Introduction to the E-ARK Content Type Specifications

Transcript v.02

[1] Standalone training from the eArchiving Initiative

Welcome to the E-ARK Specifications training course. This unit is part of the standalone training series developed by the eArchiving Initiative. In this session, we will guide you through the concepts, structures, and practical applications of the E-ARK specifications. By the end, you'll have a clear understanding of how these standards support long-term digital preservation and ensure interoperability across different systems.

[2] Lesson Structure

Lesson 3 offers a general overview of the E-ARK Content Type Specifications. In this video, you will become familiar with the content type concept and how it extends the E-ARK Common Specification.

[3] Content Information Type Specifications (CITS)

As an interoperability standard, it must be possible to use the Common Specification for Information Packages regardless of the type and format of the content users need to handle. At the same time, each individual content type and file format can have specific characteristics which need to be considered for purposes of validation, preservation and curation.

E-ARK specifications introduce the concept of Content Information Type Specifications to facilitate such in-depth control over specific content types and formats. A Content Information Type Specification can include detailed requirements as to how content, metadata, and documentation for specific content types (for example relational databases or geospatial data) have to be handled within a CSIP (or E-ARK SIP, AIP or DIP).

A Content Information Type Specification is a mechanism used to extend the scope of the CSIP by defining additional requirements for specific Content Information Types. The OAIS Reference Model defines Content Information as "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".

Content Information Types can be regarded as categories of Content Information (e.g. relational databases, scientific data or digitised maps). A Content Information Type Specification defines the format and structure, mainly in regard to the Information Object, within an Information Package. This facilitates interoperability when exchanging specific Content Information Types.

[4] E-ARK Content Type Specifications

The Content Type Specifications build on the E-ARK Common Specification, aligning the unique needs of different data types with the shared structure of E-ARK information packages. This ensures that archives can handle a wide variety of digital content while maintaining interoperability and clarity across systems. By following these specifications, organisations can be confident that their digital objects remain accessible and understandable in the future.

The E-ARK Content Type Specifications currently include:

- CS Archival Information
- CS Preservation Metadata
- SIARD and CITS SIARD (to archive databases)
- CITS ERMS (to archive records from an electronic records management system)
- CITS Geodata (to archive geospatial data)
- CITS eHealth1 and eHealth2 (to archive healthcare related data)
- and CITS 3D (to archive 3D product model data)

Content Information Type Specifications can be domain-specific, and there may be multiple specifications to cover a particular domain. For example, archival institutions might define a Content Information Type Specification for archiving web sites with descriptive metadata in EAD format, while libraries might define a specification for archiving web sites with MARC metadata.

Pragmatically it may not be sufficient to deal only with the Information Object. For complex Content Information Types or large IPs, it may be necessary to provide explicit requirements for other metadata relevant to the specific content type. For example, the ERMS Content Information Type Specification prescribes a method for referencing data (i.e. computer files) from descriptive metadata in ERMS format, ensuring package integrity. Stating these requirements in a general specification allows archival institutions receiving SIPs, including ERMS extracts or whole systems to understand and validate potentially complex information packages.

In the next part, we'll take a closer look at the different content types, exploring them one by one.

[5] CS Archival Information

The *E-ARK Common Specification for Archival Information* sets out how key archival metadata—such as finding aids, details about creators and institutions, and functional descriptions—should be packaged and exchanged in a standardised way. It builds on international archival standards like ISAD(G), ISAAR(CPF), and others, and connects them with widely used formats such as EAD and EAC-CPF. This ensures that metadata can travel consistently across different archival systems, while keeping its authenticity, context, and structure intact.

The specification also explains how to place and map metadata within an information package. This allows archives not only to transfer digital objects together with their descriptions, but also to exchange descriptive information on its own—for example, when analogue collections are being shared. By harmonising structure and rules for descriptive metadata, the specification supports long-term preservation, smooth interoperability, and better access to archival information.

[6] CS Preservation Metadata

The E-ARK Common Specification for Preservation Metadata provides a standardised framework for describing and managing preservation metadata using the international PREMIS data model. Its purpose is to ensure the authenticity, integrity, and long-term usability of digital objects within Information Packages, supporting their transfer, storage, and exchange between systems. Built on the OAIS Reference Model, the specification sets out how preservation metadata should be structured and embedded, with a particular focus on XML encoding and conformance with PREMIS version 3.0.

The specification defines the use of key PREMIS entities—objects, events, agents, and rights—and explains how they should be applied in SIPs, AIPs, and DIPs. It emphasises interoperability, requiring PREMIS metadata to be referenced in METS files and recommending controlled vocabularies, persistent identifiers, and detailed rights and event descriptions. By establishing these common rules, the specification provides a robust foundation for institutions to implement consistent, interoperable preservation workflows across diverse archival contexts.

[7] SIARD

The SIARD specification defines the *Software Independent Archival of Relational Databases* format, designed to ensure the long-term preservation and accessibility of relational databases. Originally developed by the Swiss Federal Archives, and now maintained by the DILCIS Board, it relies on open international standards such as XML, SQL:2008, Unicode, URIs, and ZIP to guarantee interoperability and sustainability over time.

The format packages database structure, data, and metadata in a single archive file, ensuring that information remains authentic, verifiable, and reusable, even when the original database systems are obsolete.

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. It is assumed that a database in SIARD format is archived as part of such an information package together with other documents.

Version 2.2 extends earlier releases by adding support for files stored outside the database, in line with SQL/MED, and introduces scalability features for handling large objects.

[8] CITS SIARD

The CITS SIARD (Content Information Type Specification for Relational Databases using SIARD) defines how relational databases should be packaged within E-ARK Information Packages for long-term preservation. It builds on the SIARD format, ensuring that database content, structure, and large objects (LOBs) can be preserved and exchanged in a consistent and interoperable way. The specification clarifies boundaries between the SIARD format itself and the packaging requirements under E-ARK, explaining how to include SIARD files, external LOBs, and accompanying documentation in CSIP-compliant packages.

The document sets out detailed requirements for package and representation METS files**,** including mandatory attributes for identifying relational database content, supported SIARD versions, and rules for handling database dumps and external LOBs. It also highlights the importance of submission agreements in SIPs to capture expectations for database preservation, and it recommends including documentation such as data dictionaries, entity–relationship diagrams, and legal context information.

[9] CITS ERMS

The **CITS ERMS (Content Information Type Specification for Electronic Records Management Systems)** defines how records and their metadata should be packaged and exchanged for long-term preservation and interoperability. It builds on the Common Specification for Information Packages (CSIP) and OAIS principles, ensuring that information exported from ERMS platforms can be validated, preserved, and reused across different archives and systems. Two extraction approaches are supported: one based on preserving relational databases using SIARD, and another using XML to represent records and aggregations with explicit semantic metadata, enabling more flexible access, indexing, and cross-system integration.

The specification provides clear rules for how ERMS content and metadata are placed in a CSIP package, supported by XML schema. Metadata requirements are detailed extensively, covering control, descriptive, administrative, and provenance elements, while allowing local extensions through controlled vocabularies or external schemas such as METS and PREMIS. Overall, CITS ERMS enables archives, producers, and software developers to manage electronic records in a standardised, authentic, and interoperable manner for long-term digital preservation.

[10] CITS Geodata

The **E-ARK Content Information Type Specification (CITS) for Geospatial Data** provides detailed guidance on how to package, describe, and preserve digital geospatial information within the E-ARK framework. Building on the European **INSPIRE directive**, it ensures that spatial data—such as maps, geographic datasets, and spatial metadata—can be stored in **CSIP-compliant Information Packages**, making them interoperable across archival systems.

The specification defines how geospatial content and metadata must be structured and embedded, ensuring that data integrity, discoverability, and long-term accessibility are maintained. By aligning with both **OAIS principles** and established geospatial standards, it allows repositories and data producers to exchange, validate, and reuse geospatial information consistently. In practice, this CITS acts as a bridge between the technical world of geographic information systems and the archival requirements of long-term preservation, enabling sustainable management of complex spatial datasets.

[11] CITS eHealth1

There are two healthcare-related e-ARK content types. The eHealth1 specification is built upon work done by the Directorate for Health in Norway on the establishment of a standard for electronically extracting health records from healthcare provider Electronic Medical Records (EMR) systems and their submission to a central health archive. The use cases envisaged for the central health archive are:
a) to provide records to next of kin in compliance with open information regulation and
b) to harvest the vast amount of historical healthcare-related data for medical research.

The CITS takes input from the Norwegian specification on the structure of extractions from submitting systems and puts this into a framework that is compliant with the EARK common and package specifications.

The eHealth1 specification also recommends the use of international domain standards for descriptive metadata. The specification is accompanied by two METS profiles for the Root and Representations within the packages.

[12] CITS eHealth2

The **CITS eHealth2 specification** defines how exports from cancer registries should be packaged and archived using the E-ARK framework. Its purpose is to ensure that cancer registry data—covering cases, mortality, population data, and life tables—can be preserved, shared, and reused in a reliable, standardised way. The specification builds on the Common Specification (CSIP) and the E-ARK SIP, DIP, and AIP models, using XML schemas and Schematron rules to guarantee interoperability and authenticity across systems and institutions.

The document sets out requirements for structuring archival information packages. These packages include a **representation folder** with core data and supporting context, a **documentation folder** with agreements and export reports, and a **metadata folder** with descriptive and preservation metadata, often based on EAD and PREMIS. By doing so, it supports international data aggregation efforts—such as those led by ENCR, JRC, and the CONCORD programme—while ensuring long-term usability. Ultimately, the CITS eHealth2 specification enables cancer registry exports to be integrated into broader health data infrastructures, while preserving their scientific and policy value

[13] CITS 3DPM

The Content Information Type Specification for 3D Product Model data aims to define the necessary elements required to preserve the accessibility and authenticity of 3D Product Model data over time and across changing technical environments. The specification builds on the international standard for long-term archiving of Product Model data (LOTAR) and facilitates conformance to the standard within an E-ARK packaging framework. In order to achieve this the specification elevates the level (and adjusts the cardinality) of some of the requirements set out in the Common Specification (CSIP) and package specifications (namely SIP and AIP) and adds new requirements for the package structure, descriptive metadata, preservation metadata and accompanying METS files. It also introduces new requirements for authentication. The specification sets out general principles that underpin the specific requirements and further context for the requirements and principles can be found in the accompanying guideline to this document.

Use of 3D data is widespread across many domains, with a plethora of applications and data formats. This 3D Product Model content specification limits its scope to the area of 3D digital product data such as computer aided design (CAD) or product data model (PDM) data where there is a current international standard for the long term archiving of this class of data in the LOTAR “Long Term Archiving and Retrieval of digital technical product information”.

[14] New CITS

The total number of Content Information Type Specifications is unlimited, and the long-term commitment of the DILCIS Board is to keep the overall environment open and inclusive. As such, interested bodies are welcome to develop their own Content Information Type Specifications.

[15] New CITS

It is hoped that many Content Information Type Specifications will be developed with the wider community to create new specifications for domains of interest to them. The DILCIS Board aims to work with the community following these principles.

[16] DILCIS Board

The E-ARK specifications are maintained and developed by the DILCIS Board (Digital Information LifeCycle Interoperability Standards Board). The DILCIS Board is an international group of experts committed to maintain and sustain a set of **interoperability specifications** which allow for the **transfer, long-term preservation, and reuse of digital information** regardless of the origin or type of the information. The Board is responsible for ensuring that the specifications remain up to date, practical, and aligned with the needs of the digital preservation community. By coordinating feedback from practitioners, archives, and system providers, the DILCIS Board supports the long-term sustainability of the specifications and promotes their adoption across Europe and beyond.

[17] DILCIS Board – Content Types

Each content type has its own page on the DILCIS Board website. Please note that most content type specifications are supported by an accompanying guideline document.

The Guidelines for the content type specifications support the use of the E-ARK specifications by offering additional explanations and practical insights beyond what is contained in the formal standards. Their purpose is to make the specifications easier to apply in practice, providing clarifications, examples, and references to related standards and resources.

[18] Thank you

Thank you for watching this overview! To explore the details further, we encourage you to explore the E-ARK content type specification documentation in detail.