

# Long Term Information Retention

**Sam Fineberg (HP Software)**

**Simona Rabinovici-Cohen (IBM)**

With lots of help from other members of the SNIA LTR TWG including Mary Baker, Roger Cummings, John Marberg, Gene Nagle, Michael Peterson, Don Post and Bob Rogers

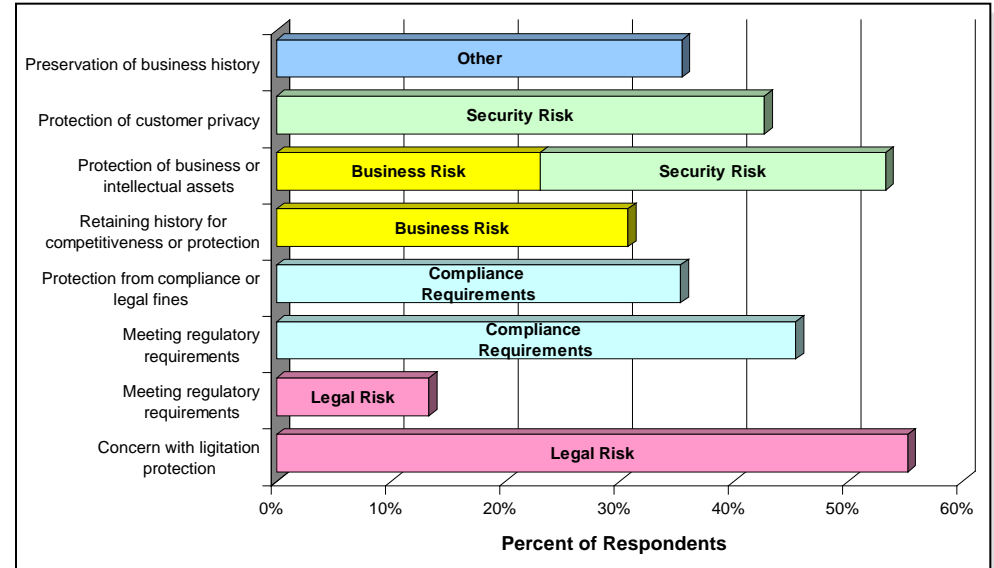
- Introduction to digital preservation
- Preservation Technologies
- SNIA SIRF: Self-contained Information Retention Format
- EU ENSURE: Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value
- Summary

# Avoiding the Digital Dark Age

- ❑ More and more critical information is created digitally and never sees paper
  - ❑ Documents, Web Pages, videos, music, photos, ...
  - ❑ Medical data, business records, historical documents, ...
  
- ❑ We have known for years that digital information is easy to lose
  - ❑ But preservation is hard, expensive, and poorly defined
  - ❑ Governments and libraries are just starting to grapple with the problem, businesses have largely ignored it
  
- ❑ As a consequence, most businesses and individuals are in danger of losing information, and may not even know it is happening
  
- ❑ **We are at risk of losing decades of digital content before we ever get around to preserving it.**

# 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.

## Key findings

- 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

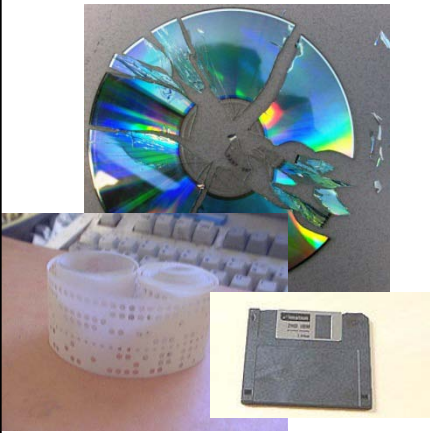


# Threats to long-term assets

- ❑ 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

# Even preserving the bits is hard

- ❑ Large scale & long time periods
  - ❑ Extremely unlikely that a single copy of a large corpus can be completely error free
  - ❑ 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

# SNIA's effort to address preservation

- ❑ 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

- ❑ 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

Photo courtesy Oregon State Archives

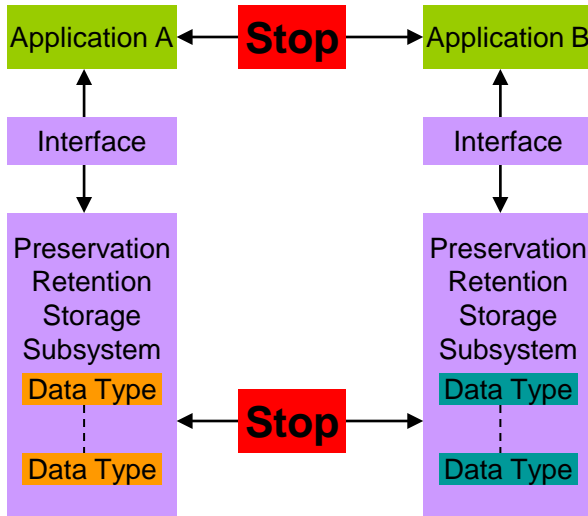


- SIRF is the digital equivalent
  - Logical container for a set of (digital) preservation objects and a catalog
  - The SIRF catalog contains metadata related to the entire contents of the container as well as to the individual objects
  - SIRF standardizes the information in the catalog

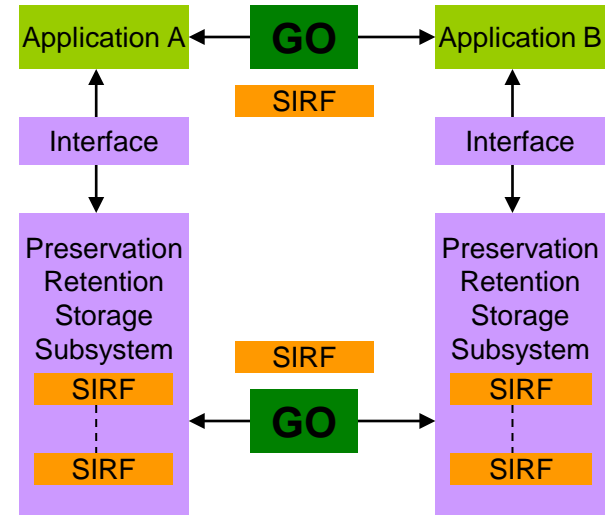


# Problems SIRF addresses

## Without SIRF



## With SIRF



Sets of linked objects moved individually; referential integrity and context may be lost

Only original application that created the objects can read and interpret them

Export and import needed to migrate objects

Preservation Objects cannot be sustained long-term

Sets of linked objects moved between systems maintaining referential integrity and full context

Any SIRF compliant application can read and interpret the objects

Objects migrated without export and import

Preservation Objects can survive longer

- ❑ Introduction to digital preservation
- ❑ Preservation Technologies
- ❑ **SNIA SIRF: Self-contained Information Retention Format**
- ❑ **EU ENSURE: Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value**
- ❑ Summary

# Self-contained Information Retention SDC

## Format (SIRF)

STORAGE DEVELOPER CONFERENCE  
SNIA ■ SANTA CLARA, 2011

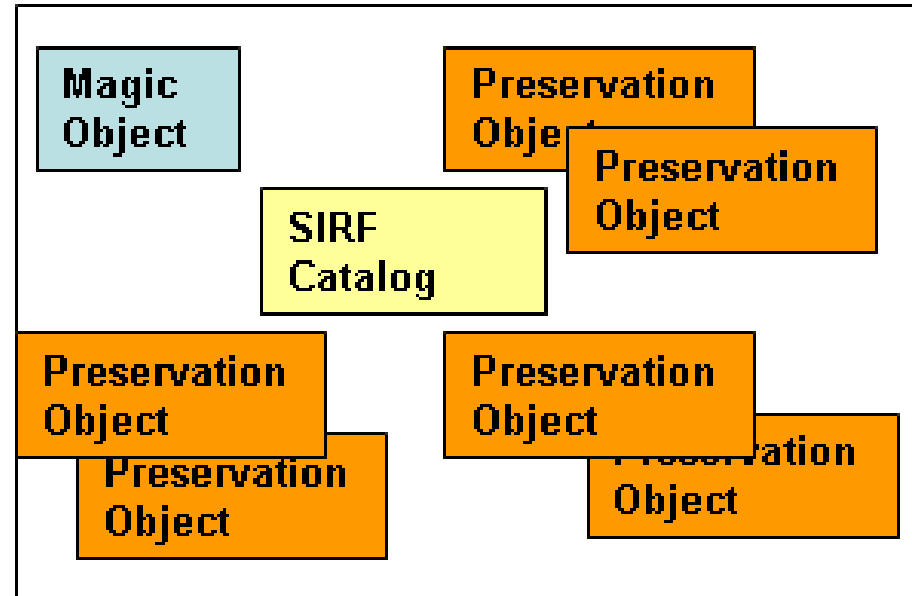
- ❑ SIRF is a logical data format of a **storage container** appropriate for long term storage of digital information
  - ❑ A storage container may comprise a logical or physical storage area considered as a unit.
    - ❑ Examples: a file system, a tape, a block device, a stream device, an object store, a data bucket in a cloud storage
  
- ❑ Required Properties
  - ❑ **Self-describing** – can be interpreted by different systems
  - ❑ **Self-contained** – all data needed for the interpretation is in the container
  - ❑ **Extensible** – so it can meet future needs



# SIRF Components

A SIRF container includes:

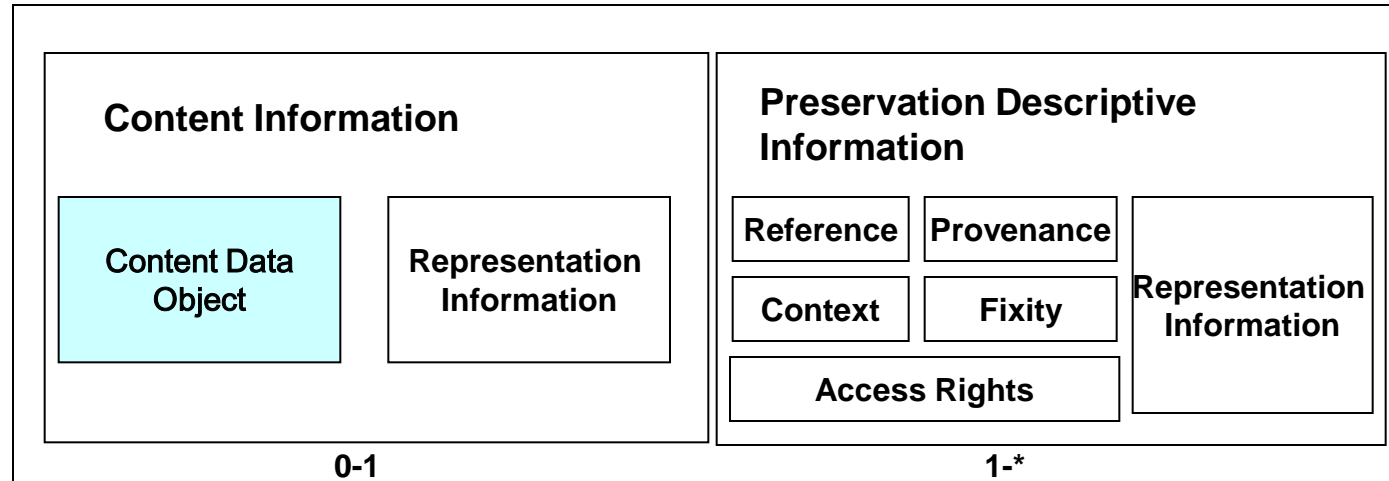
- ❑ A **magic object**: identifies SIRF container and its version
- ❑ Numerous **preservation objects** that are immutable
- ❑ A **catalog** that is
  - ❑ Updatable
  - ❑ Contains metadata to make container and preservation objects portable into the future without external functions



# What is a Preservation Object?

- ❑ A Preservation Object (PO) 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
  - ❑ result of migration is the creation of a separate and new version of the original PO that is linked to previous version
  - ❑ Audit log records the changes so authenticity is verifiable
  
- ❑ An example of a PO is OAIS Archival Information Package (AIP)

# OAIS Archival Information Package (AIP)

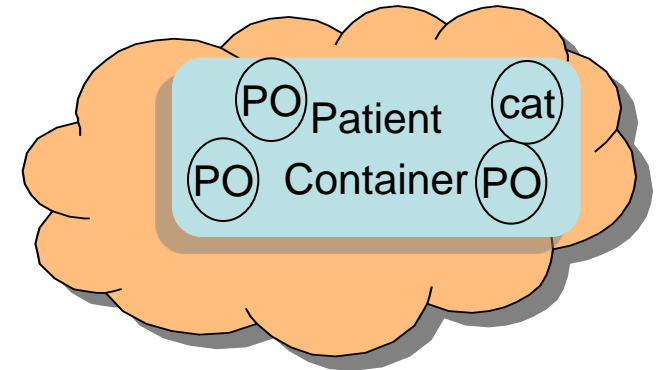


Domain specific packaging formats for POs

- ❑ XML Formatted Data Unit (XFDU)
- ❑ VERS Encapsulated Object (VEO)
- ❑ Metadata Encoding and Transmission Standard (METS)
- ❑ Preservation metadata: Implementation Strategies (PREMIS)

# SIRF and CDMI

- ❑ Cloud Data Management Interface (CDMI) specifies a 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 we have a cloud container named "PatientContainer" that is SIRF-compliant
    - ❑ each encounter is a preservation object
    - ❑ each image is a preservation object
    - ❑ the container has a catalog object



- ❑ We can read the various preservation objects and the catalog object via CDMI REST API as follows:
  - GET <root URI>/<PatientContainer>/encounterJan2001
  - GET <root URI>/<PatientContainer>/chestImage
  - GET <root URI>/ PatientContainer>/catalog

# Layered Approach

- ❑ A SIRF container assumes an underlying object interface
  - ❑ Example object layers
    - ❑ Advanced: OSD, Cloud, XAM
    - ❑ Lower level: UDF, CDFS, FAT, LTFS
  
- ❑ SIRF metadata is defined at 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 Level I – Work in Progress

The SIRF catalog includes metadata such as:

## □ General information:

- Spec ID and version
- SIRF level
- Container ID
- Provenance
- Audit log object ID

## □ For each Preservation Object:

- IDs
- Children's ID
- Dates
- Packaging format
- Fixity
- Retention
- Preservation profile
- Audit log object ID
- Extension

# Hierarchical Representation Example

- 1 containerInformation (1-1: M, NR)
  - 1.1 sirfSpecification (1-1: M, NR)
    - 1.1.1 sirfLevel (1-1: M, NR)
    - 1.1.2 sirfIdentifier (1-1: M, NR)
  - 1.2 containerIdentifier (0-\*: O, R)
    - 1.2.1 containerIdentifierType (1-1: M, NR)
    - 1.2.2 containerIdentifierValue (1-1: M, NR)

...

- 2 objectInformation (1-\*: M, R)
  - 2.1 objectIdentifier (1-\*: M, R)
    - 2.1.1 objectIdentifierType (1-1: M, NR)
    - 2.1.2 objectIdentifierValue (1-1: M, NR)

...

- 2.7 objectFixity (0-\*: O, R)
  - 2.7.1 objectDigestAlgorithm (1-1: M, NR)
  - 2.7.2 objectDigest (1-1: M, NR)
  - 2.7.3 objectDigestOriginator (0-1: O, NR)

...

- Methodology inspired by the PREMIS specification - <http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>

# objectIdentifier Table

Item	2.1 objectIdentifier
Components	2.1.1 objectIdentifierType 2.1.2 objectIdentifierValue
Definition	A designation used to uniquely identify the object within the container in which it is stored.
Rationale	
Repeatability	Repeatable
Obligation	Mandatory
Creation / Maintenance notes	
Usage notes	Identifiers must be unique within the container. They may be preexisting, and in use in other digital object management systems.

# SIRF Status and Resources

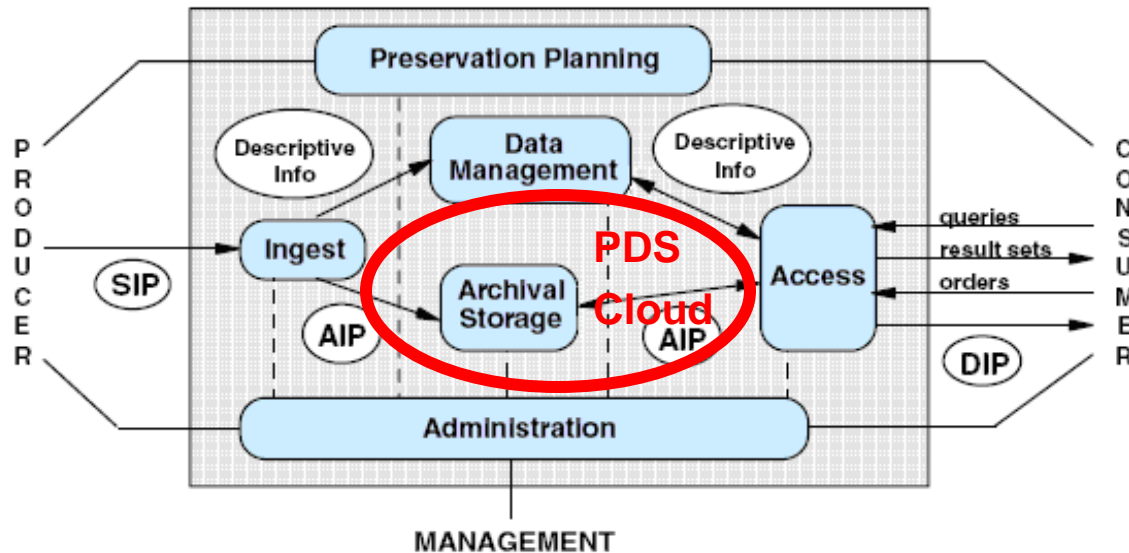
- ❑ SIRF use cases and requirements document is released for public review
  - ❑ 4 generic uses cases and 5 Workload-based use cases
  - ❑ For each use case, derived functional requirements
  - ❑ [http://www.snia.org/tech\\_activities/publicreview](http://www.snia.org/tech_activities/publicreview)
  
- ❑ Published the paper “Towards SIRF: Self-contained Information Retention Format”, in the Proceedings of the Annual International Systems and Storage Conference (SYSTOR), May 30-June 1, 2011, Haifa, Israel
  
- ❑ Work in progress on SIRF level 1 specification – need help!
  
- ❑ More information on LTR and SIRF: <http://www.snia.org/ltr>

- Introduction to digital preservation
- Preservation Technologies
- SNIA SIRF: Self-contained Information Retention Format
- **EU ENSURE: Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value**
- 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
- ❑ Demonstrated with three use case: Health Care, Clinical Trials and Finance
- ❑ Contributions to standards is a goal of the project

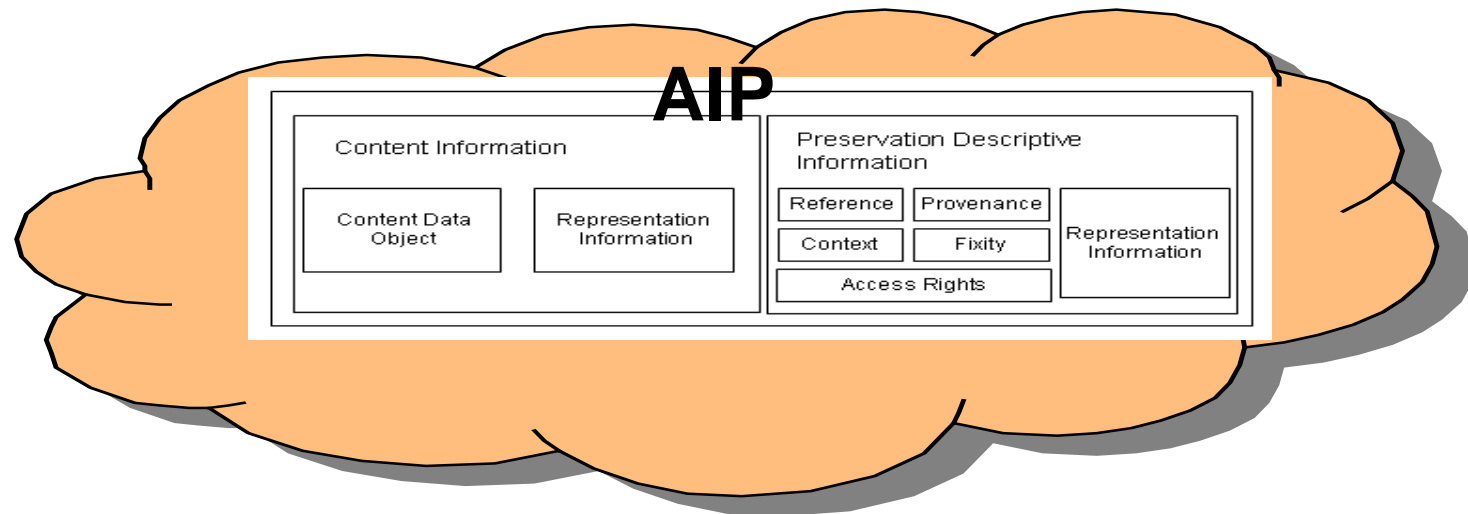


- ❑ Provides preservation-aware storage services for ENSURE
- ❑ Based on OAIS Archival Storage entity but provides more automation of preservation processes
- ❑ Built on top of multiple clouds concurrently, while taking advantage of each one's special capabilities
- ❑ Includes a SIRF Handler component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



# PDS Cloud Functionality

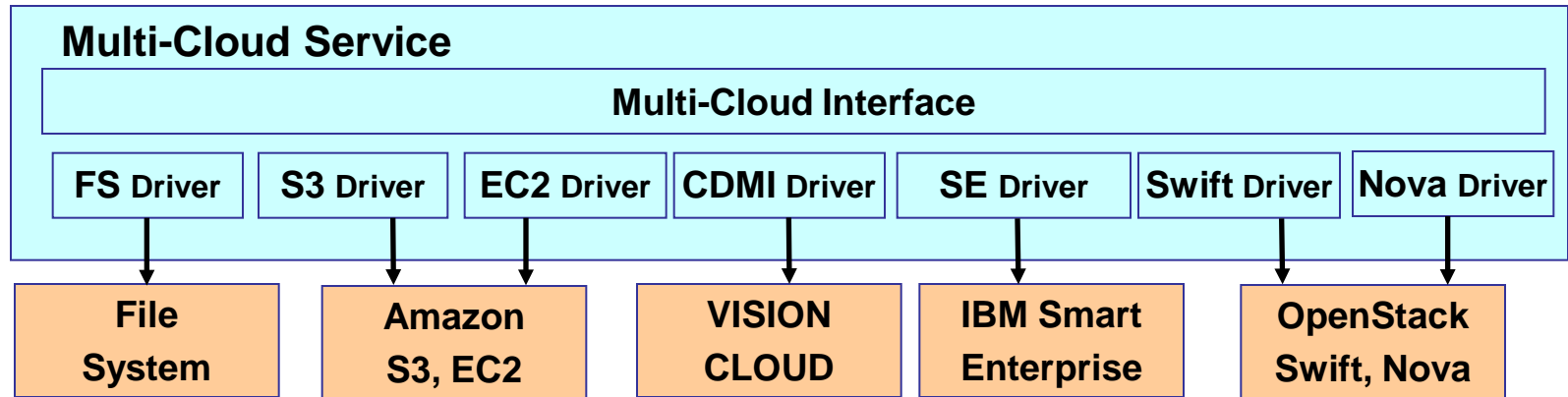
- Map OAIS AIP and the links among AIPs to the cloud data model



- Advanced fixity and audit service with multiple integrity (fixity) checks and updatable algorithms

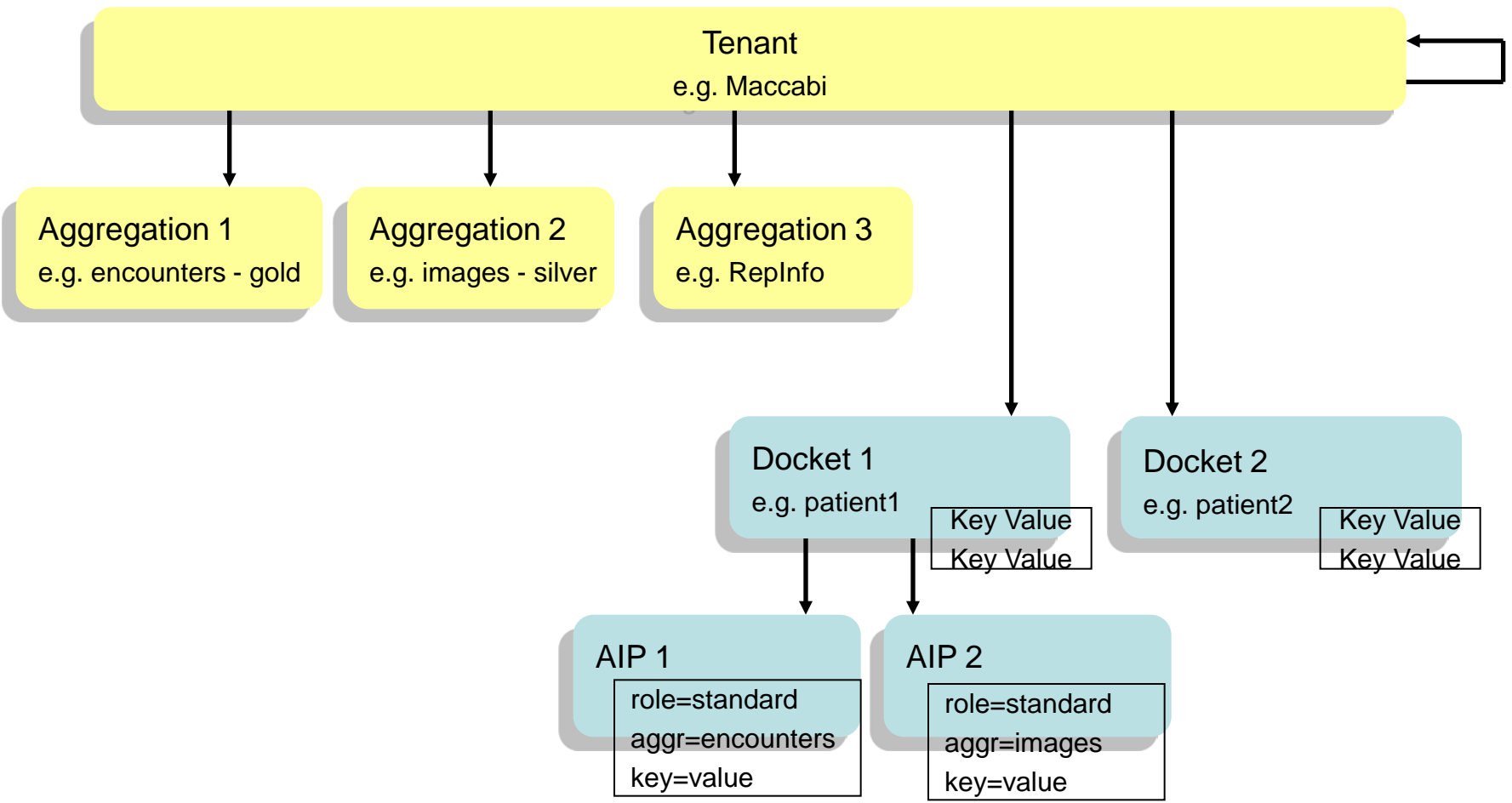
# PDS Cloud Functionality (cont)

- ❑ Support preservation object copies over multiple clouds

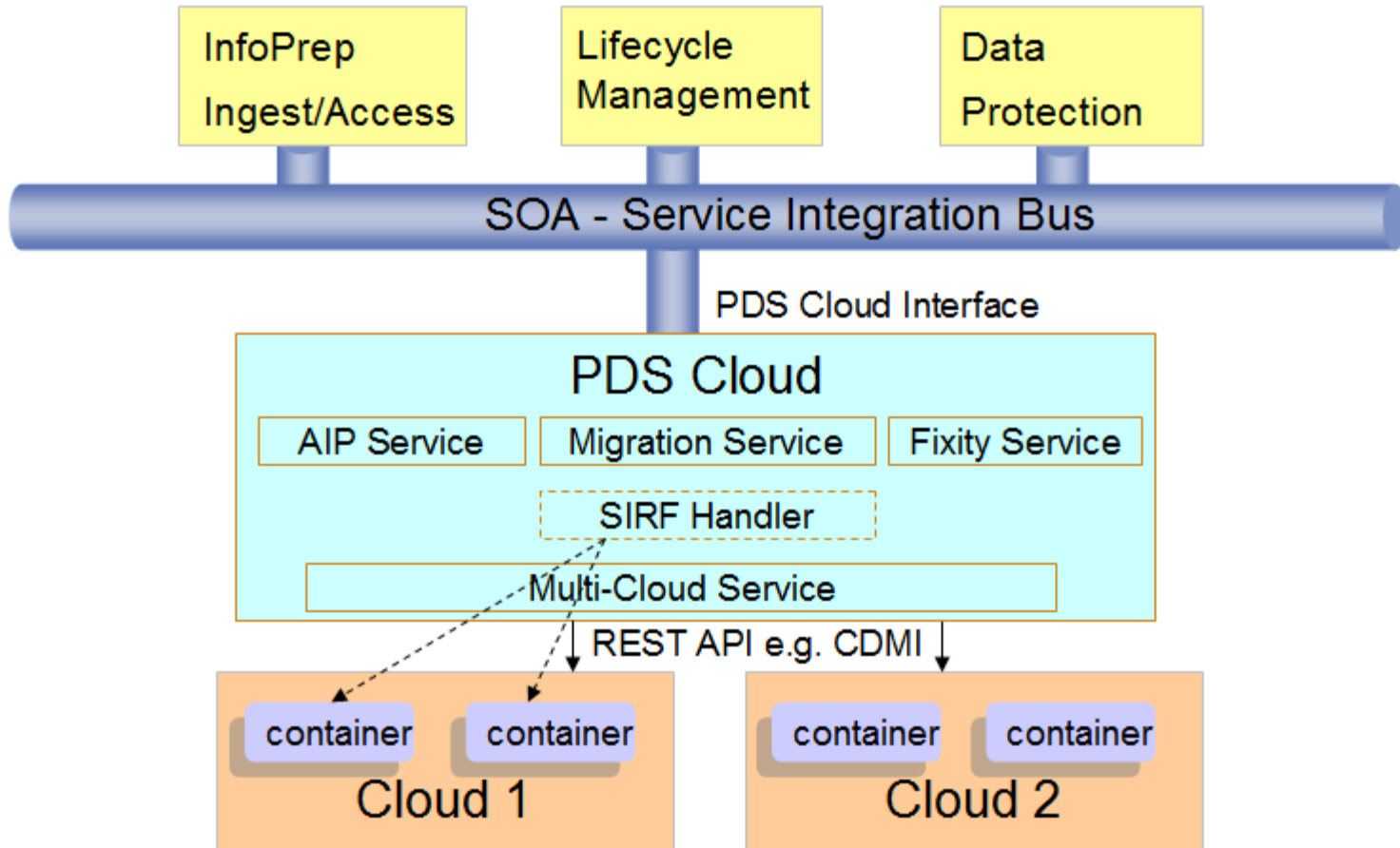


- ❑ Virtual appliance creation, provisioning and migration
- ❑ Support for computation close to the data (storlets)

# PDS Cloud Data Model



# PDS Cloud and SIRF in ENSURE



Note: SIRF Handler is for future implementation

- ❑ Digital preservation is an important problem that is only growing in importance over time
  
- ❑ 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
  
- ❑ PDS Cloud may provide a SIRF implementation in the cloud for the ENSURE EU research project



# 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](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!**

