SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

SNIA 100 Year Archive Survey 2017

Thomas Rivera, CISSP

Data Security & Privacy Consultant

Co-chair, Data Protection & Capacity Optimization (DPCO)Committee – SNIA

Secretary, Board of Directors – SNIA

Secretary, Cybersecurity & Privacy Standards Committee – IEEE

Participating SNIA Groups

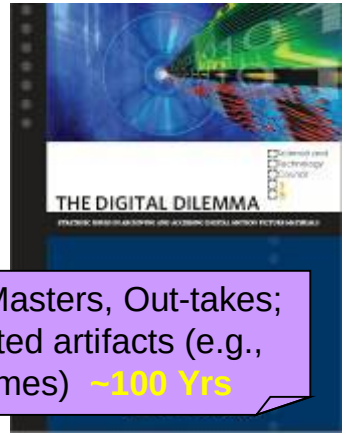
- Long-term Retention Technical Working Group (LTR TWG)
  - SNIA's LTR TWG developed a SNIA standard for a logical container format called the Self-contained Information Retention Format (SIRF)
    - This new standard enables long-term hard disk, cloud, and tape-based containers a way to effectively & 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 of Big Data

- Regulatory compliance and legal issues
  - Sarbanes-Oxley, HIPAA, FRCP, 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
- Domains that have Big Data require preservation

Media & Entertainment



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

Scientific & Cultural

Satellite data is kept **for ever**

Digital art is kept **for ever**

Healthcare

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

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

SDC 17

Goals of Digital Preservation

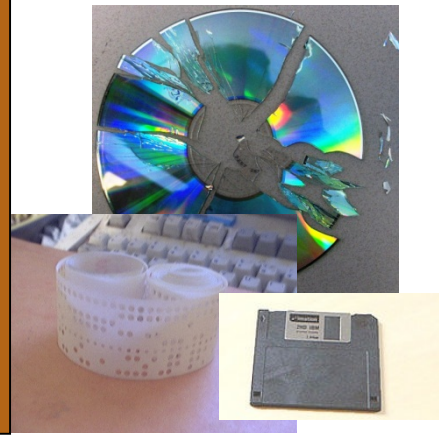
- Digital assets stored now should remain
  - Accessible
  - Undamaged
  - Usable
- For as long as desired – beyond the lifetime of
  - Any particular storage system
  - Any particular storage technology
- And at an ***affordable cost***

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



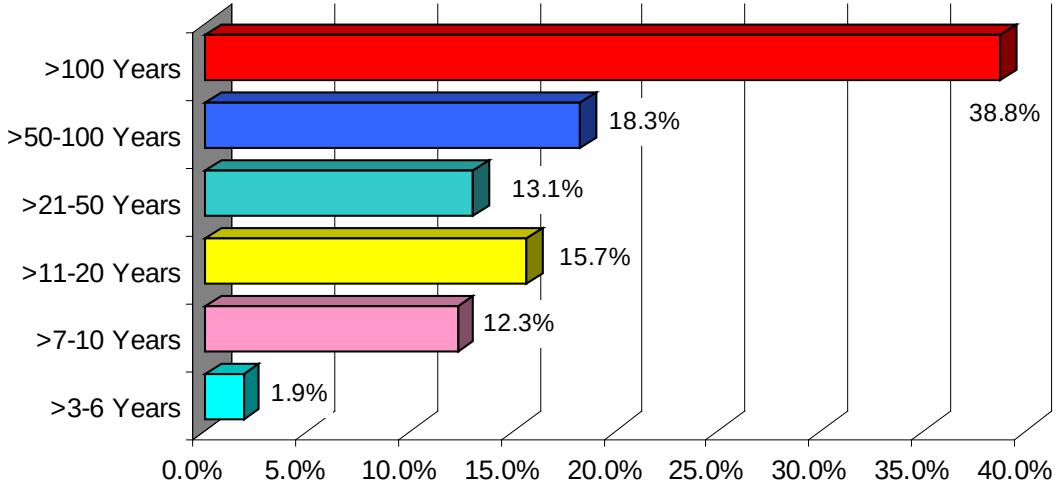
- Media/hardware obsolescence
- Software/format obsolescence
- Lost context/metadata

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 (Cont.)



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

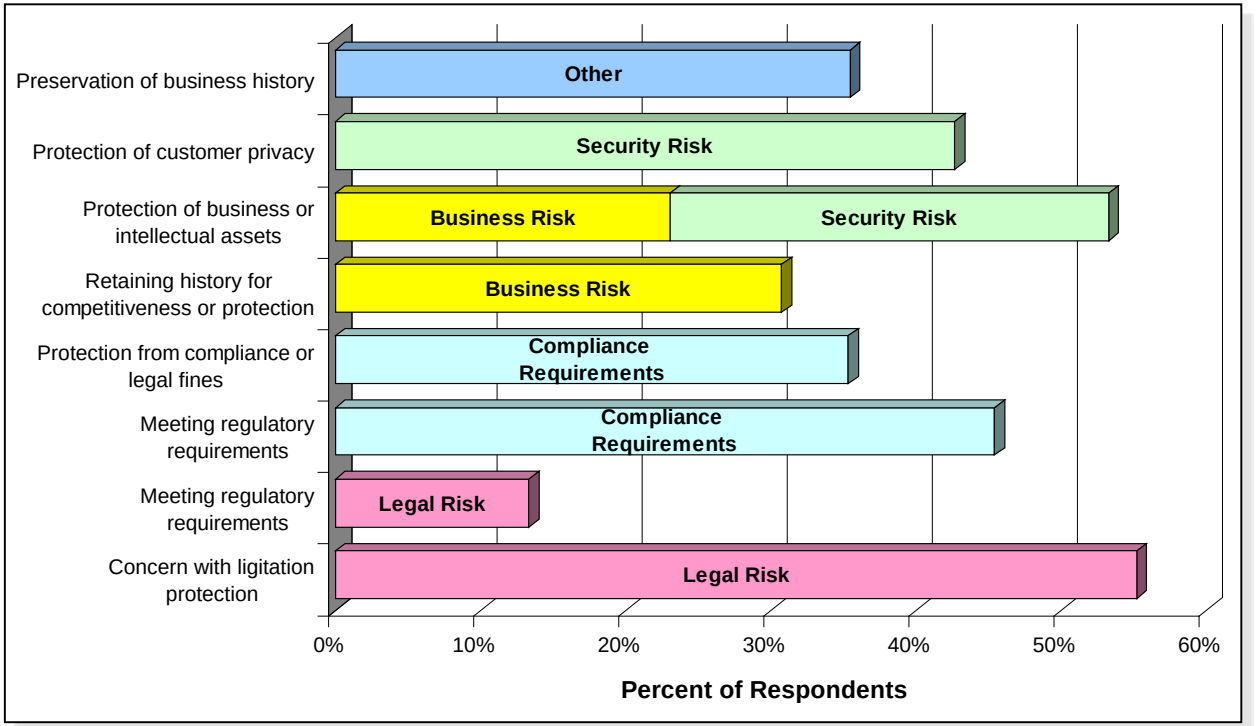
What does Long-Term Mean?

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

2007 SNIA Archive Survey (Cont.)

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 (Cont.)

- 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

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
  - Focus will be on IT best practices, not just business requirements
  - We will seek to survey the IT staff charged with long term retention requirements
  - Need to assess the impact of the cloud
- The goal of this new survey is to assess
  - 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.

Emerging Changes

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

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

Please Take the Survey!

The Long Term Retention Technical Working Group and the Data Protection Committee are pleased to announce the release of the 2017 survey!

- We are now seeking the help of IT and archiving professionals to complete the new survey to help us understand
  - How archive practices have evolved in the last ten years
  - What type of information is now being retained and for how long
  - Changes in corporate practices
  - Impact of technology changes (e.g., cloud)

- Results to be published in early 2018

Please take the survey at:
www.surveymonkey.com/r/100yrarchivesurvey

For Further Information

- Self-contained Information Retention Format (SIRF) For Future Semantic Interoperability
  - http://ceur-ws.org/Vol-1306/paper2.pdf
- SIRF specification:
  - http://www.snia.org/tech_activities/standards/curr_standards/sirf
- More information on SIRF & SNIA LTR activities (including these slides):
  - http://www.snia.org/ltr
- OpenSIRF is available at:
  - http://github.com/opensirf
- LinkedIn SNIA Archive Survey group:
  - https://www.linkedin.com/groups/8590697

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)