Long Term Information Retention

Sam Fineberg (HP Software and Solutions),

Simona Cohen (IBM), Mary Baker (HP Labs), Roger Cummings (Symantec),

and the other members of the SNIA Long Term Retention TWG

Outline

- Introduction to digital preservation
- Preservation Technologies
- SIRF
  - Background
  - Goals
  - Use cases

Why preserve digital content?

- To protect our business
- For future business or historical needs
- Because our life is online

- SNIA 100 Year Archive Survey
  - 68% had to retain data more than 100 years
  - 83% had to retain data more than 50 years
  - Less than 20% were satisfied that they could access their retained data more than 50 years in the future

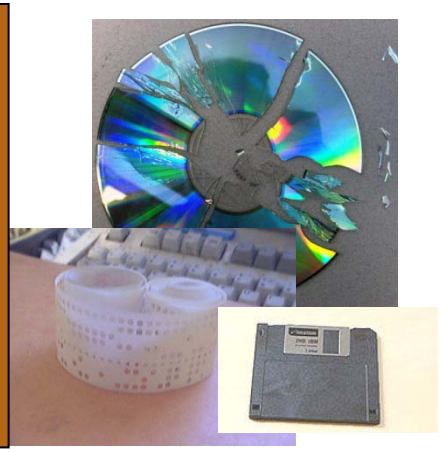
Goals of digital preservation

- Digital content should remain
  - accessible, usable, undamaged
  - for as long as desired
- Beyond the lifetime of
  - any storage system
  - any storage technology
- and at an *affordable cost*

Threats to long-term assets

- Large-scale disaster
- Human error
- Media faults

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



- Component faults
- Economic faults
- Attack
- Organizational faults

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

Even preserving the bits is hard

- Large scale & long time periods
  - 1 petabyte, 50 years, 50% chance no damage
    - That's a bit half-life of 10^{17} years
  - Even improbable events will have an effect
- Now try to keep
  - The bits usable - physical preservation
  - The information reusable - logical preservation

Practices vary by time

- Can't predict what will change – only know it will
- This means processes are key
  - Must be evolvable
    - Current processes get us to the next step
    - At that point we will likely need new processes to take over
  - Must not destroy what we are trying to protect
  - Standards make evolution easier
- A good archive is almost always in motion
  - Digital preservation is not a static activity!

Practices vary by context

- What do we preserve?
  - Bits? Applications? Logical connections? Context? Etc.?
  - Depends on customer domain
    - Example: digital copy of old book
      - words? wear and tear on the paper? political context?
  - Can't always predict the eventual use
    - Affordability may force some decisions
- What do we use?
  - Techniques
    - Virtual machines? Emulation? Canonical formats?
    - Self-describing formats? Standardized data models?
    - Loss-tolerant formats? Format migration?
    - Preservation of ancient equipment?
  - Yes: all could play a role for different domains

Preservation storage formats

- We can't predict the future
  - Systems will change, formats will change, systems will fail
- A preservation storage format must
  - Facilitate storage of preservation objects
  - Map to a wide variety of storage devices and technologies
  - Be resilient to failures and change
- Key properties
  - **Self contained** – to ensure objects are complete
  - **Self describing** – so software can interpret it
  - **Extensible** – so it can meet future needs

SIRF: Self Contained Information Retention Format

- Designed to emulate organizational practices developed for preserving physical objects
  - Archivists and records managers of physical items avoid processing individual items (e.g. documents, objects, records, etc.).
  - Instead, they gather together a group of related items, known as a series, collection, or record group.
  - Once assembled, the series is placed in a physical container, marked with a name and reference number
  - Information about the series will be included in a "finding aid" such as an online page that conforms to a defined schema

- SIRF is the digital equivalent to the physical container
  - Logical container for a set of (digital) preservation objects
  - Can also contain catalogs and metadata related to the entire contents of the container as well as to the individual objects.
  - Makes it easier to provide the processes needed to address threats to digital content

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

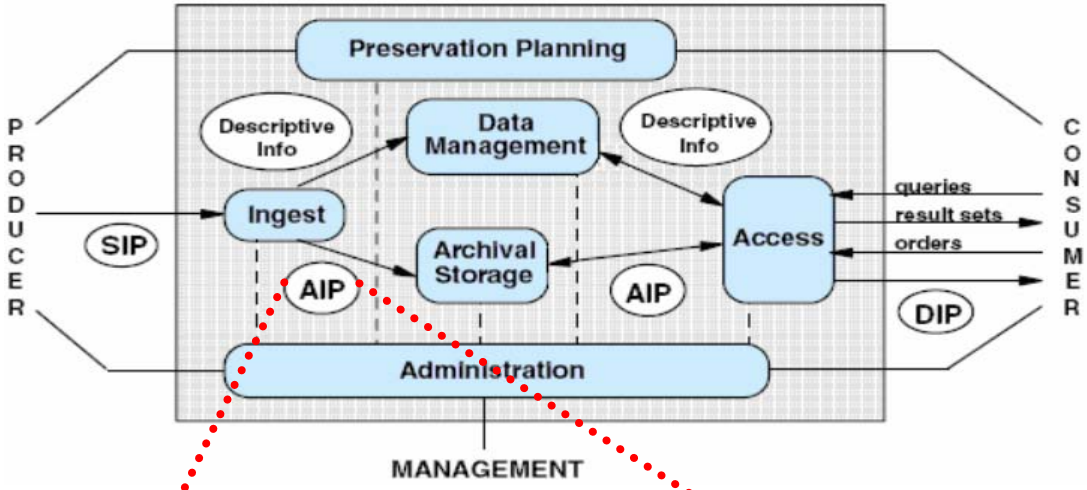
- Attributes of a PO
  - may be subject to physical and logical migrations
  - may be dynamic and change over time
  - an updated PO is a new version of the original, and its audit log records the changes that have occurred so authenticity may be verified

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

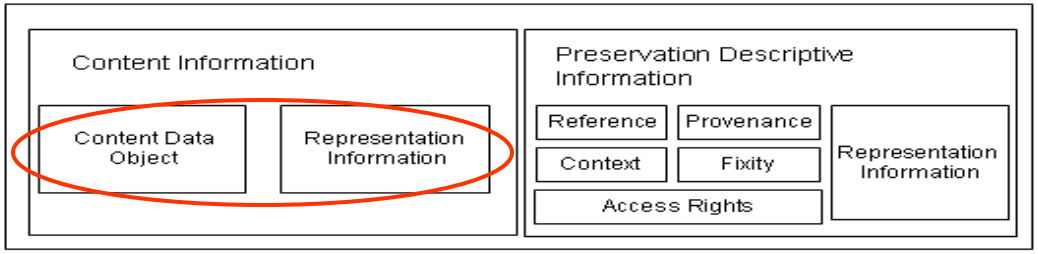
Open Archival Information System (OAIS)

- 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

*** OAIS Functional Model**

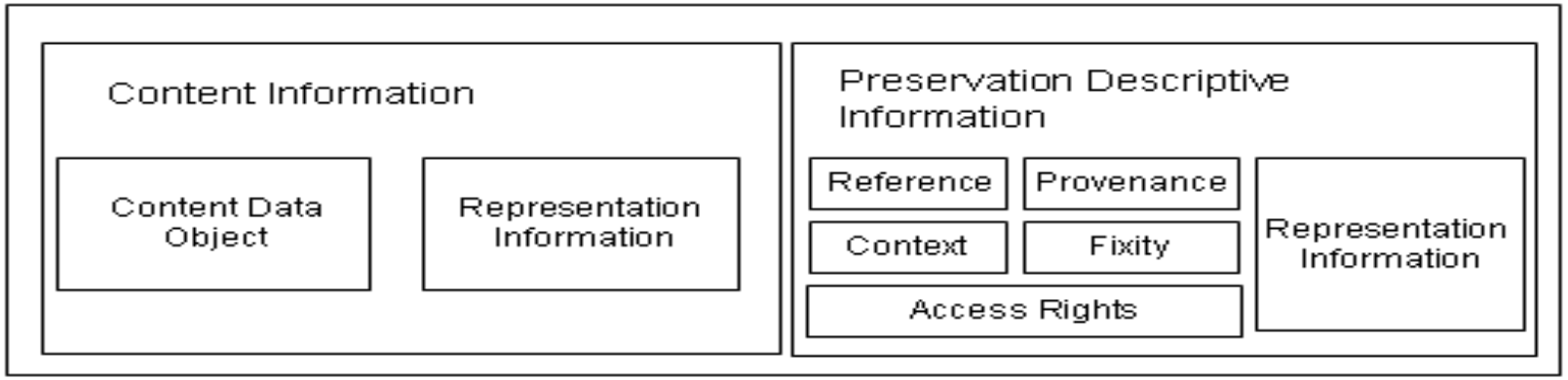


Preservation Object (AIP)



* Figure taken from the OAIS spec

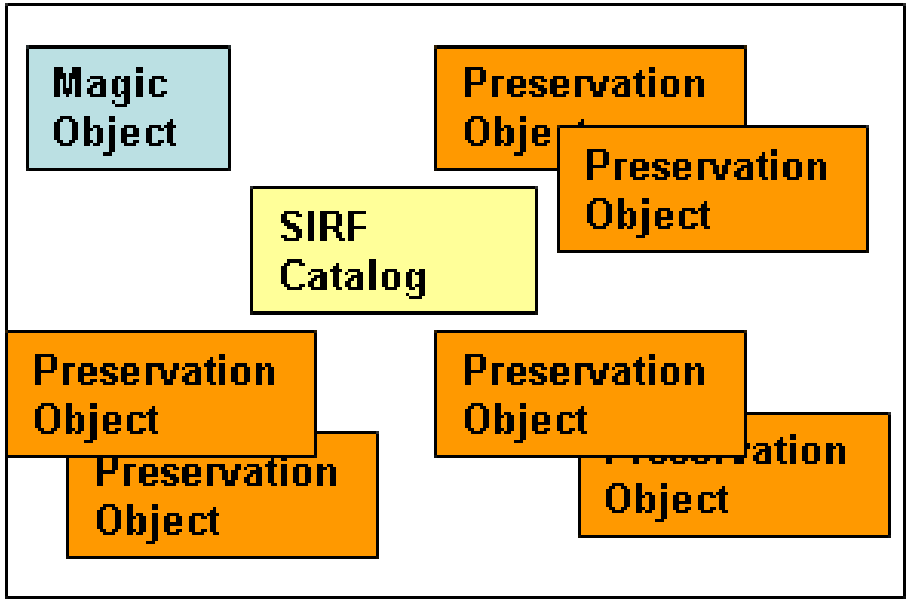
OAIS AIP Logical Structure



- **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.
- **Access Rights** - the information that identifies the access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control.

SIRF Containers

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



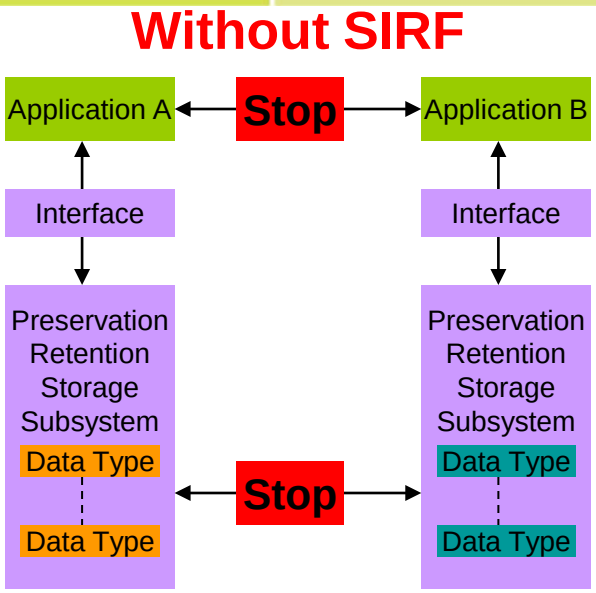
Layered Approach

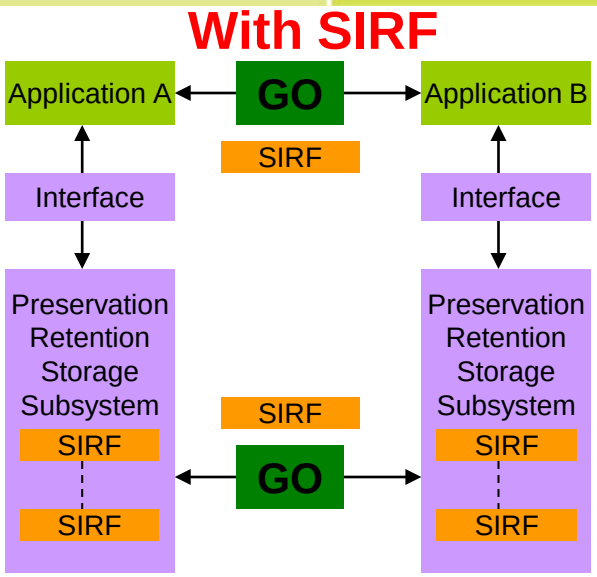
- SIRF is a logical data format
  - Assumes the underlying layer includes an object interface layer
  - Mountable units
    - Advanced: OSD, Cloud, XAM
    - Lower level: UDF, CDFS, FAT, LTFS
- SIRF defines two levels
  - Level 1 catalog (L1) – unique metadata, not in the preservation objects, that is 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

Existing Preservation Standards

- Generic formats
  - Bagit
  - JHOVE
- Domain specific packaging formats
  - XML Formatted Data Unit (XFDU)
  - VERS Encapsulated Object (VEO)
  - Metadata Encoding and Transmission Standard (METS)
  - Preservation metadata: Implementation Strategies (PREMIS)
- SIRF can be used with many of these, but it is unique because it
  - Preserves collections of objects and their relationships
  - Includes generic metadata that can be extended with domain specific information for fast access
  - Can be mapped to and physically migrated between a wide variety of underlying storage systems

Problems SIRF addresses





Sets of linked objects moved individually; referential integrity and context may be lost	Sets of linked objects moved between systems maintaining referential integrity and full context
Only original application that created the objects can read and interpret them	Any SIRF compliant application can read and interpret the objects
Export and import needed to migrate objects	Objects migrated without export and import
Preservation Objects cannot be sustained long-term	Preservation Objects can survive longer

SIRF – Status and methodology

- SIRF is being developed by the SNIA Long Term Retention TWG
  - http://www.snia.org/tech_activities/workgroups

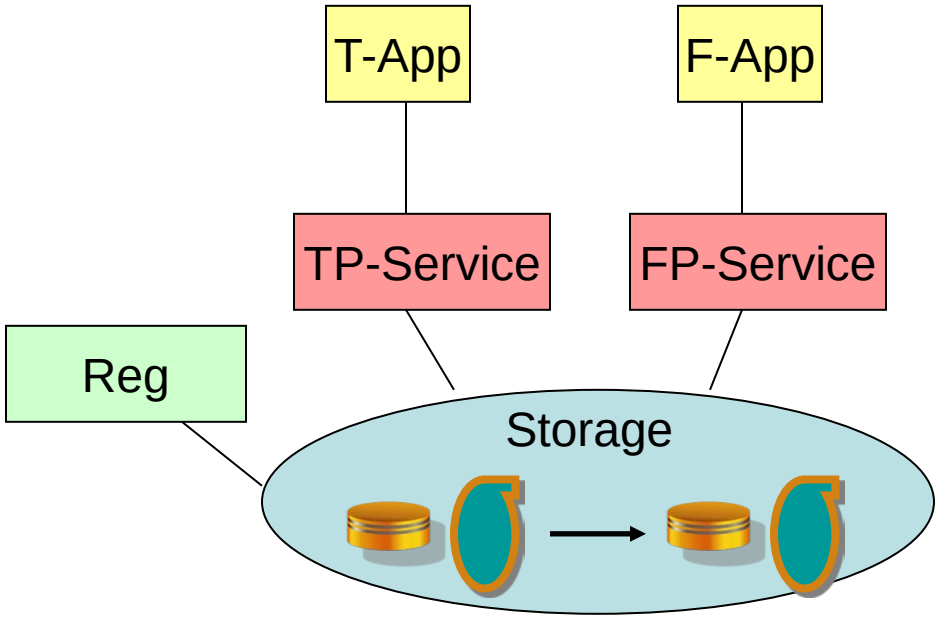
- The TWG just released a detailed use case and requirements document
  - Get your copy today!
- We are actively seeking input from the community before we begin drafting the specification

SIRF Use Cases – Methodology

- Define Actors involved in SIRF
- Define use cases and flows among the actors
  - Generic uses cases
    - Unlinked to specific type of data or application
    - Technological changes in the environment
  - 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

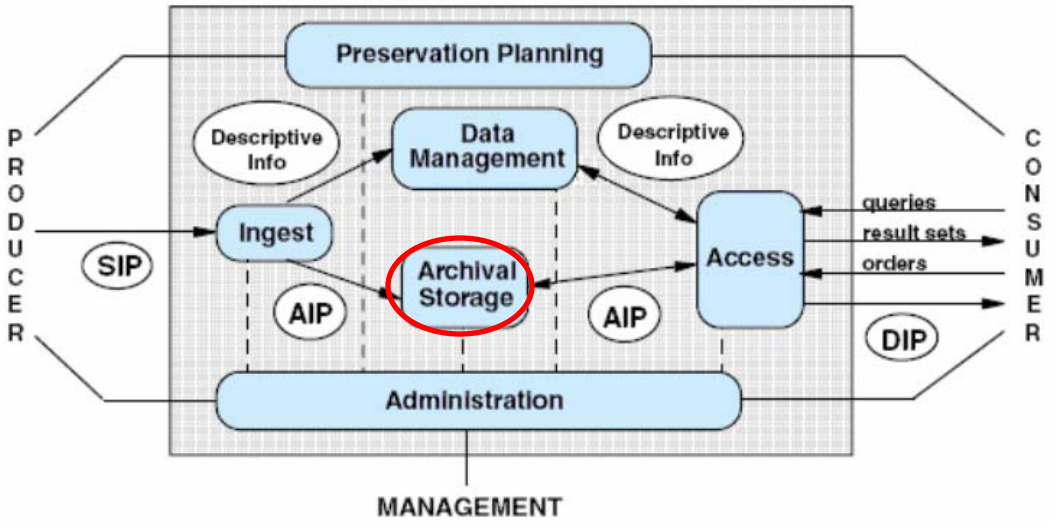
Use Case Model

- Human actors
  - Archive Employee
  - Consumer
  - Preservation Manager
  - Producer
  - System Administrator
  - Auditors
- 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



Mapping OAIS Functional Model to SIRF Actors

OAIS Function Model:

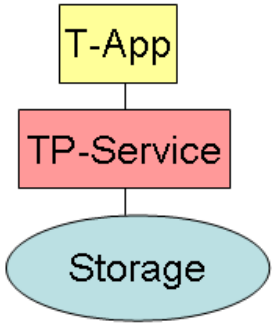


Mapping:

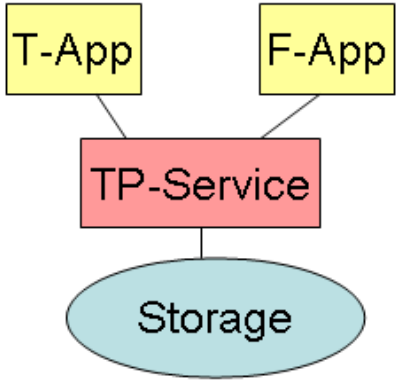
SIRF Actors	OAIS Functional Model
Storage	Archival Storage
TP-Service, FP-Service	Data management, Ingest, Access, Administration, Preservation Planning
T-App, F-App	Producer, Consumer
Reg	-

Generic use cases

- Ingest and Access with Same Application



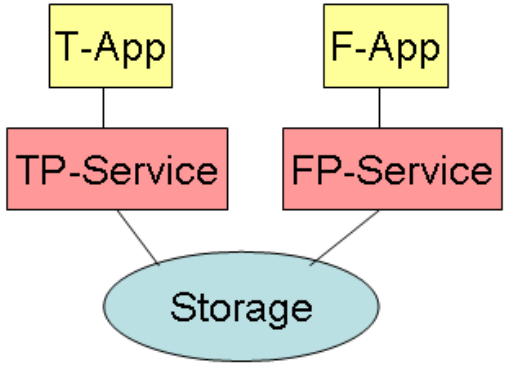
- Ingest and Access with Different Applications



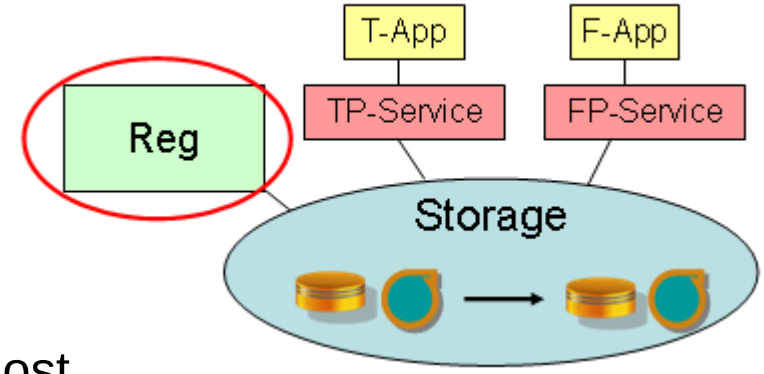
- Requirements
  - Support for standard interfaces e.g. NFS, CIFS, XAM
  - Agnostic to media, platform, vendor
  - Support multiple versions of preservation objects
  - Support multiple data models and multiple formats for the raw data

Generic Use Cases (cont.)

- Ingest and Access with Different Preservation Services



- Storage Format Change



- Requirements
  - Self-contained data so nothing is lost
  - Persistent globally unique identifiers for POs so references work
  - Self-describing via meta-language that itself should be changeable
  - SIRF Representation Information in preservable external registry

Workload Use Cases: eDiscovery

- Flow:
  1. T-App ingests a Preservation Object today via TP-Service.
  2. Time passes and the data becomes subject to eDiscovery.
  3. Potentially responsive preservation objects are identified using provenance, context and content information stored with preservation objects.
  4. Identified objects are put on “legal hold,” preventing deletion or modification.
  5. Identified objects are copied from the preservation system and collected to a case repository for processing, review, and analysis.
  6. At some future date “legal hold” is removed. The object may become subject to other legal holds or retention /disposition policies at any time.

- Main Requirements include support for:
  - retention holds on POs that prevent their deletion or modification
  - verification of document provenance and authenticity
  - verification of completeness and correctness
  - storing audits, potentially records of modification, access, etc.
  - identification, collection, and preservation of POs relevant to a legal matter

Workload Use Cases: eMail archive

- 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 include support for:
  - links between objects that are immutable as the objects
  - auditable time stamps that are immutable and created by known authority
  - "special" POs such as a PO that includes address book, search results, etc.
  - organizational unique metadata

Workload Use Cases: Consumer archive on cloud

- Flow:
  1. A user creates a genealogy container for his genealogy-related documents on a cloud that provides SLAs for preservation
  2. The user uses T-App to ingest a genealogy-related document via TP-service on the cloud
  3. TP-service on the cloud ingests the PO plus transforms the data to pdf/a and ingests the resulting PO version to the same genealogy container
  4. Time passes and the grandchildren would like to get that document
  5. FP-service will verify the grandchildren identity and will provide appropriate credentials to access the genealogy container and the document
  6. F-App access via FP-Service the latest version of the document and renders the pdf/a document

- Main Requirements include support for:
  - transformations of preservation objects
  - secured access to the data that is updatable over time
  - cloud containers to be SIRF-compliant, so containers can be migrated
  - Verification of document provenance and authenticity

Workload Use Cases: BioMedical bank

- Flow:
  1. T-App ingests via TP-service a PO that includes a standardized DICOM image of the leg of a patient that is a minor
  2. Time passes and the patient who is now an adult, scheduled an appointment to check a new problem he has encountered in the leg
  3. The identified Preservation Objects including the PO with the DICOM image will be a-priory brought from an offline media to an online media to be timely accessible for the appointment
  4. F-App at point of care access the identified PO that includes the DICOM image via FP-Service
  5. More time passes and a researcher from an adjacent academic medical research center wants to access that image for research purposes. According to HIPAA regulations, the researcher can get just a de-identified image
  6. F-App access the de-identified PO via FP-Service

Main Requirements include support for:

- hierarchical storage management
- masking of sensitive data
- verification of document provenance and authenticity

Use Case 9: Merged cloud repositories

- Flow:
  1. T-App ingests via TP-service a PO in a cloud that is provided by company “FirstCloud”
  2. T-App also ingests via TP-service a second PO in a second cloud provided by company “SecondCloud”
  3. Time passes and the two companies “FirstCloud” and “SecondCloud” are merged and their two cloud repositories are combined. This is possible as the POs identifiers are globally unique
  4. F-App access via FP-Service the two POs in the combined cloud provided by the merged company

Main Requirements include support for:

- SIRF-compliant cloud containers, so they can be interpreted by other clouds
- persistent **globally** unique identifiers for the preservation objects

Summary

- Digital preservation is an important problem that is only growing in importance over time
  - Long term is different from short
  - Digital preservation is different from preserving physical objects
  - Best practices exist, but it is not a solved problem
- Solutions are being developed, but they are
  - Domain specific
  - Based on assumptions about future needs
- The SNIA LTR-TWG is trying to improve the state of the art by developing SIRF
  - An extensible storage format, not for a specific domain
  - Suitable for long term preservation
  - Storable on a wide range of media and technologies
  - SNIA is seeking input on SIRF

About the SNIA LTR TWG

- This presentation has been developed by members of the SNIA Long Term Retention Technical Working Group (LTR TWG)
  - http://www.snia.org/tech_activities/workgroups
- 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!