LTR TWG & the Cloud

Roger Cummings
Co-Chair, LTR TWG

LTR TWG Introduction

- TWG full chartered in mid 2008
  - 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.

Program of Work Overview

- Divided into five logical areas with
  1. Liaisons
     a. Internal – OSD, Security TWGs
     b. External – SAA, IETF LTANS, IEEE Mass Storage, LoC, JEITA, AIIM, SMPTE
  2. Reference Model and related materials
     a. Overview of OAIS
     b. Knowledgebase of long-term retention & preservation
     c. Standards survey
     d. Storage arch model appropriate for long-term preservation of digital information
     e. Metadata gap analysis
     f. Technical authority for submitting terms to SNIA Dictionary

Program of Work Overview (Contd)

3. Logical Preservation
   a. Identify set of generic & workload-based use cases for a logical container format
   b. Define logical container format
   c. Interoperably standardize container format
   d. Propose early revision of c) as SNIA Architecture
   e. Define compliance strategy & interop guidelines
   f. Create info base on why, how and when logical preservation fails
   g. Develop use cases for SIRF as interchange format in cloud-based repositories
4. Education
   a. Create SNW tutorials
   b. Get materials presented @ other appropriate conferences
   c. Create bidir briefing program between the TWG and leading exponents
5. Bit Preservation
   a. Create info base on why, how and when phys storage loses integrity
   b. Create good practice defs & processes to maximize physical preservation at low cost.

Physical (Bit) Preservation

- Making sure that the bits are preserved indefinitely into the future
  - Longer than any storage device or any storage device technology will exist
  - Inevitably includes a migration process
- Based on seminal paper
  - M. Baker, et al., “A Fresh Look at the Reliability of Long-term Digital Storage.” EuroSys’06.
- One of the authors (Mary Baker of HP Labs) participates in the TWG
  - Has co-presented the SNW tutorials for the last 2 years

Logical Preservation

- Making sure that information is USABLE indefinitely into the future
  - Extending existing practices into the digital domain
- Based on work done in the EU CASPAR project
  - See http://www.casparpreserves.eu
  - Goes forward from OAIS model
  - Simona Cohen, of IBM Haifa Labs, who was a leading participant of CASPAR, is co-Chair of the LTR TWG
- Creating Self Contained Information Retention Format (SIRF)
  - Digital equivalent of the standard archive box
  - SIRF "Use Cases & Requirements" document in SNIA Public Review
    - Please give us feedback and help us get more!

SIRF: Self Contained Information Retention Format

- Designed to emulate best 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

- 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

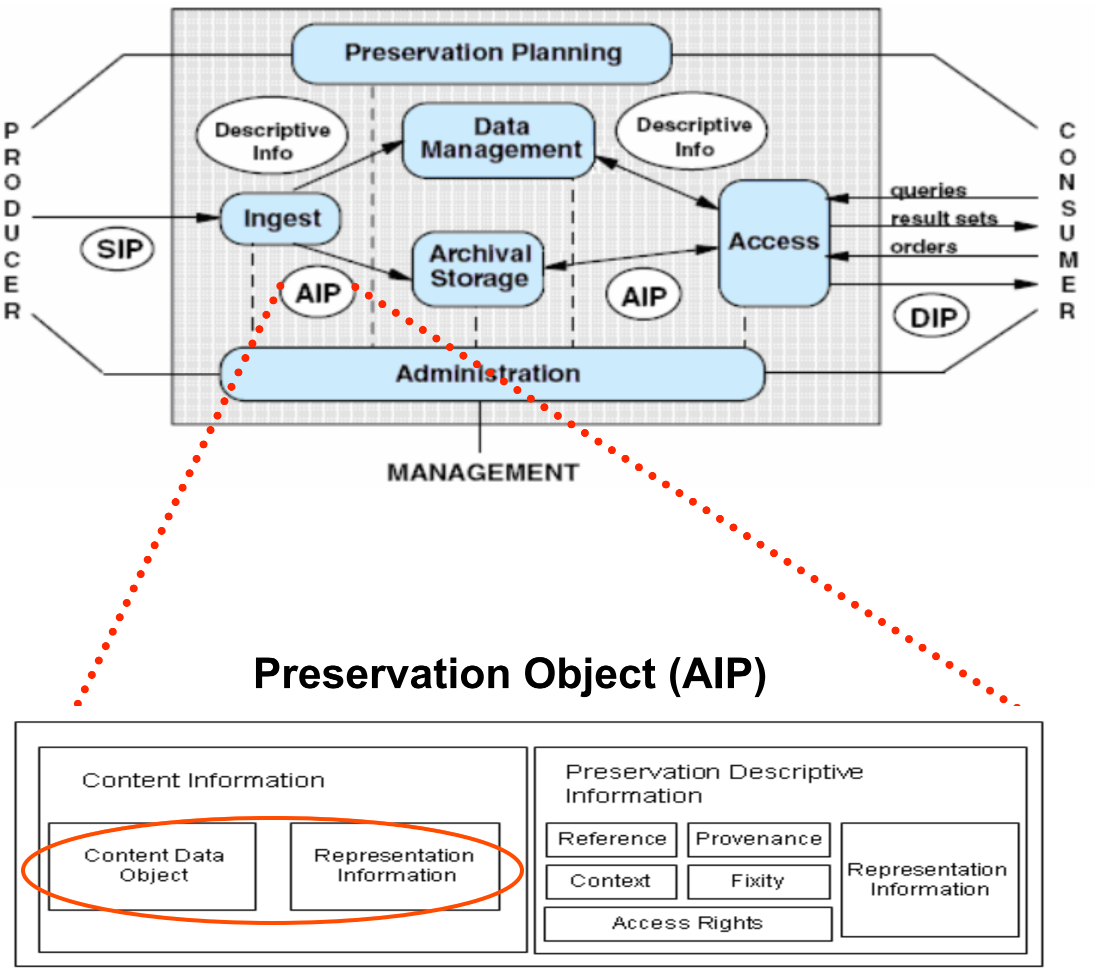
Open Archival Information System (OAIS)

SNIA
Advancing storage & information technology

- 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



Preservation Object (AIP)

UC7: Consumer Archive on the Cloud

- An individual wants to preserve his family photos and documents on a cloud that provides preservation services, so that forthcoming generations will be able to access that data and study their roots.

Use Case 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 with the original document as well as transforms the document to a standardized format believed to be more sustainable such as 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 validate the grandchildren identity and will provide appropriate credentials to access the genealogy container and the document.
6. F-App which is a future application executing on technology used at that future time, access via FP-Service the latest version of the document and renders the pdf/a document.

Derived Requirements

- Support for transformations of preservation objects e.g., support various versions of the PO and the tree structure they create
- Support for managing identifiers over time
- Support secured access to the data that is updatable over time e.g., when a security mechanism becomes weak
- Support cloud containers to be SIRF-compliant, so containers can be migrated to other clouds with all the required preservation information
- Verification of document provenance and authenticity, regardless of migrations whether logical or physical

Available LTR TWG information

- Recent BrightTalk webinar "Effective Use of Open Standards for Digital Archiving", Peter Van Garderen, Artefactual
  - See http://www.brighttalk.com/webcast/24671
- SYSTOR Conference 2010 (Haifa, IL) keynote "Bread Crumbs Don't Last" , Mary Baker, HP
  - See http://www.research.ibm.com/haifa/conferences/systor2010/program.shtml
- ECMA STAR Conference (Lausanne, CH) keynote "Long Term Retention Work @ SNIA", Roger Cummings, Symantec
  - See http://mmspl.epfl.ch/webdav/site/mmspl/shared/star2010/ppt/star2010_cummings.pdf
  - See http://www.ustream.tv/recorded/6476974

Available LTR TWG Information

- IEEE MSSTC 2010 "Standardization Efforts Related to Long-Term Retention and Preservation", Don Post, IMERGE, et al
- SNIA SDC 2010 (Santa Clara, CA) "Long Term Information Retention", Sam Fineberg, HP, et al
  - See http://www.snia.org/events/storage-developer2010/presentations/general_session/SamFineberg-Long_Term_Information_Retentionv9.pdf
- Presented "Retaining Information for 100 years" tutorial @ SNW USA for last two years getting very good marks & feedback
  - Latest version available @ http:/www.snia.org/ltr
  - Webinar version coming soon

LTR & the Cloud

- The LTR TWG work on physical and logical preservation is very applicable to cloud-based archiving
  - SIRF is storage format agnostic
  - SIRF will be able to utilize all types of metadata
  - The fundamental archival processes have to take place even in cloud-based repositories
- Major question is how a cloud-based archive relates to OAIS
  - Is the cloud portion just a scalable, cost-effective storage platform?
  - Does a complete service-based archive exist in the cloud?
    - e.g. The Dspace (see http://www.dspace.org) ingest interfaces become web services?