Funded by the European Commission

# M2.3g Procedure for the E-ARK CITS Interoperability assessment framework


E-ARK3: AGREEMENT No LC-01390244 CEF-TC-2019-3 eArchiving

# Cover Sheet

**Document Status:**

| Status |
|---|
| Draft |

**Document Approver(s)**

| Name | Role |
|---|---|
| Fulgencio Sanmartin/Adelina Cornelia Dinu | DG CNECT Business Owner |

**Document Reviewer(s)**

| Name | Role |
|---|---|
| DILCIS Board | Owner of the procedure |
| eArchiving Building Block users | eArchiving Building Block specification creators |

**Summary of Changes:**

| Version | Date | Created by | Short Description of Changes |
|---|---|---|---|
| V0.1 | 2021-09-30 | Miguel Ferreira | Draft for transformation to correct format handed over |
| V0.2 | 2021-10-13 | Karin Bredenberg and Jaime Kaminski | Draft of correct format |
| V1.0 | 2021-10-15 | Karin Bredenberg and Jaime Kaminski | Version 1.0 published |
| | | | |
| | | | |

| Glossaries of terms |
|---|
| E-ARK vocabs: http://evoc.dlmforum.eu/E-ARK/group/5568370c3448e76821b3942f/list<br>CEF Glossary: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Glossary |

# M2.3g Procedure for the E-ARK CITS Interoperability assessment framework

## Contents

# Procedure for the E-ARK CITS Interoperability assessment framework (version 1.0)


## Summary

This document defines a set of criteria and corresponding scores to capture how detailed or precise a particular E-ARK Content Information Type Specification (CITS) is regarding interoperability. The main objective is to ensure interoperability between different implementations of the same CITS. The more complete in scope and technically detailed a specification is, the closer it will get to the goal of semantic interoperability between systems, organisations and countries.


## E-ARK Information Packages

E-ARK information packages adopt an onion-like structure with the Common Specification for Information Packages (CSIP) as the outermost layer.

The general SIP, AIP, and DIP specifications add submission, archiving, and dissemination information to the CSIP specification. These two information package layers are not discussed in this document.

The third layer of the structure represents specific Content Information Type Specifications, such as the eHealth1 or the Geo-data specification (see DILCIS board Website for more information on CITS). Additional layers for business-specific specifications and local implementations of any specification may be added.

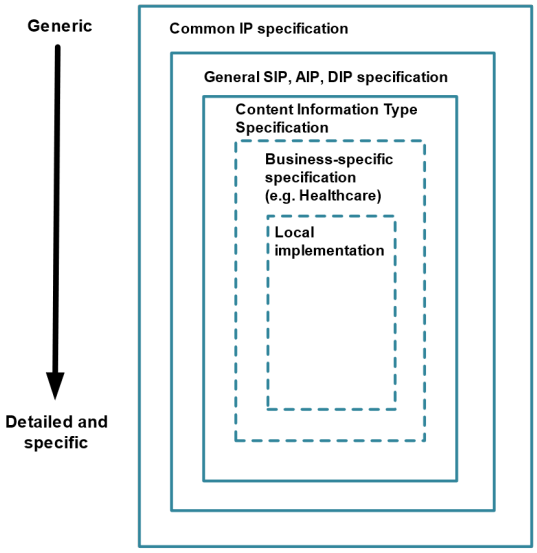


Figure 1: The onion-like structure for the E-ARK specifications.

Every layer in the data model structure inherits metadata entities and elements from the layers outside it. A flexible schema has been adopted to increase uptake. This allows for extension points where the schema in each layer can be extended to accommodate additional requirements on the next, more specific layer until any local implementation adds specific entities or metadata elements to satisfy specific local needs. Local or business-specific implementations must ensure that they remain compliant with the CITS (i.e. that they only add additional constraints instead of rewriting or changing the meaning of the constraints posed by any of the other layers).

For example, extension points can be implemented by:

- Embedding external extension schemas (in the same way as supported by METS [http://www.loc.gov/standards/mets/] and PREMIS [http://www.loc.gov/standards/premis/]). These support both increasing the granularity of existing metadata elements by using more detailed data structures and adding new types of metadata.
- Adding new folders to the package to store specific information (e.g. a particular type of package documentation).
- Tighter definition of data formats or additional representations (e.g. clearly specifying the allowed data formats to be included in each type of representation present in the package).
- Including additional metadata schemas for standards more appropriate for the local implementation (e.g. the use of the HL7 FHIR schema for patient personal information).
- Adding further controlled vocabularies to control the values of specific fields of information.

## Interoperability assessment framework

The E-ARK CITS Interoperability Assessment Framework tries to assess the coverage of a particular CITS across the three main package elements:

- **Structure**: where metadata and content files should be located within a package.
- **Metadata**: what metadata should be contained in a package and which schemas/conventions should be used to encode that metadata.
- **Data**: how should the data be packaged, what file formats should be used, etc.

Each of these elements is assessed separately, and a scoring system is used to denote the level of detail/specificity included in the CITS.

The element scores are assigned using a progressive scale, starting with zero, denoting no changes/improvements over the original Package Specification (SIP, AIP or DIP). To achieve a higher score, it is necessary to fulfil all of the previous rating level criteria, meaning that it is not possible to qualify for a three-star rating without satisfying the conditions for the one and two-star ratings.

The following table enables CITS authors and implementers to quickly understand the scope and level of specificity of a particular CITS. For authors, it is of most use as a set of considerations that should be considered while writing a new CITS. For adopters/implementers, it serves as an overview of a CITS' scope and specificity. The more complete and specific a CITS is, the easier it is to preserve data and the more interoperable between different implementations it will be. On the other hand, complete and detailed specifications will be harder to implement.

Table 1: The interoperability assessment framework scoring table.

| Interoperability score | (S) Structure | (M) Metadata | (D) Data |
|---|---|---|---|
| 0<br>Package level interoperability | **No requirements on structure**<br><br>The CITS is compliant with the CSIP and does not impose any additional restrictions on the package structure | **No requirements on metadata**<br><br>The CITS does not include any restrictions on metadata | **No requirements on data**<br><br>The CITS does not include any restrictions on the type of data included in the package |

| Interoperability score | (S) Structure | (M) Metadata | (D) Data |
|---|---|---|---|
| 1 ☆ Content level interoperability | **Defines the mandatory folders**<br><br>The CITS clearly identifies the "folders" that should be present in the package | **Defines mandatory metadata schemas**<br><br>The CITS clearly identifies all metadata schemas that should be used and describes the mandatory metadata elements that should be present in the package | **Defines the mandatory representations and file format families**<br><br>The CITS clearly identify the minimum set of representations to be included and the file format families that are expected in the package |
| 2 ☆☆ Semantic interoperability | **Defines the mandatory files**<br><br>The CITS clearly identifies the files and the naming of the files expected to be present in each package "folder" | **Defines the mandatory semantics of metadata elements**<br><br>The CITS includes detailed instructions on the semantics of the mandatory metadata elements, including the vocabularies and allowed values for each of those elements | **Defines the mandatory file formats**<br><br>The CITS includes clear instructions on which file formats are allowed to be used in the package |
| 3 ☆☆☆ Relationships interoperability | **Relationships between packages are defined**<br><br>The CITS specifies how relationships between packages should be implemented (e.g. hierarchical organisation of packages in the repository or archive collection information as defined by OAIS) | **Relationships between metadata are defined**<br><br>The CITS identifies the relationships between different metadata elements from the same or different metadata schemas (e.g. element equivalence relationships or crosswalks) | **Relationships between data are defined**<br><br>The CITS clearly specifies how relationships between data files may be established (e.g. links between a SIARD file and external BLOBs, Web page index file and its other constituent files such as images, CSS, Javascript files, etc.; relationships between files within the same representation (linking done by using the same filename for example), links between documentation and data, links to external "objects" that are necessary to render or understand the data |

## Self-assessment and publication

Each CITS author should assess their CITS before submitting it for review and publication to the DILCIS Board. This exercise enables authors to improve their CITS before submitting it, resulting in a higher quality specification. This document includes examples of self-assessment reports for the SIARD and Archival Description CITS.

During the review process, the DILCIS Board will analyse the self-assessment report of the submitted CITS and accept the suggested score or provide comments for improvement.

The resulting score will be published alongside the CITS on the DILCIS Board Website.

## Scoring system

The scoring system uses a visual indicator (star) to represent how interoperable a CITS is regarding each of the three dimensions: Structure, Metadata and Data.

Example:

| Dimension | Interoperability score |
|---|---|
| Structure | ☆☆ |
| Metadata | ☆☆☆ |
| Data | ☆ |

For shortening purposes, a second design can be used to represent the assessment score. The alternative representation is based on a coding scheme that combines a letter corresponding to dimension under assessment and a number representing the actual score under that dimension.

The previous example can be represented as a score of S2M3D1, meaning that it achieved a rating of two for structure, three for metadata and one for data specifications.

## Appendixes

The interoperability assessment framework has been used to create the following two examples written from an assessment perspective.

### Appendix 1 - SIARD CITS self-assessment report example

The SIARD CITS combines the package structure and metadata recommendations from the CSIP and SIP with the SIARD specification, https://github.com/DILCISBoard/SIARD. Together these provide best practices for the long term archiving of relational database content with associated metadata.

#### *Scope*

The first issue was establishing the scope of the assessment, that is, what elements of content, particularly metadata, should be considered. It is unreasonable and counterproductive to expect a CITS to cover all aspects of content and metadata. The SIARD CITS places restrictions on the content file formats and aspects of technical metadata, e.g. schemas. Many other metadata elements would be better covered by a dedicated CITS (e.g. ERMS data or package relationships).

#### *Structure*

The assessment of structure was relatively straightforward for the first two-star categories. The SIARD CITS covers the placement of particular content and metadata files within the information package. The first two grades build on each other, e.g. it is very difficult to define the location of mandatory files (two stars) without referring to particular folders (one star).

The three-star criteria, Relationships Interoperability, feel weak. The relationship between packages might be best described in the CSIP or a “meta” CITS that deals with this subject alone? If that is not the case, then we can expect a proliferation of different package relationship rules for different content types.

The SIARD CITS has been given a two-star rating for structure as it provides clear guidance on a set of mandatory files and folders.

#### *Metadata*

Metadata proved the most challenging aspect to assess, and the criteria may require some adjustment. The main issue is that the SIARD CITS appears to satisfy some elements of the two-star criteria (definition of mandatory semantics) by restricting some of the metadata elements in the package METS file, e.g. CONTENTINFORMATIONTYPE. The difficulty is that it could be done without defining mandatory metadata schemas for much of the package metadata. Clarity of scope is needed (i.e. which metadata types to assess). Offering small refinements/restrictions to the CSIP vocabularies should not be awarded a rating. A CITS should primarily concern itself with domain metadata (e.g. for SIARD, that is, metadata specific to database preservation, not more general metadata). Note that this might include metadata of any category (i.e. technical, descriptive, etc.), as long as the use recommendations restrict themselves to “domain concerns”.

The SIARD CITS received a two-star metadata rating. It prescribes the SIARD XML metadata format and offers some further semantic restrictions of SIARD metadata.

### *Data*

The most straightforward of the three criteria as the CSIP has little to say about data. The criteria here felt clear, and the assessment was easy to apply. The only issue is scope/coverage here (i.e. how much evidence is required to satisfy particular criteria?). It is possible that a CITS might make some stipulations about file formats but leave others as simply a format family (e.g. raw database file for SIARD). This is effectively offering strict format choices for some files and more general prescriptions for others. Whether this is worth one star or two becomes a judgment call. It might also be possible for a CITS to describe relationships between data files without prescribing a format. That is to satisfy the three-star criteria without satisfying the criteria for two stars.

The SIARD CITS was judged to be three stars for data. It offers format recommendations for much of the data, though not all (two stars). It also describes the relationships between segmented data files and data files to large binary objects.

### *Interoperability score*

| Dimension | Interoperability score |
|---|---|
| Structure | ☆☆ |
| Metadata | ☆☆ |
| Data | ☆☆☆ |

Structure: 2, Metadata 2, Data 3 (S2M2D3)

## Appendix 2 - Archival Description CITS self-assessment report example

The archival description CITS is a light document that offers recommendations for specific types of descriptive metadata:

- Archival Information (Finding Aid): ISAD-G as Encoded Archival Description (EAD);
- Archival Creator: ISAAR(CPF) as Encoded Archival Context for Corporate Bodies, Persons and Families (EAC-CPF)
- Archival Institution: ISDIAH as Encoded Archival Guide (EAG)
- Functions of Archival Creator: EGAD as Records in Context (RiC)

### *Scope*

The Archival Information CITS covers a few specific metadata formats, schemas, and the location of particular metadata files. It has very little to say about the structure and nothing about data.

### *Structure*

There is clear guidance on where to put particular metadata files (i.e. in the CSIP metadata folder). I am hovering between one star and two for structure. The CITS clarifies where the metadata files should be placed but says nothing about most other files.

### *Metadata*

The CITS gives clear guidance on the metadata schemas required for each MD type and the METS attribute values required. I'm unconvinced that this is enough for two stars and am again left hovering between one and two stars.

### *Data*

There are no recommendations regarding content. Indeed it raises the prospect of a metadata only IP, so no rating.

### *Interoperability score*

| Dimension | Interoperability score |
|---|---|
| Structure | ☆ |
| Metadata | ☆ |
| Data | |

Structure: 1, Metadata 1, Data 0 (S1M1D0)