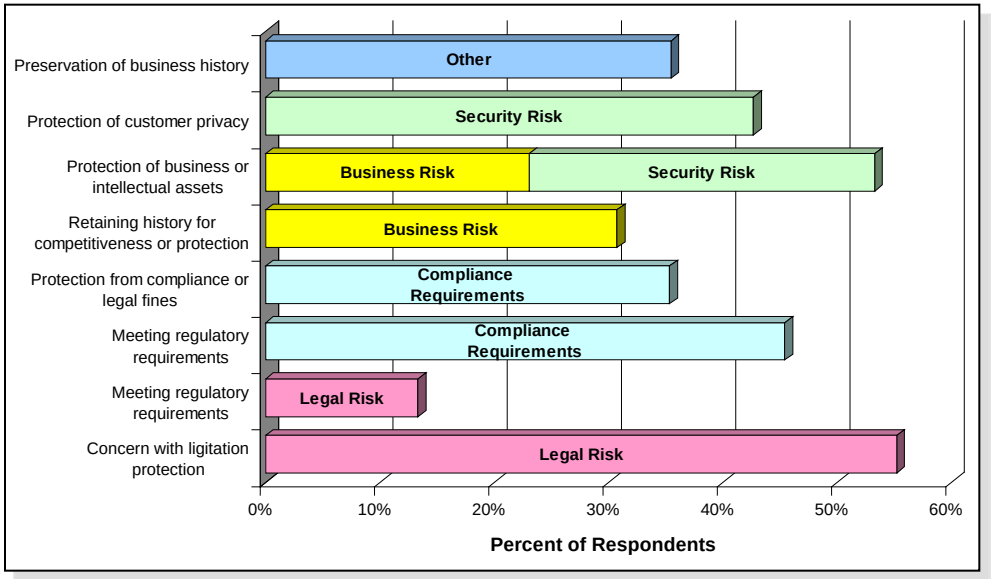
Towards SIRF: Self-contained Information Retention Format

Simona Rabinovici-Cohen, Mary Baker, Roger Cummings, Sam Fineberg, John Marberg
Storage Networking Industry Association (SNIA)
Long Term Retention (LTR) Technical Working Group (TWG)

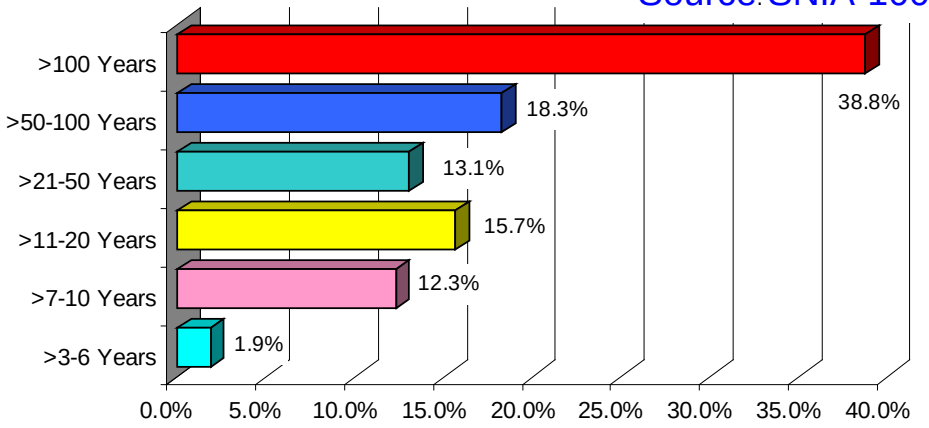
SNIA Survey from 2007

SNIA
Advancing storage & information technology

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.



What does Long-Term Mean?
More than 20 years retention is required by 70% of polls.

100 YEAR Archive Task Force

Result of the survey

- Formation of the Long Term Retention (LTR) TWG
- Goals of digital preservation
  - Digital assets stored now should remain accessible, usable, undamaged
  - For as long as desired – beyond the lifetime of any particular storage system & any particular storage technology (or any application!!)
  - And at an affordable cost (or a range of cost/performance)
- LTR TWG Program of Work addresses both “bit preservation” and “logical preservation”
  - Both are absolutely necessary to retain usability of information
  - Cannot make either reliable enough by itself @ reasonable cost
  - Migration is a potentially affordable approach for both

Migration

- Move a set of information from an old device or technology growing less reliable (e.g. LTO-2 tape) ……
- … or from an application no longer supported or in general use (e.g. WordPerfect 4.2)…..
- to a new device and/or a new format
- Requirements for migration
  - Preserve not only all the data but all related metadata too
  - Maintain provenance, authenticity & integrity
  - Be auditable and traceable
- Need a “container” to encapsulate all of the related information … and a way to automate much of migration

An Analogy

- Standard archival box
  - Archivists gather together a group of related items, known as a collection
  - Collection is placed in a physical box container
  - The box is labeled with information about its content e.g., name and reference number, date, contents description, destroy date
    - And there's an online (XML) finding aid
  - When contents migrated they're added to box
- SIRF is the digital equivalent
  - Logical container for a set of (digital) preservation objects and a catalog
  - The SIRF catalog contains metadata related to the entire contents of the container as well as to the individual objects
  - SIRF standardizes the information in the catalog

Photo courtesy Oregon State Archives

Self-contained Information Retention Format (SIRF)

SNIA
Advancing storage & information technology

- SIRF is a logical container format appropriate for long-term storage of digital information
  - Preserves collections of objects and their relationships
  - Includes generic metadata that can be extended with domain specific information
  - Can be mapped to and physically migrated between a wide variety of underlying storage systems
- SIRF use cases and requirements document is released for public review
  - http://www.snia.org/tech_activities/publicreview
- More information on SIRF is available at
  - ***http://www.snia.org/ltr***

Icon designed by Tammy Dekel (IBM Haifa)

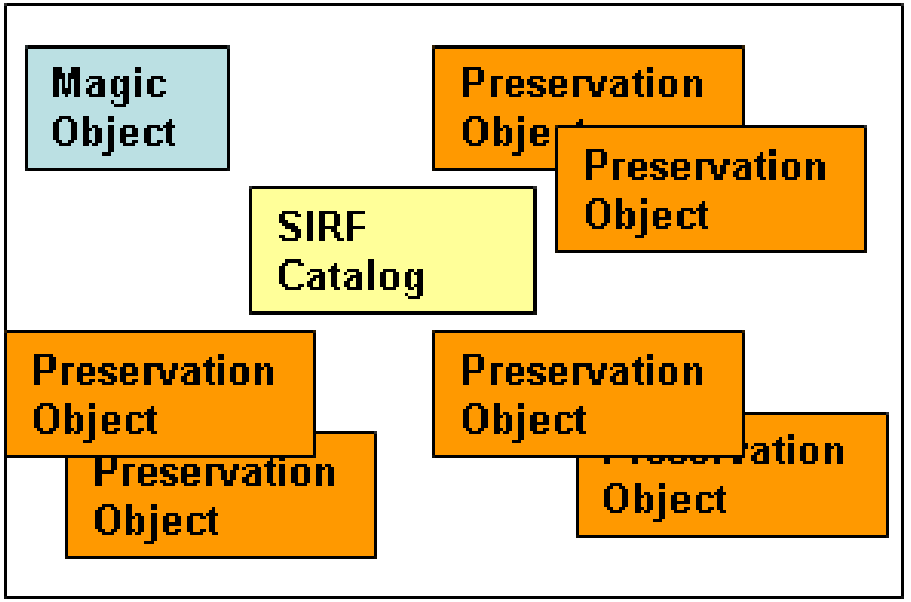
What is SIRF?

- SIRF is a logical data format of a storage container.
  - A storage container may comprise a logical or physical storage area considered as a unit.
    - For example, a storage container may comprise a mountable data storage unit, a file system, a tape, a block device, a stream device, an object store, a data bucket in a cloud storage
  - SIRF contains a set of preservation objects to be understood in future
- Required SIRF Properties
  - Self-describing – can be interpreted by different systems
  - Self-contained – all data needed for the preservation objects interpretation is in the container
  - Extensible – so it can meet future needs

SIRF Components

A SIRF container includes:

- A **magic object**: identifies SIRF container and its version
- Numerous **preservation objects** that are immutable
- A **catalog** that is
  - Updatable
  - Contains metadata to make container and preservation objects portable into the future without external functions



What is a Preservation Object?

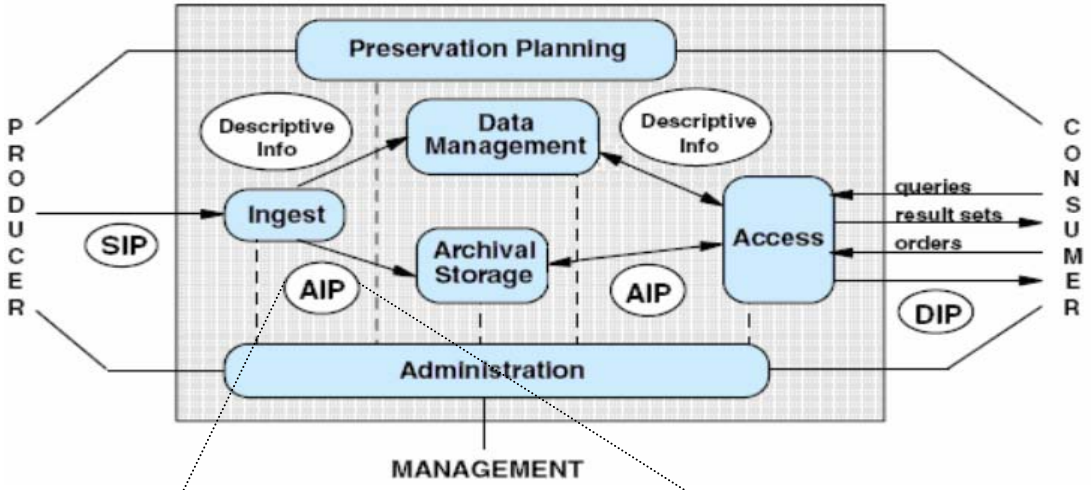
- SIRF Containers Store Collections of Preservation Objects (POs)
- A Preservation Object is
  - the raw data to be preserved,
  - plus additional embedded or linked metadata, and
  - includes everything needed to enable the sustainability of the information encoded in the raw data for decades to come
- Attributes of a PO
  - may be subject to physical and logical migrations
  - may be dynamic and change over time
  - an updated PO is a new version of the original, and its audit log records the changes that have occurred so authenticity may be verified
- Several examples of a PO
  - e.g. OAIS Archival Information Package (AIP)
    - An AIP includes recursive representation information that enables future interpretation of the raw data

Open Archival Information System (OAIS)

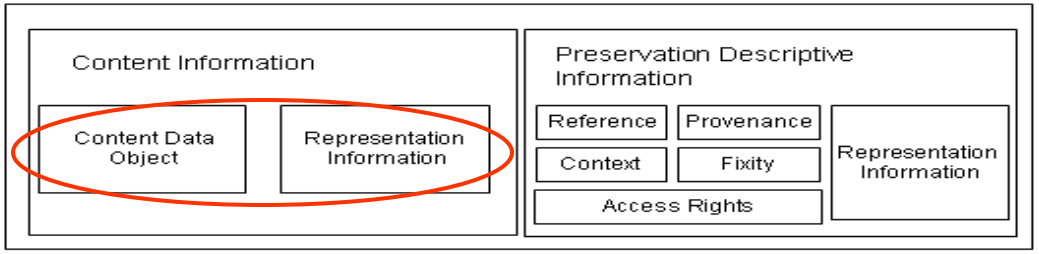
- ISO standard reference model (ISO:14721:2002)
- Provide fundamental ideas, concepts and a reference model for long-term archives
- Includes a functional model that describes all the entities and the interactions among them in a preservation system
- Archival Information Package (AIP) - a logical structure for the preservation object that needs to be stored to enable future interpretation

* Figure taken from the OAIS spec

*** OAIS Functional Model**



AIP



Layered Approach

- SIRF is a logical data format
  - Assumes the underlying layer includes an object interface layer
  - Examples
    - Advanced: OSD, Cloud, XAM
    - Lower level: UDF, CDFS, FAT, LTFS

- SIRF defines two levels
  - Level 1 catalog (L1) – unique metadata, not in the preservation objects, that is mandatory to make preservation objects portable into the future
  - Level 2 catalog (L2) – information that is probably also in the preservation objects, that is needed for fast access to the preservation objects

SIRF Use Cases and Functional Requirements Methodology

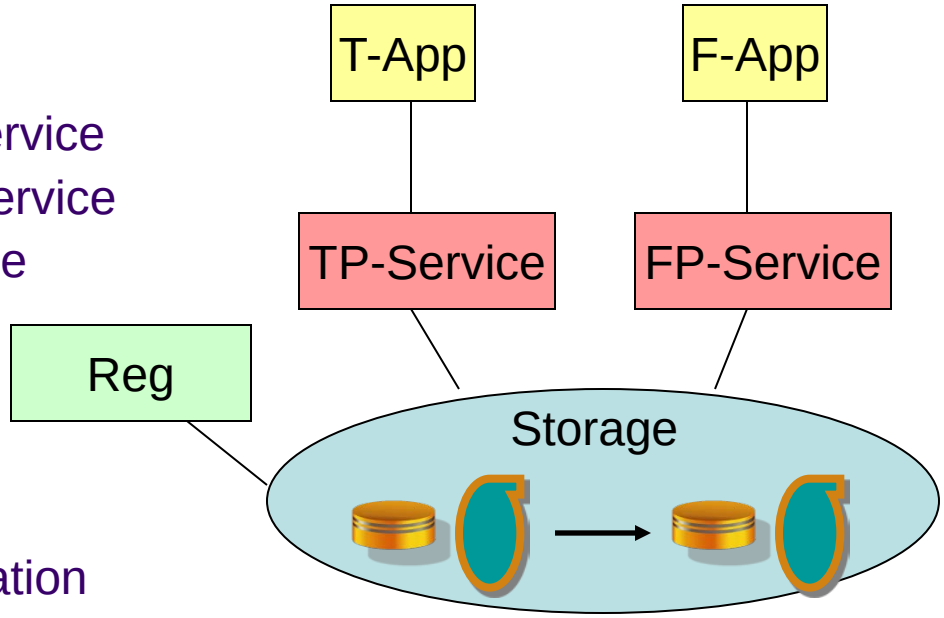
SNIA
Advancing storage & information technology

- Define Actors involved in SIRF
- Define use cases and flows among the actors
  - 4 generic uses cases
    - Unlinked to specific type of data or application
    - Technological changes in the environment
  - 5 Workload-based use cases
    - Specialized for concrete workloads
    - Additional non-technological changes in the environment
- For each use case, find the derived functional requirements
- Aggregate all functional requirements and map use cases to them
- Categorize the functional requirements
  - general requirements, format requirements, data model requirements, performance requirements, etc.
- Prioritize the functional requirements
  - Some of the requirements may conflict each other

SIRF Actors

Non-human actors:

- Storage - Storage subsystem
- TP-Service - Today's preservation service
- FP-Service - Future's preservation service
- T-App - Today's application e.g. Office
- F-App - Future's application
- Reg – Registry



- The storage persists sets of preservation objects

Example Use Case : eMail archive

SNIA
Advancing storage & information technology

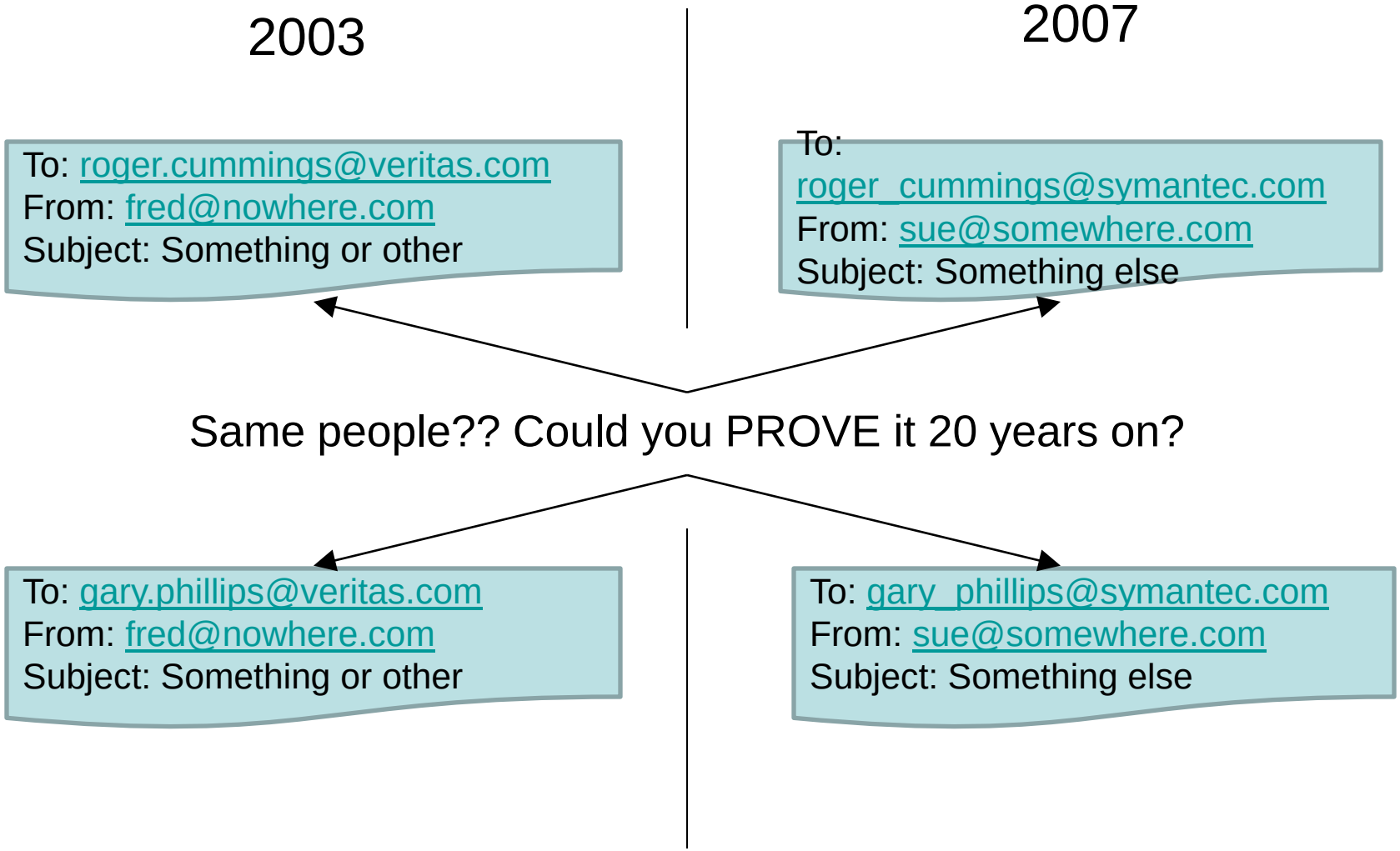
Flow:

1. T-App ingests an e-mail thread today via TP-Service. This includes ingesting a collection of several interrelated Preservation Objects (POs) - thread PO, message POs, attachments POs, PO for the address book, POs for organizational processes, POs for data leakage policies
2. Time passes and the organization changes scope, name, undergoes a merger, etc. As a result, FP-Service creates a set of new version POs for the address book and the organizational processes
3. More time passes and F-App searches the repository and creates POs for the search results to raise performance of future searches. Those new POs may contain soft links to the thread, messages and attachments created in step 1

Main Requirements:

- Support for time stamps (required quality is work-in-progress)
- Support for "special" POs e.g. address book PO, search results PO
  - For lack of a better name, we call these "special" POs - secondary catalog
- Support for hard links and soft links
- Generic support for organizational unique metadata

Real Life Example Problem



Derived SIRF Requirements – a sample

SNIA
Advancing storage & information technology

- Support for verification of document provenance and authenticity
  - Regardless of migrations whether logical or physical.
- Support methodology for verification of completeness and correctness
- Support for retention holds that prevent POs being modified or deleted
- Support for links between POs that are as immutable as the objects themselves
  - Either “soft” or “hard” links
- Support for “special” POs, auditable time stamps

Current Status of work on SIRF

SNIA
Advancing storage & information technology

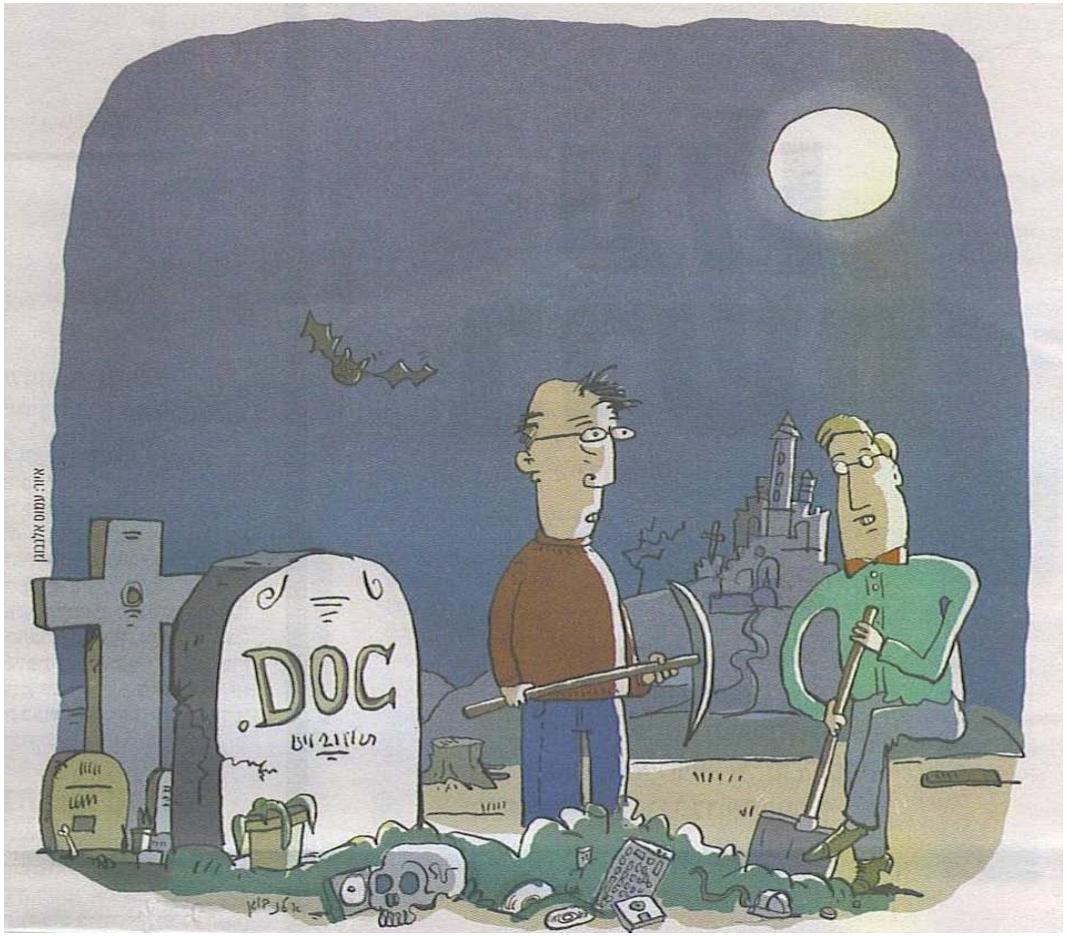
- Use cases and requirements document published
  - Five representative use cases described in paper out of set of 9
  - But we'd still be please to consider more (see URL on slide 7)
- Work of the detailed specification of SIRF is just beginning
  - Laying out the catalog structure
  - Identifying what metadata needs to be preserved where to meet the requirements
  - Taking advantage of the work in the EU CASPAR project, and anticipating the work in the EU ENSURE project
  - And also looking @ the cloud storage environment

Summary

- SIRF container is key to information retention work in SNIA, but also of interest to cloud & compliance activities etc.
  - Not trying to re-invent the wheel, leveraging existing work to the maximum extent possible
  - When combined with bit preservation activities will provide a comprehensive set of tools to address long term information retention
- Intent of this work is to be a catalyst to get the storage industry involved in projects that address digital preservation
- And finally….

What we're trying to avoid….

… is the "Digital Dark Ages", and having to do archaeology i.e. dig things up from dark & dirty corners



Yedioth Ahronoth, Sunday June 1, 2008


Thank
You!

??

About the SNIA Long Term Retention (LTR) TWG

SNIA
Advancing storage & information technology

- The Long Term Retention (LTR) Technical Working Group (TWG) is co-chaired by IBM and Symantec

- Mission
  - The TWG will lead storage industry collaboration with groups concerned with, and develop technologies, models, educational materials and practices related to, data & information retention & preservation.

- Charter
  - The TWG will ensure that SNIA plays a full part in addressing the "grand technical challenges" of long term digital information retention & preservation, namely both physical ("bit") and logical preservation.
  - The TWG will generate reference architectures, create new technical definitions for formats, interfaces and services, and author educational materials. The group will work to ensure that digital information can be efficiently and effectively preserved for many decades, even when devices are constantly replaced, new technologies, applications and formats are introduced, consumers (designated communities) often change, and so on.

- For more information, see ***http://www.snia.org/ltr***