



Education

Bringing Light to the “Digital Dark Age” – Preserving Digital Information for the Long Term

Roger Cummings, Symantec Research Labs

Co-Author: Simona Rabinovici-Cohen, IBM Research – Haifa

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced in their entirety without modification
 - ◆ The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

➤ Bringing Light to the “Digital Dark Age” – Preserving Digital Information for the Long Term

- ◆ Many organizations are facing the serious challenge of economically preserving and retaining access to a wide variety of digital content for dozens of years. Long-term digital information is vulnerable to issues that do not exist in a short-term or paper world, such as media and format obsolescence, bit-rot, and loss of metadata. Ironically, as the world becomes digital, we may be entering a "Digital Dark Age" in which business, public and personal assets are in ever greater danger of being lost.
- ◆ The SNIA Long Term Retention (LTR) Technical Working Group works with key stakeholders in the preservation field, to develop the Self-contained Information Retention Format (SIRF) and enable applications to interpret stored data, independent of the application that originally created it. SIRF is a logical container format for the storage subsystem appropriate for the long-term storage of digital information. SIRF consists of preservation objects and a catalog containing metadata relating to the entire contents of the container as well as to the individual preservation objects and their relationships. It makes it easier and more efficient to provide many of the processes that address threats to the digital content at a lower level of the system stack and can be performed close to the data using more robust, efficient, and automatic methods. Easier, more efficient preservation processes in turn lead to more scalable and less costly preservation of digital content

➤ Introduction

- ◆ Why we need digital preservation
- ◆ The goals of digital preservation
- ◆ Migration

➤ SNIA activities

- ◆ Self-contained Information Retention Format (SIRF)
- ◆ Preservation Objects
- ◆ Use Cases & Requirements

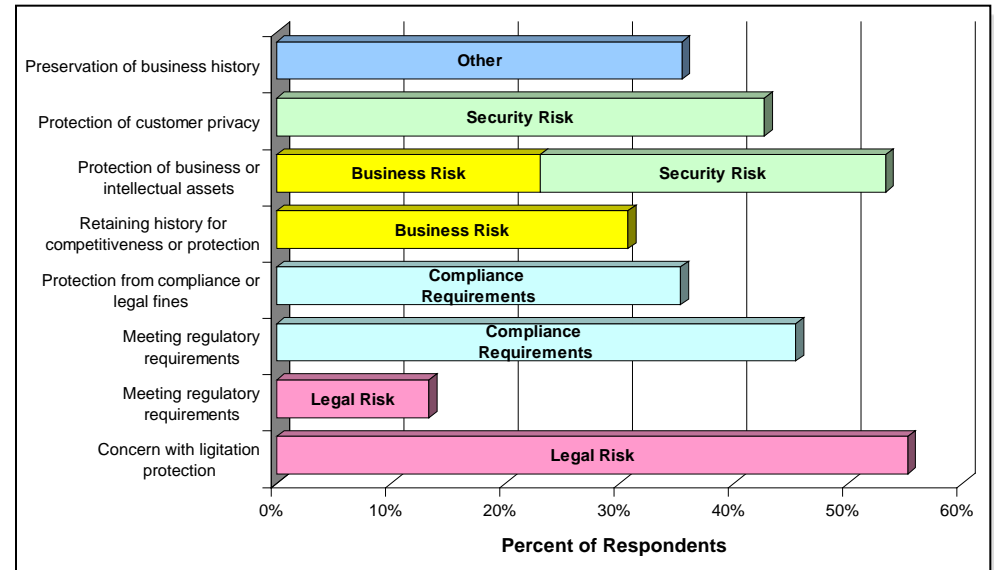
➤ EU Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value (ENSURE)

➤ Summary

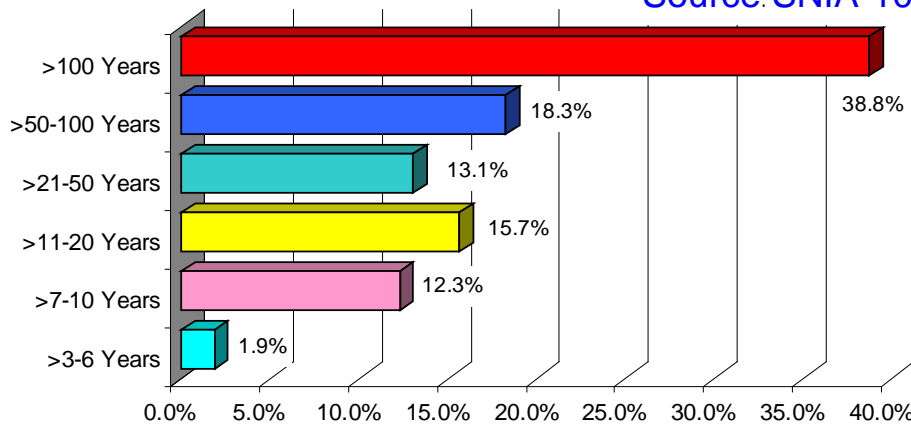
- Regulatory compliance and legal issues
 - ◆ Sarbanes-Oxley, HIPAA, FRCP, intellectual property litigation
- Emerging web services and applications
 - ◆ Email, photo sharing, web site archives, social networks, blogs
- Many other fixed-content repositories
 - ◆ Scientific data, intelligence, libraries, movies, music

SNIA Survey from 2007

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?
Retention of 20 years or more is required by 70% of respondents



- 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*

- 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

➤ Introduction

- ◆ Why we need digital preservation
- ◆ The goals of digital preservation
- ◆ Migration

➤ **SNIA activities**

- ◆ Self-contained Information Retention Format (SIRF)
- ◆ Preservation Objects
- ◆ Use Cases & Requirements

➤ EU Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value (ENSURE)

➤ Summary

- LTR TWG formed on the basis of the 2007 survey results
- TWG's 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

An Analogy

Photo courtesy Oregon State Archives



➤ 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

- Self-contained Information Retention Format (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 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 etc.
 - ◆ SIRF contains a set of preservation objects to be understood in future

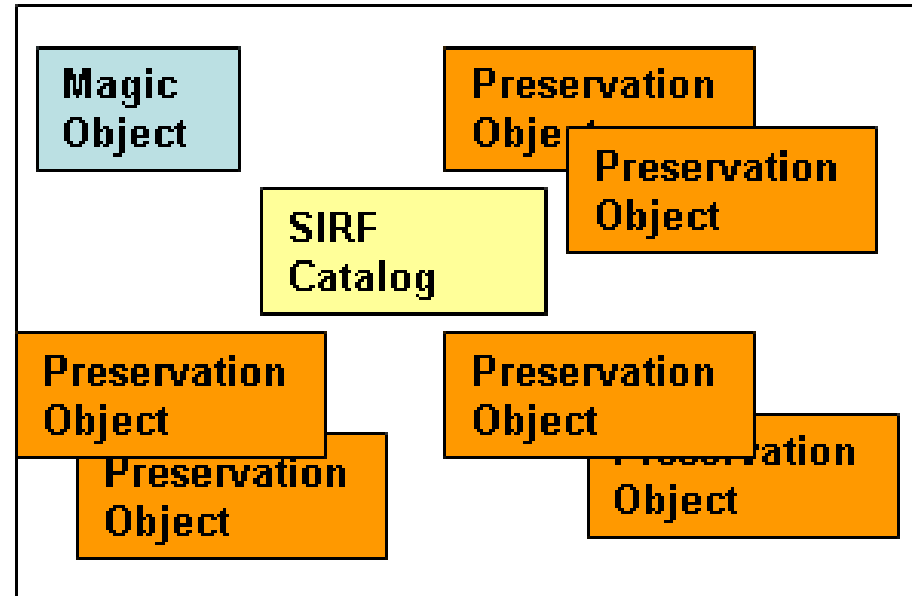
Required SIRF Properties

- 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?

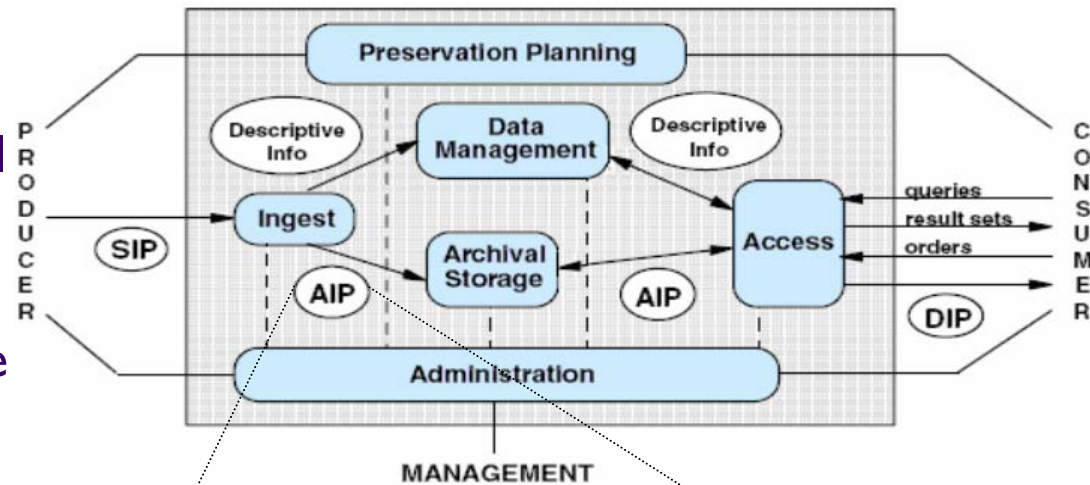
- SIRF Containers Store Collections of Preservation Objects (POs)
- A Preservation Object is
 - ◆ The raw data to be preserved,
 - ◆ Plus additional embedded or linked metadata, and
 - ◆ Plus everything needed to enable sustainability of the information contained in raw data for decades to come

- Attributes of a PO
 - ◆ May be subject to physical and logical migrations
 - ◆ May be dynamic and change over time
- Result of migrations or attribute changes is creation of a separate and **new version** of the original PO
 - ◆ New version linked to previous version
 - ◆ Audit log records the changes so authenticity verifiable
- Several examples of a PO exist
 - ◆ e.g. OAIS Archival Information Package (AIP)
 - › Includes recursive representation information that enables future interpretation of the raw data

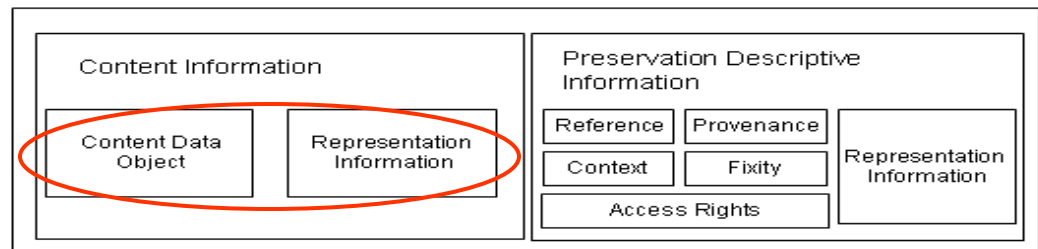
Open Archival Information System (OAIS)

- 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



AIP



* Figure taken from the OAIS std

➤ SIRF is logical data format only

- ◆ Assumes underlying layer includes object interface layer
 - Examples
 - Advanced: OSD, Cloud, XAM
 - Lower level: UDF, CDFS, FAT, LTFS

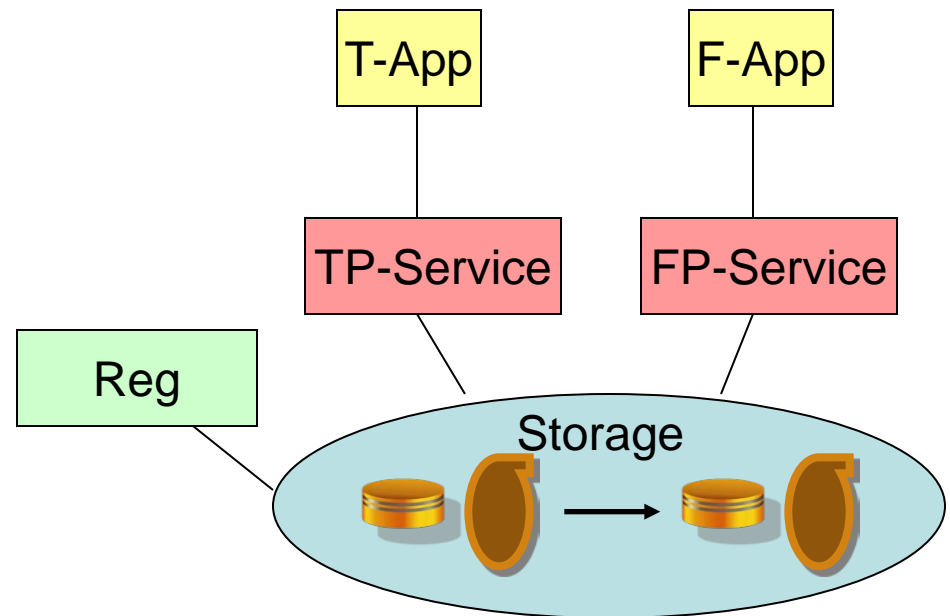
➤ SIRF defines two levels

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

- Define Actors in SIRF
- Define use cases and flows among actors
 - ◆ Four generic use cases defined
 - › Unlinked to specific type of data or application
 - › Technological changes in the environment
 - ◆ Five workload-based use cases defined
 - › Specialized for concrete workloads
 - › Additional non-technological changes in the environment

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



1. T-App ingests an e-mail thread today via TP-Service.
 - ◆ Includes ingesting a collection of several interrelated 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.
 - ◆ FP-Service creates a set of new version POs for the address book and the organizational processes
3. More time passes & F-App searches the repository and creates POs for the search results to raise performance of future searches
 - ◆ New POs may contain soft links to the thread, messages and attachments created in step 1

- For each use case, find 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

Requirements from Use Case

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

Real Life Example Problem

2003

To: roger.cummings@veritas.com
From: fred@nowhere.com
Subject: Something or other

2007

To: roger_cummings@symantec.com
From: sue@somewhere.com
Subject: Something else

Same people?? Could you PROVE it 20 years on?

To: gary.phillips@veritas.com
From: fred@nowhere.com
Subject: Something or other

To: gary_phillips@symantec.com
From: sue@somewhere.com
Subject: Something else

- 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

➤ Introduction

- ◆ Why we need digital preservation
- ◆ The goals of digital preservation
- ◆ Migration

➤ SNIA activities

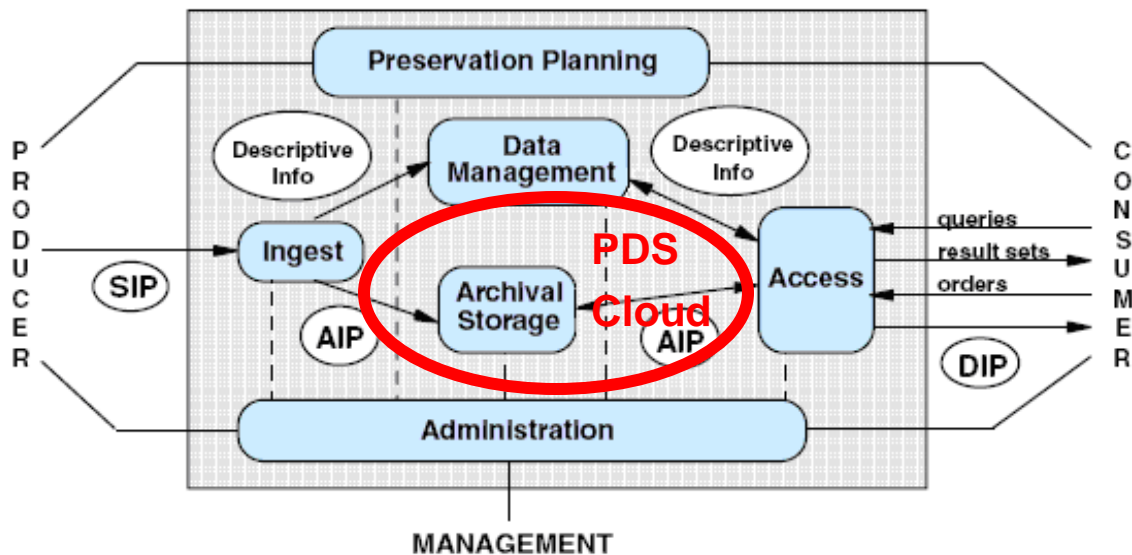
- ◆ Self-contained Information Retention Format (SIRF)
- ◆ Preservation Objects
- ◆ Use Cases & Requirements

➤ EU Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value (ENSURE)

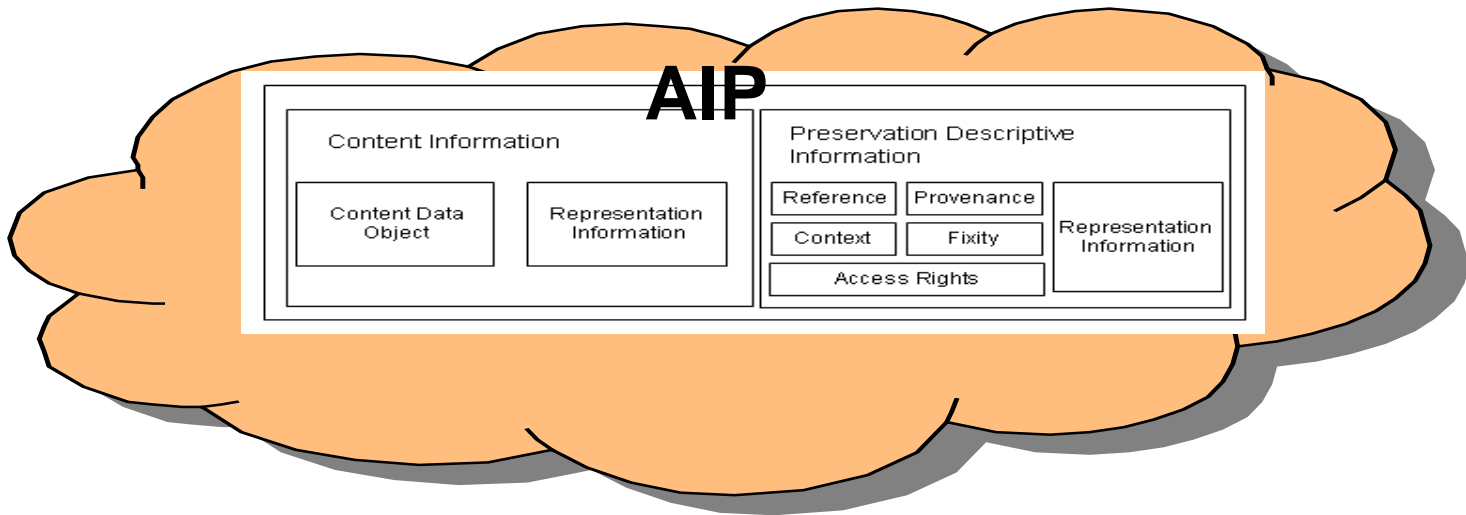
➤ Summary

- **ENSURE is FP7 EU Project in the area of preservation**
 - ◆ Three year Integrated Project (IP) started Feb. 1, 2011
 - ◆ Consortium of 13 partners (industry and academic)
- **ENSURE has a business/industry-oriented focus**
 - ◆ Drivers for preservation are both regulatory and business value
 - ◆ Past efforts on digital preservation have focused on memory institutions addressing the good of society
- **Demonstrated with three use case: Health Care, Clinical Trials and Finance**
 - ◆ Create a financially viable preservation solution
 - ◆ Handle regulation and lifecycle management
 - ◆ Expand use of emerging technologies for digital preservation
 - ◆ Maintain access and privacy rights to information and IP

- Provides preservation-aware storage services for ENSURE
- Based on OAIS Archival Storage entity but provides more automation of preservation processes
- Built on top of multiple clouds concurrently, while taking advantage of each one's special capabilities
- Includes a SIRF Handler component for future implementation

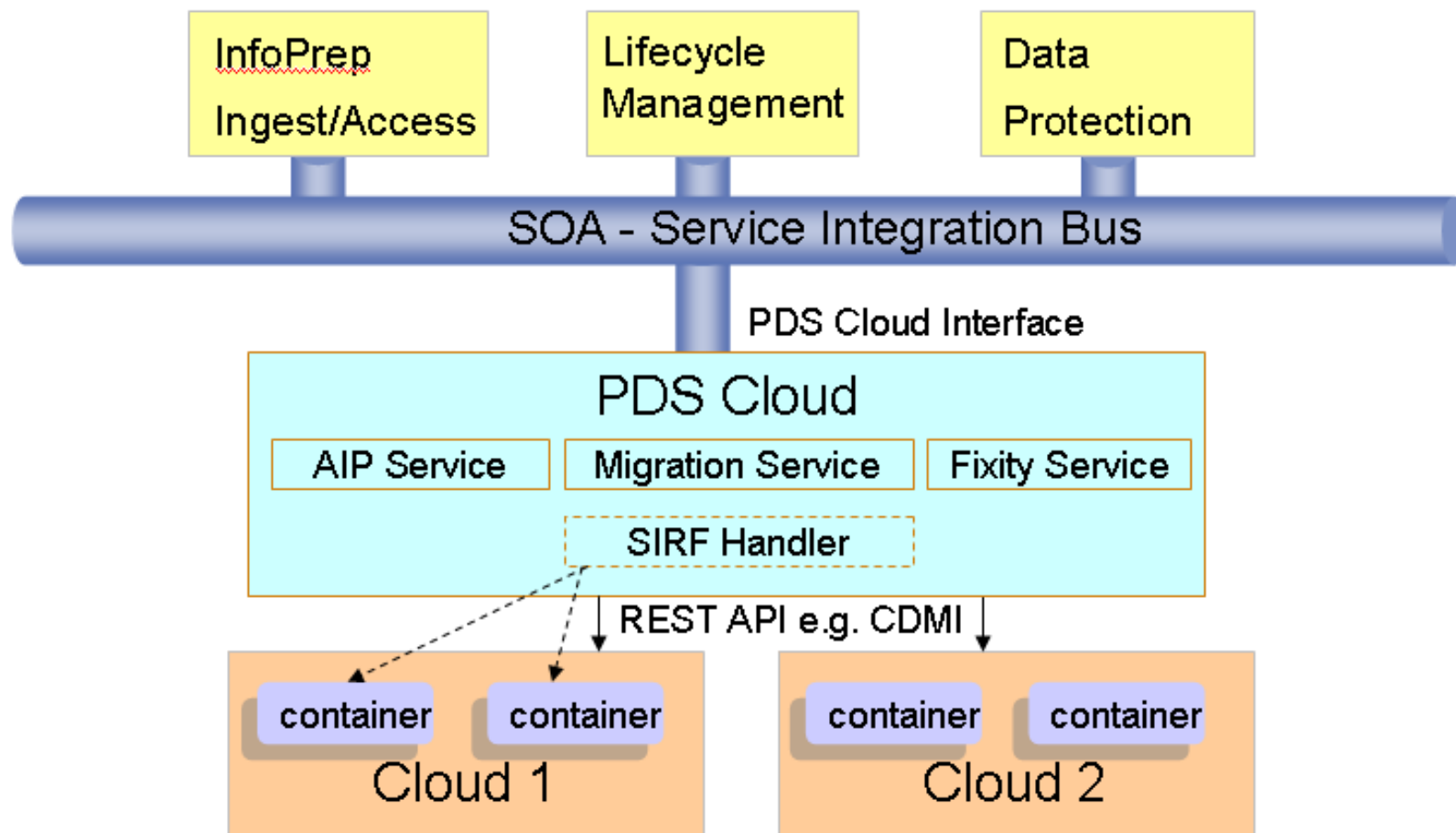


PDS Cloud Functionality



- Map OAIS AIP and the links among AIPs to the cloud data model
- Multi cloud support while considering self-containment and self-describing implications
- Migrate data in its entirety including all its metadata, provenance, context, representation information
- Support multiple integrity (fixity) checks with updatable algorithms
- Support preservation object copies over multiple clouds and reduce the data lock-in problem

PDS Cloud and SIRF in ENSURE



Note: SIRF Handler is for future implementation

- Cloud Data Management Interface (CDMI) specifies standard API for clouds
- CDMI API can be used to access the various preservation objects and the catalog object in a SIRF-compliant container
- Example
 - ◆ Assume a cloud container named "Patient X" that is SIRF-compliant.
 - › Container holds several medical records of this patient where each medical record is a PO.
 - › Additionally, container holds a catalog object.
 - ◆ Each medical record (PO) & catalog object can be read via CDMI REST API:
 - › GET <root URI>/<ContainerName>/<DataObjectName>
 - › GET <root URI>/Patient X/catalog
 - › GET <root URI>/Patient X/MedicalRecord I

➤ Introduction

- ◆ Why we need digital preservation
- ◆ The goals of digital preservation
- ◆ Migration

➤ SNIA activities

- ◆ Self-contained Information Retention Format (SIRF)
- ◆ Preservation Objects
- ◆ Use Cases & Requirements

➤ EU Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value (ENSURE)

➤ Summary

Summary

- Best practices will vary over time
- We can't predict what will change – we only know it will
 - ◆ Ability to evolve is most important aspect of digital preservation
- This means processes are key
 - ◆ Must ensure our preservation processes are evolvable
 - ◆ Current processes are the first step in an iterative solution
 - › They get us to the next step
 - › At that point we will likely need new processes to take over
 - ◆ Widely understood standards make process evolution easier
- A good archive is almost always in motion
 - ◆ Migrating, auditing, re-keying, etc.
 - ◆ *Digital preservation is not a static activity!*
 - ◆ You can't just “do it and be done with it”

- SIRF container is key to long term information retention, 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

Hands-On
LAB COMPUTERWORLD SNIA
SNW



Advanced Data Protection and Reduction

- Please send any questions or comments on this presentation to SNIA: tracktutorials@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.**

- SNIA Education Committee

**Mary Baker
Simona Rabinovici-Cohen
Roger Cummings
Sam Fineberg**

**John Marberg
Don Post
Bob Rogers**

For further information

- SIRF use cases and requirements document is released for public review
 - ◆ http://www.snia.org/tech_activities/publicreview
- More information on SIRF (& other SNIA LTR activities) is available at
 - ◆ <http://www.snia.org/ltr>
- More information on ENSURE is available @:
 - ◆ http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&CAT=PROJ&RCN=98002