INTERNATIONAL STANDARD

ISO/IEC 23681

First edition
2019-05

Information technology — Self-contained Information Retention Format (SIRF) Specification

Reference number
ISO/IEC 23681:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This document was prepared by SNIA (as SIRF Specification V1.0) and drafted in accordance with its editorial rules. It was adopted, under the JTC 1 PAS procedure, by Joint Technical Committee ISO/IEC JTC 1, *Information technology*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Many organizations now have a requirement to preserve and maintain access to large volumes of digital content indefinitely into the future. Regulatory compliance and legal issues require preservation of email archives, medical records and information about intellectual property. Web services and applications compete to provide storage, organization and sharing of consumers' photos, movies, and other creations. And many other fixed-content repositories are charged with collecting and providing access to scientific data, intelligence, libraries, movies and music. A key challenge to this need is the creation of vendor-neutral storage containers that can be interpreted over time.

Archivists and records managers of physical items such as documents, records, etc., avoid processing each item individually. Instead, they gather together a group of items that are related in some manner — by usage, by association with a specific event, by timing, and so on — and then perform all of the processing on the group as a unit. The group itself may be known as a series, a collection, or in some cases as a record or a record group. Once assembled, an archivist will place the series in a physical container (e.g., a file folder or a filing box of standard dimensions), mark the container with a name and a reference number and place the container in a known location. Information about the series will be included in a label that is physically attached to the container, as well as in a "finding aid" such as an online catalog that conforms to a defined schema and gives the name and location of the series, its size, and an overview of its contents.

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. One of the major needs to make this strategy possible is a digital equivalent to the physical container — the archival box or file folder — that defines a series, and which can be labelled with standard information in a defined format to allow retrieval when needed. Self-contained Information Retention Format (SIRF) is intended to be that equivalent — a storage container format for a set of (digital) preservation objects that also provides a catalog with metadata related to the entire contents of the container as well as to the individual objects and their interrelationship. This logical container makes it easier and more efficient to provide many of the processes that will be needed to address threats to the digital content. Easier and more efficient preservation processes in turn lead to more scalable and less costly preservation of digital content.

SIRF components, use cases and functional requirements were defined in [1] SIRF use cases and functional requirements, working draft — version 0.5a and further described in [2] "Towards SIRF: Self-contained Information Retention Format." This document goes one step further and details the actual metadata, categories and elements in the container's catalog. The document also describes how the SIRF logical format is serialized for storage containers in the cloud and for tape based containers. The SIRF serialization for the cloud is being experimented with OpenStack Swift object storage, and the implementation is offered as open source in the OpenSIRF initiative[3].

Creating and maintaining the SIRF catalog requires executing data-intensive computations on the various preservation objects including fixity checks, data transformations. This can be done efficiently via executing computational modules — storlets — close to where the data is stored. The benefits of using storlets include reduced bandwidth (reduce the number of bytes transferred over the WAN), enhanced security (reduce exposure of sensitive data), costs savings (saving infrastructure at the client side) and compliance support (improve provenance tracking). The Storlet Engine[4] (see "Storlet Engine for Executing Biomedical Processes within the Storage System") is an engine to support such storlets computations in secure sandboxes within the storage system, and can be used to create and maintain SIRF containers.

Information technology — Self-contained Information Retention Format (SIRF) Specification

1 Scope

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.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 17826:2012, *Information technology — Cloud Data Management Interface (CDMI)*

ISO/IEC 20919:2016, *Information technology — Linear Tape File System (LTFS) Format Specification*

Self-contained Information Retention Format (SIRF) use cases and functional requirements, working draft — version 0.5a, SNIA, September 2010, http://www.snia.org/tech_activities/publicreview/SIRF_Use_Cases_V05a_DRAFT.pdf

JSON. ECMA-404, The JSON Data Interchange Standard. http://json.org

OpenSIRF. http://github.com/opensirf

OpenStack Swift. http://swift.openstack.org

PREMIS. PREservation Metadata: Implementation, Strategies, http://www.loc.gov/standards/premis

Rabinovici-Cohen S., Henis E., Marberg J., Nagin K. Storlet Engine for Executing Biomedical Processes within the Storage System", Proceedings of the 7th International Workshop on Process-oriented Information Systems in Healthcare (ProHealth), September 2014, Eindhoven, the Netherlands

Rabinovici-Cohen S., Baker M.G., Cummings R., Fineberg S., Marberg J. Towards SIRF: Self-contained Information Retention Format", Proceedings of the Annual International Systems and Storage Conference (SYSTOR), May 30-June 1, 2011, Haifa, Israel. https://www.research.ibm.com/haifa/projects/storage/datastores/papers/systor56-rabinovici-cohen.pdf

Rabinovici-Cohen S., Cummings R., Fineberg S. Self-contained Information Retention Format for Future Semantic Interoperability", Proceedings of the 4th International Workshop on Semantic Digital Archives (SDA), September 2014, London, UK

W3C Prov Model Primer http://www.w3.org/TR/prov-primer/

3 Terms and definitions

No terms and definitions are listed in this document.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

4 Business Case

While no one wants to lose their digital content, the cost of maintaining integrity and access is significant, in both money and effort. And unlike paper based content, the lifespan of digital content can be very short unless if proactive steps are being taken to protect it. The use of a storage container format like SIRF adds little expense and greatly increases the sustainability of data. However, this is not adequate unless if the cost of preserving content is less than the (potential) cost of losing it.

In a business context, there are three major reasons why content is preserved. These are: to preserve history, to mitigate risk or meet a legal mandate, and for future value of information. One or more of these may apply, and the amount an entity is willing to spend will differ depending on how well these reasons are aligned with the business goals of an organization.

One of the main reasons why people and organizations preserve content is to preserve history. In the case of an individual, it may be photos, videos, and other content preserving one's life history. In a business context, libraries, national archives, historians, and others have a primary mission to preserve history.

This should mean that information under the control of these organizations or individuals would be well protected. However, reality dictates that:

1) Organizations (and individuals) have limited resources, and they have to make choices how much to invest in preservation. Typically, this happens with the organization's knowledge of what trade-offs are being made. However, in the case of digital content, the choice not to invest in preservation will typically result in loss.

2) Organizations cannot preserve everything as the cost would be prohibitive. However, it is difficult to predict what will be of historic value. Sometimes important content is lost simply because its value was not known at the time.

3) Lack of skilled and experienced personal may also result in loss, especially in situations where simple solutions do not exist.

Another often cited reason for preserving data is for "risk mitigation", or in some cases for "legal mandate". These are closely related reasons because legal mandate is often looked at through the lens of legal risk. In other words, a mandate that is not enforced or whose penalty is small is less of a risk than a mandate with a larger penalty.

For example, consider government entities like a national or state archive. Legislators require those entities to keep records for a prescribed period of time. The penalty for losing those records may include loss of job, loss of funding, or even criminal penalties. Since preservation is directly funded, there is little excuse for losing information. However, even in those cases, underfunding, and lack of expertise may result in loss. Further, when those agencies preserve historical information, the mandate and funding will not allow them to keep everything. Therefore, even these archivists are required to make choices about what to keep.

Another often cited legal mandate is in healthcare, where medical organizations are required to retain information for the lifetime of a patient. This seems like a difficult requirement, especially since records are often maintained in private doctors' offices and other places that may not exist 50 years or 75 years into the future. Anecdotal evidence shows that medical records are not maintained that long. So, why is this happening? It is because records retention is expensive, and there are no penalties for losing information. That is not to say that doctors and hospitals don't try, rather they won't spend the necessary money to ensure that records are not lost.

The private sector is similar to healthcare. Preservation is seen as a cost, so it only makes sense where the return on investment is positive. For example, businesses are very good at keeping recent tax records, because they know that if they don't, the tax authority may levy fines. In the US, businesses retain emails, because they know judges will fine them if they don't. They even implement searchable archive systems, because they know that it will save money when they are forced to produce documents responsive to a lawsuit. Note that this is different outside of the US, where judges often do not levy fines or force companies to spend large sums of money on legal discovery.

When private sector companies do preserve information, they are typically focused on risk mitigation, not preservation. If a company can legally delete information, they often will, because it eliminates the chance that it can be used against them. If information is purely a risk, and a company is not in the business of preservation, why would they keep it? The obvious answer is that they will keep information that is valuable, and other information will not be kept. Value can result from future revenue, cost savings, technical advantage, etc.

Regarding future value of information, one obvious example is in the entertainment industry. Movies, TV shows, music, and other content can be re-sold and repurposed decades after its creation. This can result in many dollars in revenue. So not surprisingly, organizations like the Motion Picture Experts Group are at the leading edge of digital preservation. Entertainment companies spend significant amounts of money retaining their content so that they will have it available to repurpose. However, this does not mean they can retain everything. With the advent of digital movie production, the amount of data that can be generated during the creation of a single film is immense. Therefore, even here where future value is tangible, some hard choices need to be made.

In other industries, the mandate may not be as clear. Are design documents from an existing product valuable? What about a retired product? What about research leading up to a product design? These things may be needed for risk mitigation, but what about for use in future products? Companies make decisions about these kinds of Intellectual Property content every day, and more times than not, the data is either actively destroyed or lost due to inaction. The reason for this is actually because its value (beyond risk mitigation) is unknown, so it's not clear how much a company should invest in retaining the information.

So, how does SIRF help? SIRF brings down the expense of preservation, because data can remain accessible even if the software that created the data no longer exists. This is because the stored data is designed to be understandable, and does not need specialized software to interpret it. SIRF reduces the complexity of logical and physical migration, making it easier for businesses to justify. By using SIRF today, it becomes possible to retain more information, and to retain information with a lower perceived future value. This is unlike proprietary and undocumented formats, which become useless soon after a business stops paying for support.

5 Specification Overview

5.1 Container Components

Figure 1 illustrates the SIRF container, which includes the following components:

- A magic object that identifies whether this is a SIRF container and gives its version. The magic object is independent of the media and has an agreed defined name and a fixed size. It also includes the means to access the SIRF catalog (for example, the catalog's location).

- Preservation objects that contain the actual data to be preserved. An example preservation object can be the OAIS Archival Information Package (AIP). The container may include multiple versions of a preservation object and multiple copies of each version, but each specific preservation object is generally immutable.

- A catalog that is updateable and contains metadata needed to make the container and its preservation objects portable and accessible into the future without relying on metadata external to the storage subsystem.

While the semantics of traditional storage systems include only limited standardized metadata about each object, SIRF provides for the rich metadata needed for preservation and ensures its grouping with the data. This rich metadata is defined in the catalog in a logical format to allow its serialization for different storage technologies.

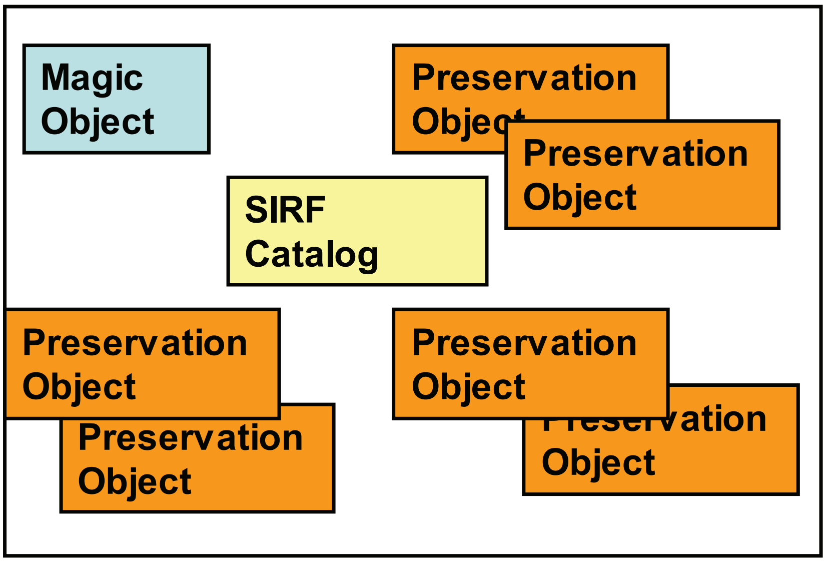

Figure 1 — SIRF Components

5.2 SIRF Catalog

The SIRF catalog is an object that includes metadata about the Preservation Objects (POs) in the container and their relationship. It has a well-defined standardized format so it can be understandable in the future. The SIRF catalog is separated from the metadata contained in the POs themselves because a strict standardized format is difficult to impose on the POs that are generated by different applications and domains. Additionally, The SIRF catalog level 1 includes metadata that is not included in the PO, e.g., fixity value of the whole PO. Including this metadata within the PO changes the fixity value of the PO making this metadata inherently incorrect.

The SIRF catalog includes metadata related to the whole container as well as metadata related to each preservation object within the container. Both types of metadata are organized into categories.

The metadata for the whole container includes the following categories:

1 Container Information

1.1 Specification

1.2 Container ID

1.3 State

1.4 Provenance

1.5 Audit Log

The metadata for all the preservation objects set is aggregated under:

2 Objects Set

The metadata for each preservation object includes the following categories:

3 Object Information

3.1 Object IDs

3.2 Related Objects

3.3 Dates

3.4 Packaging Format

3.5 Fixity

3.6 Retention

3.7 Audit Log

3.8 Extension

Examples of SIRF Catalogs are shown in Annex A: (informative) XML schema for the SIRF catalog, Annex B: (informative) Sample XML catalog, and Annex C: (informative) Sample JSON catalog.

5.3 Metadata Units

This document describes the specific metadata units in the various categories. Each category includes several elements in which each element may be composed of several attributes. The document also provides a hierarchical representation of the metadata units. The notation of the hierarchical representation is based on PREMIS[5] (PREservation Metadata: Implementation, Strategies) when possible and includes the following notions:

Repeatability (R): A metadata unit designated as “Repeatable” (R) can take multiple values. It does not mean that a repository must record multiple instances of the metadata unit. Similarly, a metadata unit can be designated as Not Repeatable (NR).

Optionality (O): Whether a value for the metadata unit is optional (O) or mandatory (M). Values for mandatory metadata units are required while values for optional metadata units are encouraged but not required.

SIRF serialization for CDMI/LTFS/Swift specify how a CDMI container, LTFS Tape or Swift container can become also SIRF-compliant. A SIRF-compliant CDMI container, LTFS Tape or Swift container enables future CDMI/LTFS/Swift clients “understand” containers created by today’s CDMI/LTFS/Swift clients although the properties of the future client is unknown to us today. By “understand”, it means that clients can identify the preservation objects in the container, the packaging format of each object, its fixity values, etc. (as defined in the SIRF catalog). This document also includes sections on SIRF serialization for CDMI (Section 6), SIRF serialization for LTFS (Section 7) and SIRF serialization for OpenStack Swift (Section 8).

SIRF includes metadata about the storage container, to help “understand” the container information in the future. No single technology will be usable over the time spans mandated by current digital preservation needs. SNIA CDMI, Swift and LTFS technologies are among best current choices, but are good for perhaps 10 years to 20 years. SIRF provides a vehicle for collecting all of the information that will be needed to transition to new technologies in the future, and it can be serialized for the future technologies as they come.

6 Container Information Metadata

6.1 Specification Category

This category includes information about the SIRF specification used for this container. As the **specification may evolve over time and distinct containers may use different SIRF specifications, it is** denoted in the SIRF catalog the specification used, and the SIRF level. SIRF level 2 specification defines more detailed metadata in the container.

The elements of the Specification category are:

— **Specification ID** (containerSpecificationIdentifier) — the specification identifier e.g., "SIRF-1.0"

— **Specification Version** (containerSpecificationVersion) — the specification version e.g., "1.0"

- **SIRF Level** (containerSpecificationSirfLevel) — the SIRF level that should be "1"

Hierarchical Representation

1 containerInformation (1-1: M, NR)

1.1 containerSpecification (1-1: M, NR)

1.1.1 containerSpecificationIdentifier (1-1: M, NR)

1.1.2 containerSpecificationVersion (1-1: M, NR)

1.1.3 containerSpecificationSirfLevel (1-1: M, NR)

6.2 Container ID Category

This category includes the container unique identifier and it has just one element:

- **Container ID** (containerIdentifier) — the container unique identifier, e.g., the tape id or cloud container id

The Container ID element is composed of the following attributes:

- **Container Identifier Type (containerIdentifierType)** — a designation of the naming authority and the domain within which the object identifier is unique.
- **Container Identifier Locale (containerIdentifierLocale)** — the locale of the identifier based on the Internet Assigned Numbers Authority (IANA).
- **Container Identifier Value (containerIdentifierValue)** — a Unicode/UTF-8 string for identifier actual value.

Hierarchical Representation

1 containerInformation (1-1: M, NR)

1.2 containerIdentifier (1-1: M, NR)

1.2.1 containerIdentifierType (1-1: M, NR)

1.2.2 containerIdentifierLocale (1-1: M, NR)

1.2.3 containerIdentifierValue (1-1: M, NR)

6.3 State Category

The state metadata is an indication of the progress of any activities that are to be carried out against a container. For example, if a container holds many preservation objects, state may indicate whether all of the objects intended for a container have been included or not. Or, state may indicate an in-process migration of a container.

The elements of the State category, as shown in Table 1, are:

- **State Type** (containerStateType)
- **State Value** (containerStateValue)

Table 1 — State category types and values

Type	Accepted Values	Use
INITIALIZING	TRUE	TRUE when the container is being initialized, i.e., the magic object, container provenance and/or catalog are being created
READY	ACTIVE FINALIZED	ACTIVE when the container is ready to receive new POs or catalog changes FINALIZED when the container is closed, read-only, and cannot receive more POs
NOT READY	DESTROYED ERROR	DESTROYED when the container has been destroyed and can no longer be read or modified. ERROR when there's a failure that cannot be specified using any other state type
MIGRATING	TRUE	TRUE when the data stored in a container are being moved to/from another container

The container state transitions occur depending on the current state and the action that is being taken in the container. Figure 2 shows the possible state transitions.

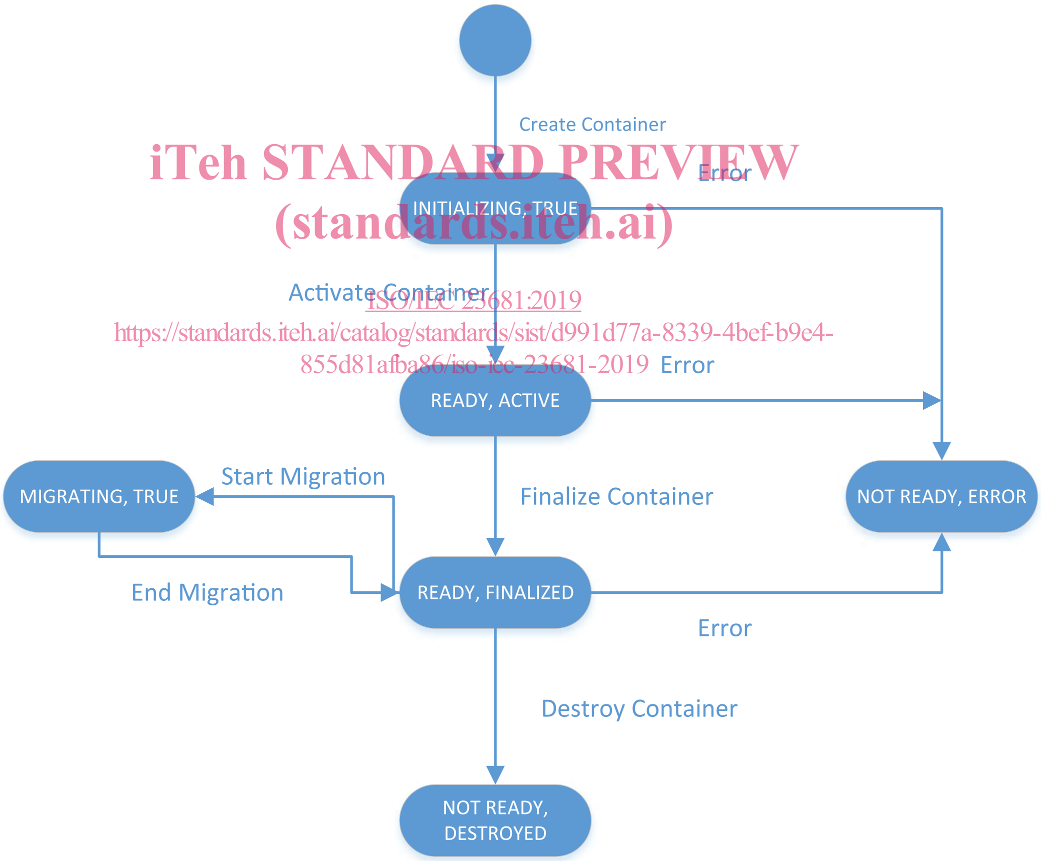

Figure 2 — Possible container states and transitions

Hierarchical Representation

1 containerInformation (1-1: M, NR)

1.3 containerState (1-1: M, NR)

1.3.1 containerStateType (1-1: M, NR)

1.3.2 containerStateValue (1-1: M, NR)