SNIA®

SNIA Long Term Retention for Medical AI Applications

Simona Rabinovici-Cohen
IBM Reseach – Haifa
simona@il.ibm.com

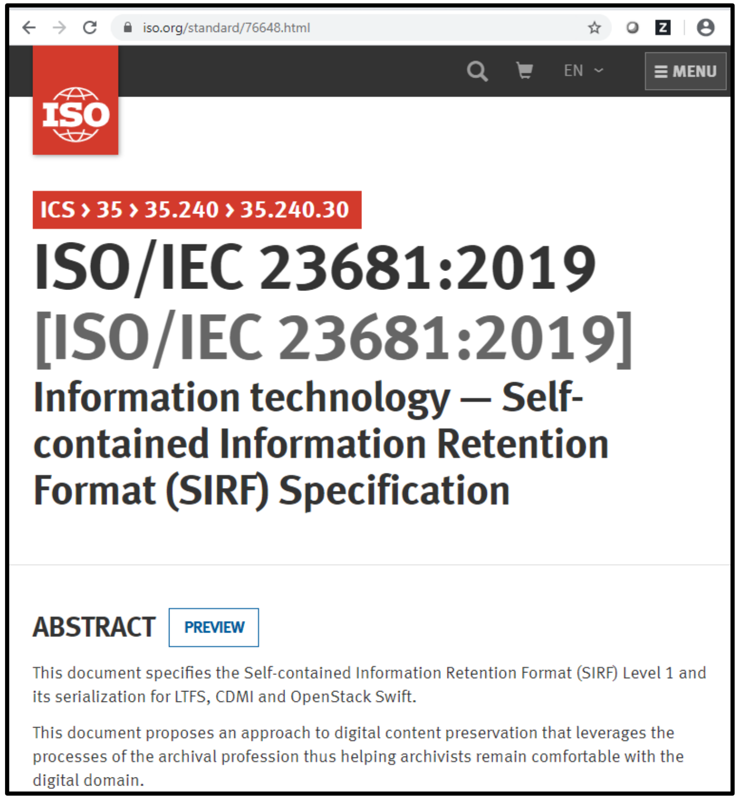
SDC EMEA, February 5, 2020

Introduction

- SNIA Self-contained Information Retention Format (SIRF) became an ISO/IEC 23681:2019 standard.
  - SIRF defines a storage container for long term retention and preservation
  - Results from a long journey of SNIA Long Term Retention technical working group

- Medical analytics such as in the BigMedilytics project requires a lot of scarce data that should be retained for the long term

- In this presentation, I'll discuss the use of SIRF for medical artificial intelligence (AI) via Medical AI Bank



ICS › 35 › 35.240 › 35.240.30

ISO/IEC 23681:2019
[ISO/IEC 23681:2019]
Information technology — Self-contained Information Retention Format (SIRF) Specification

ABSTRACT PREVIEW

This document specifies the Self-contained Information Retention Format (SIRF) Level 1 and its serialization for LTFS, CDMI and OpenStack Swift.

This document proposes an approach to digital content preservation that leverages the processes of the archival profession thus helping archivists remain comfortable with the digital domain.

Outline

- **Digital preservation challenges**
- The SIRF standard
- Breast cancer pilot in BigMedilytics
- SIRF for medical AI
- Summary

Need for Digital Preservation of Big Data

- 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
- Domains that have Big Data require preservation

Healthcare

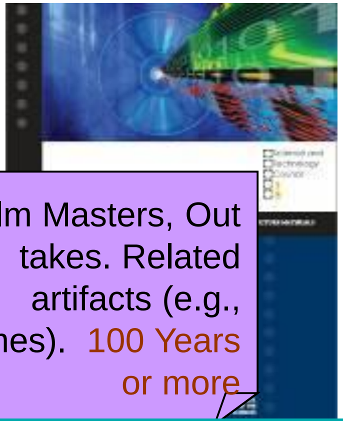
X-rays are often stored for periods of 75 years

Records of minors are needed until 20 to 43 years of age

Scientific and Cultural

Satellite data is kept for ever

We would like to keep digital art for ever



Film Masters, Out takes. Related artifacts (e.g., games). 100 Years or more

M&E

Goals and Threats of Digital Preservation

- Digital assets stored now should remain
  - Accessible
  - Undamaged
  - Usable

- For as long as desired – beyond the lifetime of
  - Any particular storage system
  - Any particular storage technology

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

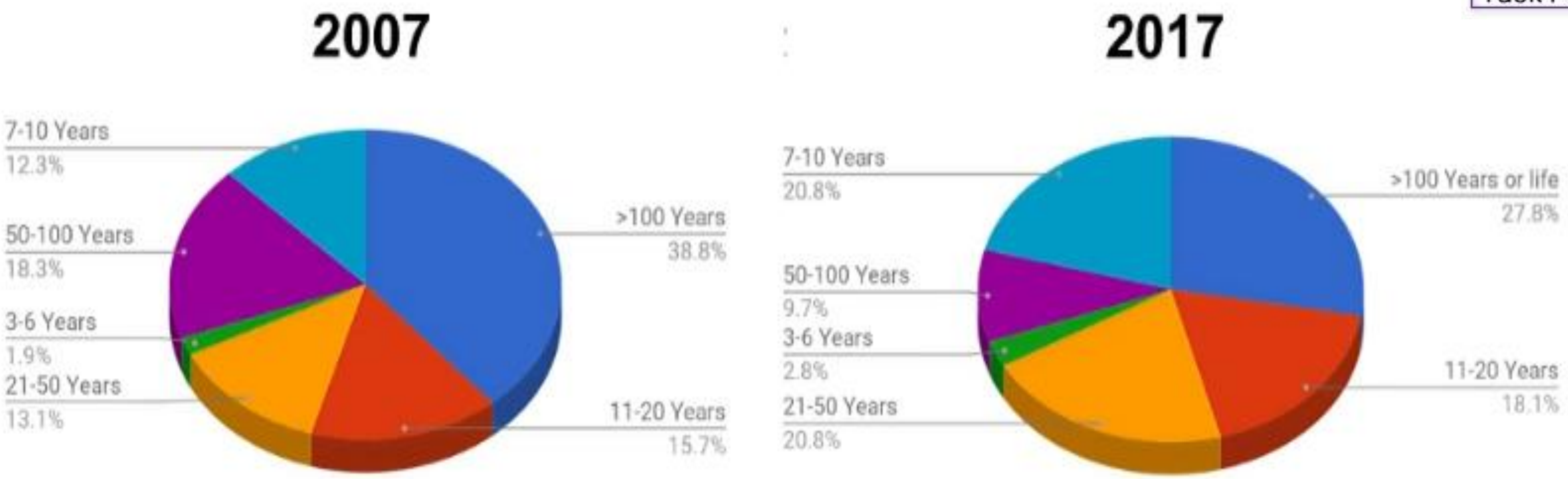
- Faults
  - Economic faults
  - Organizational faults
  - Human error
  - Attack

Requires both physical preservation and logical preservation and at an *affordable cost*

What is Long Term?

The 100-Year Archive Survey Results 2007 and 2017

100 YEAR Archive Task Force



Solutions

- Solutions are now becoming available
  - Standards – OAIS, VERS, MoReq, …
  - Storage formats - SIRF, OpenAXF, PREMIS, BagIt….
  - Software – Fedora, LOCKSS, DSPace, Arkivum, iRods, Rosetta, ….
  - Cloud Services – Preservica, Duracloud, Chronopolis, Dternity, Glacier, ….
- But, their usage is still limited
  - Primarily used in government agencies, libraries, and highly regulated industries
- Why
  - Lack of education or understanding?
  - Lack of need, will, funding, etc.? Lack of penalties?
  - Short term focus?

Outline

- Digital preservation challenges
- **The SIRF standard**
- Breast cancer pilot in BigMedilytics
- SIRF for medical AI
- Summary

SIRF: Self-contained Information Retention Format

An Analogy

- Standard physical archival box
  - Archivists gather together a group of related items and place them 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
- 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

Photo courtesy Oregon State Archives

SIRF Properties

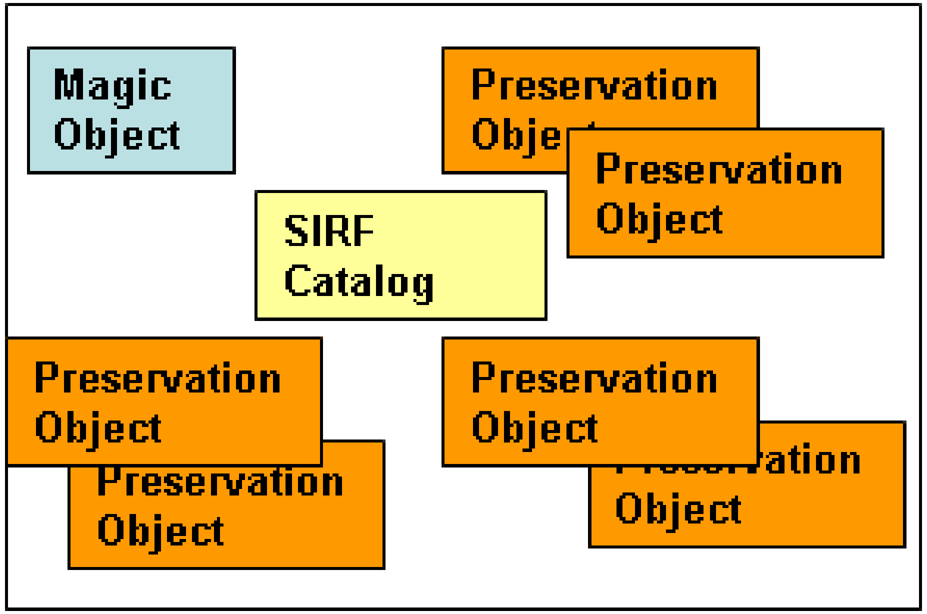
- SIRF is a logical data format of a **storage container** appropriate for long term storage of digital information
  - A storage container may comprise a logical or physical storage area considered as a unit.
    - Examples: a file system, a tape, a block device, a stream device, an object store, a 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
- **Preservation objects** (PO) which are immutable
- A **catalog** that is
  - Updatable
  - Contains metadata to make container and preservation objects portable into the future without external functions



SIRF is inspired by the Open Archival Information System (OAIS) - ISO 14721:2003

SIRF Categories

The SIRF catalog includes metadata organized in a hierarchy of categories, elements and attributes. The categories are:

- Container information:
  - Specification
  - Container ID
  - State
  - Provenance
  - Audit Log
- For each Preservation Object:
  - Object IDs
  - Related Objects
  - Dates
  - Packaging Format
  - Fixity
  - Retention
  - Audit Log
  - Extension

PO Information – IDs Category

- Elements:
  - **PO name (objectName)** – non unique identifier e.g. file name
  - **PO version ID (objectVersionIdentifier)** – unique identifier that identifies the specific version of the PO
  - **PO logical ID (objectLogicalIdentifier)** - a unique identifier that identifies the various versions that originate from the same ancestor
  - **PO parent ID (objectParentIdentifier)** - a unique identifier that identifies the parent PO from which this PO version was created. Parent PO shares the same logical ID as the current PO, but has different version ID.

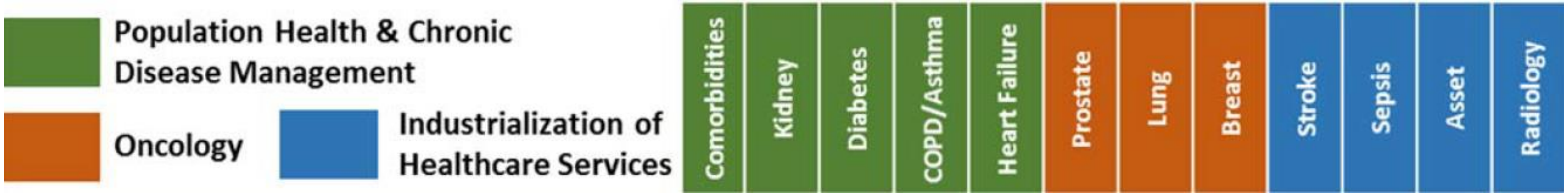
Outline

- Digital preservation challenges
- The SIRF standard
- **Breast cancer pilot in BigMedilytics**
- SIRF for medical AI
- Summary

BigMedilytics EU project

BigMedilytics is an EU project on medical Big Data

- Aims to transform Europe’s Healthcare by using state-of the-art Big Data technologies to:
  - reduce costs
  - improve patient outcomes
  - deliver better access to healthcare facilities
- A Private Public Patnership project (PPP)
- Includes 12 pilots in 3 themes:



🟩 Population Health & Chronic Disease Management

🟧 Oncology

🟦 Industrialization of Healthcare Services

Comorbidities, Kidney, Diabetes, COPD/Asthma, Heart Failure, Prostate, Lung, Breast, Stroke, Sepsis, Asset, Radiology

Breast Cancer Pilot

Goals: Improve treatment response for breast cancer by using AI to analyze MG, US, and MRI images along with structured clinical data. Reduce costs by tailoring treatment for the individual patient.

IBM

- Pilot lead
- Image analytics
- Clinical data analytics
- Deep learning

Institute Curie

- Patient data (images and clinical)
- Clinical knowledge
- Data hosting

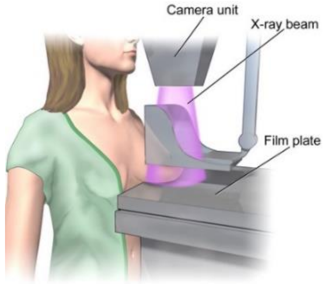
VTT

- Image analysis
- Interpretable features

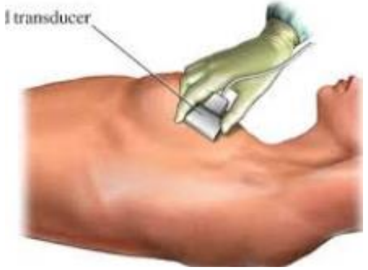
Breast Cancer Pilot

- Neoadjuvant Chemotherapy Treatment (NACT) is a treatment option in breast cancer
  - Decision today is made based on: tumor size (T3, T4), patient preference for breast conservation, hormonal receptors, HER2
  - Less than half of treated patients achieve pathological complete response with no evidence of cancer cells
  - Failed treatment delays a more effective treatment

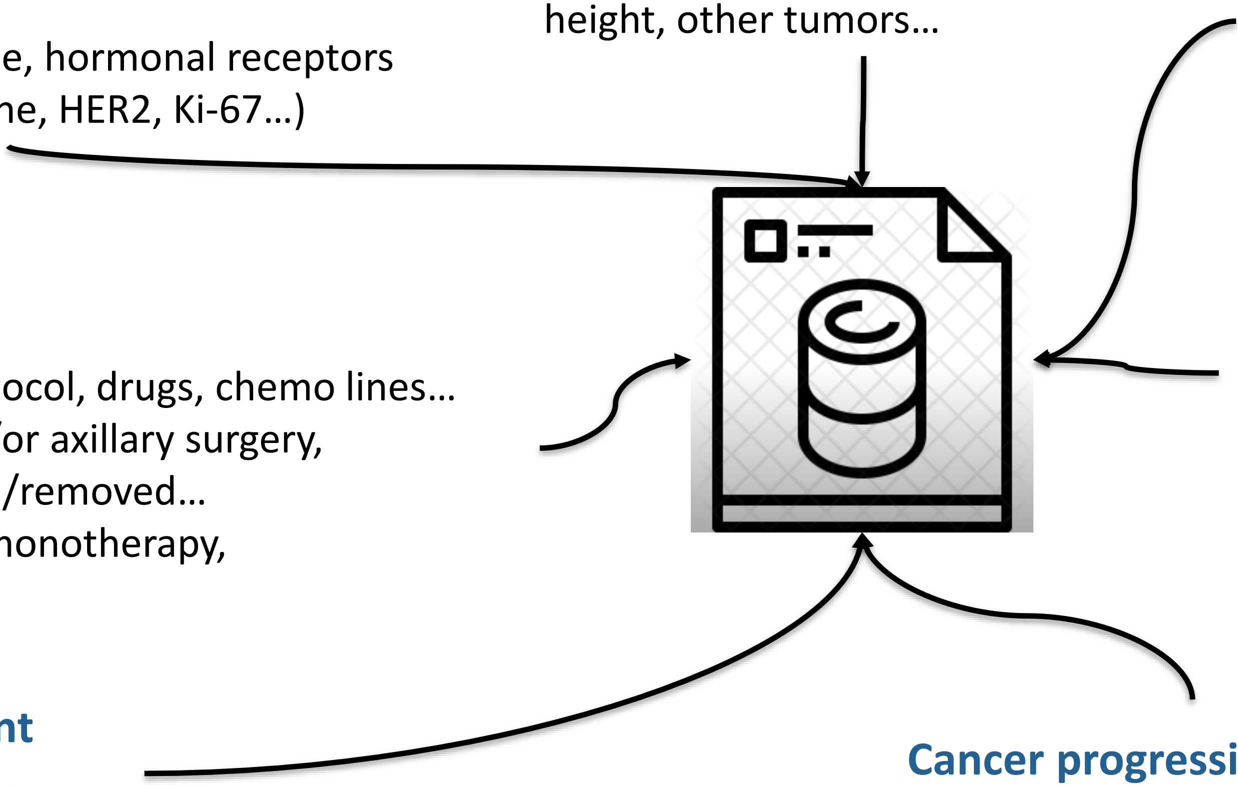
- Radiomics can improve NACT response prediction
  - Extract large amount of quantitative features from multi modal medical images
  - Apply advanced deep learning and computer vision algorithms for precision medicine applications



Mammogram



Curie Heterogenous Data Collection



Patient data

Age at diagnosis, sex, weight, height, other tumors…

Clinical data

Side (Left, Right), grade, hormonal receptors (estrogen, progesterone, HER2, Ki-67…)

Imaging

MRI, MG, US…

Treatment

Chemotherapy: Date, protocol, drugs, chemo lines…
Surgery: Date, breast and/or axillary surgery,
lymph nodes invaded/removed…
Other: Radiotherapy, hormonotherapy, immunotherapy

Other

Genetic
Transcriptomic

Response to treatment

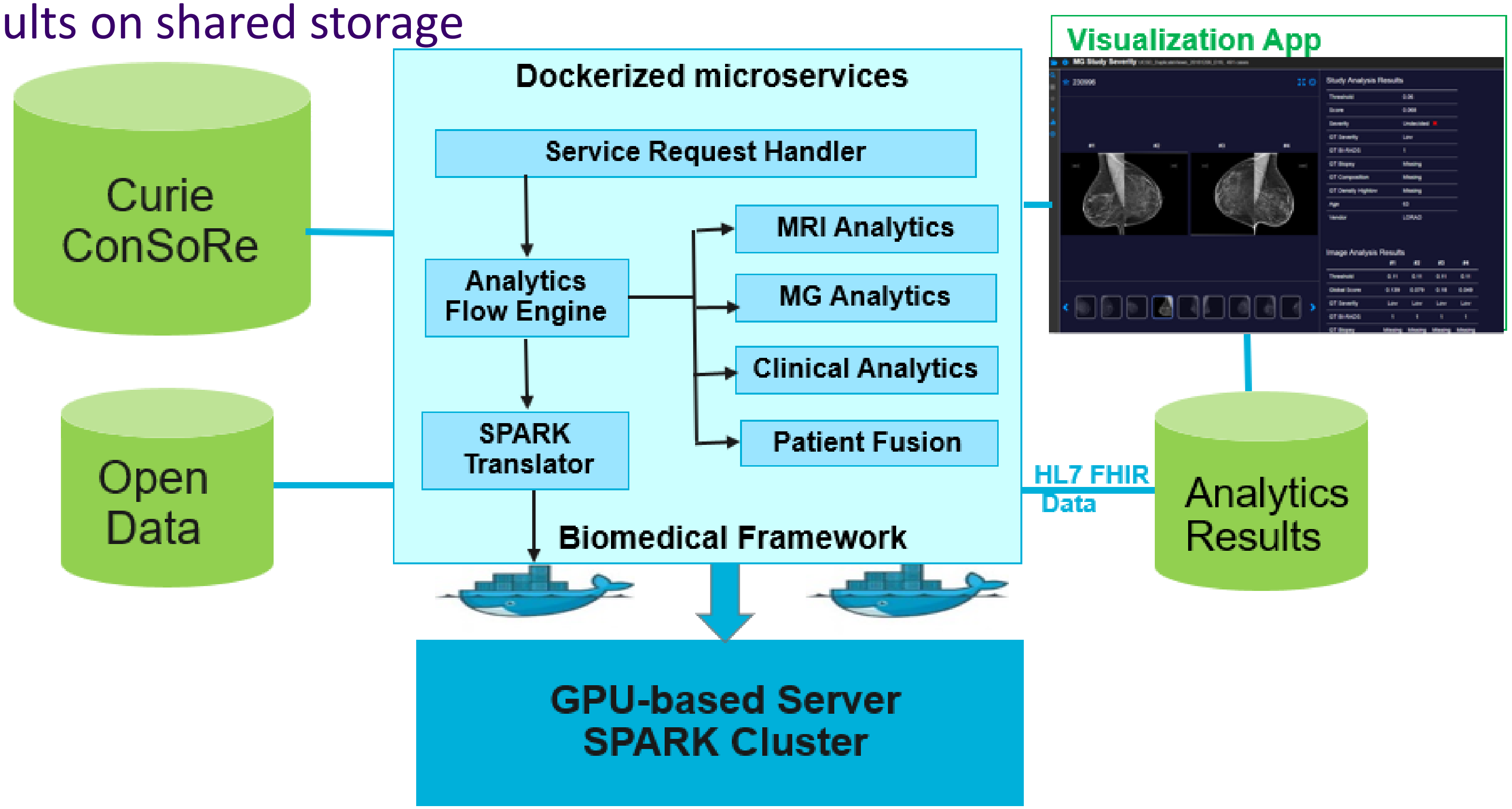
Several scoring methods:
Chevalier, Sataloff, RCB, ypTN

Cancer progression

Possible relapses/metastasis

Pilot Architecture

- Model-to-Data paradigm
- Suitable for on-prem or on-cloud
- All data and results on shared storage



Outline

- Digital preservation challenges
- The SIRF standard
- Breast cancer pilot in BigMedilytics
- **SIRF for medical AI**
- Summary

Medical AI Bank

- For medical AI we need big data and some of it annotated

- But getting the medical data is difficult
  - The data is scarce and distributed
  - Needs preprocessing
    - e.g. no standardized protocol for MRI scan acquisition
  - Adding annotations is expensive
  - Adhere to privacy regulations e.g. GDPR

- The need for Medical AI Bank
  - Analytics-ready data that is preserved for future research
  - Based on medical standards (DICOM, HL7 FHIR, ICD-10, UMLS,…)
  - Can get individual’s data donations after his lifetime
  - Includes storage containers with SIRF serialization

Goals of SIRF Serialization for Cloud/FS

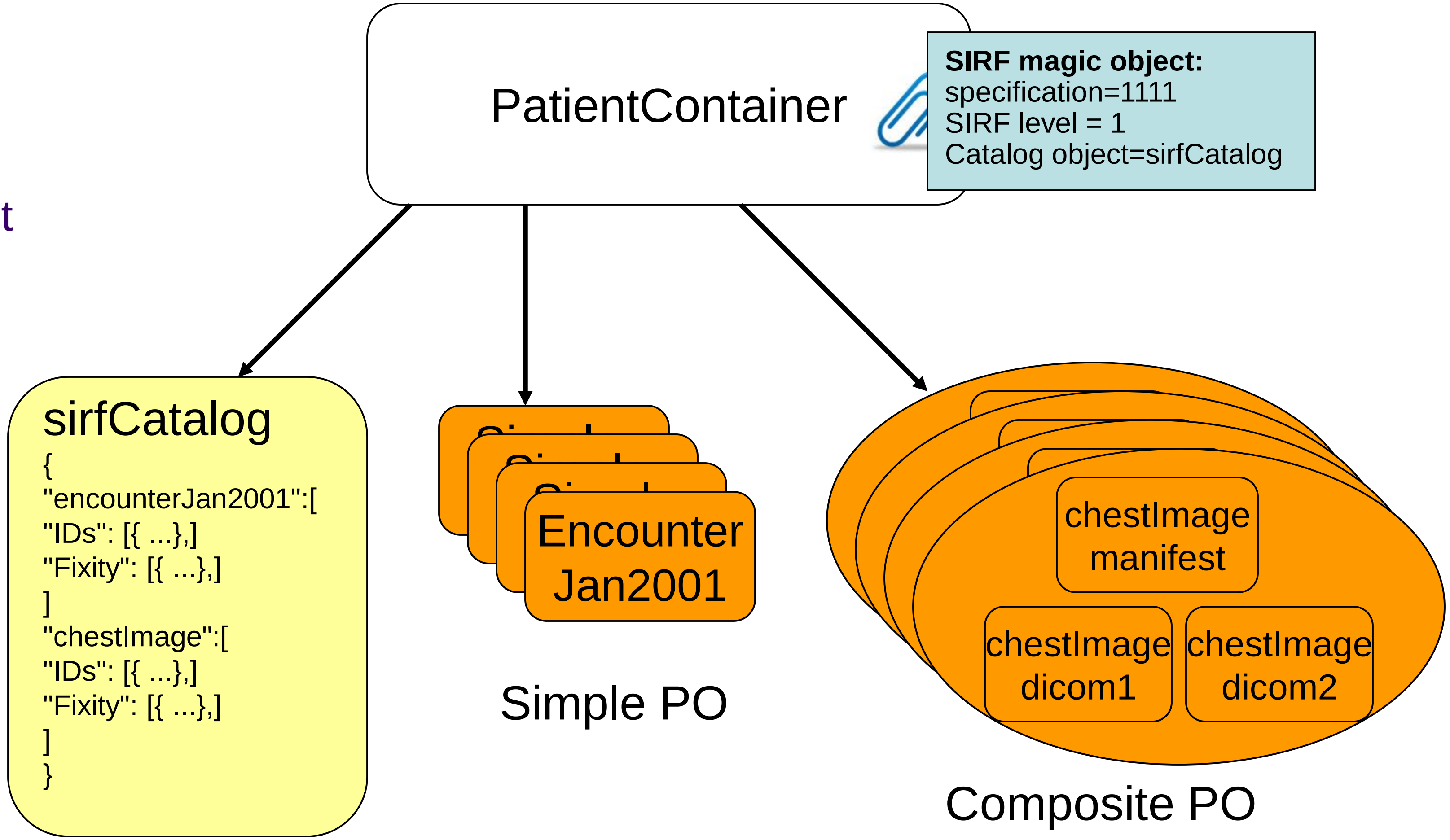
- SIRF serialization for Cloud/FS specifies how a SNIA Cloud Data Management Interface (CDMI) cloud container or Linear Tape File System (LTFS) tape also becomes SIRF-compliant

- A SIRF-compliant cloud or file system containers enable a future storage client to "understand" containers created by today' s storage client
  - The properties of the future client is unknown to us today
  - "understand" means identify the preservation objects in the container, the packaging format of each object, its fixities values, etc. (as defined in the SIRF catalog)

- OpenSirf is an open source implementation of SIRF serialization for the cloud

SIRF Serialization for Cloud

- •SIRF magic object is mapped to the CDMI container metadata
- •SIRF catalog is an object in the CDMI container formatted in JSON
- •SIRF Simple/Composite PO is mapped to CDMI data object/set of data objects

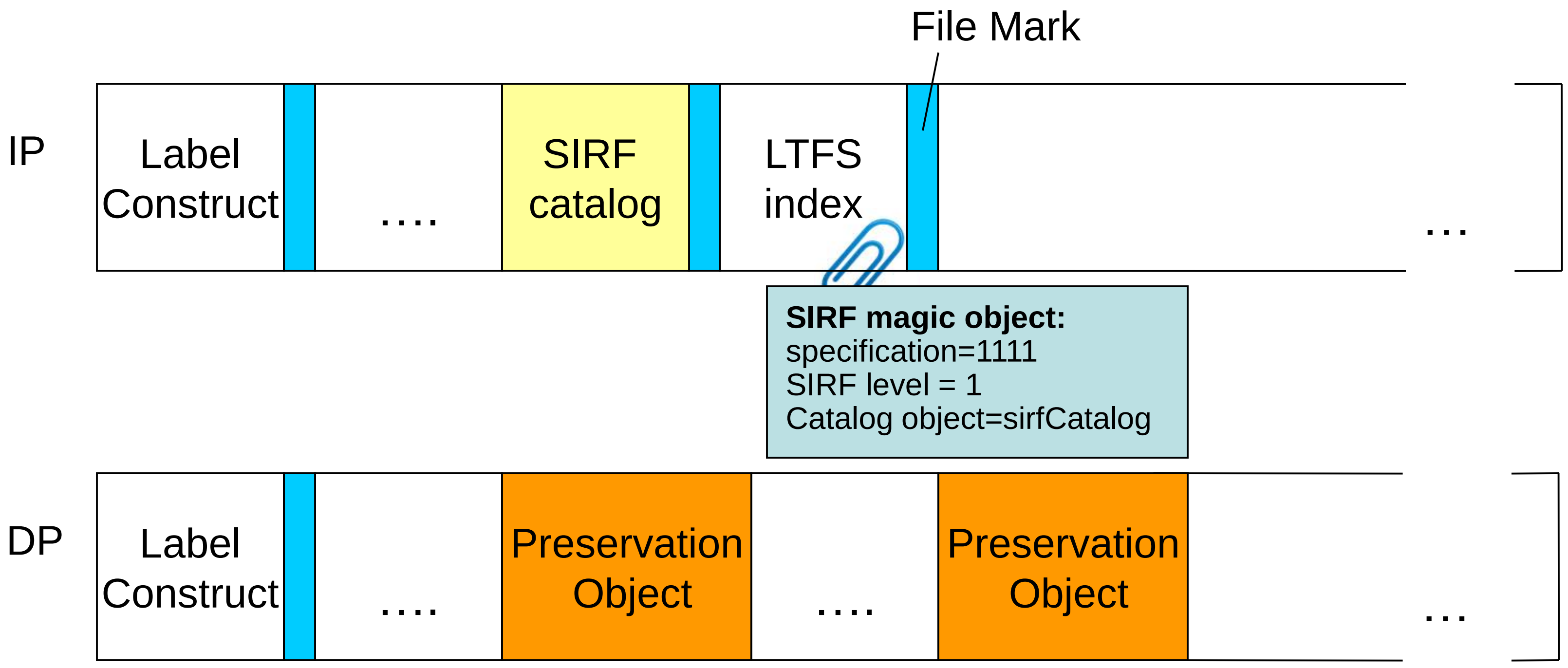


SIRF magic object:
specification=1111
SIRF level = 1
Catalog object=sirfCatalog

sirfCatalog

```
{
"encounterJan2001":[
"IDs": [{ ...},]
"Fixity": [{ ...},]
]
"chestImage":[
"IDs": [{ ...},]
"Fixity": [{ ...},]
]
}
```

SIRF Serialization for LTFS Tape



SIRF magic object:
specification=1111
SIRF level = 1
Catalog object=sirfCatalog

- SIRF magic object is mapped to extended attributes of the “LTFS index” root directory
- SIRF catalog resides in the index partition and formatted in XML
- SIRF Simple/Composite PO is mapped to a LTFS file/set of files

Outline

- Digital preservation challenges
- The SIRF standard
- Breast cancer pilot in BigMedilytics
- SIRF for medical AI
- **Summary**

Summary

- Need to retain not only information of interest but ALL other information to make it fully usable in the future
  - Put it all in the SIRF “digital box”, preserve that as a unit
  - SIRF includes metadata about the storage container, to help “understand” the contents of the container in the future

- Medical AI Bank is a vision that requires digital preservation
  - Utilize SIRF for collecting all of the information that will be needed to transition to new technologies in the future
  - SIRF can be serialized for the future technologies as they come

For further information

- SIRF specification
  https://www.iso.org/standard/76648.html
  http://www.snia.org/tech_activities/standards/curr_standards/sirf

- More information on SIRF & SNIA LTR activities
  http://www.snia.org/ltr

- OpenSIRF is available at:
  http://github.com/opensirf

- BigMedilytics EU project
  https://www.bigmedilytics.eu/

Thank You

Muchas Gracias

Thank You

תודה רבה

Merci Beaucoup

Kiitos Paljon

Vielen Dank