Common Specification and the Information Package 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 2 takes a closer look at the E-ARK Information Package Specifications. In this video, you will learn about the requirements of the Common Specification and explore the specific features of the SIP, AIP, and DIP specifications. By the end of this lesson, you'll understand both what the packages share in common and what makes each type unique.

[3] OAIS Information Packages

The Open Archival Information System, or OAIS, is the internationally recognised framework for long-term digital preservation. It doesn't describe a single technology, but rather a reference model – a kind of blueprint that defines the functions and responsibilities of a trustworthy digital archive.

[4] OAIS Information Packages

At the centre of the OAIS model are the information packages. These come in three main forms: the **Submission Information Package**, or SIP, which is created and delivered to the archive by the producer; the **Archival Information Package**, or AIP, which is the internal version preserved and managed within the archive; and the **Dissemination Information Package**, or DIP, which is what the archive provides to users, the consumers. Together with OAIS processes, these packages ensure that digital materials are not only stored, but remain understandable and usable over time, no matter how technologies evolve.

[5] E-ARK Information Packages

Although the OAIS model is widely used, it's important to remember that it does not prescribe the exact internal structure of information packages. In other words, OAIS tells us that there should be SIPs, AIPs, and DIPs, and what they need to achieve, but it leaves open how they are actually organised, documented, or exchanged in practice. This flexibility is useful, but it also means that without further guidance, archives and institutions might each implement their packages in different ways, making interoperability difficult.

To address this, more detailed and structured approaches have been developed. One of the most significant is the **E-ARK specifications**, produced under the European Commission's **eArchiving Initiative**. These specifications build on the OAIS model and define common, standardised formats for information packages. By following them, organisations can ensure that their archival packages are consistent, interoperable, and easier to share across systems and countries. In this way, E-ARK provides the practical detail that complements the OAIS reference model.

[6] E-ARK Common Specification

As we have seen in Lesson 1, the E-ARK specifications cover all three types of information packages defined in the OAIS model: Submission Information Packages (SIPs), Archival Information Packages (AIPs), and Dissemination Information Packages (DIPs). However, much of the structure and content of these packages is the same. To avoid

unnecessary repetition, E-ARK has created a Common Specification, which defines all the shared elements. The individual SIP, AIP, and DIP specifications then build on this common foundation, describing only the differences and package-specific requirements. This approach ensures consistency, reduces complexity, and makes the specifications easier to use in practice.

[7] Common Specification for Information Packages (CSIP)

The latest version of the Common Specification is available on the DILCIS Board website, where you can view it online in HTML format or download it as a PDF.

[8] Common Specification for Information Packages (CSIP)

The Common Specification for Information Packages, or CSIP, defines the base profile and shows how METS is used to package OAIS-compliant information packages. This ensures interoperability between repositories and supports the long-term preservation of digital content. The CSIP serves as a foundation that can be used on its own, or extended through the E-ARK SIP, AIP, and DIP specifications, or even with custom extensions tailored to local needs.

[9] Common Specification for Information Packages (CSIP)

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. The main purposes of CSIP are to:

- Establish a common understanding of the requirements which need to be met 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.
- And 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.

[10] CSIP – Document Structure

The CSIP specification is divided into two main parts. The first explains the purpose of a common specification and sets out the core principles needed for interoperability, such as structure and metadata use. The second part focuses on implementation, describing the required folder and file structure, the detailed metadata rules, and possible extensions. Appendices then provide practical examples of conformant packages to support real-world use.

[11] Information flow between live and archival systems

Most systems that create information are not designed for long-term preservation. This means data often has to be moved between business, records, and archival systems. To do this effectively—while protecting authenticity and integrity, and keeping the information discoverable—we need common principles for exchange. The CSIP provides these principles through a shared set of interoperability specifications.

[12] Standards behind the specifications

The E-ARK specifications are grounded in internationally recognised standards that ensure consistency, interoperability, and long-term preservation. At their core is the OAIS Reference Model, ISO 14721, which defines the fundamental concepts of digital preservation. Alongside this, XML provides a structured way to encode data, while METS is used to organise and package digital objects. PREMIS adds a framework for recording preservation metadata, and archival description standards such as EAD, EAC-CPF, and Records in Context support detailed documentation of content and context.

On the right, you can see how these standards move through different levels of adoption. They begin as international standards, like METS or EAD. These are then adopted at the European level in the form of Common Specifications. From there, individual countries may adapt them into national standards, such as Sweden’s FGS, and

finally, they are implemented locally by organisations for specific use cases. This layered approach ensures both consistency across borders and flexibility to meet local needs.

[13] CSIP – Principles

This section introduces a set of high-level principles that guide the Common Specification. Their main purpose is to ensure interoperability, so that information packages can be exchanged, identified, validated, and reused across different systems. Each principle is numbered, comes with a requirement level such as ‘must’ or ‘should’, and is supported by a short explanation of its purpose and background.

[14] CSIP – Principles

The General Principles of the Common Specification set the foundation for how information packages should be structured and managed. They emphasise consistency, interoperability, and clarity, ensuring that packages can be created, exchanged, and understood across different systems and organisations. These principles also highlight the importance of using common metadata, maintaining authenticity, and supporting both immediate access and long-term preservation.

[15] CSIP – Principles

The remaining principles are divided into three main sections: Identification of the Information Package, Structure of the Information Package, and Information Package Metadata.

The principles of identification ensure that every information package can be uniquely and persistently referenced. They require that packages include identifiers within their metadata, and that these identifiers remain stable over time, even if the package is moved or transformed. This guarantees authenticity, traceability, and reliable exchange between systems.

The structural principles define how the package must be organised. They require a clear and consistent folder hierarchy, the use of standardised file naming, and separation of content from metadata. This common structure makes it easier to validate, exchange, and process information packages across different organisations and tools.

The metadata principles focus on ensuring that information packages carry enough descriptive, structural, and technical information to be both understandable and preservable. They require the use of XML-based standards, mandate that metadata be complete and consistent, and recommend using established schemas such as METS and PREMIS. This guarantees that the content can be properly managed, discovered, and reused over time.

[16] CSIP – Folder Structure

Section 4 of the CSIP specification defines the required structure of an information package. It sets out how folders and files must be organised, beginning with a root folder and including mandatory elements such as the `METS.xml` file. The section also describes where content data should be stored, how metadata is arranged in dedicated subfolders, and how optional elements like schemas and documentation can be added. Together, these rules ensure that every CSIP package follows a predictable structure, making it easier to validate, exchange, and preserve across different systems.

[17] CSIP – Use of Metadata

Section 5 of the CSIP specification focuses on how metadata is used within information packages. It defines the different categories of metadata—descriptive, structural, technical, and preservation—and explains their roles in making content understandable, manageable, and reusable over time. The section also sets out the requirement to use standardised, XML-based formats such as METS, PREMIS, and Dublin Core, ensuring interoperability across systems. By following these rules, information packages carry the essential context and documentation needed for long-term preservation and reliable access.

Section 5.1 sets out the general rules for metadata in CSIP information packages. It requires that all metadata be expressed in XML and conform to established schemas to guarantee consistency and machine readability. Metadata must be complete, accurate, and linked to the objects it describes, ensuring that information can be properly validated, discovered, and preserved. These requirements form the baseline for interoperability, making sure that packages can be understood and processed across different systems and over long periods of time.

[18] CSIP – Use of Metadata – METS

The METS requirements section explains how the Metadata Encoding and Transmission Standard, or METS, is used as the backbone of a CSIP information package. METS provides a single, central XML document that ties together all the descriptive, structural, administrative, and preservation metadata with the digital objects they describe. The specification outlines which METS sections must be included, such as the header, file section, structural map, and metadata references, and how they should be applied. It also defines mandatory, recommended, and optional elements to ensure that every package is both interoperable and adaptable. By following these requirements, METS acts as the glue that holds the package together, supporting both exchange between systems and long-term digital preservation.

[19] CSIP – Use of Metadata – METS

The Common Specification also provides complete examples of METS files in Appendix A, showing how the specification can be applied in practice.

[20] CSIP – Use of Metadata – PREMIS

The CSIP recommends using the PREMIS standard to record preservation and technical metadata for digital objects. Its application is further detailed in the E-ARK Common Specification for Preservation Metadata using PREMIS.

[21] CSIP – Descriptive Metadata

Descriptive metadata is used to capture the intellectual content of archival holdings, making it easier to find, identify, and understand information packages. In CSIP, all descriptive metadata must be placed in the ‘metadata’ folder, and it is recommended to create a dedicated subfolder called ‘descriptive’. The specification does not prescribe a single schema, so implementers are free to use standards such as EAD, Dublin Core, or others, following the guidance in the E-ARK Archival Information specification. Importantly, descriptive metadata must always be referenced from the METS file, rather than embedded directly within it.

[22] CSIP – E-ARK SIP/AIP/DIP

Up to this point, we have explored the principles and requirements of the Common Specification. The CSIP provides a shared foundation for packaging digital information, ensuring consistency and interoperability. Building on this base, the three E-ARK package types—the SIP, AIP, and DIP—introduce their own modifications and extensions to meet specific needs. In the next part of this tutorial, we will look at each of these packages in turn, focusing on how they differ from, and expand upon, the Common Specification.

[23] CSIP – E-ARK SIP/AIP/DIP

In general, the E-ARK SIP and E-ARK DIP specifications reuse and apply fully all the requirements set in the Common Specification. However, they also extend it with aspects relevant only for the respective processes. For example, the E-ARK SIP specification extends the CSIP with further requirements about recording relevant information on a submission agreement and the actors of the submission process. On the other hand, the E-ARK DIP provides possibilities for describing complex access environments needed to reuse the content of a DIP.

However, it is important to note that the E-ARK AIP format does not extend the CSIP in the same way the E-ARK SIP and E-ARK DIP formats do. The reason for this is that while the SIP and the DIP are like “snapshots” in time, the AIP needs to deal with an “evolving object” which is constantly updated by preservation actions undertaken in the

course of the object's lifecycle. As such, while the E-ARK AIP specification does implement all of the core metadata requirements defined in the Common Specification and extends these (for example it describes a means to record preservation actions about the IP), it does also extend the default structure of the CSIP. Essentially the AIP introduces a more complex structure which allows at the same time to securely hold an E-ARK SIP (which itself follows in full the CSIP) and at the same time add and modify additional representations over a series of preservation actions.

[24] E-ARK SIP

The E-ARK SIP specification focuses on defining a general structure for Information Packages delivered by the Producer to the OAIS for ingestion or updating Archival Information Packages (AIPs). It extends the CSIP with metadata necessary during the transfer moment.

The SIP specification includes a list of semantic elements that should be present in a Submission Agreement, covering project information, change management, producer/archive/designated community details, SIP content/metadata, submission session information, ingest procedures, and submission risks. However, it does not mandate a specific format or the use of these elements.

Regarding the Package Structure the SIP specification accounts for packages with zero representations, meaning they contain only metadata. This is a special type of package designed to deliver metadata updates to previously ingested packages. The structure also includes additional folders for documentation (e.g., data dictionaries) and schemas (for XML files), which can exist at the root or representation level, crucial for accurate interpretation and validation.

And there are several METS file extensions in the SIP specification related to the Submission Agreement and the Agent.

[25] E-ARK DIP

The E-ARK DIP specification defines requirements for Information Packages derived from one or more AIPs and sent to a Consumer in response to a request. The core definition of an E-ARK DIP is that it is a CSIP ready for processing by its designated Access Software.

Purpose-driven Content
An E-ARK DIP will always consist of some files and folders that are optional in the CSIP minimum structure, as there must be data to disseminate. While an AIP can fully serve as a DIP, conversion is often needed to produce new, more user-friendly representation formats, and to update or add metadata specific to access requirements.

METS file
Here as well, the specification defines several modifications to the METS file.

PREMIS for Access Information (Access Software)
The DIP specifies using PREMIS to record representation information to facilitate access, particularly focusing on how to describe and link access software.

Access Restrictions in Descriptive Metadata
The DIP specification recommends that Access Rights Information concerning the end-user MUST be available in descriptive metadata (e.g., EAD) and not in PREMIS.

[26] E-ARK AIP

The E-ARK AIP specification defines requirements for building AIPs that store information for the long term in an archive. Unlike SIP and DIP, which are "snapshots" in time, the AIP addresses an "evolving object" that is continually updated by preservation actions throughout its lifecycle.

Provenance and Life-cycle
The AIP places strong emphasis on recording provenance and life-cycle information. It tracks sequential SIP submissions, defines AIP versions for migrations or metadata updates, and introduces generations to represent physical manifestations of logical AIPs. This ensures the evolving nature of preserved information is fully documented.

METS Elements
In the AIP, some METS attributes are stricter than in the Common Specification. For example, the OBJID must remain unchanged throughout the AIP's life-cycle, ensuring persistent identification, and the profile attribute must point to the dedicated AIP profile. Other sections, like the header and descriptive metadata, also include AIP-specific rules.

Physical Container Packaging
The AIP defines clear rules for packaging content. Filenames should follow consistent naming derived from the AIP identifier, all contents must be extracted into a single folder, and if TAR files are used, they should be aggregated without compression. For large or versioned content, the AIP also supports practices like the Oxford Common File Layout.

PREMIS Object Metadata
The use of PREMIS metadata is more explicit in AIPs. It can capture format and storage information, define relationships between AIPs, and document events and agents involved in preservation actions. This ensures both technical and preservation metadata are well-structured and interoperable.

[27] CSIP – E-ARK SIP/AIP/DIP

In summary, while all three packages adhere to the foundational CSIP, the SIP introduces specific metadata for the submission process (agents, submission agreements), the AIP details how to manage the evolving nature of preserved information through versioning and preservation metadata (provenance, structural division), and the DIP focuses on making content accessible by linking representations to necessary access software and detailing access restrictions.

[28] Guidelines

The Guideline for Information Packages supports the use of the E-ARK specifications by offering additional explanations and practical insights beyond what is contained in the formal standards. It covers the Common Specification for Information Packages (CSIP) along with the specific SIP, AIP, and DIP profiles, as well as related specifications for archival information and preservation metadata. Its purpose is to make the specifications easier to apply in practice, providing clarifications, examples, and references to related standards and resources.

The document introduces the broader context of digital preservation, explains how the specifications align with the OAIS model, and highlights the role of metadata standards such as METS and PREMIS. It also outlines how specifications can be adapted to local requirements while ensuring interoperability. By bringing together explanations of structure, metadata, and implementation considerations, the guideline acts as a bridge between high-level standards and practical application.

[29] Thank you

Thank you for watching. To learn more, we encourage you to explore the E-ARK specification documentation in detail.