Specification for the E-ARK Content Information Type Specification for Relational Databases using SIARD (CITS SIARD)

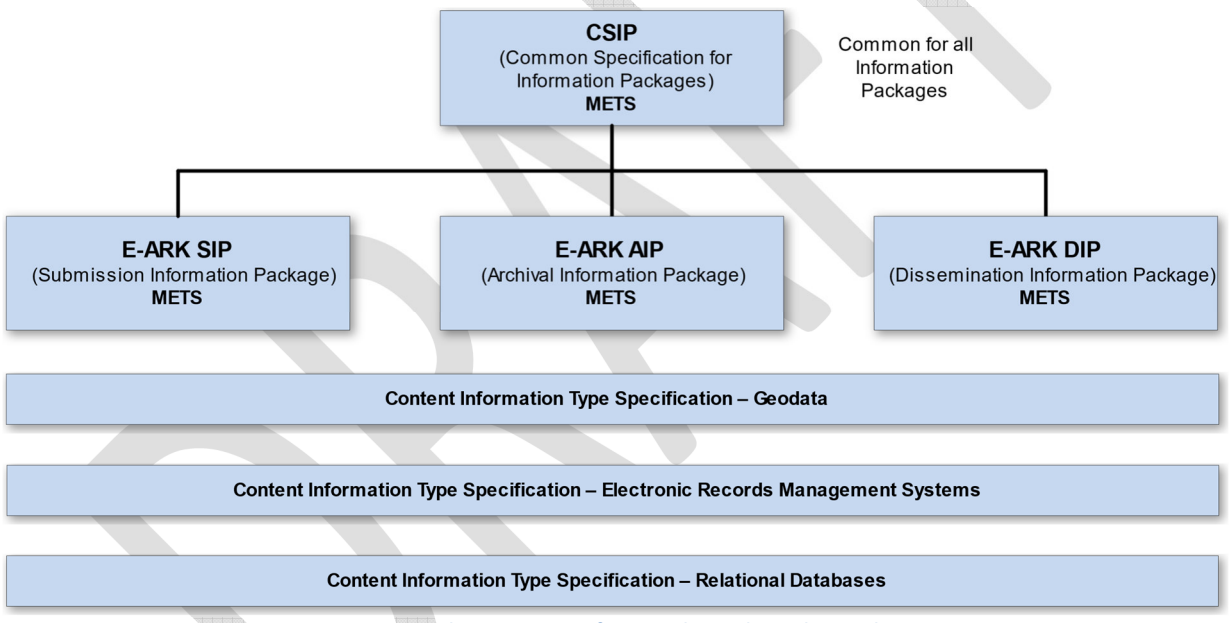
A proper front page will be created for the publication occurring after implementation of review comments.

1 Preface

The correct preface will be inserted for the publication occurring after implementation of review comments.

1.1 Aim of the specification

This E-ARK specification is part of a family of specifications that provide a common set of requirements for packaging digital information. These specifications are based on common, international standards for transmitting, describing and preserving digital data. They have been produced to help data creators, software developers and digital archives tackle the challenge of short-, medium- and long-term data management and reuse in a sustainable, authentic, cost-efficient, manageable and interoperable way.
The foundation for these specifications is the Reference Model for an Open Archival Information System (OAIS) which has Information Packages at its core. Familiarity with the core functional entities of OAIS is a prerequisite for understanding the specifications. A visualisation of the current specification network can be seen here:



The E-ARK specification dependency hierarchy

Specification	Aim and Goals
Common Specification for Information Packages	This document introduces the concept of a Common Specification for Information Packages (CSIP). Its three main purposes are to: • Establish a common understanding of the requirements which need to be met in order to achieve interoperability of Information Packages. • Establish a common base for the development of more specific Information Package definitions and tools within the digital preservation community. • Propose the details of an XML-based implementation of the requirements using, to the largest possible extent, standards which are widely used in international digital preservation.

	Ultimately the goal of the Common Specification is to reach a level of interoperability between all Information Packages so that tools implementing the Common Specification can be adopted by institutions without the need for further modifications or adaptations.
E-ARK SIP	The main aims of this specification are to: • Define a general structure for a Submission Information Package format suitable for a wide variety of archival scenarios, e.g. document and image collections, databases or geographical data. • Enhance interoperability between Producers and Archives. • Recommend best practices regarding metadata, content and structure of Submission Information Packages.
E-ARK AIP	The main aims of this specification are to: • Define a generic structure of the AIP format suitable for a wide variety of data types, such as document and image collections, archival records, databases or geographical data. • Recommend a set of metadata related to the structural and the preservation aspects of the AIP as implemented by the reference implementation eArchiving ToolBox (formerly earkweb). • Ensure the format is suitable to store large quantities of data.
E-ARK DIP	The main aims of this specification are to: • Define a generic structure of the DIP format suitable for a wide variety of archival records, such as document and image collections, databases or geographical data. • Recommend a set of metadata related to the structural and access aspects of the DIP.
Content Information Type Specifications	The main aim and goal of a Content Information Type Specification is to: • Define, in technical terms, how data and metadata must be formatted and placed within a CSIP Information Package in order to achieve interoperability in exchanging specific Content Information. The number of possible Content Information Type Specifications is unlimited.

1.2 Organisational support

This specification is maintained by the Digital Information LifeCycle Interoperability Standards Board (DILCIS Board). The DILCIS Board (http://dilcis.eu/) was created to enhance and maintain the draft specifications developed in the European Archival Records and Knowledge Preservation Project (E-ARK project) which concluded in January 2017 (http://eark-project.com/). The Board consists of eight members, but there is no restriction on the number of participants in the work. All Board documents and specifications are stored in GitHub (https://github.com/DILCISBoard) while published versions are made available on the Board webpage. Since 2018 the DILCIS Board has been responsible for the core specifications in the Connecting Europe Facility eArchiving Building Block (https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving).

1.3 Authors

A full list of contributors to this specification, as well as the revision history can be found in Appendix 1.

TABLE OF CONTENTS

LIST OF FIGURES

2 Context

2.1 Purpose

The purpose of this document is to describe the Content Information Type Specification for Relational Databases (RDB) using the format Software Independent Archiving of Relational Databases (SIARD). The specification is designed to be used for the transfer to and from archives.

2.2 Layered data model

This section introduces the structure of the data model, which is based on a layered approach for information package definitions (Figure 1). The Common Specification for Information Packages (CSIP) forms the outermost layer. The general SIP, AIP and DIP specifications add, respectively, submission, archiving and dissemination information to the CSIP specification. The third layer of the model represents specific content information type specifications, such as this CITS SIARD specification. Additional layers for business-specific specifications and local variant implementations of any specification can be added.

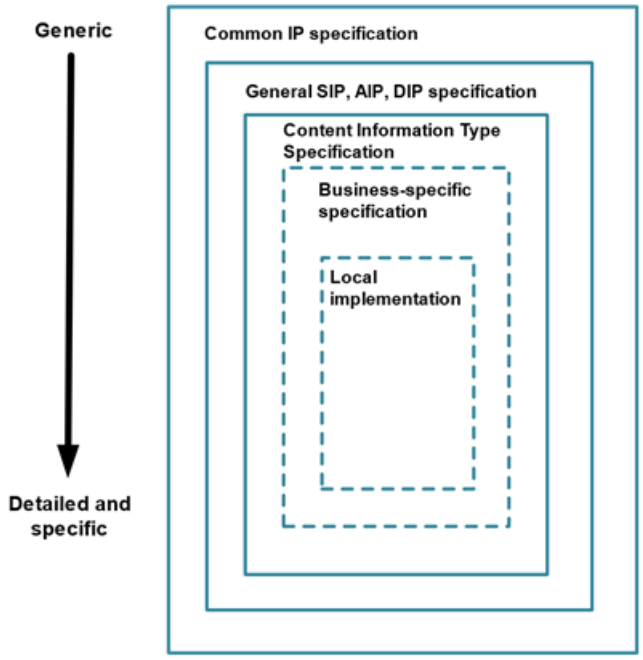


Figure 1: Data Model Structure

Every level in the data model structure inherits metadata entities and elements from the higher levels. In order to increase adoption, a flexible schema has been developed. This will allow for extension points where the schema in each layer can be extended to accommodate additional information on the next specific layer until, finally, the local implementation can add specific entities or metadata elements to satisfy very specific local needs. Extension points can be implemented by:

- Embedding foreign 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 as well as adding new types of metadata.
- Substituting metadata schemas for standards more appropriate for the local implementation.

The structure allows the addition of more detailed requirements for metadata entities, for example by:

- Increasing the granularity of metadata elements by using more detailed data structures, or
- Adding local controlled vocabularies.

For consistency, design principles are reused between layers as much as possible.

2.3 The boundaries of this specification and the SIARD-specification

SIARD is an independent format for archiving relational databases and hence has its own specification (https://github.com/DILCISBoard/SIARD), but there are areas where the SIARD specification deliberately states that packaging of the SIARD-file among other aspect is outside the scope of the SIARD specification:

> *"It should be noted that the SIARD format is only the long-term storage format for a specific type of digital documents (relational databases) and is therefore designed entirely independently of package structures such as the SIP (Submission Information Package), AIP (Archival Information Package) and DIP (Dissemination Information Package) in the OAIS model.*
> *It is assumed that a database in SIARD format is archived as part of such an information package together with other documents (externalised large object files, translation maps for external file names, database documentation, business documents relevant to the understanding of the database, etc.)."*
> - *SIARD 2.1.1, p. 7*

This CITS SIARD specification describes how to package SIARD-files and any accompanying external LOBs in CSIP package(s). This specification also describes how to package extra metadata and context documentation so that long-term preservation and dissemination can take place.

As in all classification issues, it is important to have collectively exhaustive and mutually exclusive categories, and even though the SIARD specification deliberately states that package structures are not part of the specification, then there are circumstances and scenarios where it is not clear whether an issue falls under the scope of a specification like this one or under the scope of the SIARD specification itself.

This CITS SIARD specification has been written during the eArchiving building block in the two-year E-ARK3-project, and during this period the project participants have been struggling with issues that lie in the grey area between the two specifications. In the E-ARK3-project the project participants have decided to create a Request For Comments (RFC) for an update to the SIARD specification which concentrates on large databases and external lobs. This CITS SIARD specification is therefore dependent on the outcome of that RFC and has been written as if the RFC has been adopted into the SIARD specification.

3 CITS SIARD Requirements

3.1 Folder structure and example

A visualisation of an example of a valid CITS SIARD-package is illustrated in Figure 2. The example and other examples can also be found as downloadable packages at this link: https://github.com/DILCISBoard/SIARD-CITS/tree/master/examples. The example is an information package where a database has LOBs that resides outside the .siard-file. See LOB details under section 3.7 {SIARD_lobs} – requirements.

Folder Structure of Northwind Sample Database

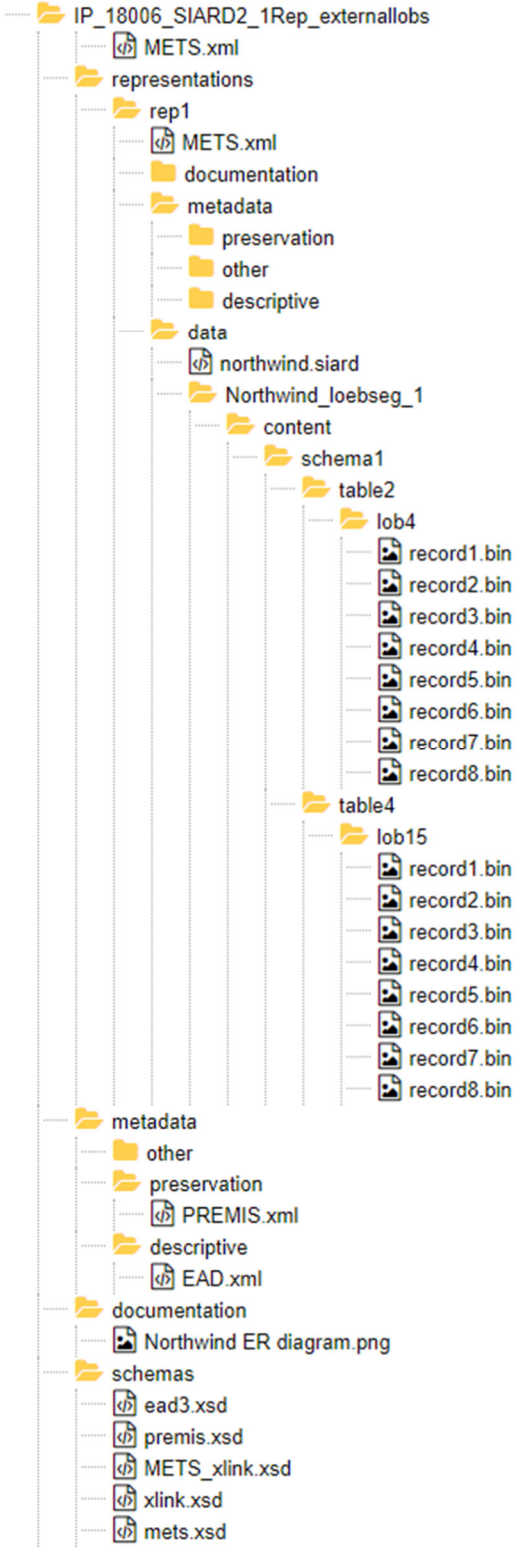


Figure 2: Information Package folder structure

3.2 Package and Representation METS

A CSIP can consist of zero to many representations, and this is an important feature that needs to be taken into consideration when packing SIARD files within CSIPs.

There can easily be different representations of the same database located within one CSIP. For example, one package could consist of:

- one representation where the native proprietary dump is located;
- one representation with SIARD-file that conforms only to an older version of the SIARD specification;
- one representation with the newest version of the SIARD specification;
- one representation where database normalisation and/or other dissemination tasks have taken place.

There can be several DIP representations. There can also be other databases and for example, geodata within the same package.

As for this specification, there always needs to be a minimum of one representation and therefore, a minimum of two METS.xml. The Package METS.xml has to be a general METS.xml describing if the package itself is mainly a CITS_SIARD package, and then the single representations need to describe what specific SIARD versions they consist of.

ID	Name and Location	Description and Usage	Card & Level
SIARD_1		There **MUST** be minimum one representation and therefore minimum one Package METS.xml and minimum one Representation METS.xml in a CITS SIARD package.	1..1 MUST

2.3 Package METS requirements

ID	Name and Location	Description and Usage	Card & Level
SIARD_2 Ref CSIP2	Type mets/@TYPE	For information packages that primarily contain relational databases, the value in Package mets/@TYPE **MUST** be "Databases" as taken from the CSIP Vocabulary for Content Category.	1..1 MUST
SIARD_3 Ref CSIP4	Content Information Type Specification mets/@csip:CONTENTINFORMATIONTYPE	For information packages that primarily contain relational databases, the value in Package mets/@csip:CONTENTINFORMATIONTYPE **MUST** be "CITS_SIARD" as taken from the CSIP Vocabulary for Content Information Type.	1..1 MUST

SIARD_4 Ref CSIP5	Other Content Information Type Specification mets/@csip:OTHERCONTENTINFORMATIONTYPE	For information packages that primarily contain relational databases, the Package METS must **NOT** have a mets/@csip:OTHERCONTENTINFORMATIONTYPE	0..0 NOT
SIARD_5 Ref CSIP6	METS Profile mets/@PROFILE	For information packages that primarily contain relational databases the value in the @PROFILE **MUST** be "https://SIARD.dilcis.eu/profile/CITS_SIARD.xml"	1..1 MUST
SIARD_6 Ref CSIP62	fileSec Representation Content Information Type Specification mets/fileSec/fileGrp[@USE='Representations']/@csip:CONTENTINFORMATIONTYPE	There **MUST** be a minimum of one mets/fileSec/fileGrp[@USE='Representations']/@csip:CONTENTINFORMATIONTYPE with the value "CITS_SIARD" as taken from the CSIP Vocabulary for Content Information Type that direct to the representation METS.xml in the representation containing a relational database.	1..n MUST
SIARD_7 Ref CSIP63	fileSec Other Content Information Type Specification mets/fileSec/fileGrp[@csip:CONTENTINFORMATIONTYPE=' CITS_SIARD ']/@csip:OTHERCONTENTINFORMATIONTYPE	For any mets/fileSec/fileGrp[@csip:CONTENTINFORMATIONTYPE with the value "CITS_SIARD" there **MUST** be a @csip:OTHERCONTENTINFORMATIONTYPE attribute with a value taken from the vocabulary {SIARD_1.0; SIARD_2.0, SIARD_2.1, Database_dump}.	1..1 MUST
SIARD_8 Ref CSIP105-CSIP112	StructMap METS pointer	For any fileGrp/@csip:CONTENTINFORMATIONTYPE with the value "CITS_SIARD" there **MUST** be a corresponding @div-representation in the StructMap-element	1..1 MUST

Example 1: Package METS element example.

3.4 Representation METS requirements

ID	Name and Location	Description and Usage	Card & Level
SIARD_9 Ref CSIP2	Type mets/@TYPE	For representations that primarily contain relational databases, the value in Package mets/@TYPE **MUST** be “Databases” as taken from the CSIP Vocabulary for Content Category.	1..1 MUST
SIARD_10 Ref CSIP4	Content Information Type Specification mets/@csip:CONTENTINFORMATIONTYPE	For representations that primarily contain relational databases and that conforms to CITS SIARD the value in Package mets/@csip:CONTENTINFORMATIONTYPE **MUST** be “CITS_SIARD” as taken from the CSIP Vocabulary for Content Information Type.	1..1 MUST
SIARD_11 Ref CSIP5	Other Content Information Type Specification mets/@csip:OTHERCONTENTINFORMATIONTYPE	For representations where mets/@csip:CONTENTINFORMATIONTYPE has the value “CITS_SIARD” then mets/@csip:OTHERCONTENTINFORMATIONTYPE **MUST** have a value taken from the vocabulary {SIARD_1.0; SIARD_2.0, SIARD_2.1, Database_dump}	0..0 NOT
SIARD_12 Ref CSIP6	METS Profile mets/@PROFILE	For information packages that primarily contain relational databases the value in the @PROFILE **MUST** be “https://SIARD.dilcis.eu/profile/CITS_SIARD_representation.xml”	1..1 MUST
SIARD_13 Ref CSIP64-CSIP79	File Pointer fileSec/fileGrp/file@csip:OTHERCONTENTINFORMATIONTYPE	If the value in mets/@csip:OTHERCONTENTINFORMATIONTYPE is {SIARD_1.0, SIARD2.0, SIARD2.1, Database_dump} then there **MUST** exist one and only one file in the fileGrp with @USE = “data” with an identical value in fileSec/fileGrp/file@csip:OTHERCONTENTINFORMATIONTYPE that is used to locate the relevant database file.	1..1 MUST

3.5 METS requirements between Package and Representation

ID	Name and Location	Description and Usage	Card & Level
SIARD_14	Type mets/@TYPE	If the value in representation mets/@csip:OTHERCONTENTINFORMATIONTYPE is {SIARD_1.0, SIARD2.0, SIARD2.1, Database_dump} then the Package METS.xml fileGrp who refers to the Package METS.xml **MUST** have the same value.	1..1 MUST

3.6 {SIARD_1.0, SIARD2.0, SIARD2.1} – requirements

ID	Name and Location	Description and Usage	Card & Level
SIARD_15		If the value in mets/@csip:OTHERCONTENTINFORMATIONTYPE is {SIARD_1.0, SIARD2.0, SIARD2.1} then there **MUST** exist a file named [databaseName].siard in representations/[RepresentationName]/data	1..1 MUST
SIARD_16		The SIARD version of the SIARD-file **MUST** be the same as the version provided in mets/@csip:OTHERCONTENTINFORMATIONTYPE and fileSec/fileGrp/file@csip:OTHERCONTENTINFORMATIONTYPE	MUST
SIARD_17		The representations/[RepresentationName]/data/[databaseName].siard **SHOULD** be a valid SIARD file	SHOULD
SIARD_18		There **SHOULD** be minimum of one validation report in the documentation folder for the validation of the SIARD-file	1..n SHOULD
SIARD_19		The file name of the SIARD file representations/[RepresentationName]/data/[databaseName].siard **MAY** be the short database identifier of the database as	

	specified in the <dbname> element of the metadata.xml file in the SIARD file, but it is not recommended.	MAY

3.7 {Database_dump} – requirements

For authenticity and possible dissemination purposes, the OAIS might want to have a representation with a proprietary database dump from the original database management system.

ID	Name and Location	Description and Usage	Card & Level
SIARD_20		If the value in mets/@csip:OTHERCONTENTINFORMATIONTYPE is “Database_dump” then there **MUST** exist a proprietary database dump in representations/[RepresentationName]/data	1..1 MUST
SIARD_21		There **SHOULD** be preservation metadata describing the proprietary database dump	1..n SHOULD

3.8 {SIARD_lobs} – requirements

A relational database can consist solely of table data, but it can as easily have large objects (LOBs). Large object (LOB) is the common description for large character content (CLOB) or large binary (BLOB) content – such as video, sound, images, word processing documents, etc.

These LOBs can be stored inside a relational database as CLOBs or BLOBs within cells or outside as external files – also called external LOBs (SQL/MED).

In the SIARD specification from SIARD2.0 and onwards, the external LOBs can be placed outside the table data within the folder structure in the .siard-file, or they can be placed outside the .siard-file. As stated in the scope of this specification, there is work being done to create an RFC to the SIARD specification that handles large databases and LOBs.

ID	Name and Location	Description and Usage	Card & Level
SIARD_22			

	If a database has LOBs outside the .siard-file then these **SHOULD** be stored in the same representation as the .siard-file	SHOULD
SIARD_23	LOBs **MAY** be stored in its own representation, and the value in mets/@csip:OTHERCONTENTINFORMATIONTYPE is "SIARD_lobs". For storage and preservation actions, the OAIS can decide to handle LOBs in its own representation. This way, there can be different representations of .siard-files that link to the same lob-representation. The complexity rises by choosing this solution, and the CSIP states: "Representation level METS files should not reference files outside of their representation". It, therefore, has to be a deliberate choice to allow this way of handling LOBs.	MAY

4 SIP requirements

4.1 Submission Agreement requirements

There should be a submission agreement in the SIP representation that has been tailored to handle the preservation of relational databases. Since no standard for submission agreements for databases exist yet, the following requirements are not yet able to be automatically validated at this specification level. It is up to the business-specific specification layer or local implementation layer (see 1.2 Layered Data Model) to set up requirements that can be automatically validated.

ID	Name and Location	Description and Usage	Card & Level
SIARD_23		There **SHOULD** be a submission agreement in the SIP representation that has been tailored to handle preservation of relational databases	1..1 SHOULD
SIARD_24		The submission agreement **SHOULD** describe how many representations of the database that the Producer has to submit.	0..1 SHOULD
SIARD_25		The submission agreement **SHOULD** describe whether the submitted representations of a database is 1:1 with the running	0..1 SHOULD

	database (Full SIARD export) or if any alterations have been made (only a subset of tables)	
SIARD_26	The submission agreement **SHOULD** list the tables that are required to be submitted to the archive and to be preserved.	0..1
SIARD_27	The submission agreement **SHOULD** list a set of SQL queries that are decided to be submitted to the archive and are to be preserved under the <views>-element in metadata.xml. The SQL queries **SHOULD** provide the most useful queries in the database for designated communities.	0..1 SHOULD
SIARD_28	The submission agreement **SHOULD** list the documentation that is decided to be submitted to the archive. See 7 Documentation requirements	0..1 SHOULD

5 AIP requirements

None yet.

6 DIP requirements

None yet.

7 Documentation requirements

There should be documentation in the representations and/or in the information package. Since no standard for documentation of databases exists yet, the following requirements are not yet able to be automatically validated at this specification level. It is up to the business-specific specification layer or local implementation layer (see 1.2 Layered Data Model) to set up requirements that can be automatically validated.

ID	Name and Location	Description and Usage	Card & Level

SIARD_29	Tables, columns/fields, keys, coded values SHOULD be explained, preferably in the metadata.xml and via code tables or the SIARD file or alternatively in the Documentation folder.	1..n SHOULD
SIARD_30	There **SHOULD** be a system diagram in the Documentation folder. Preferably an Entity/Relationship Diagramme	1..n SHOULD
SIARD_31	The (main) system-user dialogues **SHOULD** be documented, down to the identification of the database columns/fields involved in the dialogues, documented as a combination of: · Screenshots, annotated with column/field descriptions, stored in the Documentation folder. · User documentation describing the system-user dialogue stored in the Documentation folder. · Views, if available, as part of the SIARD file. · If views are not present, additional descriptions of the system (application) logic, stored in the Documentation folder.	1..n SHOULD
SIARD_32	Documentation of the legal context of the database and associated system **SHOULD** be provided in the Documentation folder.	1..n SHOULD
SIARD_33	There **MAY** be videos or screen dumps from the system as seen from the user's point of view in the Documentation folder.	1..n SHOULD

Glossary

Table 2: Glossary

Name	Description
Content Information Type (CIT)	A type of a set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information.
DBMS	Database Management System
OAIS	Open Archival Information System
RDBMS	Relational Database Management System
SIARD	Software Independent Archival of Relational Databases

Postface

AUTHOR(S)	
Name(s)	Organisation(s)
Phillip Aasvang Tømmerholt	The Danish National Archives
Anders Bo Nielsen	The Danish National Archives
Asbjørn Skjødt	The Danish National Archives
Henrik K. R. Jakobsen	The Danish National Archives
Karin Bredenberg	Kommunalförbundet Sydarkivera
Arne-Kristian Groven	Norwegian National Archives (Akrivverket)

REVIEWER(S)	
Name(s)	Organisation(s)
Janet Anderson	Highbury/DNA
Lauri Ratsep	National Archives of Estonia
Jaime Kaminski	Highbury/DNA

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the Consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Submitted Revisions History

Revision No.	Date	Authors(s)	Organisation	Description
[Version]	[Date]	[Who]	[Affiliation]	[What]
1.0	14-08-2020	Phillip Aasvang Tømmerholt	Danish National Archives	DRAFT version for internal review
1.1	31-08-2020	Phillip Aasvang Tømmerholt	Danish National Archives	DRAFT version with changes based on internal review
1.2	29-09-2020	Phillip Aasvang	Danish National	Included Layered Data Model and

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.