Long Term Information Retention Format

Sam Fineberg
HP Software
fineberg@hp.com

Simona Cohen
IBM
simona@il.ibm.com

This presentation was created with help from Roger Cummings and Mary Baker as well as the other members of the SNIA Long Term Retention TWG

Outline

- **Introduction**
  - Why we need digital preservation
  - Why we need a long term information retention format
- SIRF
  - Motivation
  - Use cases
  - Requirements
  - Status

Why we need digital preservation

- 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
- Responses to 100 Year Archive Requirements Survey
  - 68% of organizations had requirements for over 100 years
  - 83% of organizations had requirements for over 50 years

100 YEAR Archive Task Force

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
- and at an *affordable cost*

Threats to long-term digital 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

Threats to long-term digital assets

- Large-scale disaster
- Human error
- Media faults

- Component faults
- Economic faults
- Attack
- Organizational faults

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

Threats to long-term digital assets

- Large-scale disaster
- Human error
- Media faults

- Component faults
- Economic faults
- Attack
- Organizational faults

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

Threats to long-term digital assets

- Large-scale disaster
- Human error
- Media faults

- Component faults
- Economic faults
- Attack
- Organizational faults

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

Even preserving just the bits is hard

- Large scale & long time periods are a problem
- 1 petabyte, 50 years, 50% probability of no damage
  - Sounds reasonable, doesn't it?
- That's a bit half-life of 10^{17} years
  - A million times the age of the universe
  - Even improbable events will have an effect
- Now try to keep
  - The bits usable
  - The information reusable
  - The applications usable
- Preserve just the bits (physical preservation)?
  - Can't interpret the content
- Focus only on the logical aspects (logical preservation)?
  - The bits have been trashed

Key qualities of a preservation store

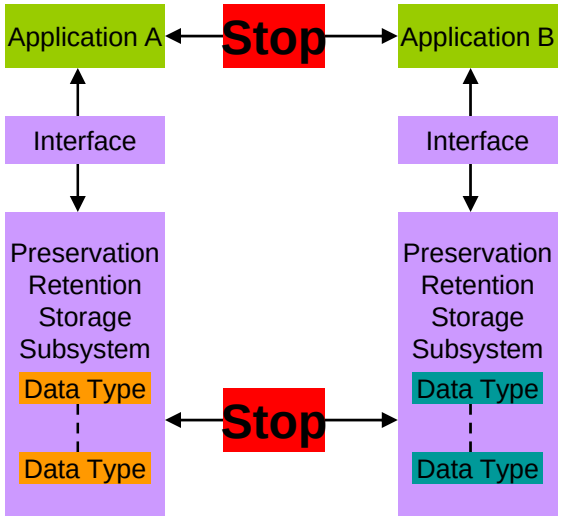
- We can't predict the future
  - Storage systems will change
  - Formats will change
  - Systems will fail
- Preservation objects need to be stored in a format that is
  - **Self contained** – to ensure objects are complete
  - **Self describing** – so software can interpret it
  - **Extensible** – so it can meet future needs

SIRF: logical container format

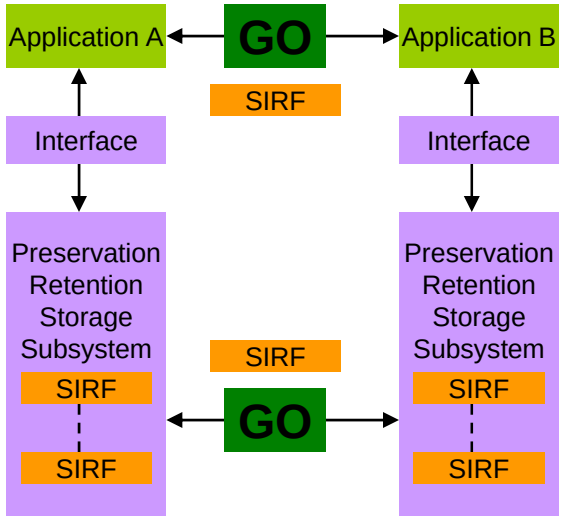
- Being developed by SNIA LTR-TWG and DMF-LTACSI
- SIRF is a format appropriate for long-term storage of digital information
  - Logical data format of a mountable unit
    - File system, block device, stream device, object store, tape, etc.
  - Includes a cluster of “interpretable” preservation objects
    - Self-describing – can be interpreted by different systems
    - Self-contained – all interpretation data contained in object cluster
  - Facilitates transparent migration for long-term preservation
    - Logical
    - Physical
- SIRF implementations may leverage other standards
  - Open Archival Information System (OAIS) ISO standard
  - Network Attached Storage (NFS/CIFS), XAM (Extensible Access Method), Object Storage (OSD)
  - Others

Problems SIRF addresses

Without SIRF



With SIRF

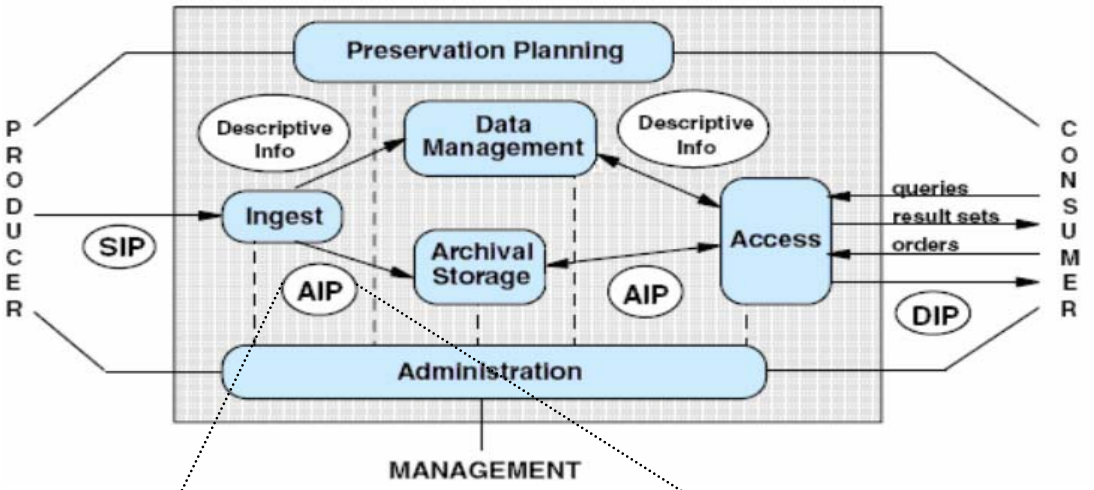


Cannot move cluster of preservation objects between systems without help	Can move cluster of preservation objects between systems by itself
Only original application that wrote the preservation objects can read and interpret them	Any SIRF compliant application can read and interpret preservation objects
Need export and import processes	No need for export and import processes
Preservation objects cannot be sustained for long-term	Preservation objects survive longer

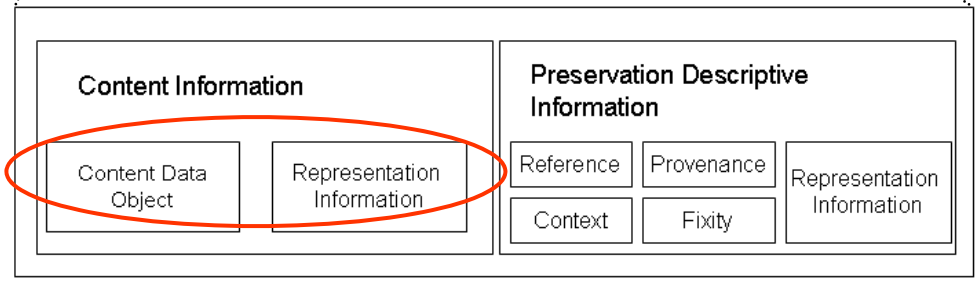
Open Archival Information System (OAIS)

- 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

*** OAIS Functional Model**

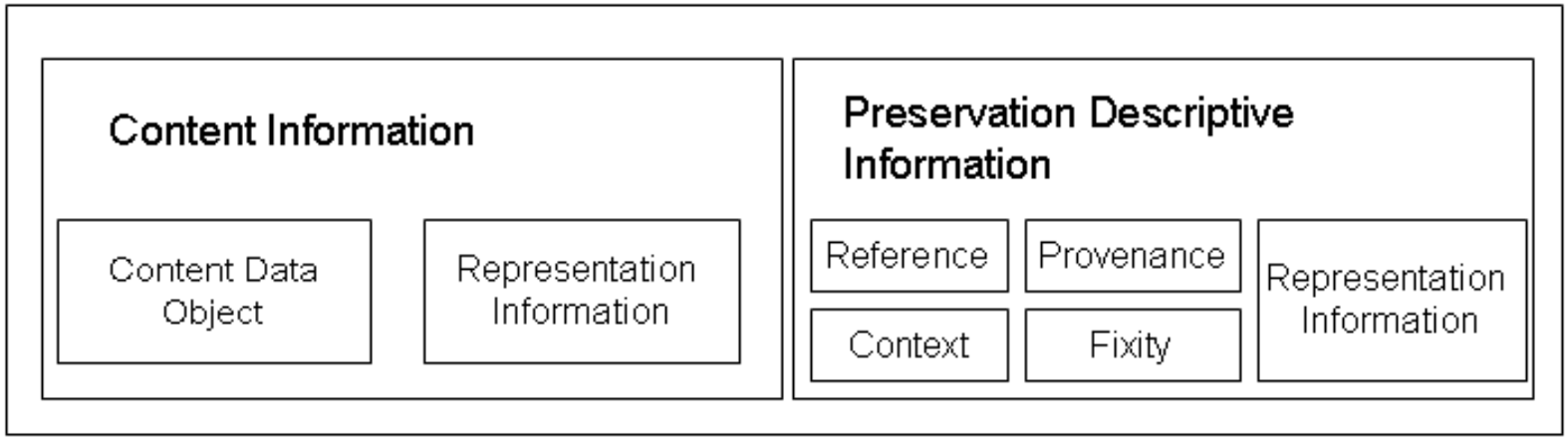


AIP



* Figure taken from the OAIS spec

OAIS AIP Logical Structure



- **Content Data Object** - the raw data that is the focus of the preservation.
- **Representation Information** – the information required to interpret the raw data to its designated community.
- **Reference** – globally unique and persistent identifiers for the content information.
- **Provenance** – the history and the origin of the content information and any changes that may have taken place since it was originated, and who has had custody of it since it was originated.
- **Context** – documents reason for creation of the content information and relationship to its environment.
- **Fixity** – a demonstration that the particular content information has not been altered in an undocumented manner.

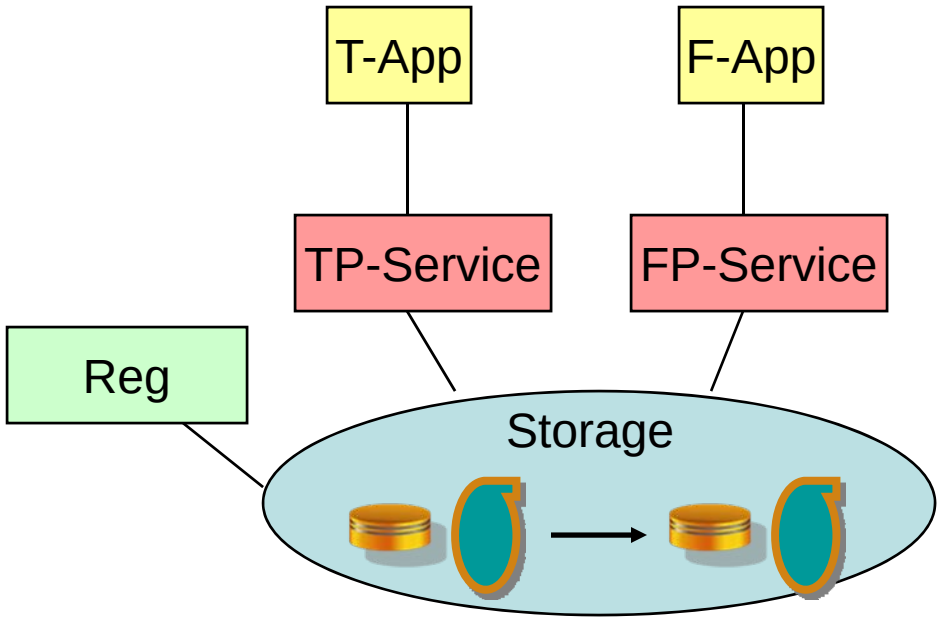
What is a Preservation Object?

- The raw data to be preserved plus additional embedded or linked metadata needed to enable the sustainability of the information encoded in the raw data for decades to come
  - The preservation object may be subject to physical and logical migrations
  - The preservation object may be dynamic and change over time
  - The updated preservation object is a new version of the original preservation object and its audit log records the changes that have occurred so authenticity may be verified

- An example of a preservation object is OAIS Archival Information Package (AIP)
  - An AIP includes recursive representation information that enables future interpretation of the raw data

Preservation Use Cases: Actors

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

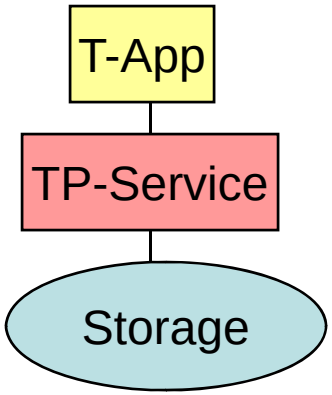


- The storage persists clusters of preservation objects

Use Case 1: ingest and access with same application

Flow:

1. T-App ingests a Preservation Object today
2. T-App access the Preservation Object today



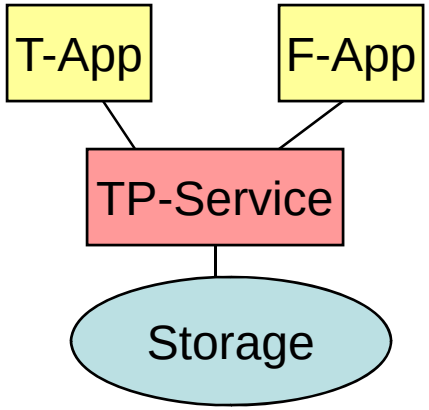
Main Requirements:

- Standard interfaces e.g. NFS, CIFS, XAM
- Agnostic to media, platform, vendor

Use Case 2: ingest and access with different applications

Flow:

1. T-App ingests a Preservation Object today
2. F-App access the Preservation Object in the future



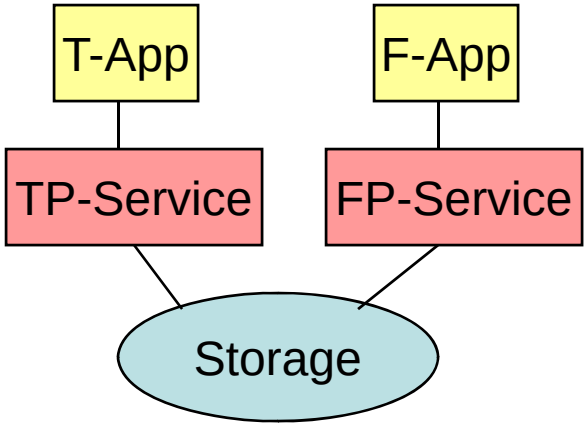
Main Requirements:

- Support migrations of preservation objects
- Support different data models and different formats for the raw data

Use Case 3: ingest and access with different preservation services

Flow:

1. T-App ingests a Preservation Object today via TP-App
2. Time pass and the preservation application migrated to a new one
3. F-App access the Preservation Object in the future via FP-Service



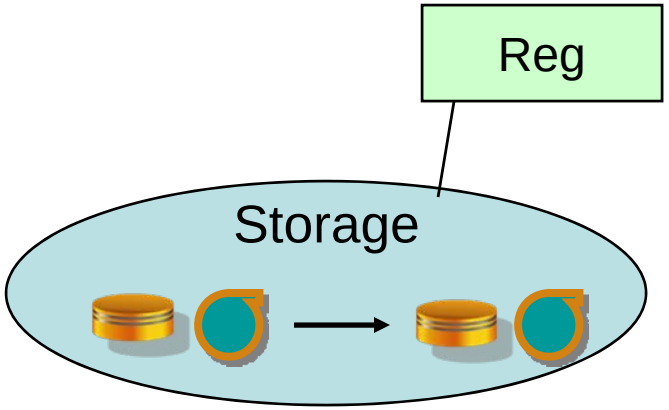
Main Requirements:

- Persistent globally unique identifiers for the preservation objects
- Self-contained data

Use case 4: Storage format is changed

Flow:

1. T-App ingests a Preservation Object today via TP-Service
2. Time pass and the preservation application migrated to a new one
3. F-App access the Preservation Object in the future via FP-Service



Main Requirements:

- Self-describing via a simple formalized meta-language that itself should be changeable
- SIRF Representation Information should be preserved in an external registry

SIRF Initial Requirements – General

- Media agnostic
  - Tape, disk, future media
  - Direct random access and serial access
  - Support mixture of storage technologies
  - Required by: all use cases
- Vendor and Platform agnostic
  - Required by: all use cases
- Support different standard storage technologies and interfaces e.g. NFS, CIFS, XAM
  - Required by: use case 1
- Extensible
  - Support additional information which may be added in the future
  - Required by: use cases 2, 3

SIRF Initial Requirements – Format

- Self-describing
  - The amount of a-priory information is small and can be acquired in stages
  - Interpretable by both humans and machines
  - Ability to do offline inspection
  - Required by: use cases 2, 3, 4
- Support self-contained data
  - Include means to represent internal links and cross references
  - Ability to reconstruct only from the container and any well defined external resources
  - Required by: use cases 3, 4
- Support different SIRF formats preserved in an external registry
  - Required by: use case 4
- Interoperability
  - Ability to migrate data between different systems without loss of information – data should be interpretable after migrations
  - Can be interpreted in the future
  - Required by: use cases 3, 4
- Support methodology for verification of completeness and correctness
  - Required by: use cases 3, 4

SIRF Initial Requirements - Preservation Object Data Model

- Support different data models for preservation objects
  - Support different object data models at one time
  - Support complex data structures like collections of objects
  - Support migrating objects from one data model to an alternative data model
  - Required by: use cases 3, 4
- Can handle any proper data format for the raw data
  - No restrictions on file formats
  - Required by: use case 2
- Enable keeping various versions of the same preservation object with their relations
  - References from new to existing preservation objects of the same version series
  - Required by: use cases 2, 3, 4
- There must be a persistent identifier for each preservation object
  - Include additional external identifiers
  - Required by: use case 3

SIRF Initial Requirements - Performance

- Performance
  - Need to have good performance even for data that includes text and binaries
  - Support large objects e.g. web archiving objects, database archiving objects
  - Do not require complete scanning for access
  - Required by: all use cases
- Enable parallel data migration
  - Enable parallel reads and writes
  - Required by: all use cases

Current Status

- SIRF status
  - The SNIA LTR TWG is currently finishing up the requirements/use case definition phase for SIRF
    - More detailed use cases and storyboards
    - Formal technical requirements and documentation
  - Work on the specification itself is expected to begin late this year
  - All SNIA members are invited to participate in the LTR TWG
- More information
  - SNIA Technical Working Groups (including the LTR TWG) – http://www.snia.org/tech_activities/workgroups
  - 100 year archive survey – http://www.snia-dmf.org/100year
  - SNW tutorial on Long Term Retention – http://www.snwusa.com/documents/presentationsF08/2008_Thursday_0830_MaryBaker.pdf
  - OAIS source doc – http://public.ccsds.org/publications/archive/650x0b1.pdf
  - XAM – http://www.snia.org/forums/xam/
  - NFS – http://www.ietf.org/dyn/wg/charter/nfsv4-charter.html

About the SNIA LTR TWG

- This tutorial has been developed, reviewed and approved by members of the SNIA Long Term Retention Technical Working Group (LTR TWG)
  - 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.
- Please join us!