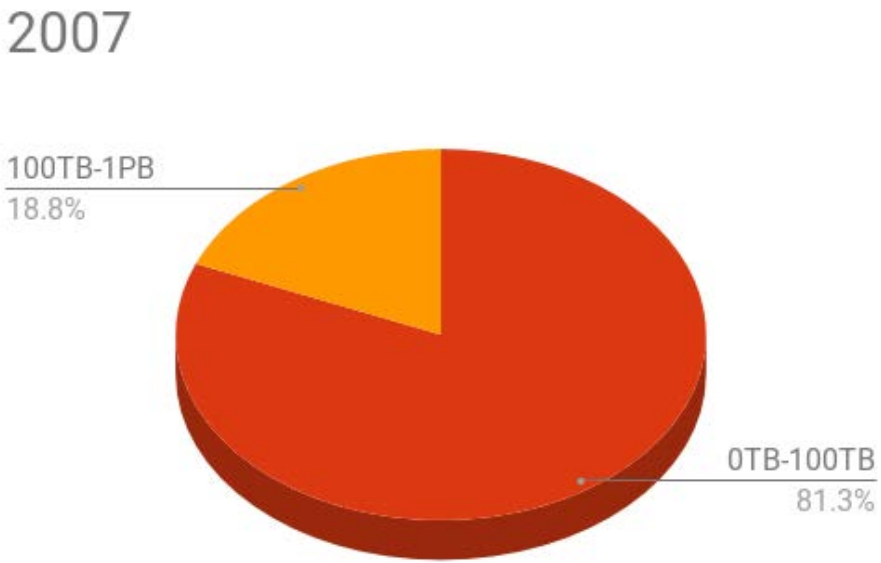
How large is your disk archive?

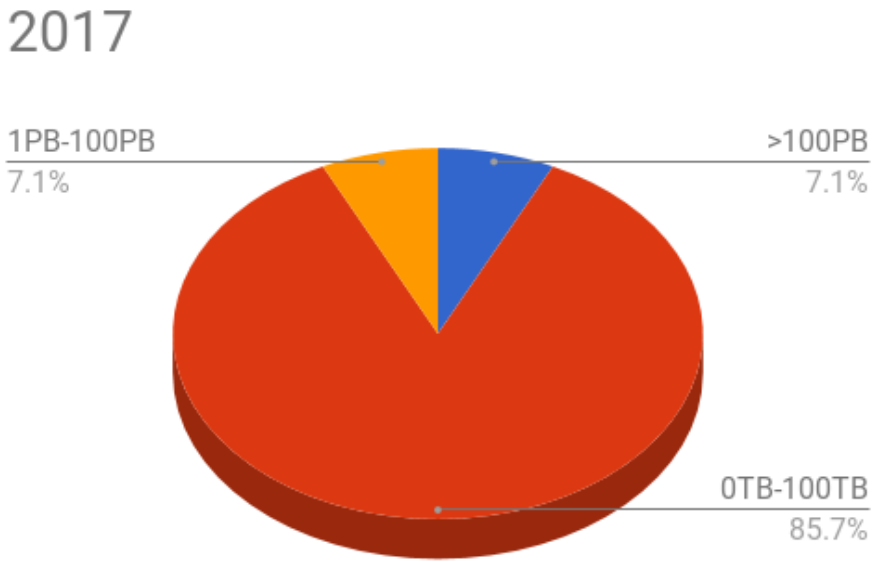




Note: In 2017 >80% of respondents reported disk usage

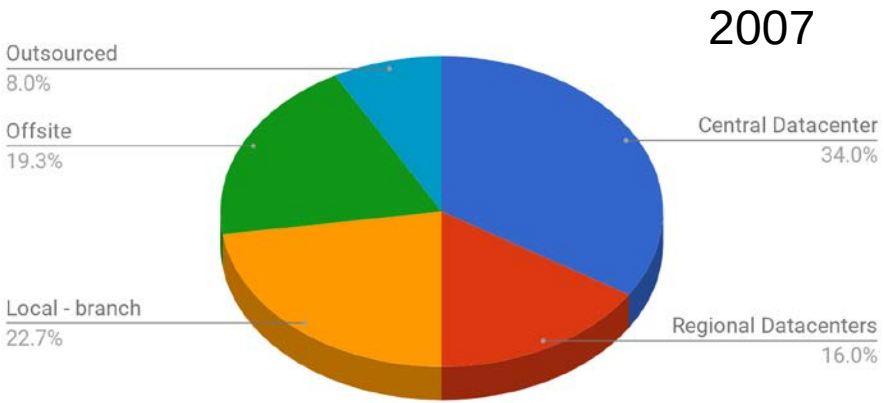
How large is your tape archive?

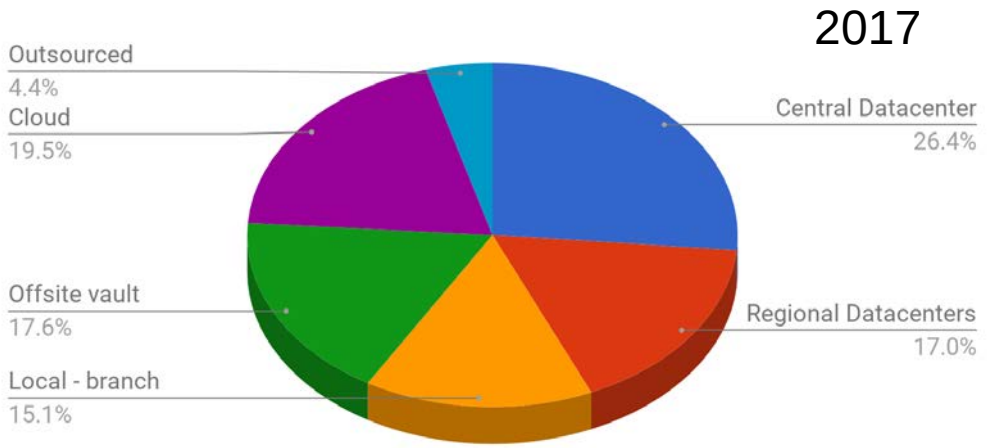




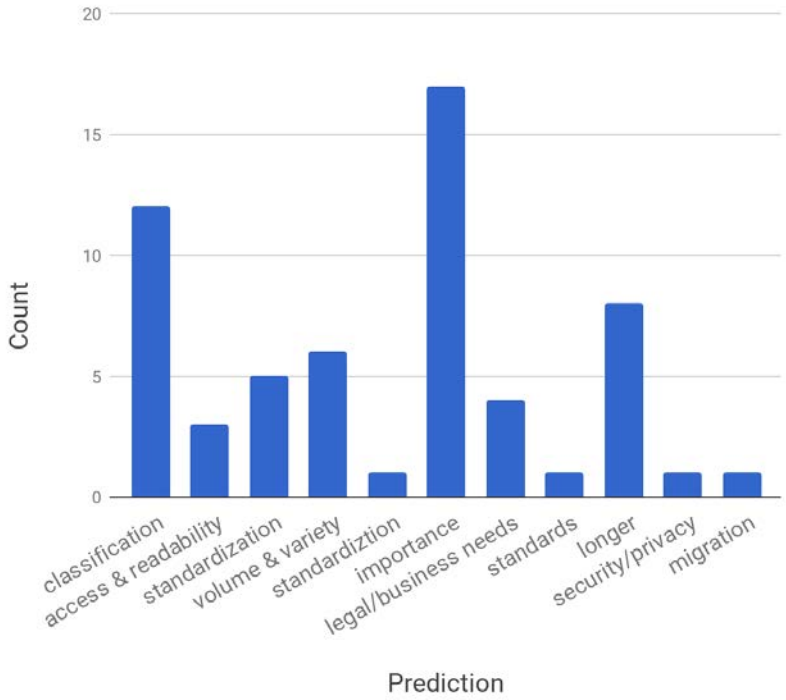
Note: in 2017 only <25% of respondents reported tape usage

Where are your long term records kept?



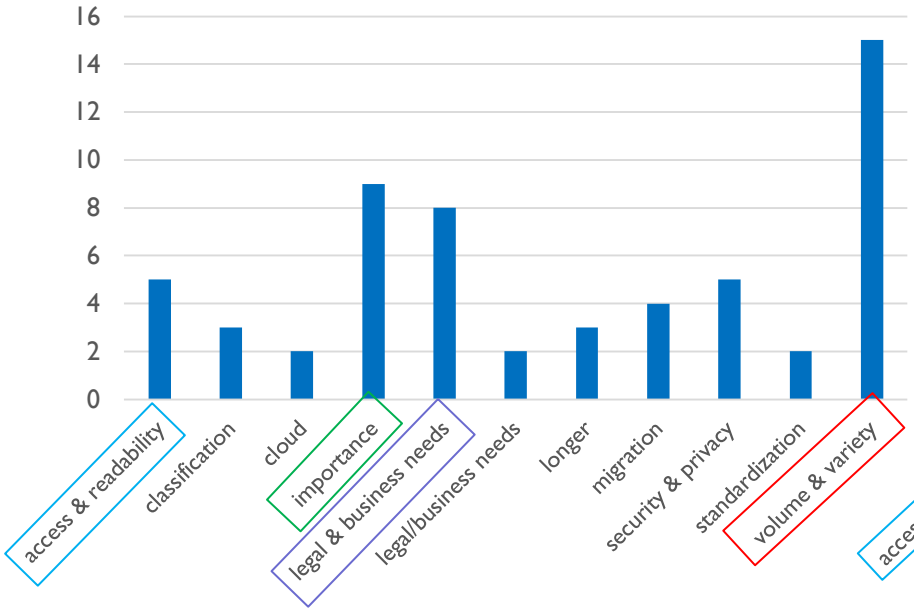


Predictions from 2007

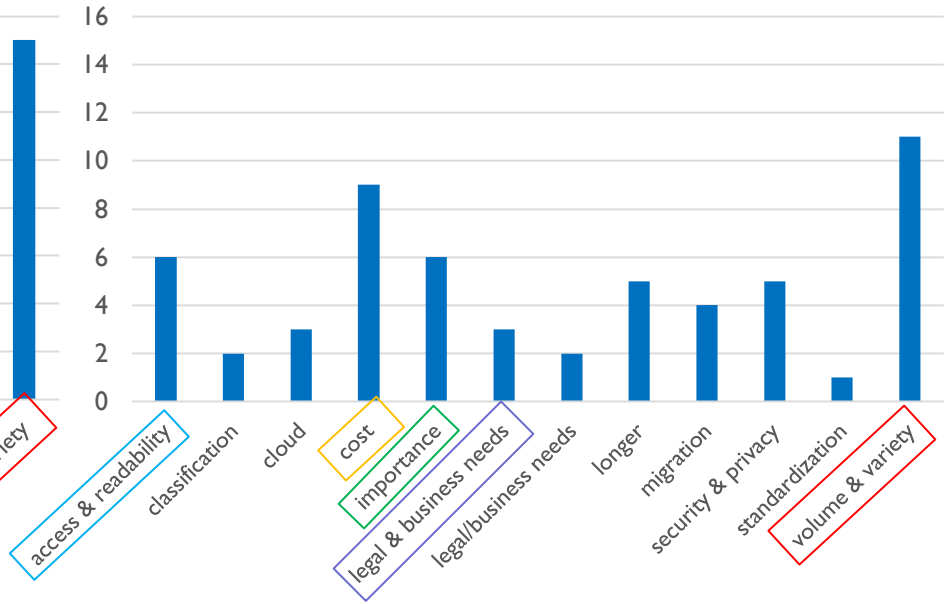


- Classification needs to improve
- Standardization is needed
- LTR will become more important
- Volume and variety will increase
- Retention times will increase
- Readability and access are concerns
- Migration will still be critical
- Security and privacy issues will increase
- Legal and business needs will continue to drive requirements

Changes and Predictions 2017



What has changed since 2007?



What do you expect to change over the next 10 years?

Key Findings

- Data growth has been explosive since 2007, with the largest users having more than 1000 times the amount of data in 2017 compared to 2007
- Respondents continue to define 50+ years as long term, but there is a growing recognition that even shorter periods of time can be considered long term for IT
- The cloud, which didn't really exist in 2007 now accounts for almost 20% or respondents long term storage
- Databases and custom business apps continue to be the biggest challenge for preservation, but there is concern for all types of applications

DPCO Active Programs

- White papers
  - Wrote the Data Protection Best Practices guide
  - Contributed to the Storage Security White Paper
- Review and edit of ISO documents to support the Security TWG
- Regular data protection and privacy Webcasts
- Expanding and clarifying SNIA dictionary definitions
- Support for SDC
  - Privacy vs Data Protection: The Impact of EU Data Protection Legislation - Thursday Sept 27 @ 10.35am
  - Combating evolving ransomware at the block level - Thursday Sept 27 @ 3.35pm

Special Thanks

The SNIA would like to thank the following individuals & groups for their contributions to the creation of this 100-Year Archive Survey

100-Year Archive Survey (2017) Contributors:

Sam Fineberg	Philip Viana
Bob Rogers	Simona Cohen
Paul Talbut	Lori Ashley
Thomas Rivera	Reagan Moore
Eric Hibbard	Mary Baker
Gene Nagle	Mark Carlson
Michael Peterson	Bill Martin

SNIA - Long Term Retention Technical Working Group (LTR TWG)

SNIA - Data Protection & Capacity Optimization (DPCO) Committee

SNIA - Security Technical Working Group (Security TWG)