



CEF Digital
Connecting Europe

Specifications Event

26th October 2021



Instructions for this live event:



This is an interactive event.
Please ask your questions
in the chat!
All answers in the end!



Please note that
this
event will be
recorded!

Why eArchiving specifications?

CSIP and related specifications, guidelines and procedures

CITS eHealth1

CITS eHealth2

CITS ERMS

CITS Geospatial

CITS SIARD and SIARD standard

Validation

Questions

Final words

S
P
E
A
K
E
R
S



Fulgencio Sanmartín
Project Officer
European Commission DG CNECT



Anja Paulič
Senior Advisor
Arhiv Republike Slovenije



Radoslav Jakub
Project Officer
HaDEA



Gregor Završnik
Geospatial preservation consultant, manager
Geoarh



Karin Bredenberg
Metadata Strategist
Sydarkivera and DLM Forum



Anders Bo Nielsen
Senior Adviser
Danish National Archives



Stephen Mackey
Product Manager
PiqI



Carl Wilson
Technical Lead
Open Preservation Foundation



https://www.europeana.eu/en/item/9200369/webclient_DeliveryManager_pid_6328451_custom_att_2_simple_viewer

Why eArchiving specifications?



Fulgencio Sanmartín, European Commission DG CNECT

- In digital preservation, doing nothing is not an option
- Actually, doing nothing (or little) *will* cost you money later
- More time, more money
more collections to digitise, more metadata to add/curate, more old formats to migrate, technologies becoming obsolete...
- Often, this cannot be avoided, but do take it into account
- Risk mitigation: stick to known, **interoperable European standard specifications**

Actrices con fondo de Deuda pública - Biblioteca Digital Memoria de Madrid, Spain - CC BY-NC.
https://www.europeana.eu/en/item/2022711/urn_repos_ist_util_pt_MH_27894



How to use eArchiving?

Build, buy or reuse the eArchiving specifications on your own, co-develop the solution or partner with other parties.

Co-develop and partner with other parties



Build
The solution from
scratch, and test it



Buy
A conformant solution
from the market

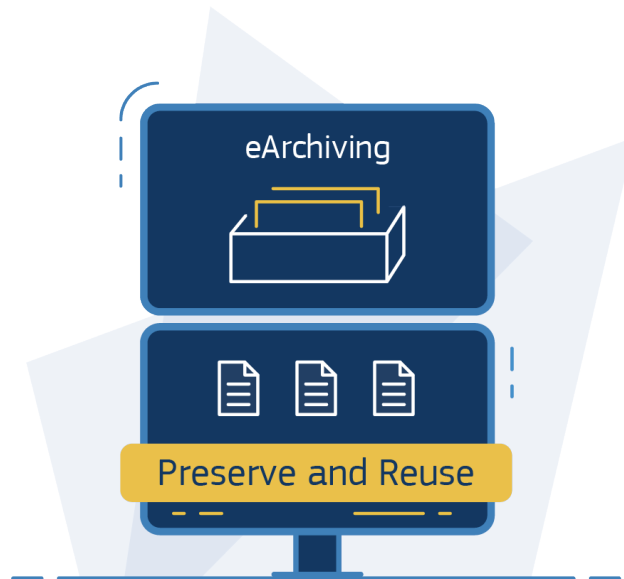


Reuse
Sample, open source,
conformant software

European Standards

What is eArchiving?

<<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving>>



eArchiving

Facilitates the preservation, migration, reuse and trust of your data.



Standard specifications

- **E-ARK** Common Specification for Information Packages:
 - Submission (**METS**)
 - Archive (with **PREMIS**)
 - **Interoperable** dissemination & re-use
- Geodata (**INSPIRE** Directive)
- **eHealth**



Sample software

- **Relational Database** archiving and reuse tools (**SIARD**)
- Pre-ingest and **SIP** creation tools
- **Long-term Repository** solutions
- Online **validation** tool



Support

- **Service Desk**
- Training and user guides
- **Knowledge Center**
- E-ARK General model



Onboarding & follow up

- Awareness raising
- **Conformance, compliance**
- Maturity Assessment Service
- **Webinars** and workshops

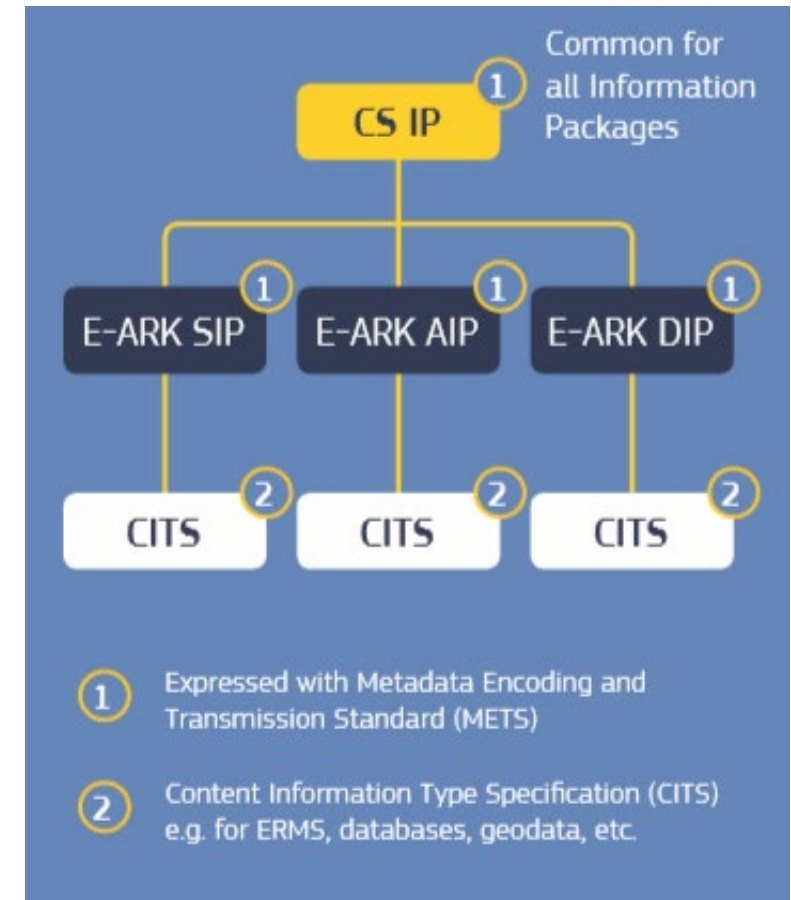
eArchiving specifications

Common Specification for Information Packages (CSIP)

1. Submission Information Packages (**SIP**), based on
 - Metadata Encoding Transmission Standards (**METS**)
2. Archival Information Packages (**AIP**), including
 - Preservation Metadata standard (**PREMIS**)
3. Dissemination Information Package (**DIP**)
 - **Interoperable** across borders and institutions

Content Information Type Specifications (CITS)

- **Databases** (**SIARD** standard)
- **Geographical data** (including **INSPIRE** Directive)
- Electronic Record Management Systems (**ERMS**)
- **eHealth 1** and **eHealth 2**



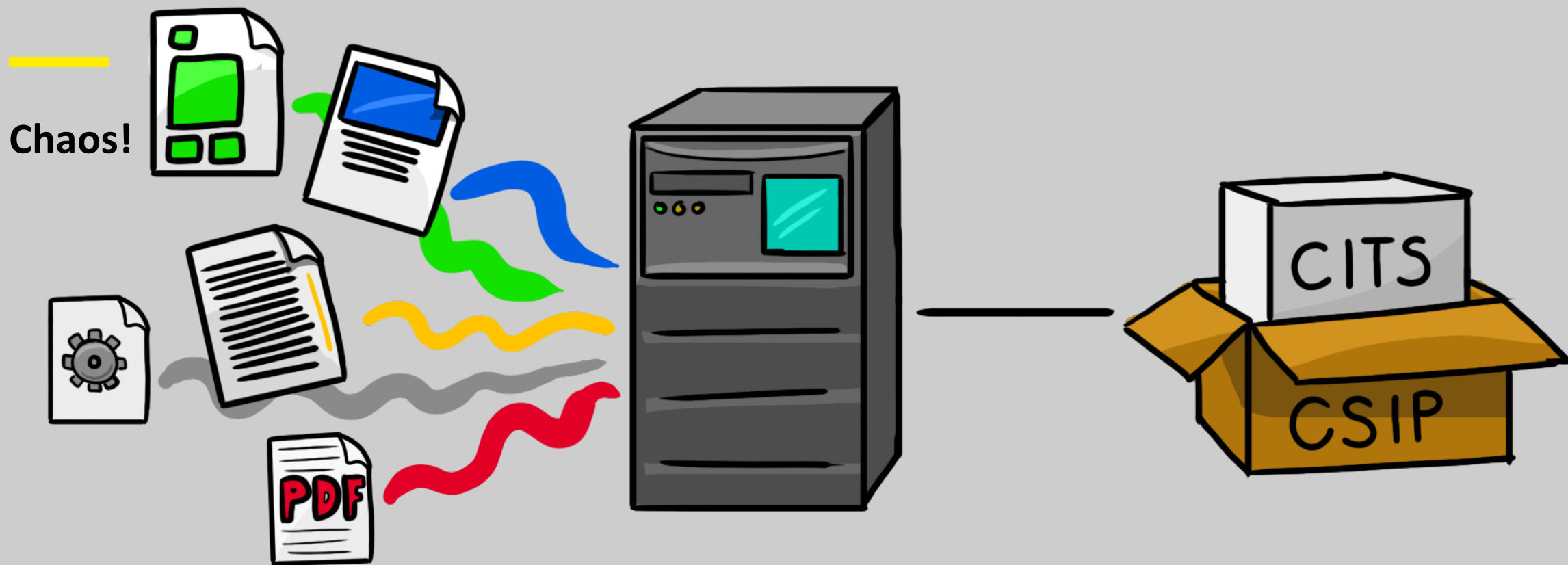


fotografi - 1911 - Kulturmagasinet, Helsingborgs museer, Sweden - Public Domain.
https://www.europeana.eu/es/item/916114/HeM_fotografi_257824

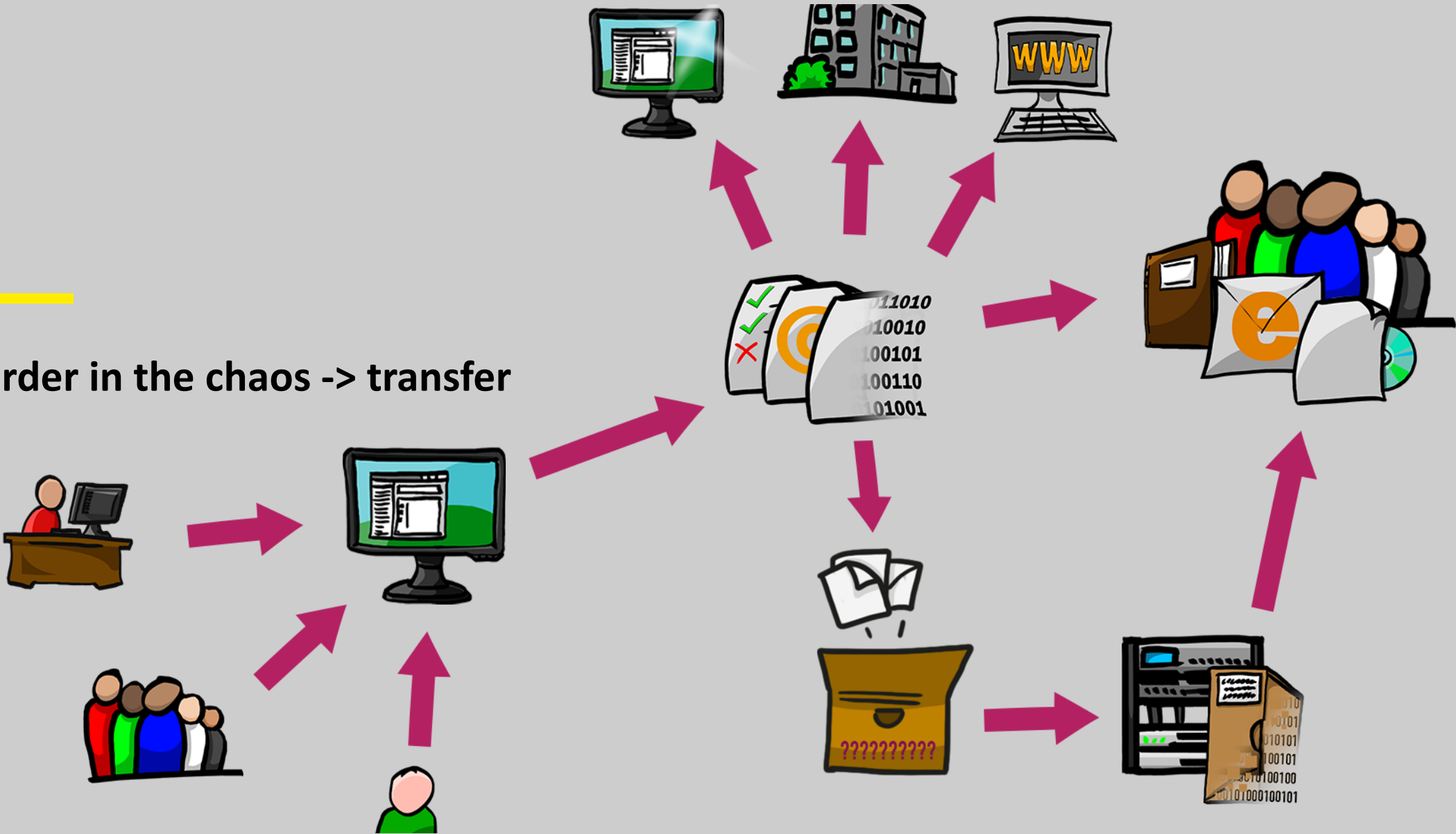
CSIP and related specifications, guidelines and procedures



Karin Bredenberg, Sydarkivera

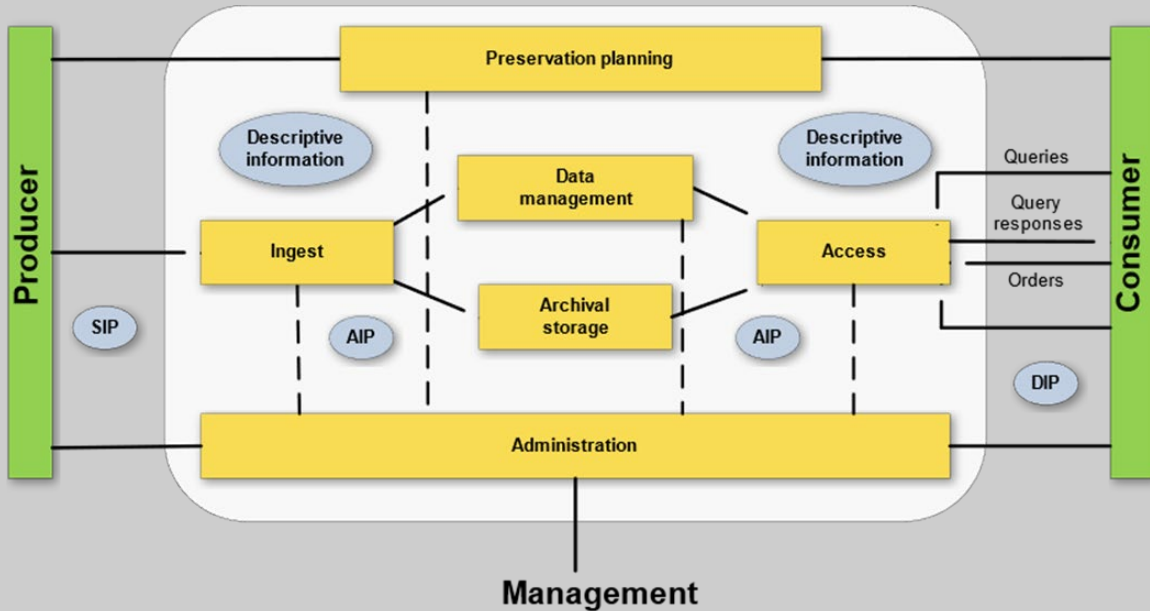
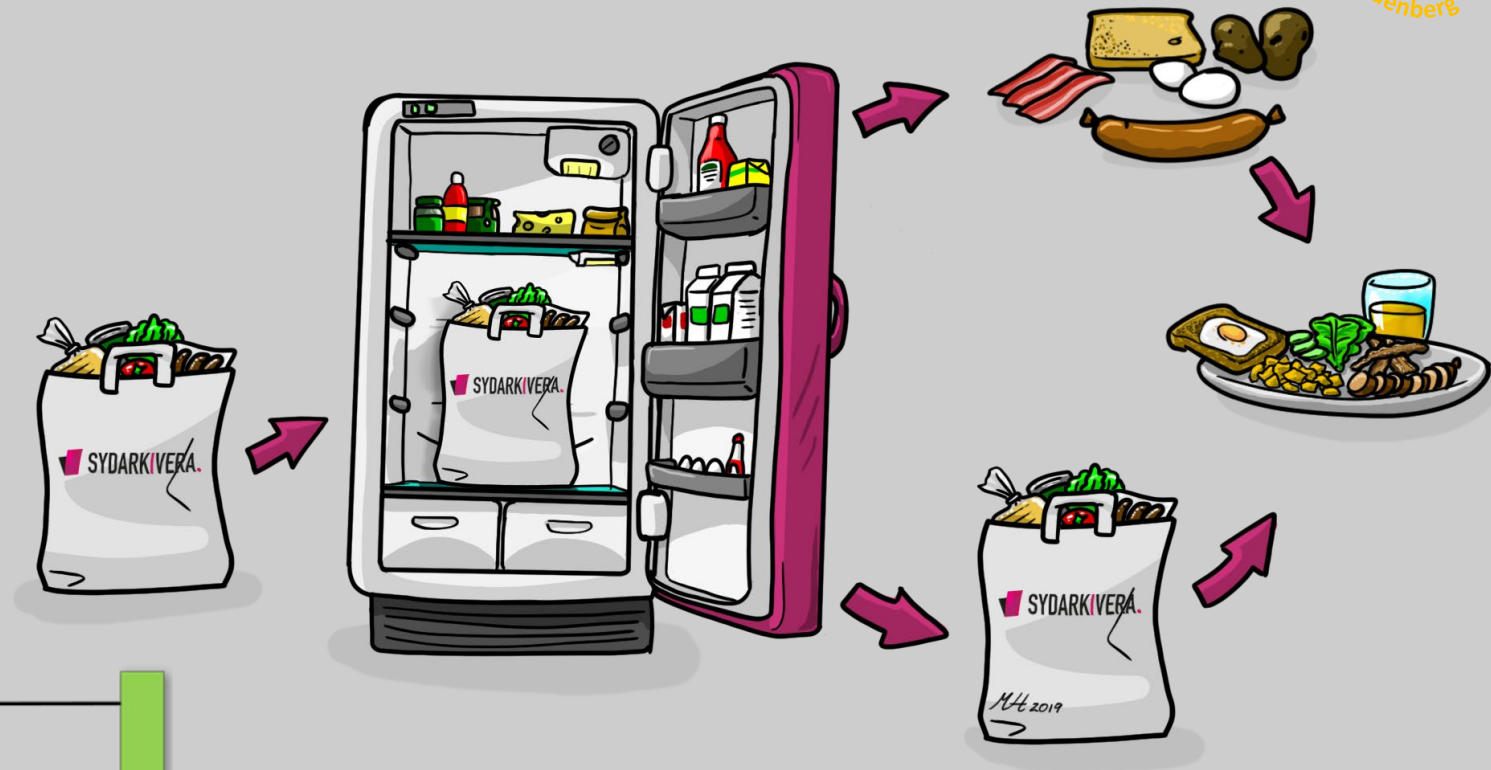


Order in the chaos -> transfer



The OAIS Reference Model

<https://www.iso.org/standard/57284.html>



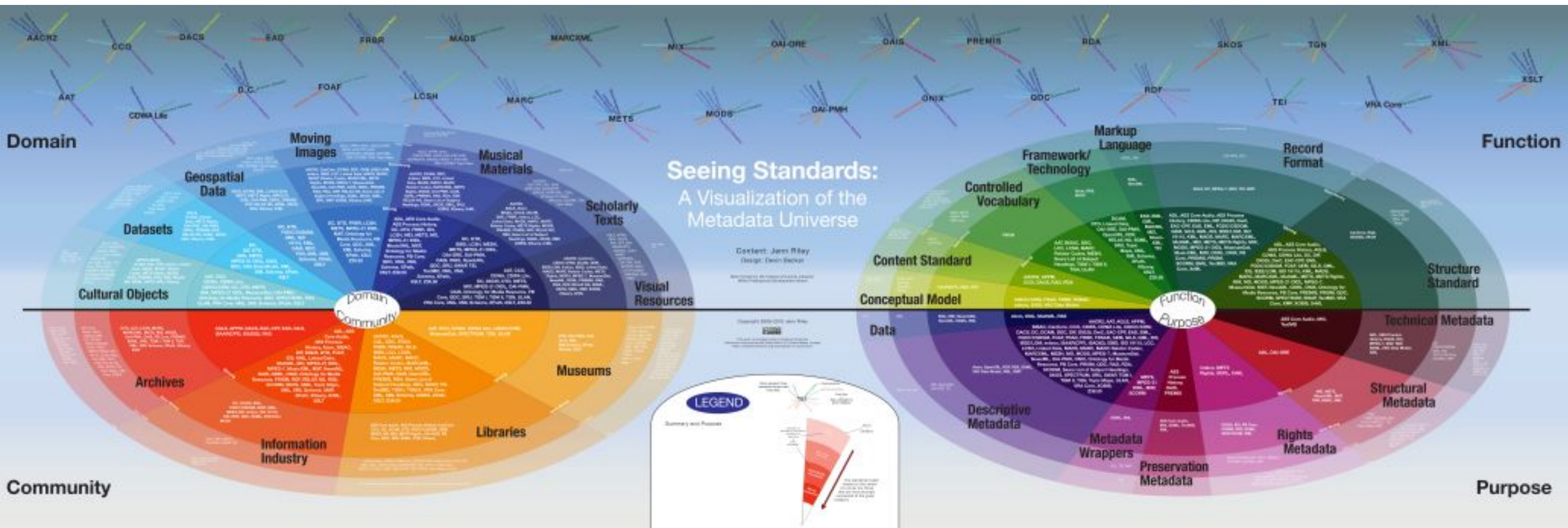
The specifications

The basics for both CS and CITS; using standards

4

There are plenty of standards to use for data transfer and conformance

<http://jennriley.com/metadatamap/>



Currently XML is used for the descriptions

We are monitoring what is happening and will evaluate alternatives as they emerge but it takes time, stability!



Extensible Markup Language (XML)

Standards currently used

For the package we use the Metadata Encoding & Transmission Standard, METS, <http://www.loc.gov/standards/mets/>



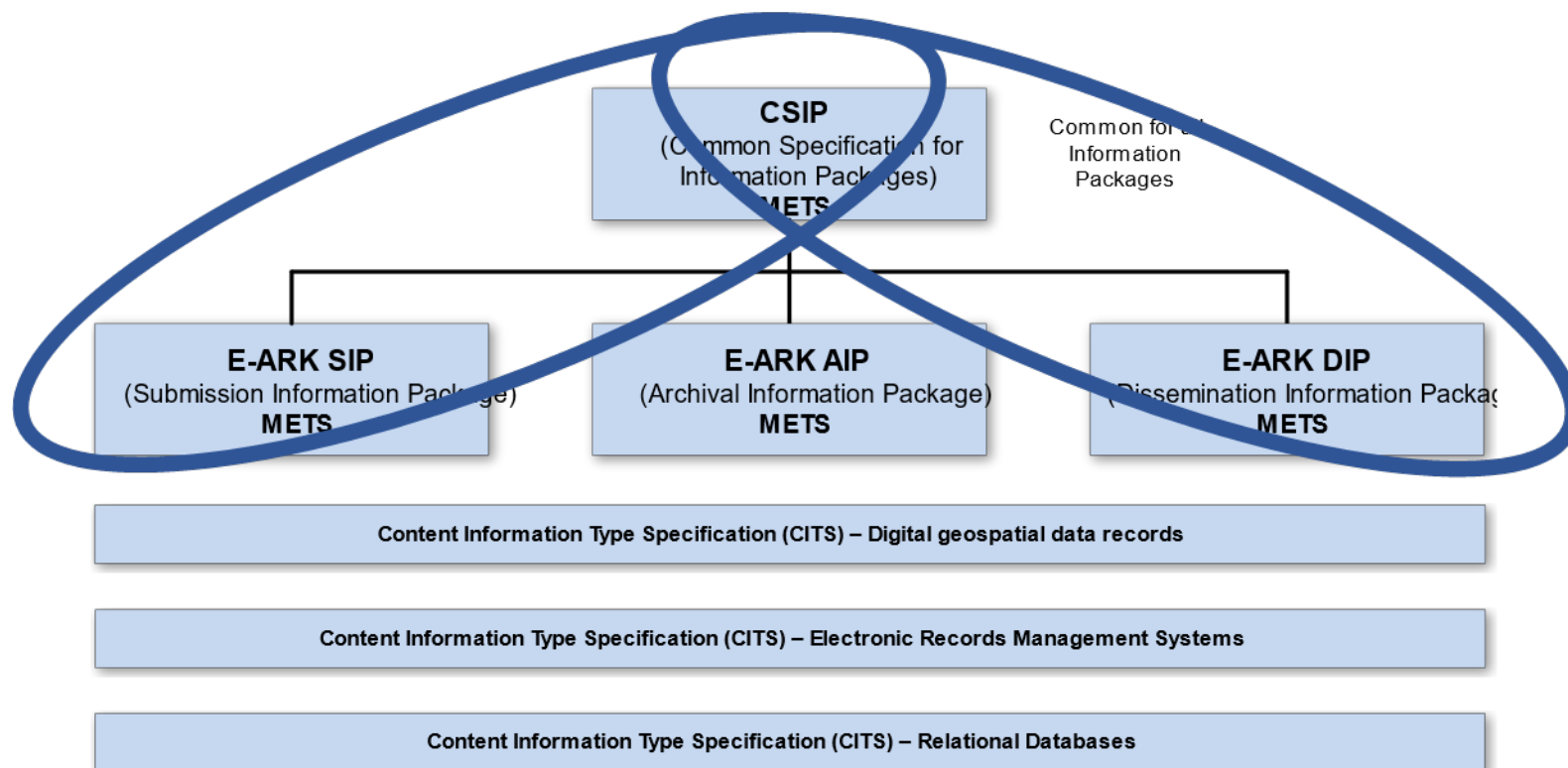
The number of standards used increases with each additional CITS



**There are two different types of specifications;
the Information Package Specifications (CS) and
the Content Information Type Specifications (CITS)**

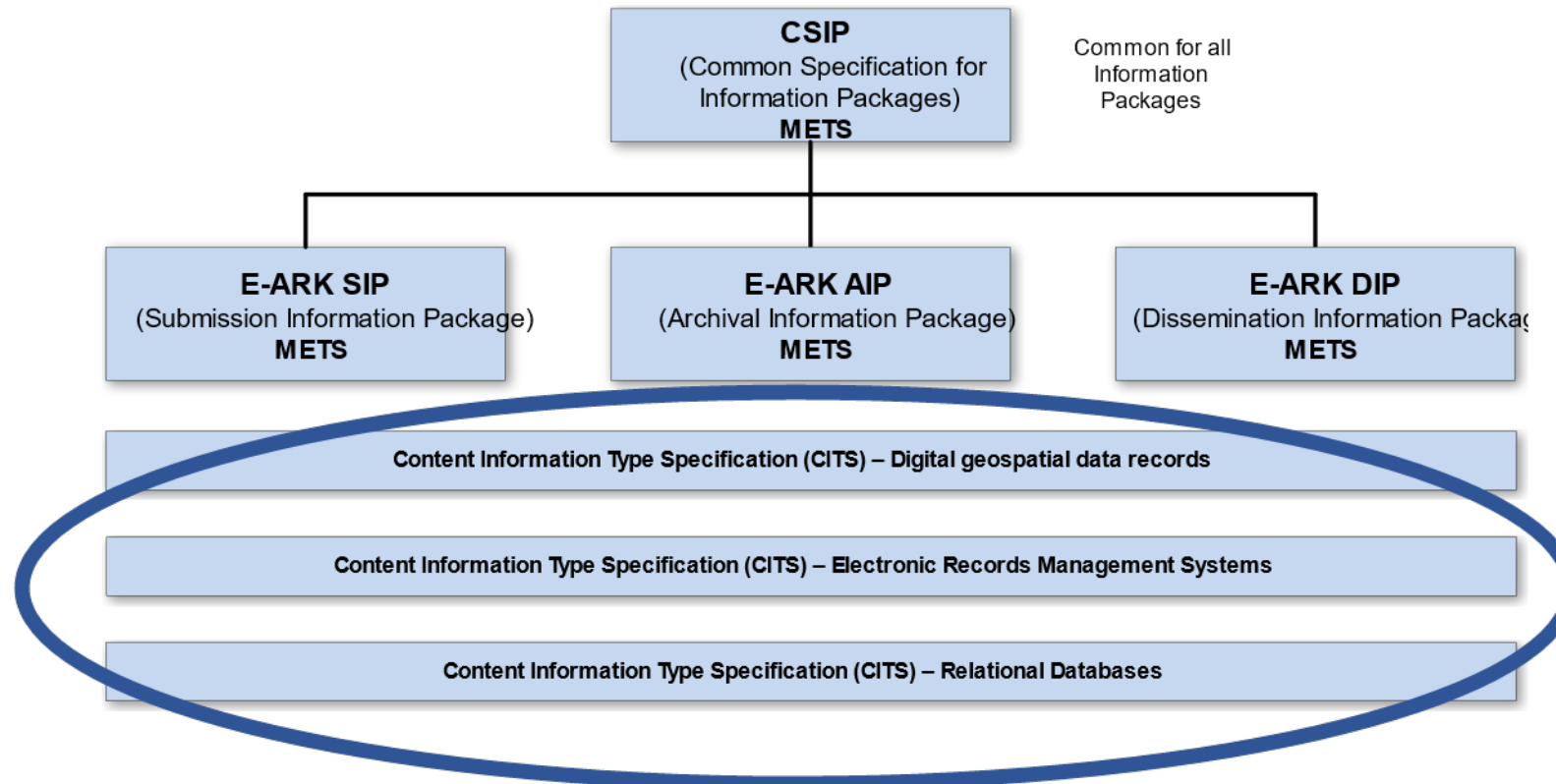
CSIP, SIP, AIP and DIP

The different Information Packages in the OAIS Reference Model



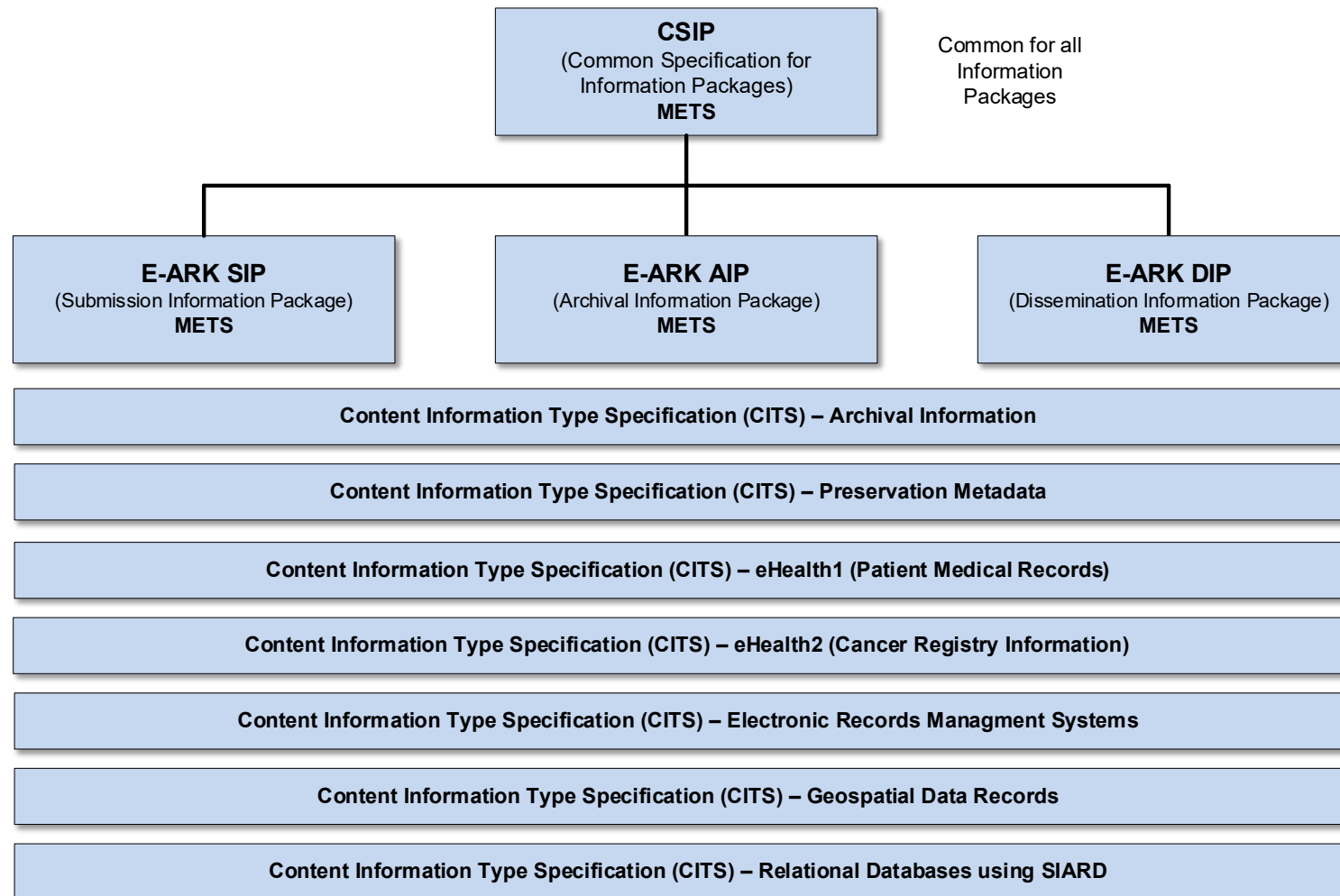
The Content Information Type Specifications (CITS)

The data/information/content in the package



The Content Information Type Specifications (CITS)

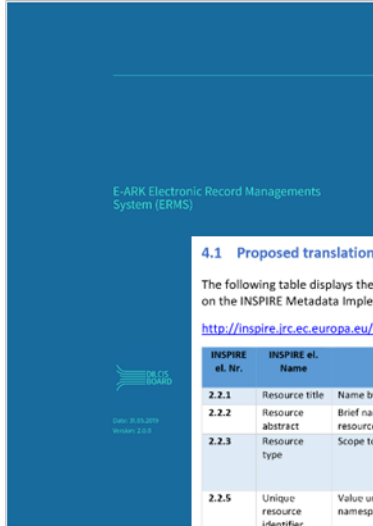
The ongoing work extends the number of CITS



Reading a specification, a text with rules and
Implementing a specification, a text with rules

Document and files for a specification

The text document and files providing the XML structure and validation rules



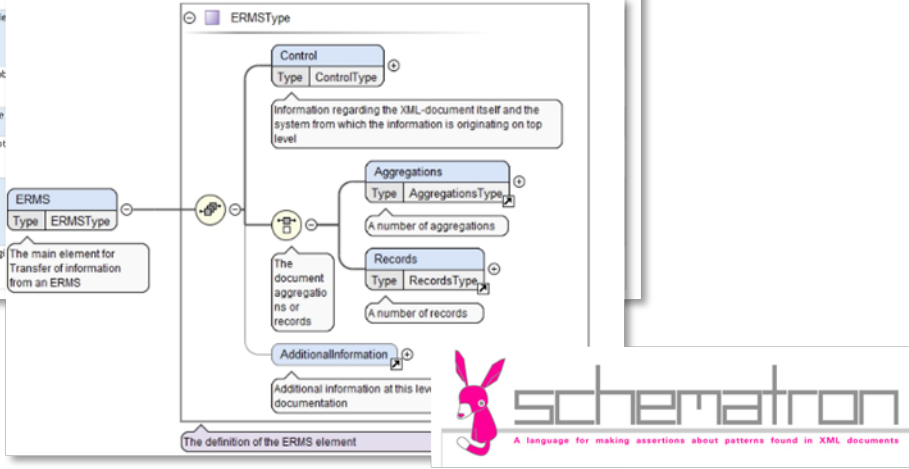
Specification

4.1 Proposed translation schema for the INSPIRE metadata descriptions for geospatial resources in ISAD(G)

The following table displays the identified counterparts of the required INSPIRE metadata elements used in the ISAD(G) structure. Initial elements are based on the INSPIRE Metadata Implementing Rules.: Technical Guidelines, based on EN ISO 19115 version 1.3. and INSPIRE Metadata Implementation Rules at: http://inspire.jrc.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf.

INSPIRE el. Nr.	INSPIRE el. Name	Explanation	Metadata data type	Proposed Cardinality	ISDG code	Comments
2.2.1	Resource title	Name by which the cited resource is known	text	1..1	3.1.2 Title	
2.2.2	Resource abstract	Brief narrative summary of the content of the resource(s)	text	0..1	3.3.1 Scope and content	
2.2.3	Resource type	Scope to which metadata apply				
2.2.5	Unique resource identifier	Value uniquely identifying an object namespace				
2.2.6	Coupled resource	Provides information about the service operates on				
2.2.7	Resource language	Language(s) used within the dataset				
2.3.1	Topic category (INSPIRE specific)	Main theme(s) of the dataset				
2.3.2	Spatial data service type	A service type name from a registry				

Requirements



XML-Schema/Schemas

Expression of requirements not possible to make in the XML-schema



Guideline



Example

The key terms when reading a specification

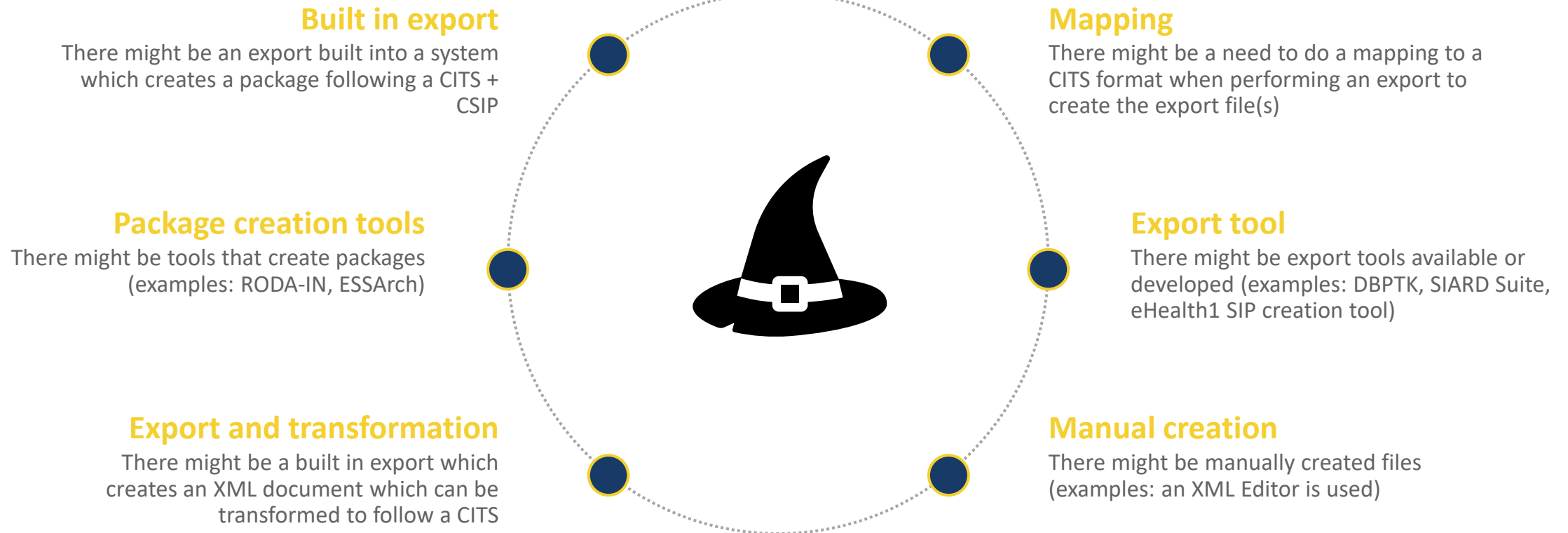
Understanding occurrence and obligation

ID	Name, Location & Description	Card & Level
CSIP1	<p>Package Identifier</p> <p><code>mets/@OBJID</code></p> <p>The <code>mets/@OBJID</code> attribute is mandatory, its value is a string identifier for the METS document. For the package METS document, this should be the name/ID of the package, i.e. the name of the package root folder. For a representation level METS document this value records the name/ID of the representation, i.e. the name of the top-level representation folder.</p>	1..1 MUST
Content Category		

- Card = Cardinality answers:
 - How many times can an element or attribute occur?
 - Is the element or attribute mandatory or optional?
- Level answers:
 - If the element or attribute is not mandatory what are the recommendations?

Creation of files following a specification

There are many ways to get the resulting content and put it into a package



The Information Package CS

The package principles and requirements

The principles of a package

What makes a package a package and how do we identify it and its content?

3.1.6. Principle 1.6:

3.1.7. Principle 1.7:

3.2. Identification of the Information Package

3.2.1. Principle 2.1:

3.2.2. Principle 2.2:

3.2.3. Principle 2.3:

3.2.4. Principle 2.4:

3.2.5. Principle 2.5:

3.3. Structure of the Information Package

3.3.1. Principle 3.1:

3.3.2. Principle 3.2:

3.2. Identification of the Information Package

3.2.1. Principle 2.1:

*The Information Package OAIS type (SIP, AIP or DIP) **MUST** be clearly indicated.*

One of the first tasks in analysing any Information Package is to identify its current status in the overall archival process. Therefore, any Information Package must explicitly and uniformly identify itself as a SIP, AIP or DIP.

3.2.2. Principle 2.2:

*Any Information Package **MUST** clearly identify the Content Information Type(s) of its data and metadata.*

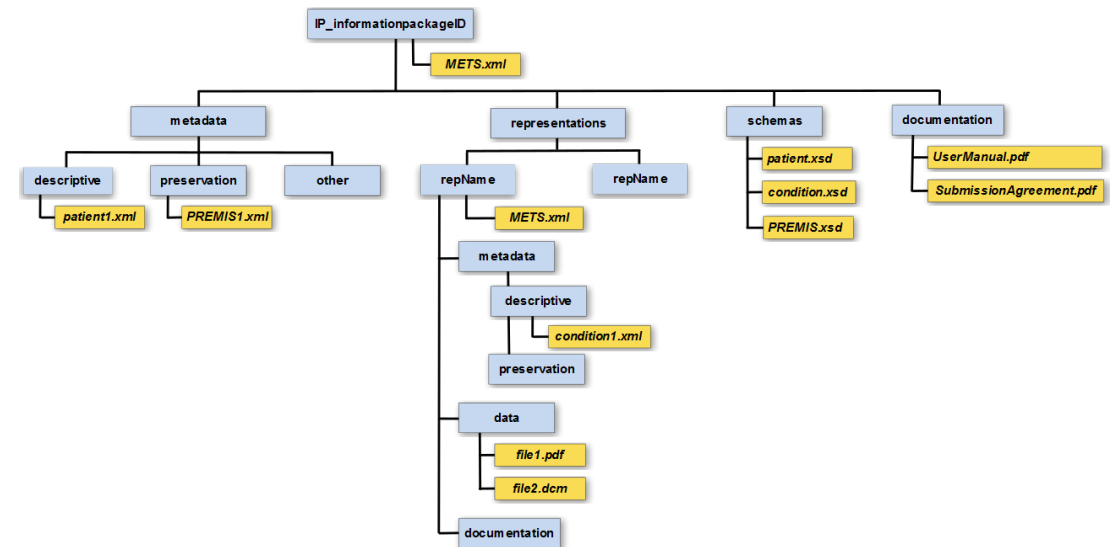
As stated in Principle 1.1, any Information Package **MUST** be able to include any kind of data and metadata. At the same time, we have introduced in earlier Sections the concept of Content Information Types which allow users to achieve more detailed control and fine-grained interoperability. As such, any CSIP Information Package **MUST** include a statement about which Content Information Type Specification(s) has been followed within the Information Package, or on the contrary, indicate clearly that no specific Content Information Type Specification has been followed.

The practical implication of principles 1.1, 2.1 and 2.2 is that, once these have been followed in implementations, it is possible to develop modular identification and validation tools and workflows. While generic components can carry out high-level tasks regardless of the Content Information Type, it is possible to detect automatically which additional content-aware modules need to be executed.

The folder structure of a package

If we do not have a manifest, we still need to be able to understand the package

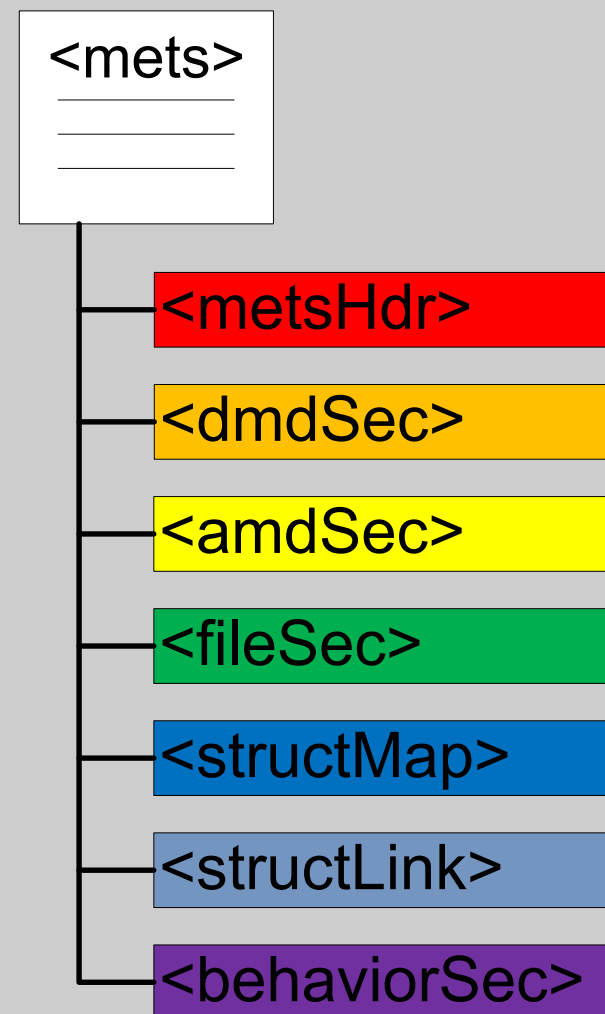
- 4. CSIP structure
- 4.1. Folder structure of the CSIP
- 4.2. Implementing the structure



METS - different packages CSIP, SIP, AIP and DIP and their relationships

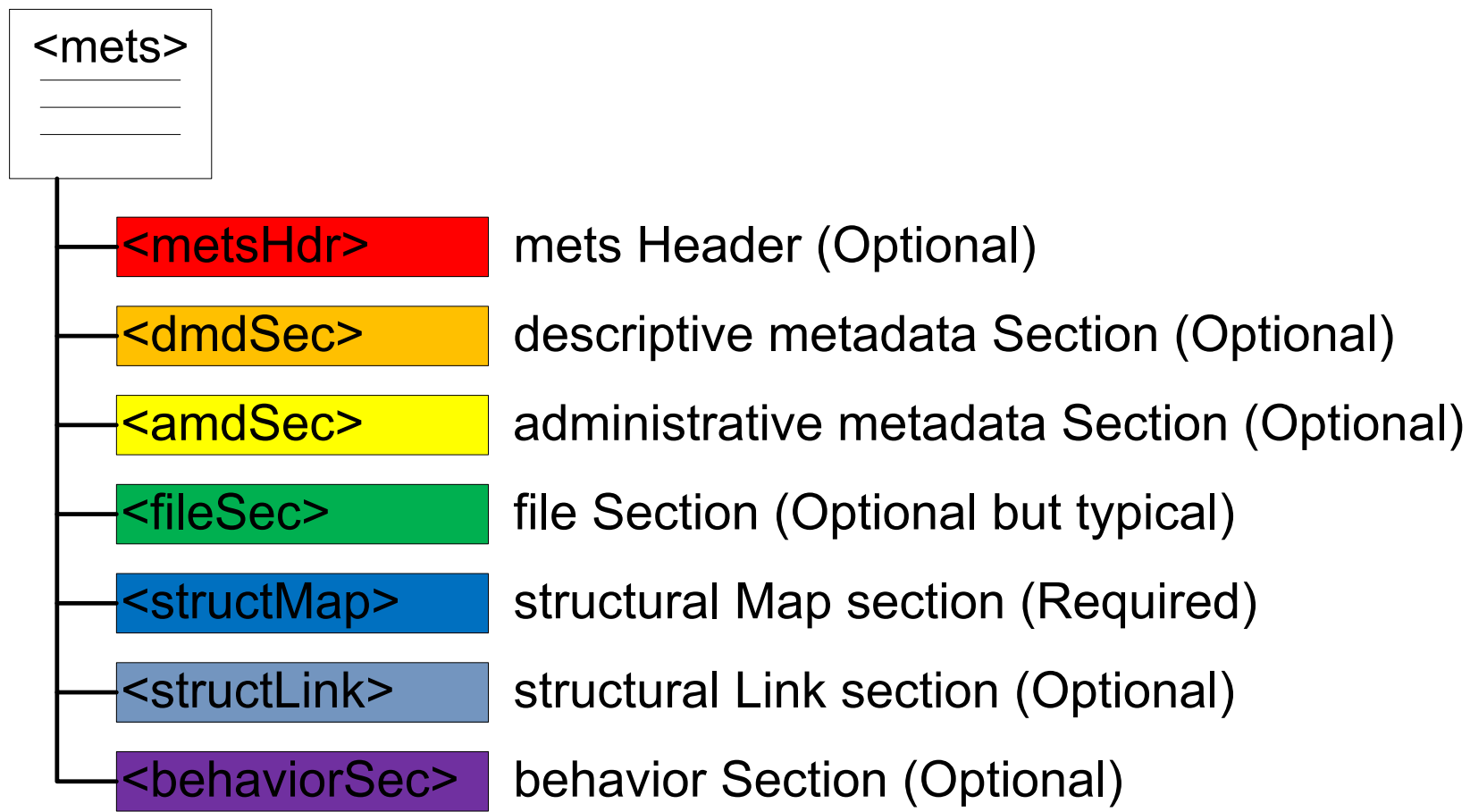
The Metadata Encoding and
Transmission Standard (METS) used in
CSIP

<https://www.loc.gov/standards/mets/>



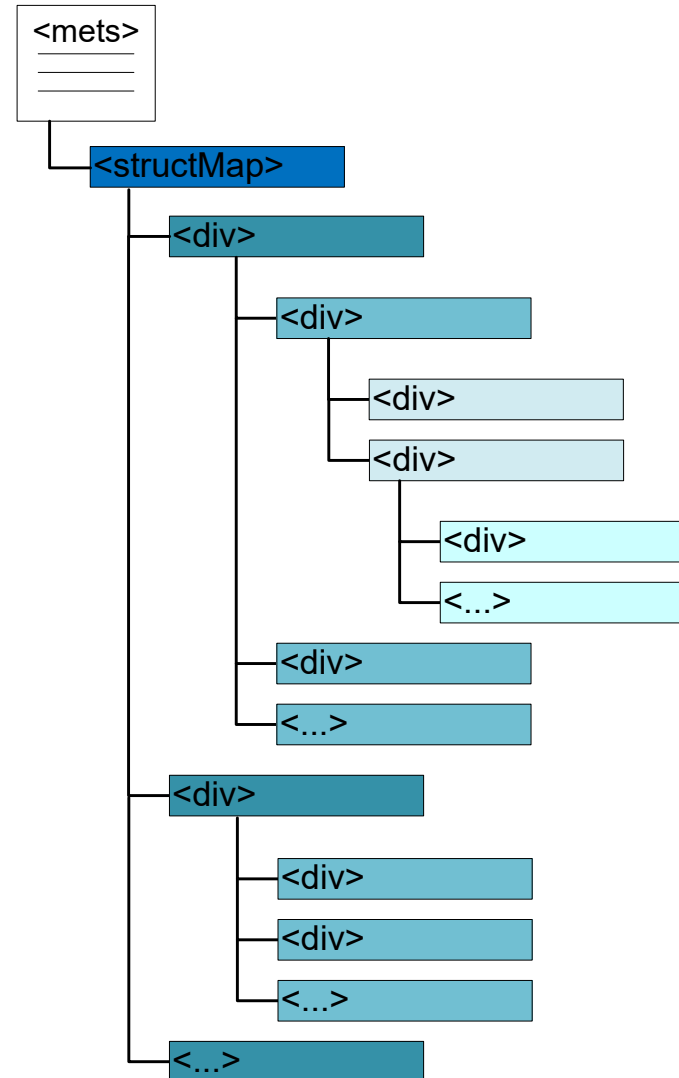
The standard METS

The sections of METS, a short overview



The standard METS and its core element

The structural map in METS

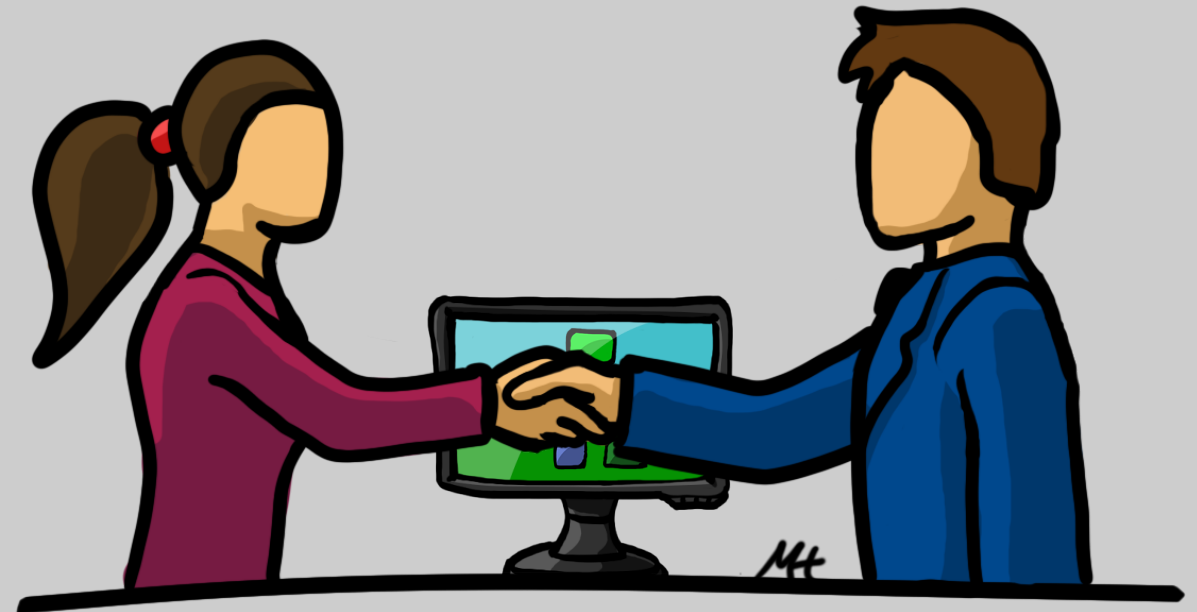


The standard METS and its core element

The structural map used in CSIP

```
<mets:structMap ID="struct-map-example-1" TYPE="PHYSICAL" LABEL="CSIP">
  <mets:div ID="struct-map-example-div" LABEL="csip-mets-example">
    <mets:div ID="struct-map-metadata-div" LABEL="Metadata" ADMID="digiprov-premis-file-1 digiprov-premis-file-2" DMDID="dmd-ead-file">
    </mets:div>
    <mets:div ID="struct-map-doc-div" LABEL="Documentation">
      <mets:fptr FILEID="file-ptr-doc">
      </mets:fptr>
    </mets:div>
    <mets:div ID="struct-map-schema-div" LABEL="Schemas">
      <mets:fptr FILEID="file-grp-schema">
      </mets:fptr>
    </mets:div>
    <mets:div ID="struct-map-reps-sub-div" LABEL="Representations">
      <mets:fptr FILEID="file-grp-rep-subdata">
      </mets:fptr>
    </mets:div>
  </mets:div>
</mets:structMap>
```

Connecting CSIP with METS



Common Specification for Information Package (CSIP)

The common elements and attributes used in the transfer described in a requirements table and a METS profile

ID	Name, Location & Description	Card & Level
CSIP1	Package Identifier <code>mets/@OBJID</code> The <code>mets/@OBJID</code> attribute is mandatory, its value is a string identifier for the METS document. For the package METS document, this should be the name/ID of the package, i.e. the name of the package root folder. For a representation level METS document this value records the name/ID of the representation, i.e. the name of the top-level representation folder.	1.1 MUST
CSIP2	Content Category <code>mets/@TYPE</code> The <code>mets/@TYPE</code> attribute MUST be used to declare the category of the content held in the package, e.g. book, journal, stereograph, video, etc.. Legal values are defined in a fixed vocabulary. When the content category used falls outside of the defined vocabulary the <code>mets/@TYPE</code> value must be set to "OTHER" and the specific value declared in <code>mets/@csip:OTHERTYPE</code> . The vocabulary will develop under the curation of the DILCIS Board as additional content information type specifications are produced. See also: Content Category	1.1 MUST
CSIP3	Other Content Category <code>mets[@TYPE="OTHER"]/@csip:OTHERTYPE</code> When the <code>mets/@TYPE</code> attribute has the value "OTHER" the <code>mets/@csip:OTHERTYPE</code> attribute MUST be used to declare the content category of the package/representation. See also: Content Category	0.1 SHOULD
CSIP4	Content Information Type Specification <code>mets/@csip:CONTENTINFORMATIONTYPE</code> Used to declare the Content Information Type Specification used when creating the package. Legal values are defined in a fixed vocabulary. The attribute is mandatory for representation level METS documents. The vocabulary will evolve under the care of the DILCIS Board as additional Content Information Type Specifications are developed. See also: Content information type specification	0.1 SHOULD
CSIP5	Other Content Information Type Specification <code>mets[@csip:CONTENTINFORMATIONTYPE="OTHER"]/@csip:OTHERCONTENTINFORMATIONTYPE</code> When the <code>mets/@csip:CONTENTINFORMATIONTYPE</code> has the value "OTHER" the <code>mets/@csip:OTHERCONTENTINFORMATIONTYPE</code> must state the content information type.	0.1 MAY
CSIP6	METS Profile <code>mets/@PROFILE</code> The URL of the METS profile that the information package conforms with.	1.1 MUST

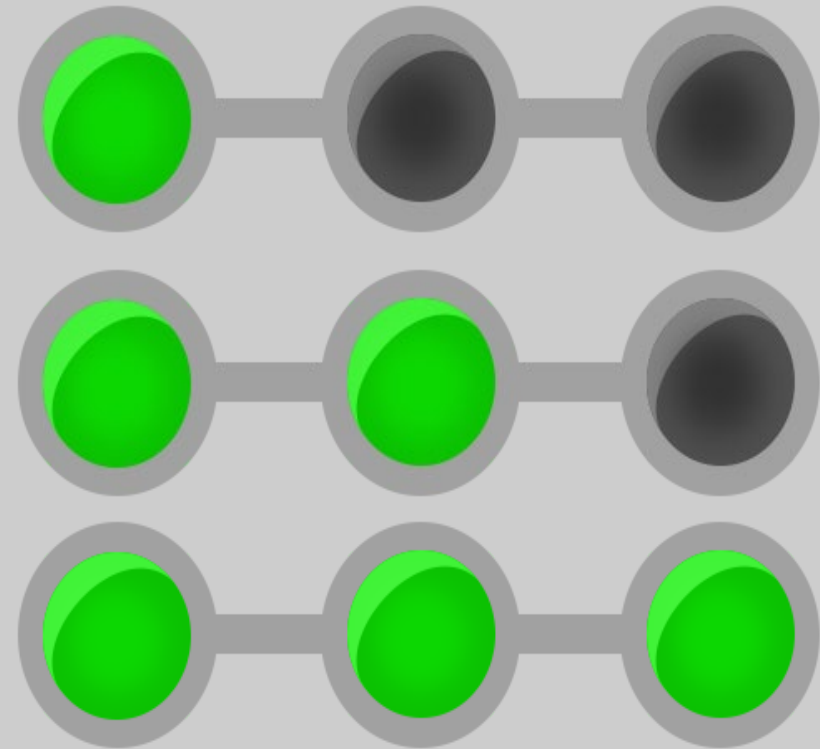
Example: METS root element showing use of `csip:OTHERTYPE` attribute when an appropriate package content category value is not available in the vocabulary. The `@TYPE` attribute value is set to OTHER.

```
<mets:mets OBJID="uuid-4422c185-5407-4918-83b1-7abfa77de182" LABEL="Sample CSIP Information Package" TYPE="OTHER" OTHERTYPE="OTHER" />
</mets:mets>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Draft of E-ARK CSIP METS Profile 2.0 -->
<METS_Profile xmlns="http://www.loc.gov/METS_Profile/v2"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:mets="http://www.loc.gov/METS/"
  xmlns:csip="https://DILCIS.eu/XML/METS/CSIPExtensionMETS"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xsi:schemaLocation="http://www.loc.gov/METS_Profile/v2 http://www.loc.gov/standards/mets/profile_docs/mets.profile.v2-0.xsd http://www.loc.gov/standards/mets/profile_docs/mets.profile.v2-0.xsd"
  <URI LOCTYPE="URL" ASSIGNEDBY="local">https://earkcsip.dilcis.eu/profile/E-ARK-CSIP.xml</URI>
  <title>E-ARK CSIP METS Profile</title>
  <abstract>This base profile describes the Common Specification for Information Packages (CSIP) and the implementation of METS
    This will enable repository interoperability and assist in the management of the preservation of digital content.
    This profile is a base profile which is extended with E-ARK implementation of SIP, AIP and DIP.
  The profile can be used as is, but it is recommended that the supplied extending implementation are used. Alternatively, an o
  <date>2020-01-08T12:00:00</date>
  <contact>
    <institution>DILCIS Board</institution>
    <address>http://dilcis.eu/</address>
    <email>info@dilcis.eu</email>
  </contact>
  <related_profile>This profile has no related profiles</related_profile>
```



The connection between the different Information Package specifications



E-ARK SIP, E-ARK AIP and E-ARK DIP

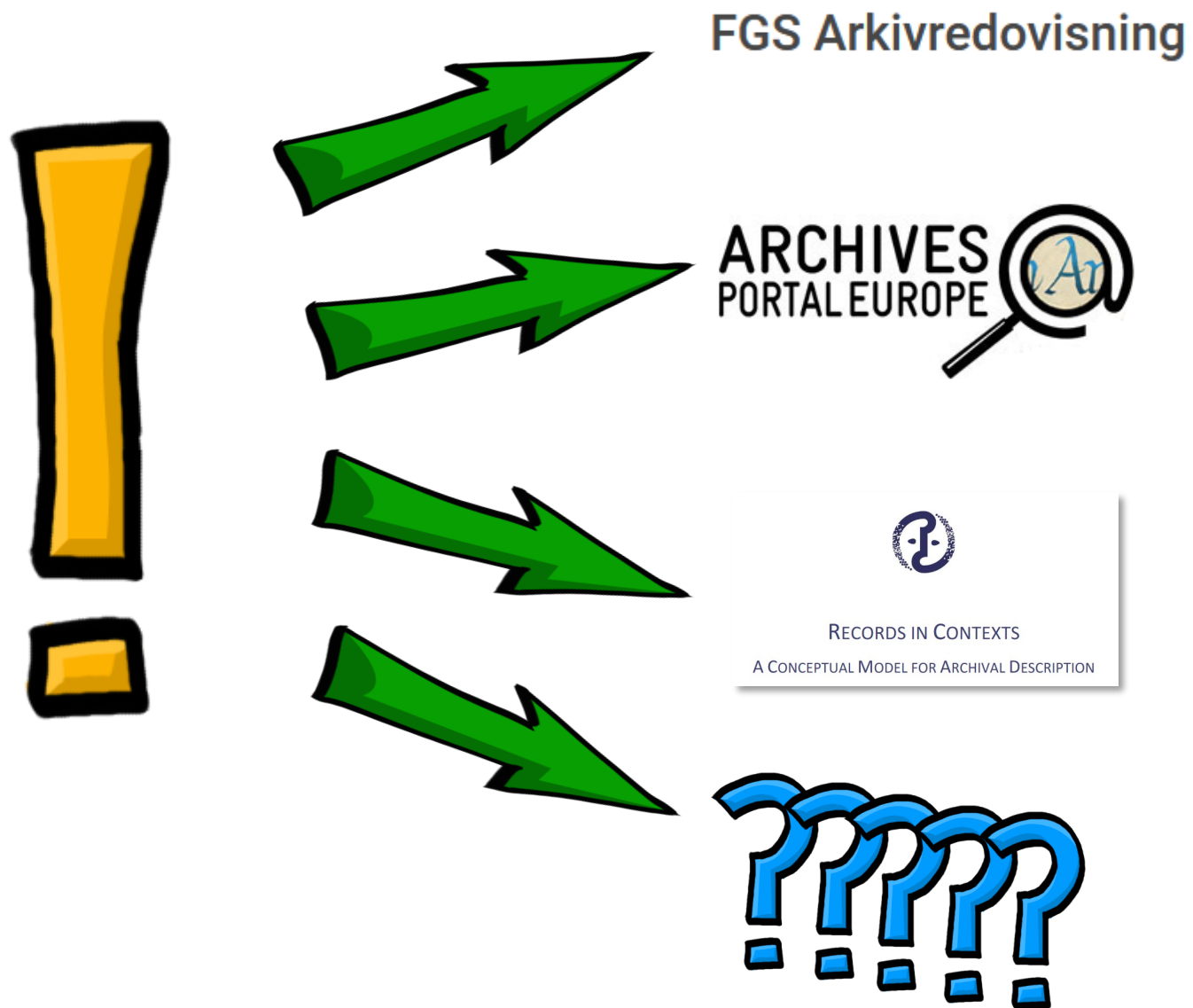
Special use cases of the CSIP

- E-ARK SIP + E-ARK DIP adds more rules to CSIP
 - Defines the Submission Information Package and the Dissemination Information Package
 - Some extra attributes
 - Defines some values to use like informing about where we are in the OAIS reference model; SIP and DIP
- E-ARK AIP description of how to handle a package in an electronic archive

Connected to the content of the package

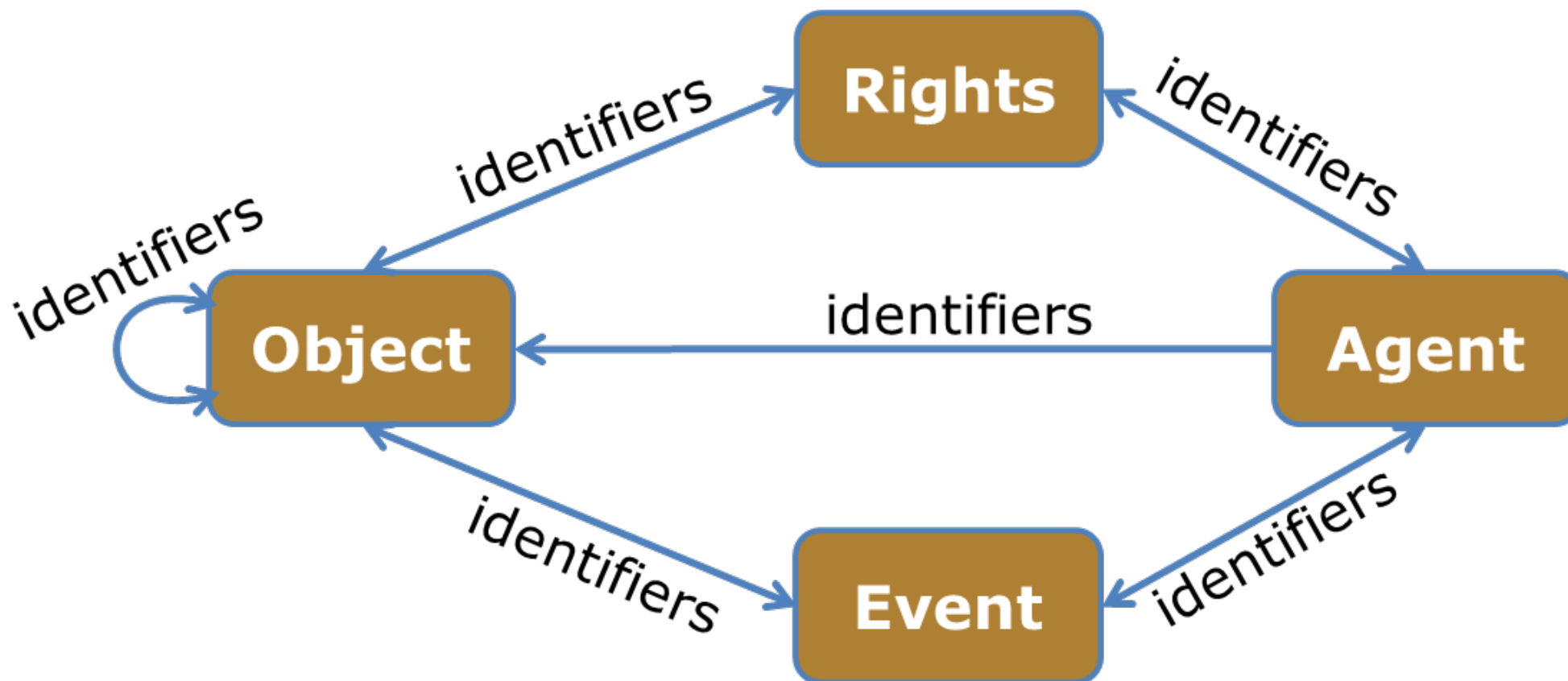
Archival Information

Information regarding the creator and finding aid



Preservation metadata

Preservation metadata for the content



Maintenance and contributions



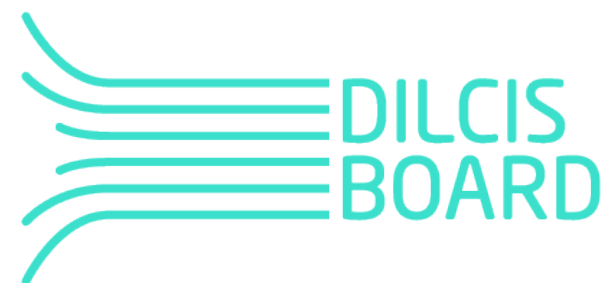
the Digital Information LifeCycle Interoperability Standards Board (DILCIS Board)



DILCIS Board

The maintainers

- <https://dilcis.eu/>
- <https://github.com/DILCISBoard>
- info@dilcis.eu



Can I or we contribute?!

4

Contributions

The users

- Request to create a CITS specification
- Request CITS endorsement of a specification
- Use the existing specifications and give us feedback in GitHub
- Participate in the reviews and give your comments!





CITS eHealth1



Stephen Mackey, Piql

CEF Telecom Call Requirement

Defined in the CEF Telecom call for proposals 2019

“ ... specifications for eHealth will be developed by the activity. One specification will be based upon the Norwegian eHealth archives transfer format of patient journals (from provider EMR systems to a central health archive). ”

Electronic Medical Record & Health Record Systems

No strict definitions



EMR

- Medical in nature, used by clinicians
- Mainly unstructured data, core composition of natural language documents, often attested, accompanied by data files and metadata
- Distributed
- Not interoperable, records do not travel easily

EHR

- Consider the whole health of the patient, take contributions from all healthcare providers
- Provide universal access for healthcare providers and patients
- Centralised
- Mainly structured data with additional data files

Use Cases for a Central Health Archive

Patient medical records

1

Transparency

To provide records to next of kin in compliance with open information regulation

3

Portability and Interoperability

Common standard for exchange of complete patient medical records

2

Research

To harvest the vast amount of historical healthcare-related data within the archive for medical research.

4

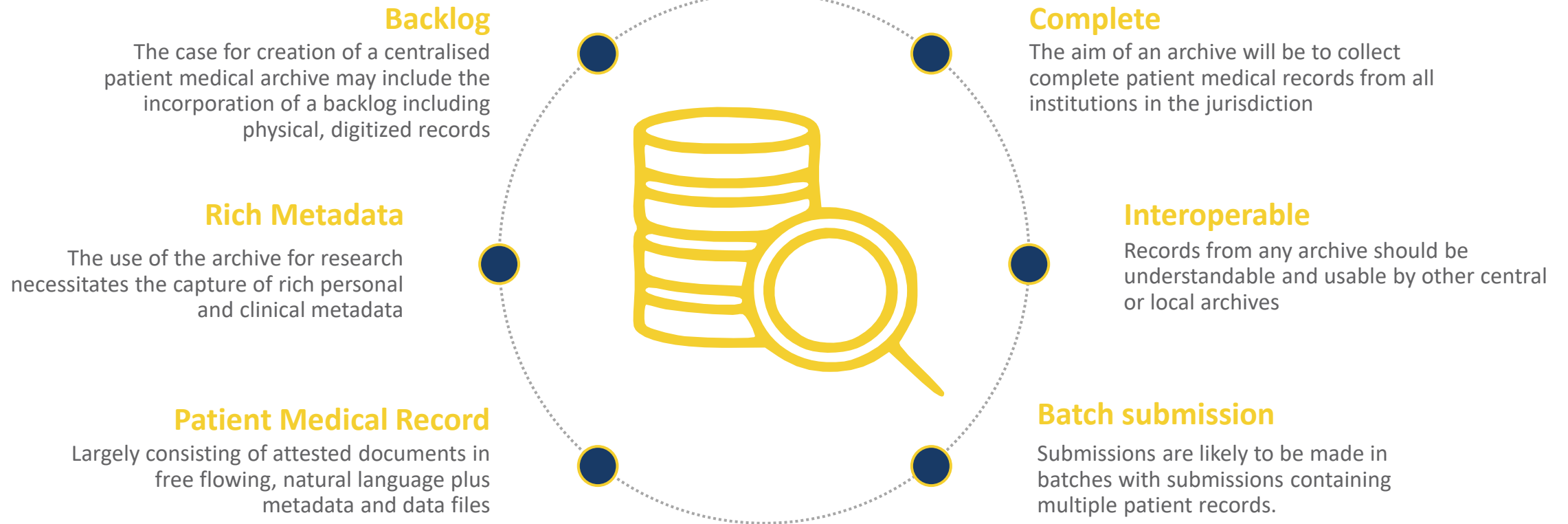
Management of EHR systems

To archive data from online, centralised EHR systems to manage storage capacities



Scope

Patient medical record archive



eHealth1 specification

Considerations

Patient Centricity

- Use cases determine an organization (taxonomy) of data by patient. All records for a patient from an institution are grouped within archival packages.
- Impractical and unnecessary to aggregate patient records from different submission institutions
- High level descriptive metadata is patient personal information (such as the FHIR patient resource)

Aggregation Structure

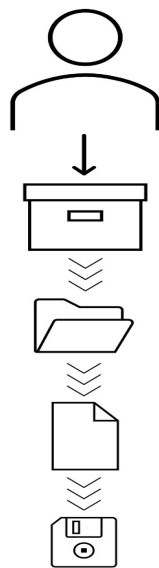
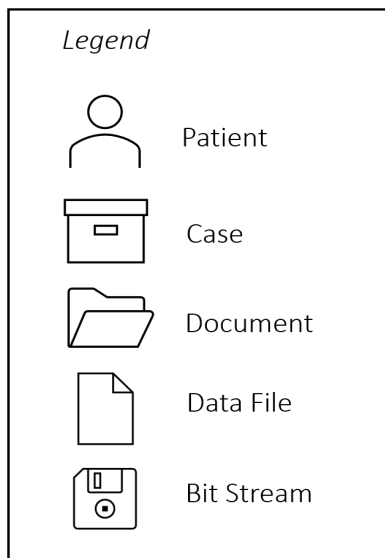
- Records are organised in a case/sub-case/document taxonomy
 - ✓ Case – documents grouped by period, condition or treatment
 - ✓ Sub-case – documents related by a specific department or treatment
 - ✓ Document – set of files with common metadata, attested
 - ✓ Data Files – documents, images, audio, video, scans. Can be encapsulated bit streams (e.g. Dicom)
 - ✓ Bitstream – data

Metadata

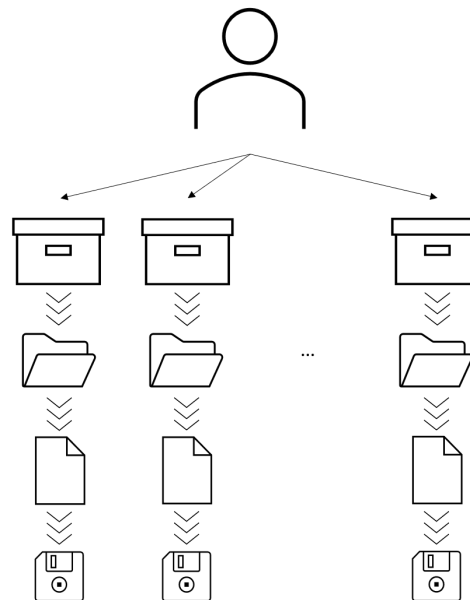
- Recommends but does not mandate the use of FHIR resources for:
 - ✓ Patient personal
 - ✓ Clinical – Condition, AllergyIntolerance, Procedure, etc
- References other international standards such as ICD, SNOMED
- Extensible and adaptable

Case structure

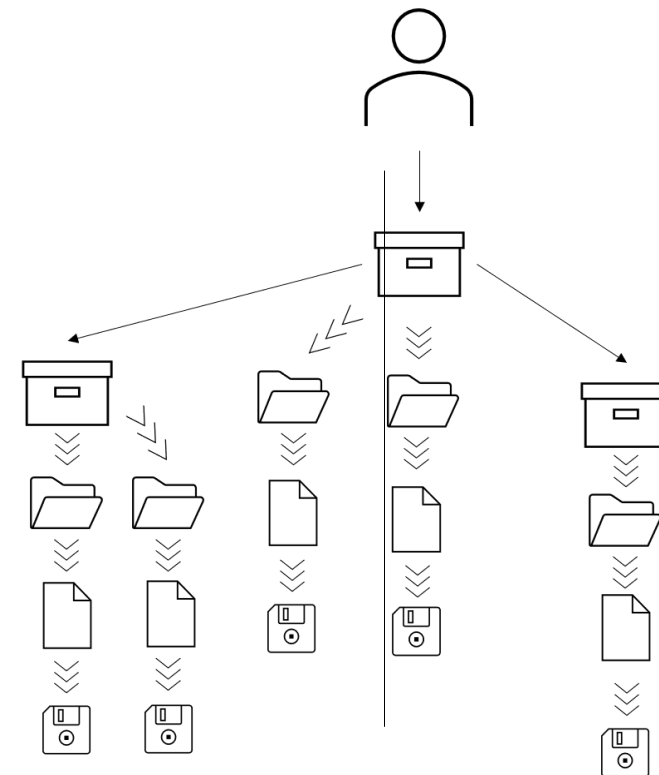
A range of use cases



Single Document



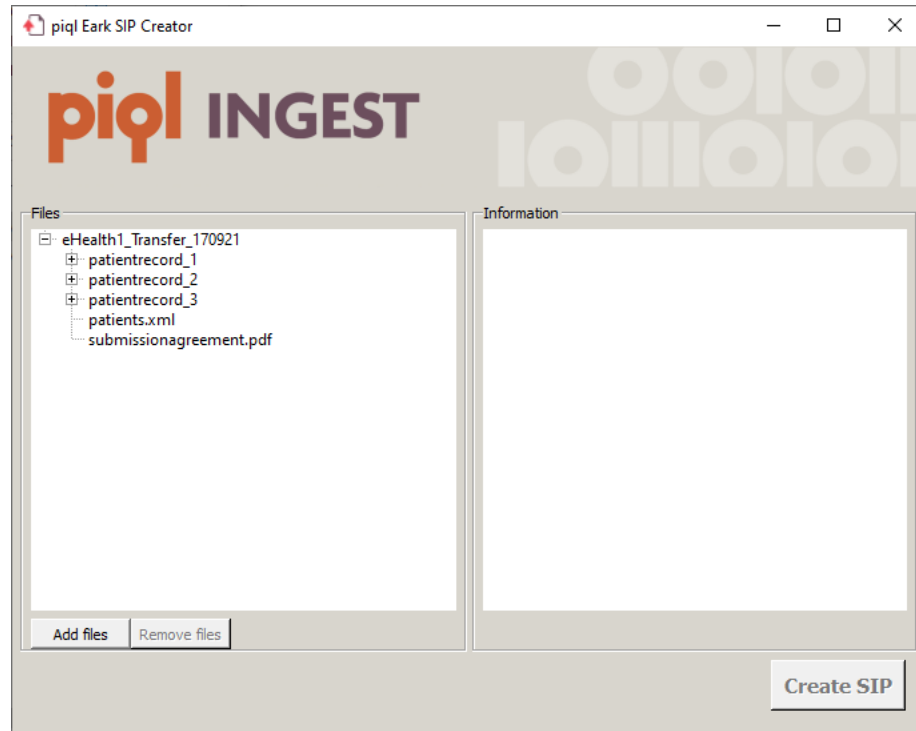
Multiple Cases, Documents



Multiple Cases, Sub-cases, Documents

eHealth1 SIP Creator

Production of compliant eHealth1 SIPs from source EMR systems



SIP Creator

- Desktop Windows application
- Produces compliant SIPs from EMR system extracts (data and metadata)
- Configurable for different metadata standards
- Transformation scripts can be embedded to map export structures and metadata
- Outputs can be validated using eArchiving validation tools

Summary

1

CITS specification for archival of patient medical records in a central health archive

Specification focuses on SIPs but requirements flow through logically into AIP and DIP structures and the specification allows for multiple (batch) submissions in individual packages.

2

Uses a patient centric and case based taxonomy for organisation of EMR sourced patient records

Future use cases could consider export from EHR systems or systems with a higher structured content. The SIP Creator tool aids production of compliant SIPs.

3

Adoption will be driven by creation of central/regional health archives

The specification allows for interoperability, portability of archival packages and cost saving through the use of open-source eArchiving tools for implementation such as the eHealth1 SIP Creator.



shorturl.at/fjpHX

CITS eHealth2

Archiving of Cancer Registry's Exports



Anja Paulič, Archives of the Republic of Slovenia

Cancer Registries and their data

Cancer

one of the most important public health issues in the developed world

Cancer Registry

an institution that systematically and continuously gathers data on cancer patients in a designated area

Cancer Registry Data

is used for assessing, planning and controlling the cancer burden in the community, its completeness and data quality are constantly improved

Cancer Registry Export

dataset, which contains at least one file with data from the cancer registry and stored separately from CR data

eHealth2

Slovenian Cancer Registry, Archives of the Republic of Slovenia and experience of Norwegian CR, JRC, other CR data aggregators



eHealth2 development

How did we do it?

Contains patient's characteristics, data on disease, treatment, vital status, mortality: standard cancer burden indicators

Cancer Registry Export Data

Based on classifications and standards
Prepared in agreement with aggregator

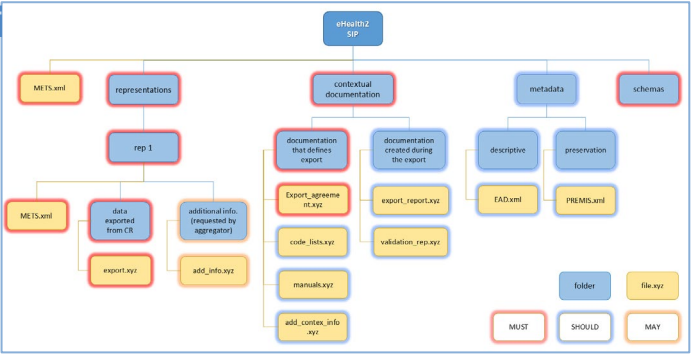


CSIP (the E-ARK Common Specification for Information Packages)

eArchiving

E-ARK SIP (the E-ARK Specification for Submission Information Packages)

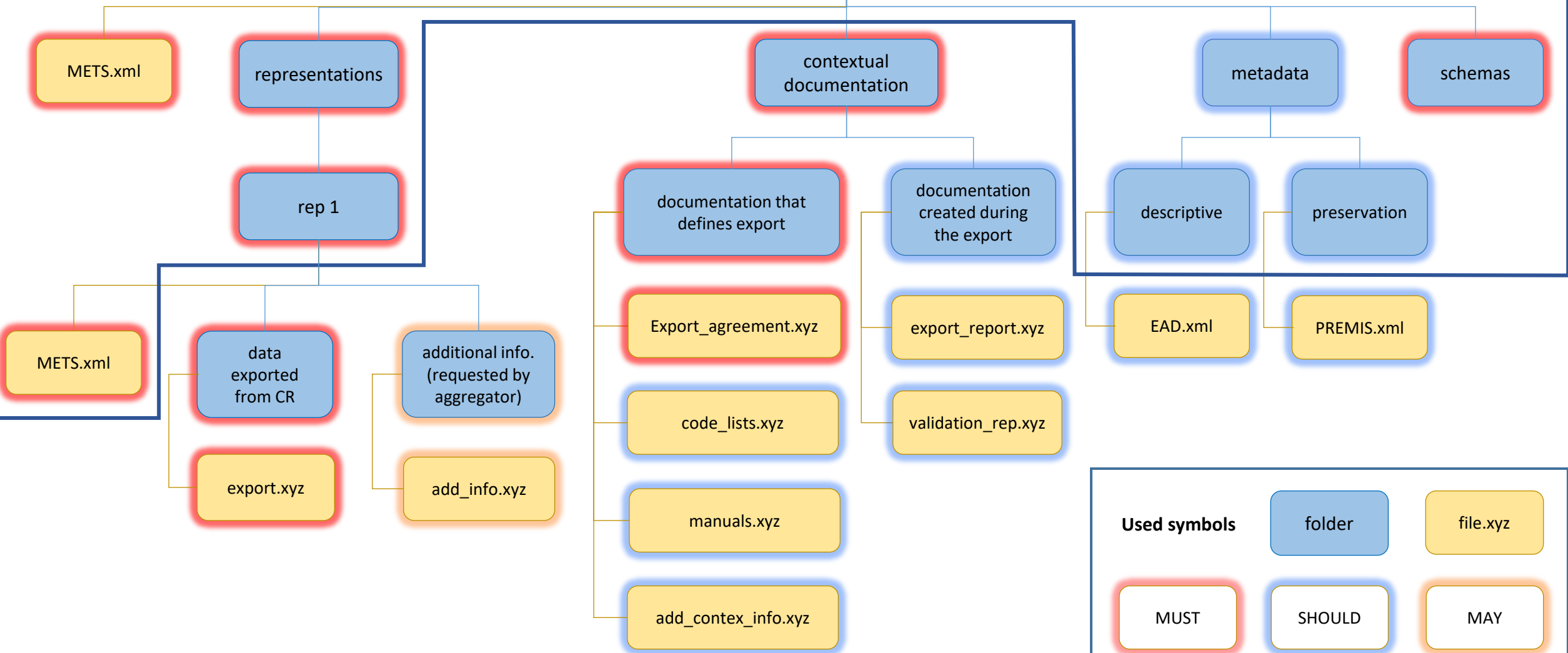
eHealth2 Content Specific Information Type Specific



eHealth2 Guidelines

implement specification in your workflow

eHealth2 SIP



eHealth2 specification

Used symbols

folder	file.xyz
MUST	MAY
SHOULD	

eHealth2 guidelines

Step by step manual for eHealth2 specification

Description of the context

Presentation of the context in which specification can be used

Identifying users

Written for different stakeholders that may have different background knowledge of topics discussed in the specification

Providing examples

Added real life examples from Slovenian Cancer Registry based on JRC data call



eHealth2 impact on stakeholders



Preserve

Retention of data that has impact on health policies



Trust

Clear and easily verifiable cancer burden related data



Usability

Specification and guidelines can be used for all kinds of cancer registry data exports



**Cancer
Registries**

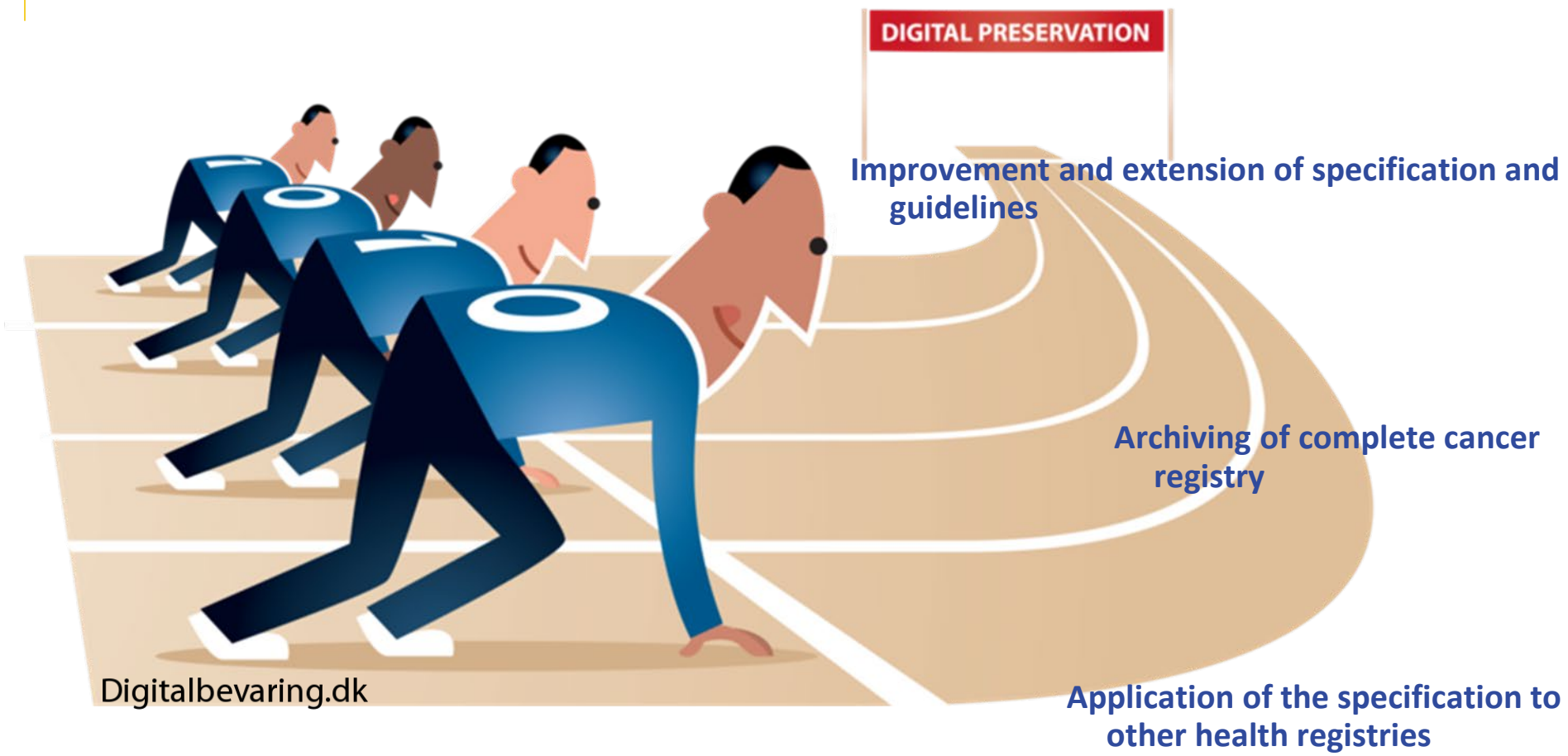


**Aggregators, Academic and
other Researchers**



**Archives and other Record
Keepers**

eHealth2 going forward



1

2

3

Digitalbevaring.dk



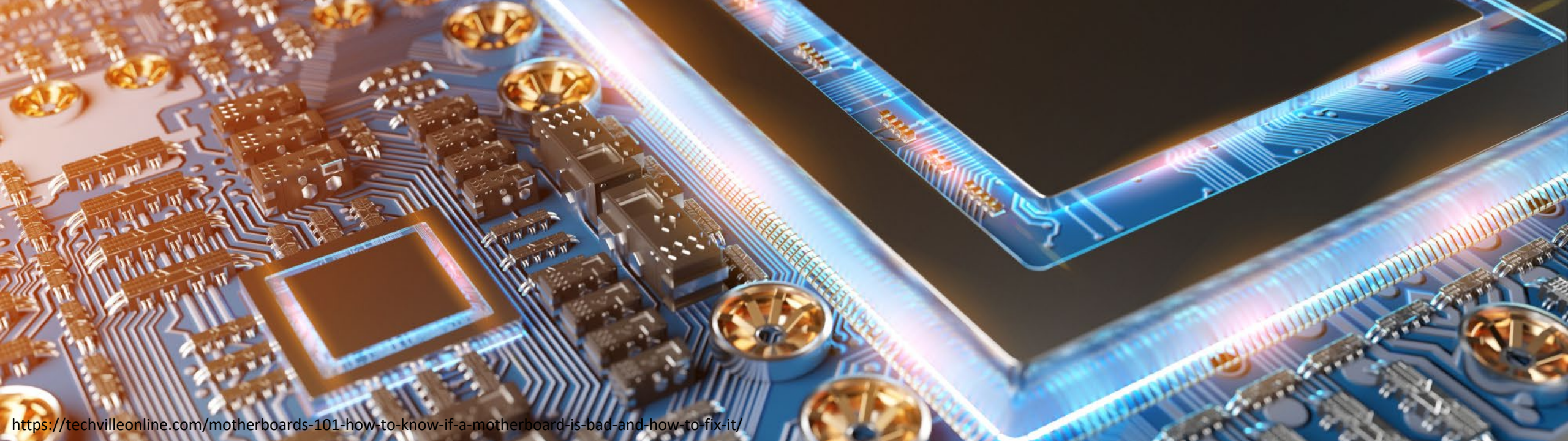
CEF Digital
Connecting Europe



Anja Paulič

Thank you!

<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving>
anja.paulic@gov.si, joze.skofljanec@gov.si
STomsic@onko-i.si, TZagar@onko-i.si, VZadnik@onko-i.si



<https://techvilleonline.com/motherboards-101-how-to-know-if-a-motherboard-is-bad-and-how-to-fix-it/>

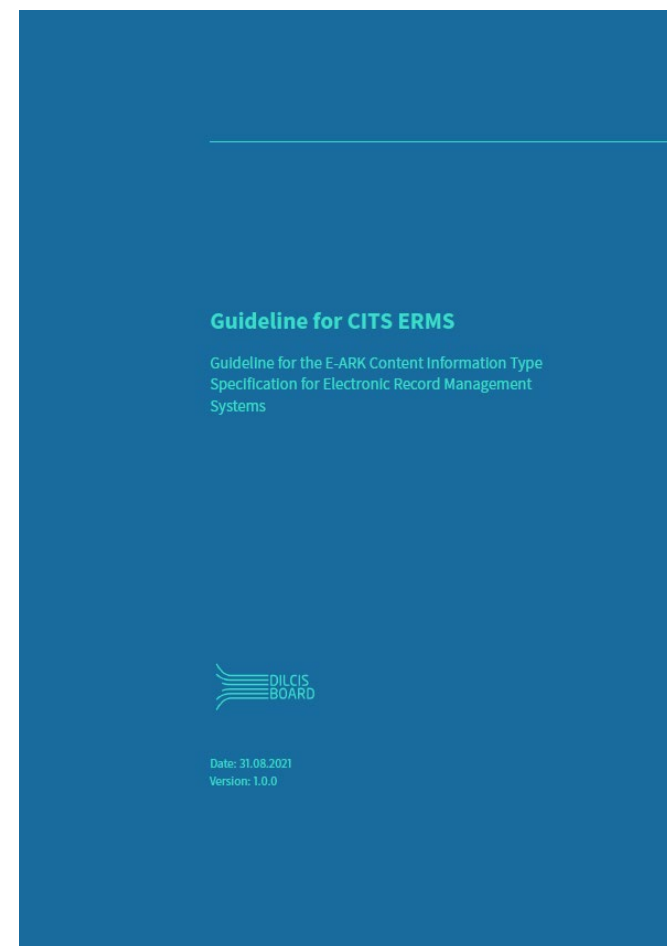
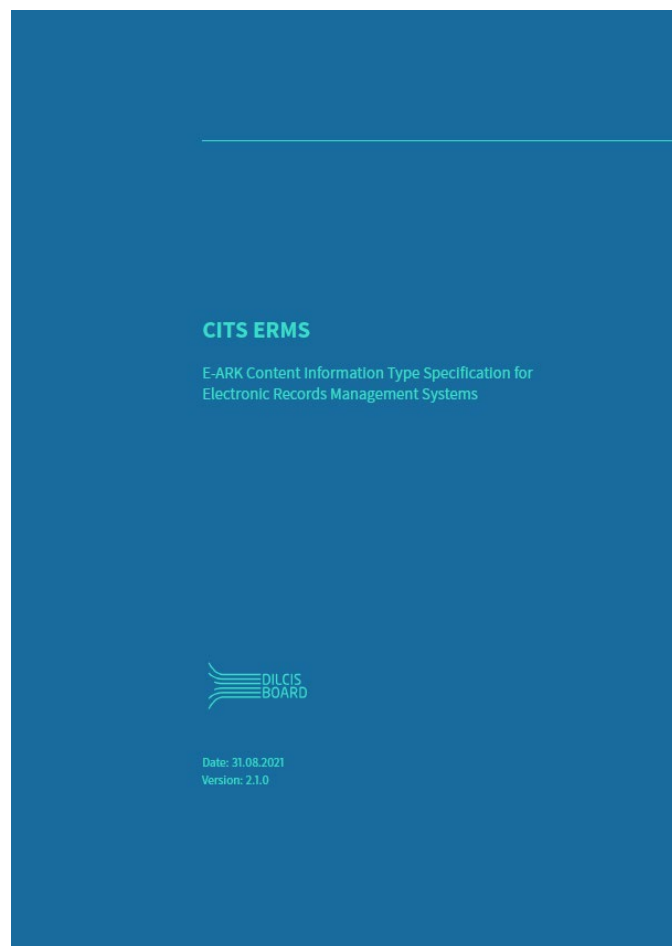
CITS Electronic Records Management Systems



Karin Bredenberg, Sydarkivera

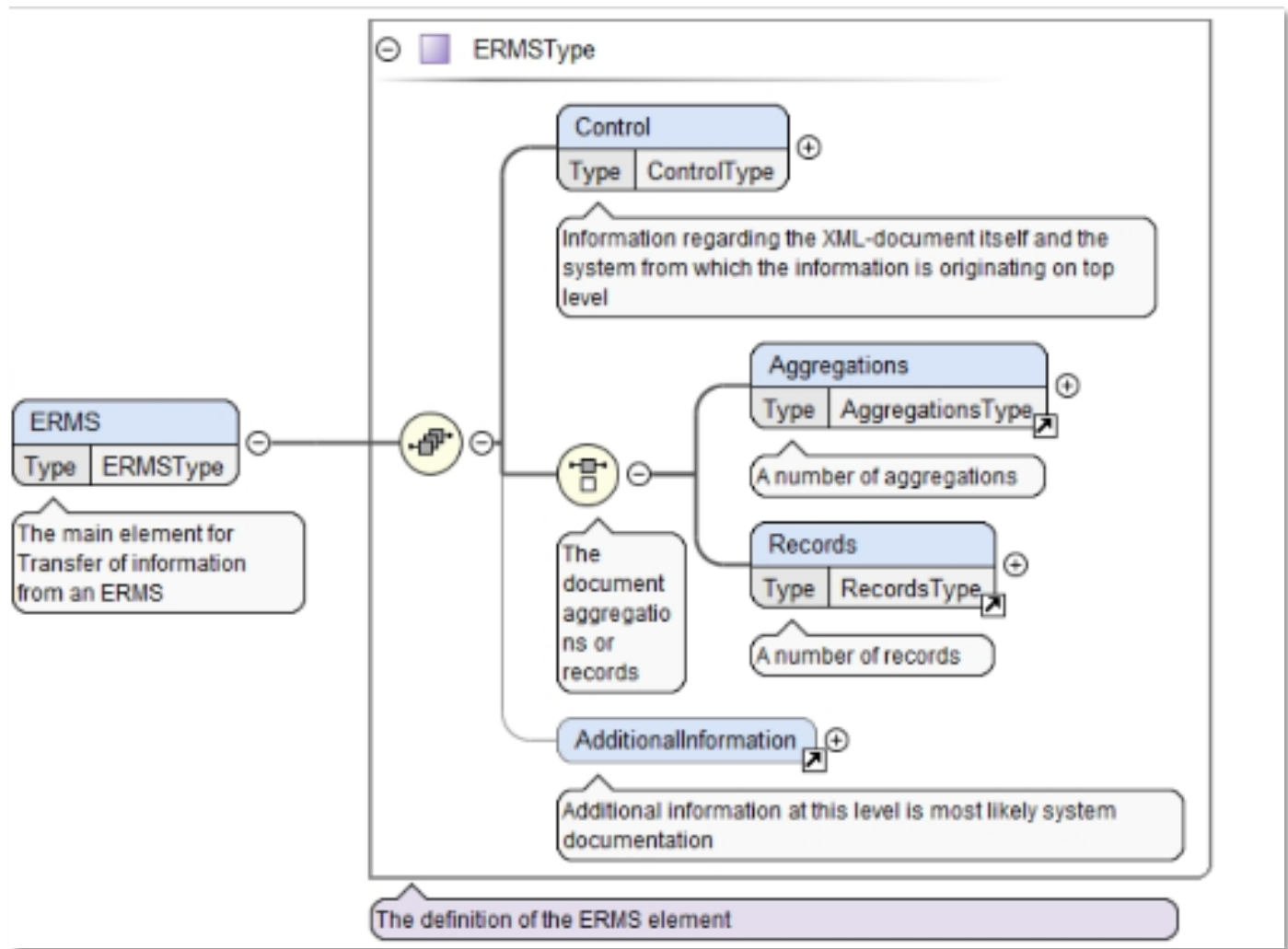
CITS ERMS Specification and Guideline

The documents



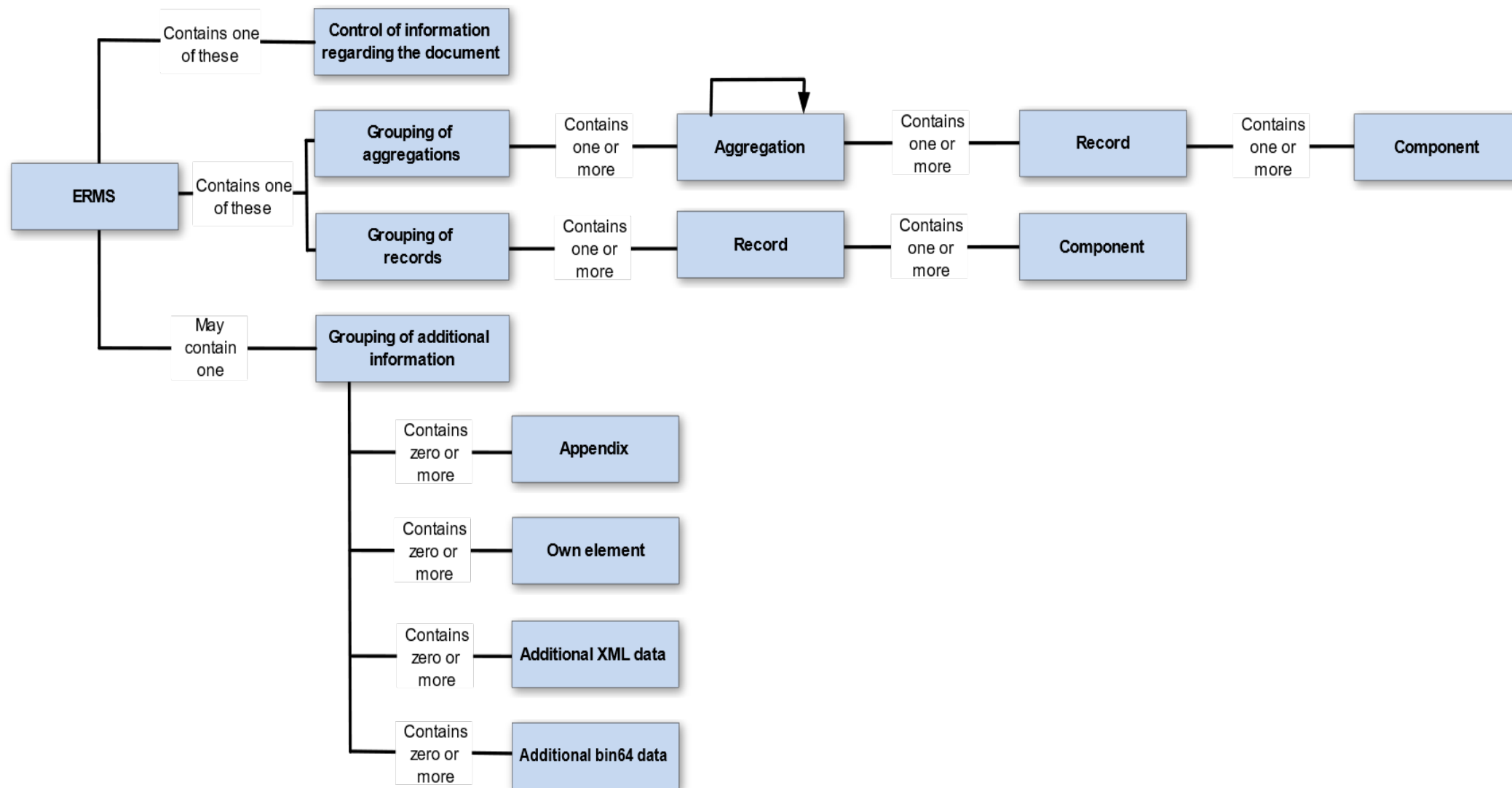
CITS ERMS XML-schema + Schematron

The supporting rule documents



CITS ERMS

The main elements



CITS ERMS and CSIP

The connection between the two

Table 1: Specific fields to use in CSIP

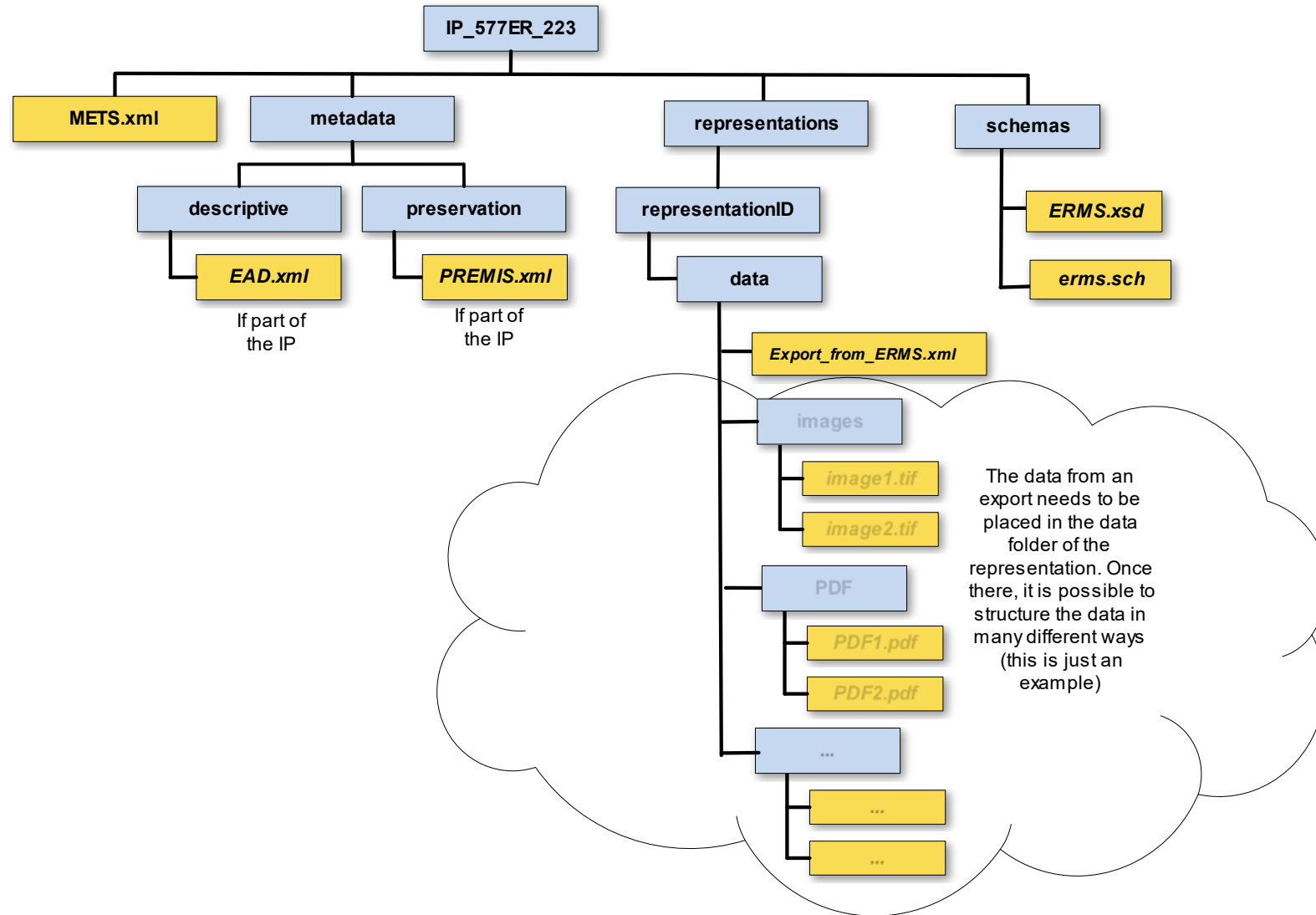
Element name	METS path	Value
General content type	mets/@TYPE	Dataset
Specific content type	mets/@csip:CONTENTINFORMATIONTYPE	ERMS
Specific content type	fileGrp/@csip:CONTENTINFORMATIONTYPE When the FileGrp describes a Representation	ERMS

3.3.2 Placement of data in a CSIP Information Package

The ERMS document is placed as a representation file following the instructions in CSIP.

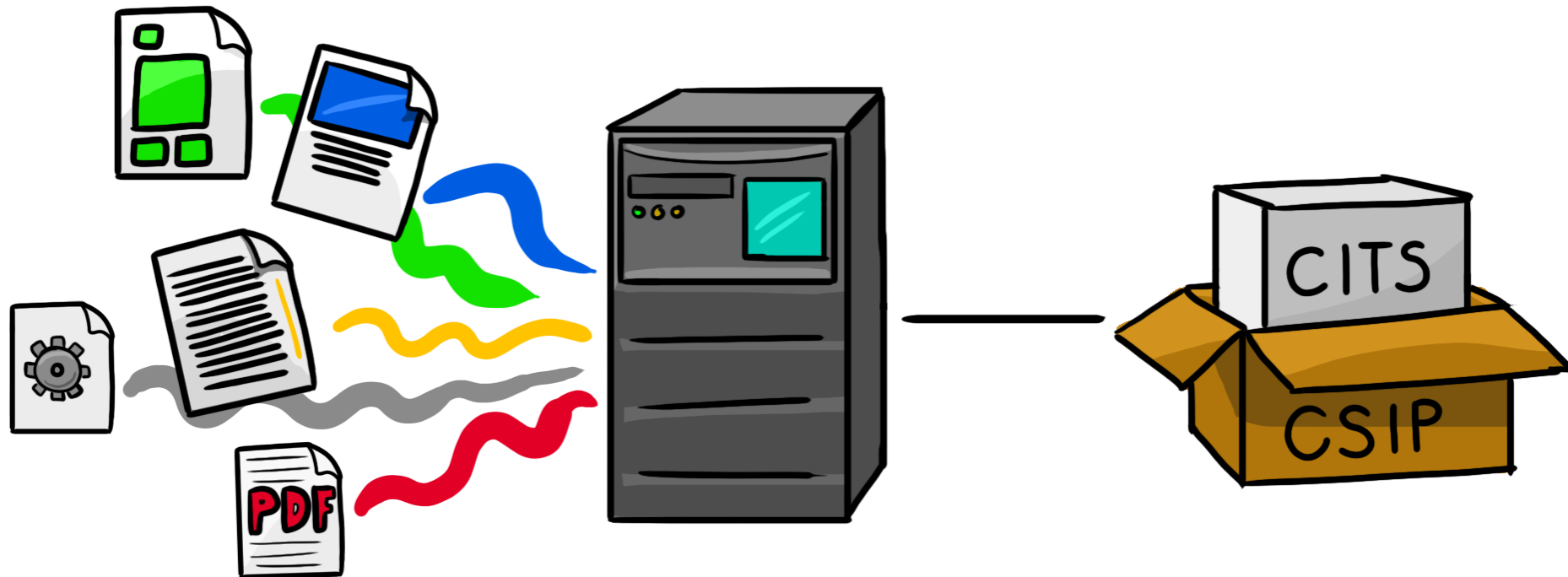
CITS ERMS placement

Where do the files go on an information package?



Using the CITS ERMS

Mapping is the first step





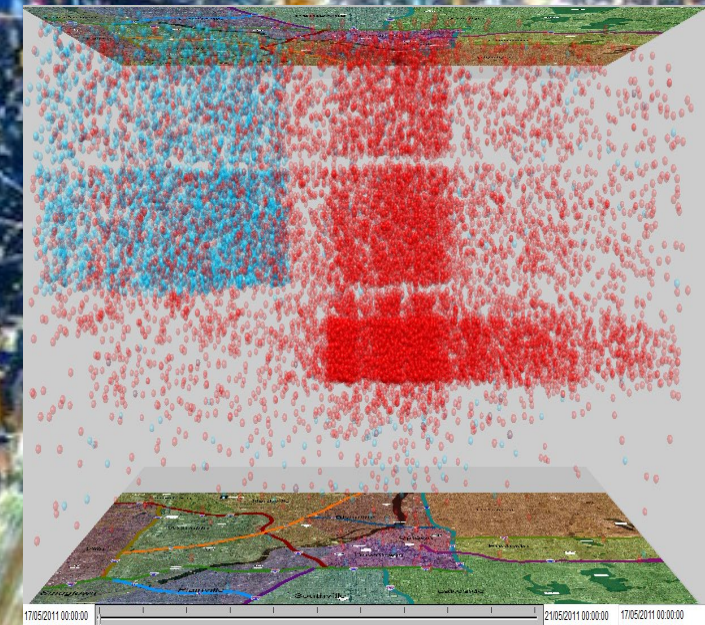
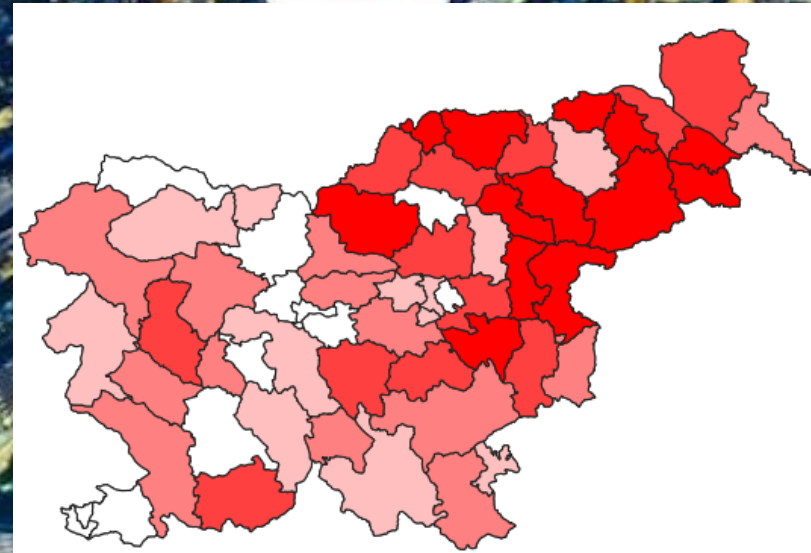
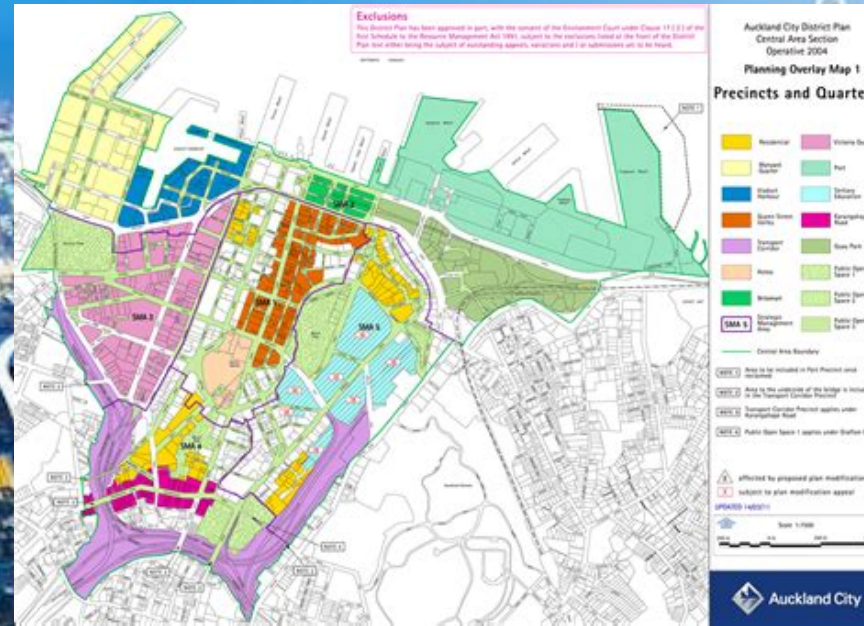
CITS Geospatial



Gregor Završnik, Geoarh, Slovenia

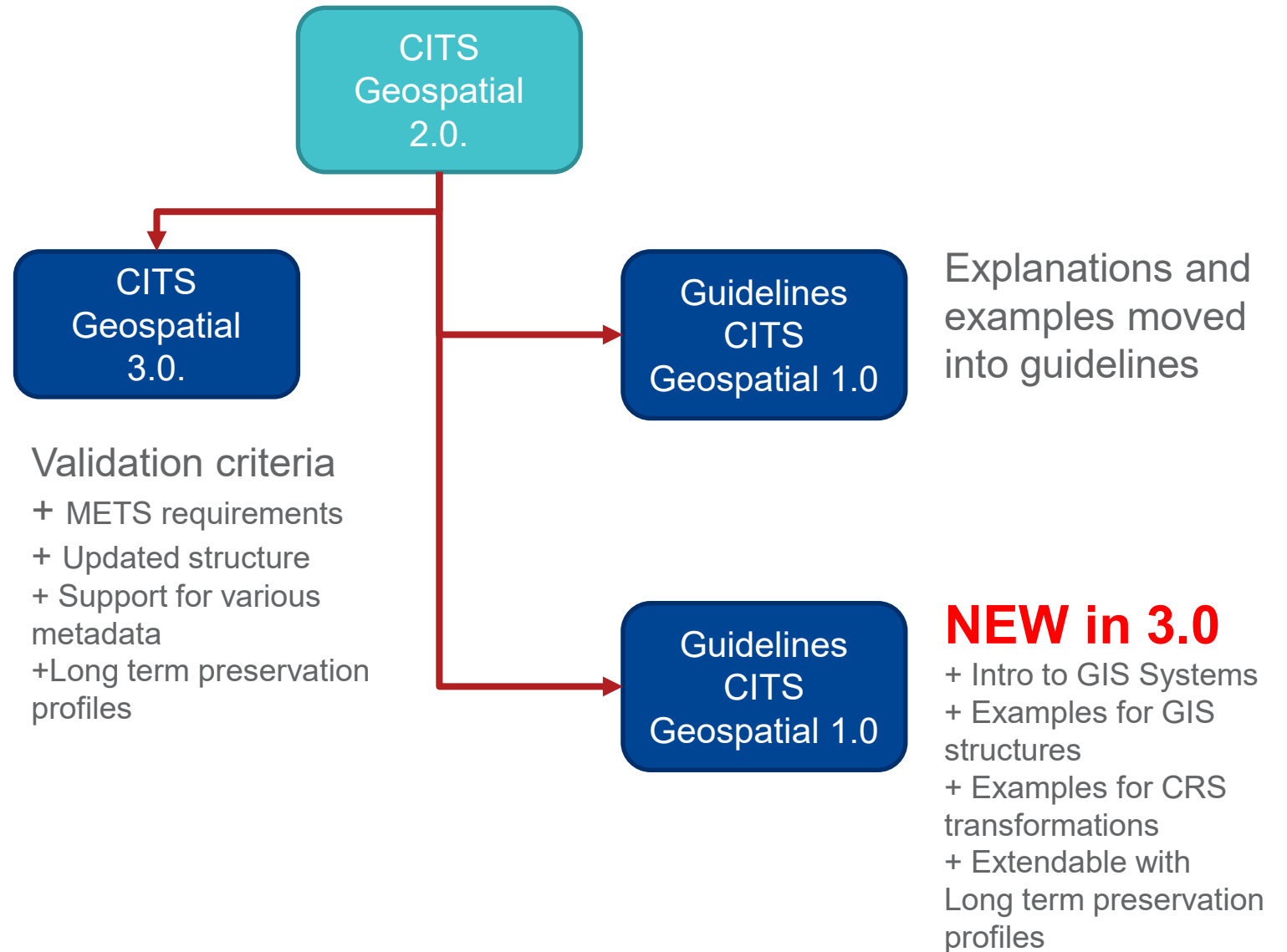
Quick intro: Why preserve Geospatial records?

- Everything happens somewhere
- Form of official records
- Common denominator for all data
- Cornerstone of Digital Europe Data Economy
 - Analysis across time
 - Basis for many inovative solutions



What is new in 3.0

- Alignment with the CSIP package structure
- Requirements redesign according to "Significant Properties"
- Introducing "Long term preservation Profiles"
- Support for Geospatial metadata Standards
- Guidelines



Validation criteria:

2. CITS Geospatial Requirements structure

2.1.

Folder structure requirements

How to package Geospatial records within the CSIP Structure.

Adds suggestions for storage of significant properties for preservation of records

2.2.

METS Requirements

Specific requirements for the METS file, showing us where the package contains Geospatial records, to support validation.

2.3.

Data Requirements

Defines a basic set of requirements specific to Geospatial records. Extended requirements can be defined in Long-term preservation format profiles.

2.4.

Documentation requirements

Specifies essential and recommended technical and contextual documentation, required for future reuse.

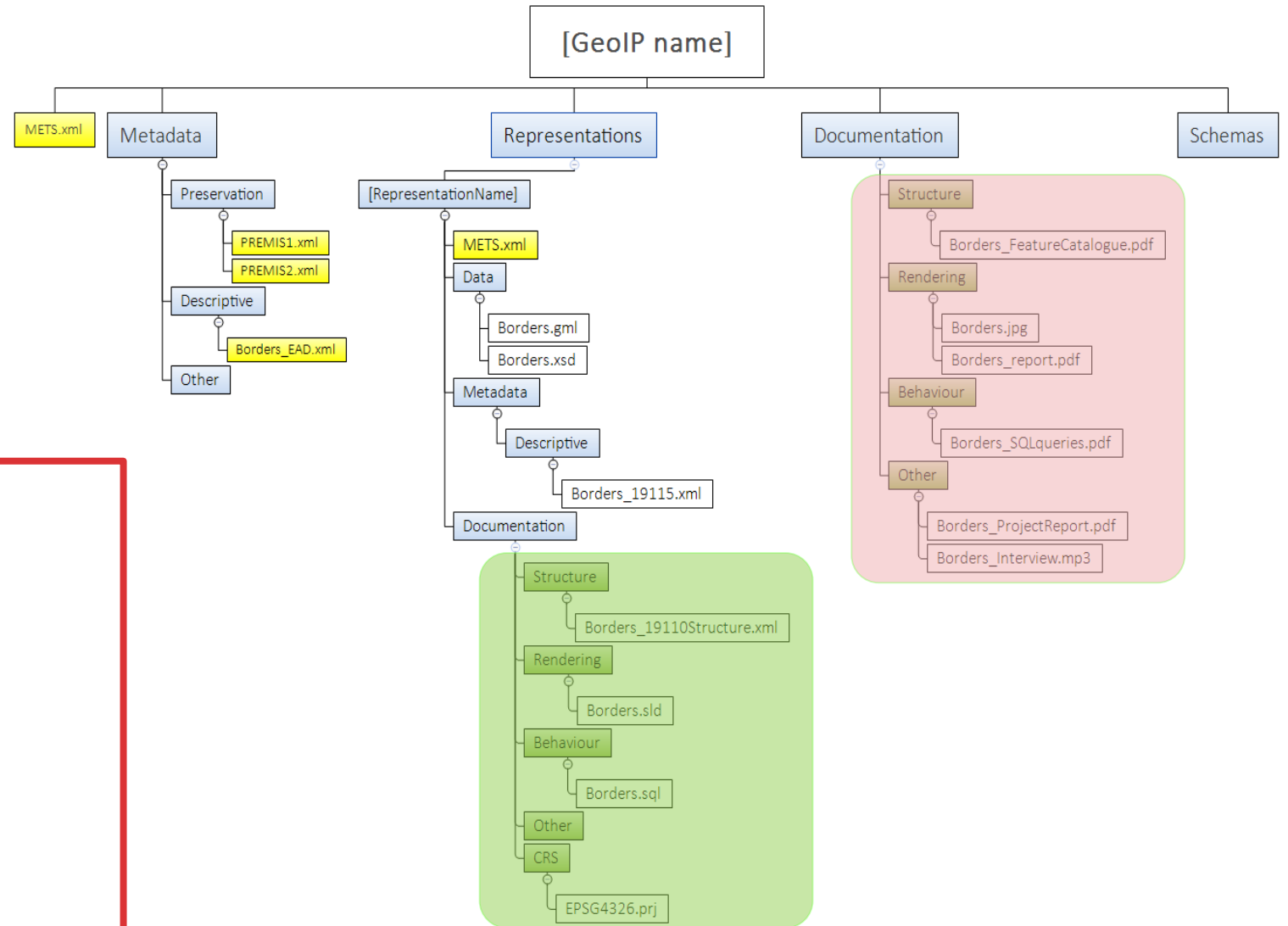
2.5.

Metadata requirements

Defines requirements for geospatial metadata elements and their placement within the information package. The aim is to support automated accessibility and findability.

2.1. Folder structure requirements

- Structure based on CSIP
- Structure supports the “Significant Properties” model
- Standardized machine-readable vs descriptive
- Additional folders are strongly recommended not mandatory



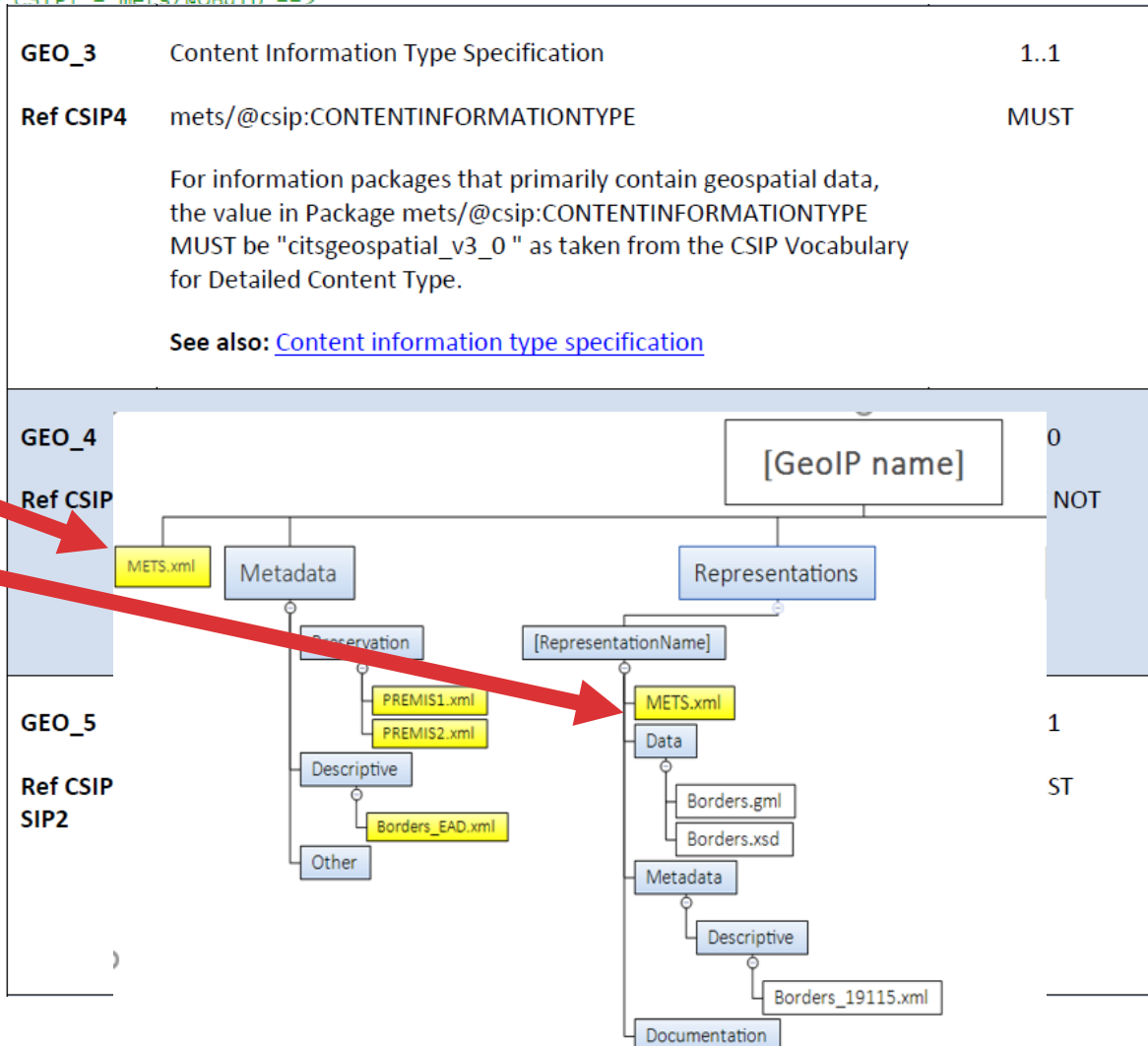
<https://significantproperties.kdl.kcl.ac.uk/>

2.2. METS file requirements

- Mets profile supports identification of the Content Type within
 - Package
 - Representation
- Vocabulary can be extended in the future

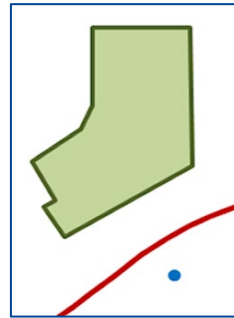
```

1 <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2 <!-- In this example IDs are carrying information to ease understanding - they look like paths to ease
3 <!-- CSIPSTR15 goes like this: We recommend including all schema documents for any structured metadat
4 <mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.loc.gov/METS/"
5     xmlns:xlink="http://www.w3.org/1999/xlink"
6     xmlns:csip="https://DILCIS.eu/XML/METS/CSIPExtensionMETS"
7     xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance schemas/XMLSchema.xsd http://www.loc.
8         http://www.w3.org/1999/xlink schemas/xlink.xsd
9         https://DILCIS.eu/XML/METS/CSIPExtensionMETS schemas/CSIPExtensionMETS.xsd"
10    OBJID="IP_18002_VectorGeodata_1Rep" TYPE="Geospatial Data" csip:CONTENTINFORMATIONTYPE="GeoData"
11    PROFILE="https://GEO.dilcis.eu/profile/CITS_Geodata.xml" csip:OAISPACKAGETYPE="SIP">
12 <!-- CSIP1 - mets/@OBJID -->
    
```

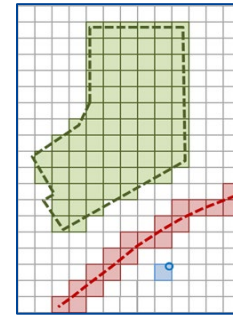


2.3. Data Requirements

- General data requirements
 - Vector requirements
 - Raster requirements
- Long term Preservation format profiles

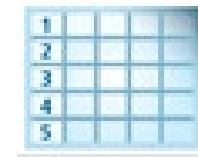
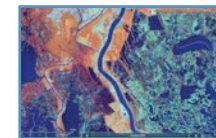


Vector



Raster

ID	Name, Location & Description	Card & Level
GEO_11	Minimum one file in a geospatial format If the value in mets/@csip: CONTENTINFORMATIONTYPE is "citsgeospatial_v3_0 ", then there SHOULD exist at least one file in a geospatial format in representations/[RepresentationName]/data	0..n SHOULD
GEO_12	Subfolders in data representations/[RepresentationName]/data If there are more geospatial records in a representation, each geospatial file MAY be placed or grouped in subfolders in representations/[RepresentationName]/data	0..n MAY
GEO_13	Long term preservation format representation The Information Package SHOULD contain at least one representation of geospatial record in a long-term preservation format, as defined by the Archive or in the Long-term Preservation Format Profile (See chapter 3.3.5.)	0..n SHOULD

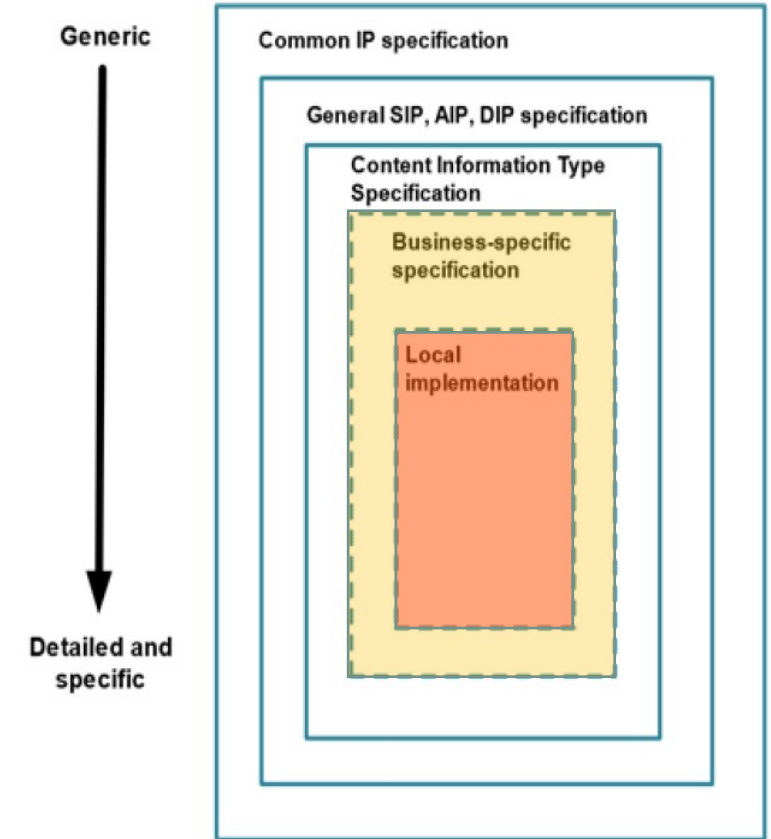


???? GML, SHP, KML, TIFF, JPEG2000, .gpkg, csv ????

2.3.5. Why Long-term Preservation format profiles?

- Geospatial records vary in type (formats) and in content
- Local implementations might require specific validation rules.
- Examples provided as Appendix to Guideline:
 - Profile for Geospatial Vector data using GML 3.2.1.
 - Profile for Geospatial Raster data using TIFF baseline 6

ID	Description of requirement	M/O
D_5.2-1	<p>GML files larger than 1 GB MUST be subdivided into smaller GML files</p> <p>Recommendation <i>It is recommended that GML files larger than 1 GB are subdivided into smaller GML files because GML files larger than 1–2 GB are impossible to produce, test, correct or visualise in a GIS.</i></p>	M
D_5.2-2	Geometries and attributes from the same geospatial vector dataset SHOULD be kept together within the same GML file	O



2.4. Documentation requirements

- Requirements on how to document the geodata in the IP
 - Structure of geospatial records
 - Rendering and visualization
 - Behaviour
 - Coordinate reference systems
 - Other

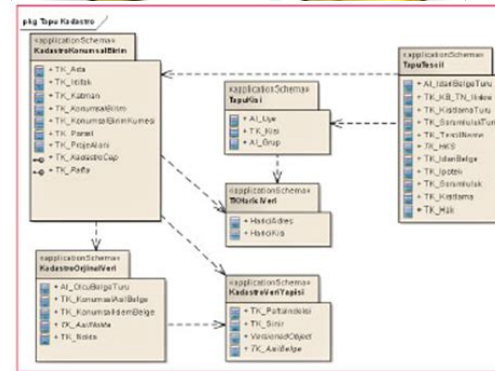
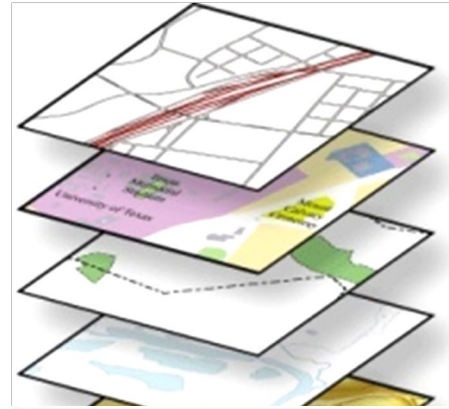
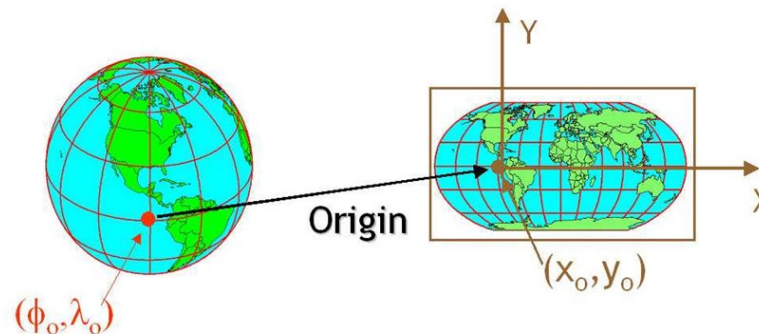


Figure 3: Data groups (packages) of the application schema



Source: <http://ayresriverblog.com>

	Point	Line	Polygon	Associative	Ordered	Quantitative	Selective
Size	● ● ●	— — —	■ ■ ■		●	●	●
Value	● ● ●	— — —	■ ■ ■		●		●
Texture	● ● ●	— — —	■ ■ ■	●	●		○
Color	● ● ●	— — —	■ ■ ■	○			●
Orientation	— — —	— — —	■ ■ ■	●			○
Shape	● ● ●	— — —	■ ■ ■	●			


● Strong
○ Weak

Ali, Amr. (2017). Framework Development of Cybercartography for Mobile Environment. 6. 14-25. 10.5923/j.ajgis.20170601.02.



2.5. Geospatial Metadata

- Geospatial metadata enhances findability and reuse
- List of mandatory elements
- Support for standardized machine readable data
 - ISO 19115-1
 - ISO 19115-2
 - ISO 19165-2
- Support for proprietary machine-readable metadata



INSPIRE GEOPORTAL

Enhancing access to European spatial data

Data set Metadata ▲

Resource Title
Natura 2000 🗺️

Resource Abstract
Fulfilment of the obligation under Article 5 of the Decree on Special Protection Areas (Natura 2000 Areas) and informing the public of NATURA 2000 sites. The data set is part of the inventory of sites of importance for biodiversity conservation. Natura 2000 is a European network of Special Protection Areas (SPAs) proclaimed in the Member States of the European Union with the basic objective of preserving biodiversity for future generations. Special areas of conservation are therefore intended for the conservation of animal and plant species and habitats that are rare or endangered on the European level owing to human activity. The data set is official data for valid NATURA 2000 sites adopted by the Decree on Special Protection Areas (Natura 2000 areas) (Official Gazette of the Republic of Slovenia Nos 49/04, 110/04, 59/07, 43/08, 8/12, 33/13, 35/13 – Corr., 39/13 – Decision, US, 3/14, 21/16 and 47/18). The data set shows the NATURA 2000 sites designated under the Birds Directive (Council Directive 79/409/EEC of 2 April 1979 on the conservation of wild birds) – SPAs, and the Habitats Directive (Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora) – pSCI/SAC sites. 🗺️

Lineage
Na podlagi biogeografskega seminarja, ki je bil junija 2014 v Ljubljani (zaključki), je Evropska komisija 3. decembra 2014 potrdila slovenski predlog območij. Odločitev o posodobitvi seznama območij v alpski biogeografski regiji in v celinski biogeografski regiji je Komisija objavila 23. januarja 2015 v Uradnem listu EU. Vlada je območja NATURA 2000 potrdila z Uredbo o posebnih varstvenih območjih (območjih Natura 2000).

Unique Resource Identifier
Code: A4BB6417-4C82-44FF-801A-9590224AEB8F
Namespace: SI.ARSO.NATURA

Spatial Data Theme
[Zavarovana območja](#)


Topic Category
environment
biota

Reporting Tags
Priority Dataset
Spatial Scope

Conditions Applying To Access And Use
Obvezna navedba vira: Vir: Agencija RS za okolje ali vir: ARSO

Limitations On Public Access
<http://inspire.ec.europa.eu/metadata-codelist/LimitationsOnPublicAccess/noLimitations>

Geographic Bounding Box



Leaflet | Credits: © OpenStreetMap contributors | EC-GISCO, © EuroGeographics for the administrative boundaries (disclaimer)

Responsible Party
Organisation name
Agencija RS za okolje
E-mail gp.arso@gov.si

Metadata Point Of Contact
Organisation name
Agencija RS za okolje
E-mail ursa.mezan@gov.si

Metadata Language
slv

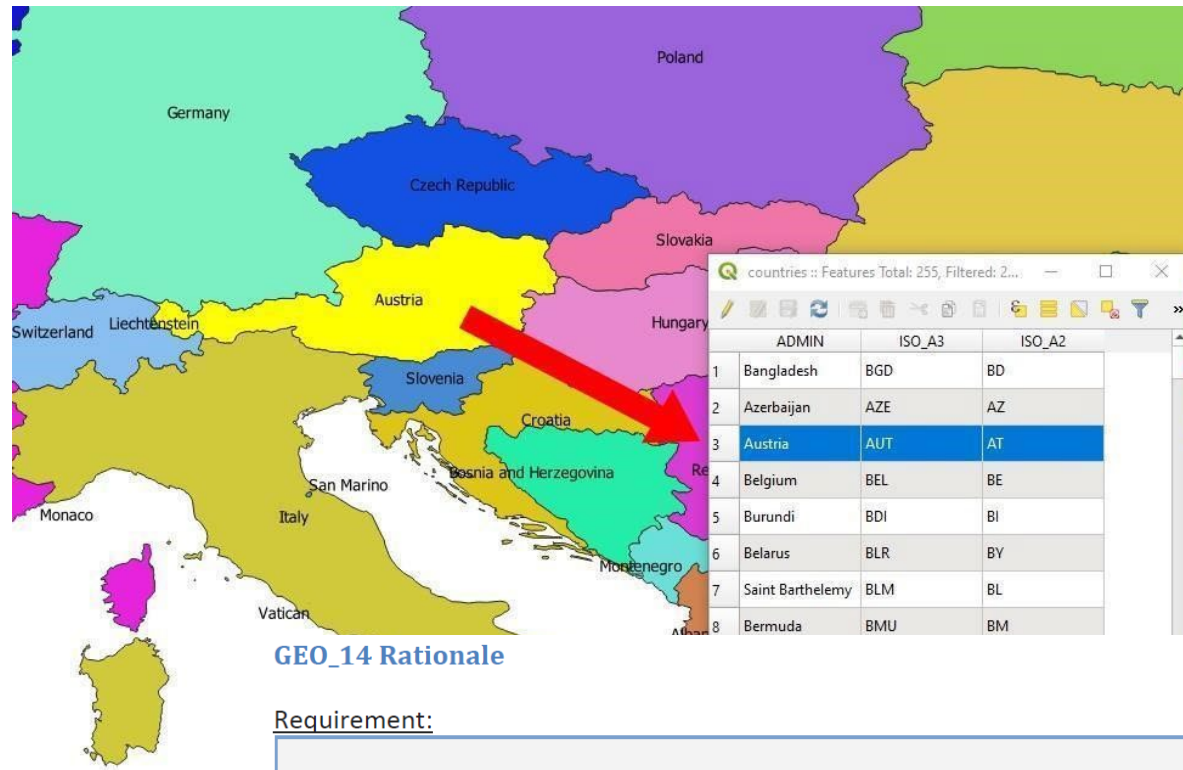
Metadata Date
2020-05-12T10:29:49

fileIdentifier
c93002fa-8064-4b7c-866b-2648ca1c403e

Download metadata
[application/vnd.iso.19139+xml](#)

Guidelines for CITS Geospatial

- Introduction to Geospatial records
- Introduction to “Significant Properties” concepts for Geodata
- Rationalizations of all CITS Requirements
- Examples for many requirements



GEO_14 Rationale

Requirement:

GEO_14	Original format representation	The Information Package MAY contain a separate representation of the same data, containing geospatial data in its original format	0..1 MAY
--------	--------------------------------	--	-------------

Description:

This requirement allows an additional representation in the IP with the geospatial data in the original format.

Example:

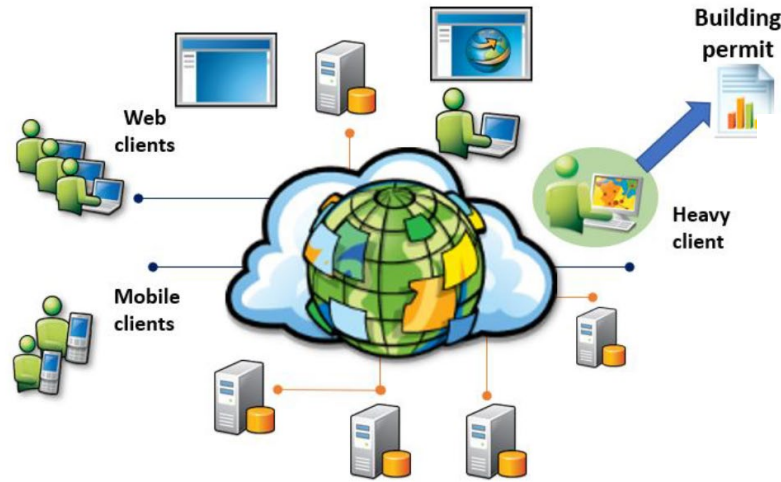
Figure 2 in chapter 3.1.2 shows an IP with two representations. One representation contains a Long-Term Preservation vector data format (GML321), and the other contains a representation of the original format in an ESRI shapefile format (SHP).

Rationale:

Original formats are often richer and easier to use than the preservation format and suitable for dissemination in the short term. However, it does not ensure the long-term preservation of the data. Geospatial data in original format can also be used for validation on submission mitigating loss of data and significant properties during migration to preservation format. The idea is that the users could use this representation until the original formats becomes obsolete.

Guidelines for CITS Geospatial **with GIS**

- Introduction to GIS Systems
- GIS Preservation strategies
- Rationalizations of all CITS Requirements
- With Examples based on standards
 - OWS Context
 - WKT2 for CRS
 - SLD
 - Etc.



GEO_33 Rationale

Requirement:

GEO_33	Rendering configuration	A standardised machine-readable rendering configuration for one or more geospatial datasets MAY be provided in the Information Package	0..n MAY
GEO_33a Ref GEO_33	Placement of rendering configuration	If a standardised machine-readable rendering configuration for one or more geospatial datasets exists, it SHOULD be provided in representations/[RepresentationName]/documentation/rendering	0..n SHOULD

Description:

This requirement recommends that rendering configurations are documented in a standardised machine-readable format to support dissemination automation.

Example:

An example of Standardised machine-readable formats for the rendering of geospatial records are SLD²⁰ files. KML²¹ files also have some of that capability:

SLD files example

SLD is an OGC²² (Open Geospatial Consortium) standard for symbology and is the OGC Styled Layer Description XML format (SLD files). If the producer cannot provide the archive with SLD files, these can be recreated from the description provided in the **Documentation** in an open-source GIS application like QGIS²³. Raster files can have a colour map associated with the pixel value. The SLD standard is used for rendering geodata in OGC web services and, therefore, could be used as an appropriate input for an easier DIP creation in the future. An example of an SLD file is shown in figure 13.

```
<StyledLayerDescriptor xmlns="http://www.opengis.net/sld"
  xmlns:ogc="http://www.opengis.net/ogc"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  version="1.0.0"
  xsi:schemaLocation="http://www.opengis.net/sld StyledLayerDescriptor.xsd">
  <NamedLayer>
    <Name>Simple Point</Name>
    <UserStyle>
      <Title>SLD Cook Book: Simple Point</Title>
      <FeatureTypeStyle>
        .
      </FeatureTypeStyle>
    </UserStyle>
  </NamedLayer>
</StyledLayerDescriptor>
```



Summary

New CITS for Geospatial is

- More extendable
- Supports multiple standards and formats
- Guidelines for easier implementation

What is next:

- Collaboration is key
- More user needs



Digitalbevaring.dk

Questions? Answers in the end

Gregor Završnik
gregor@greoarh.si

E-ARK Programme

LinkedIn: www.linkedin.com/groups/8343650/

Twitter: #EARKProject

Ready to get started?

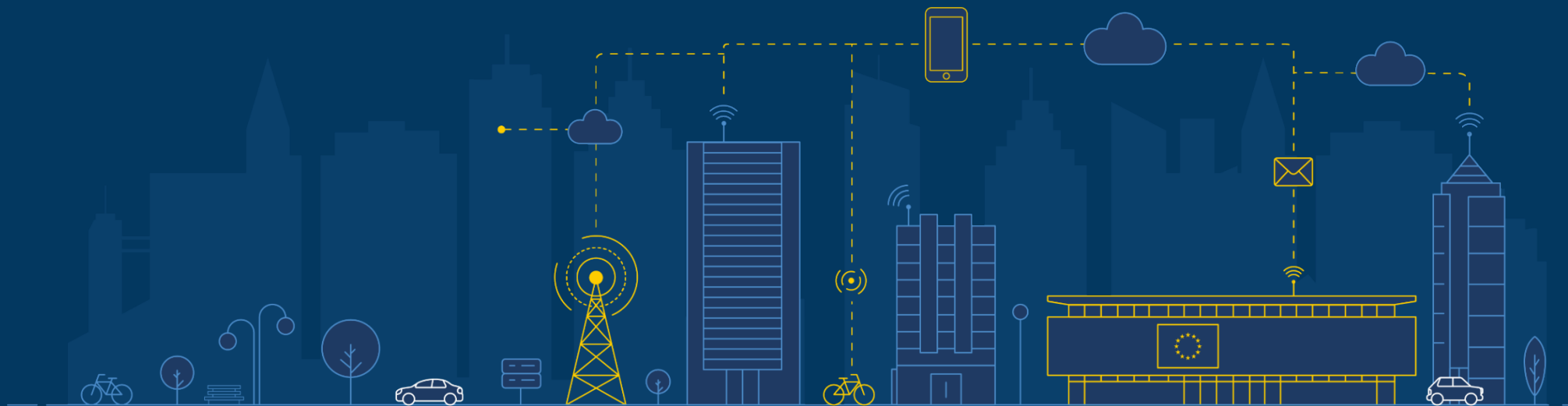
Find out more at:

ec.europa.eu/cefdigital

Contact us:

cef-building-blocks@ec.europa.eu

Thank you!



DIGITAL PRESERVATION

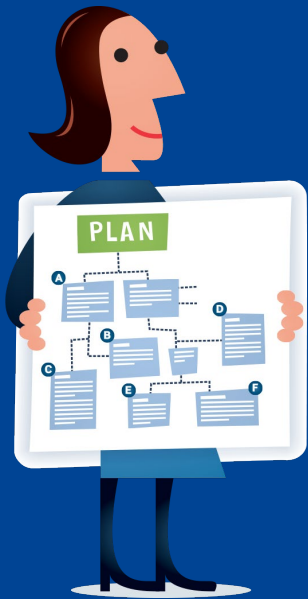
CITS SIARD and SIARD standard



Anders Bo Nielsen, Danish National Archives

Agenda

CITS SIARD and SIARD



- CITS SIARD and SIARD - relational database archiving
- What is CITS SIARD and what is SIARD?
- SIARD
 - based on open standards, contains its own database metadata, table spec. and table content
- Development of the SIARD file format
- Large objects - ISO SQL/MED
- SIARD 2.2
- CITS SIARD

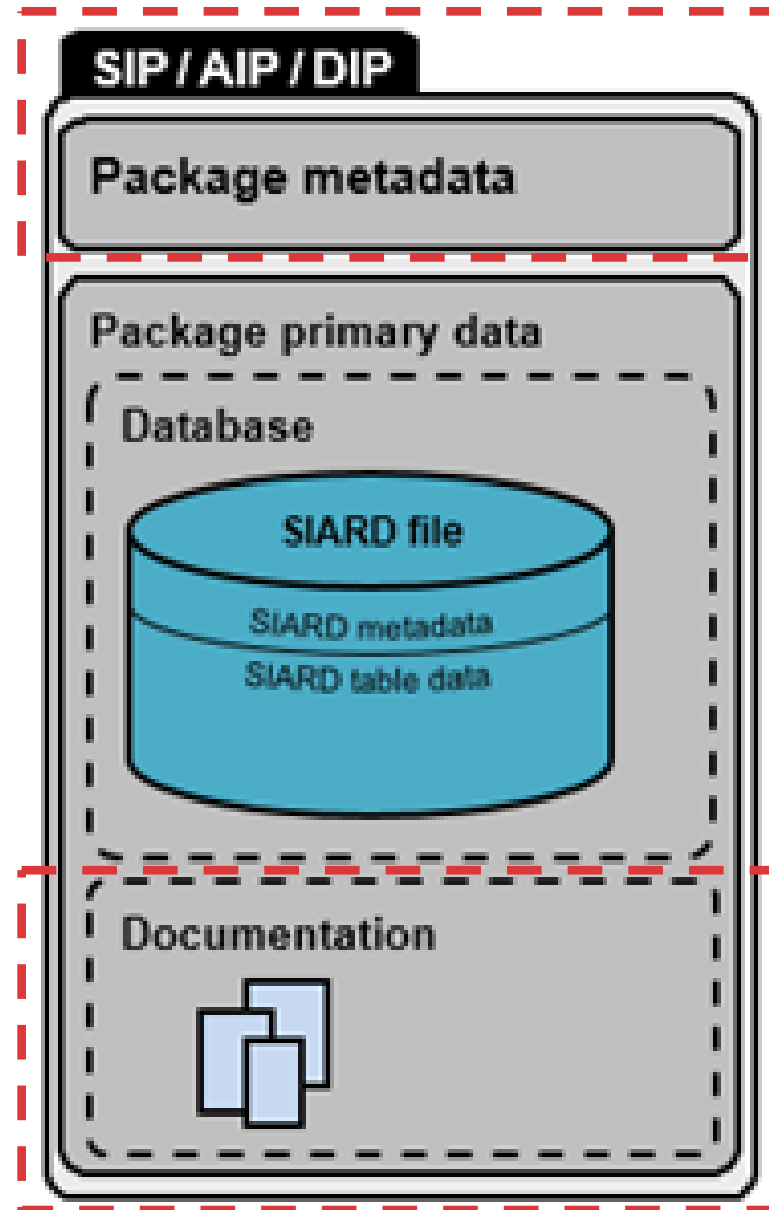
- Relational Database Archiving Interest Group

CITS SIARD

...is a packaging specification



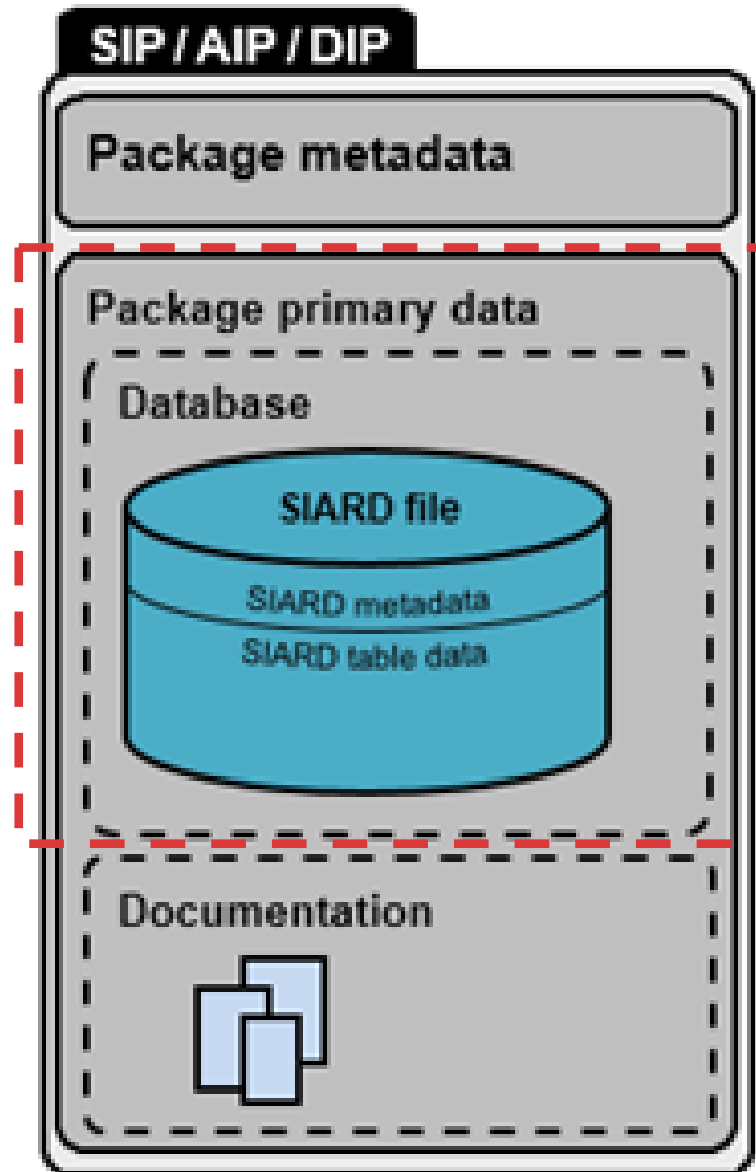
www.digitalbevaring.dk



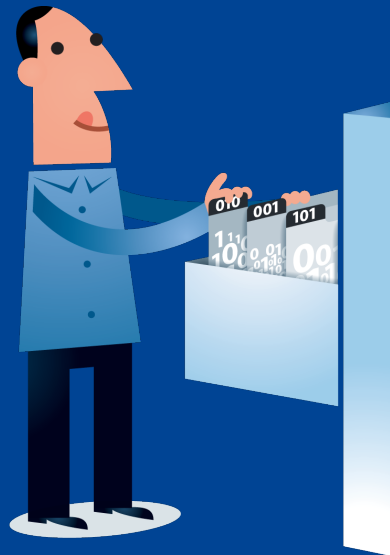
SIARD

Is an open file format for archived RDB

An archived relational database in SIARD format contains its own metadata describing the document more precisely in addition to the actual table data



Relational database archiving - why we need SIARD



- The relational database (RDB) is still the most widespread and dominant form of organising structured digital information
- Limited interoperability of proprietary RDBMS
- Proprietary SQL dialects despite ISO SQL
- Need for persistent data format outside the RDB Management System (RDBMS)

What is SIARD?

Software Independent Archival of Relational Databases

SIARD-2.1.1-Formatspezifikation

Seite 1 von 75

Name	SIARD-2.1.1-Formatspezifikation
Kategorie	Standard
Rolle/grad	Implementiert
Version	2.1.1
Status	Stabile Version
Beschluss am	2019-05-15
Ausgabedatum	2019-05-15
Ersetzt Version	eCH-0166 Version 2.1
Voraussetzungen	Keine
Beilagen	metadata.xsd, ech-0165_oe.siard ¹
Sprachen	Deutsch (Original), Französisch (Übersetzung), Englisch (Übersetzung)
Autoren	<p>Marcel Büchler, Schweizerisches Bundesarchiv, marcel.buechler@bar.admin.ch</p> <p>Luis Faria, KEEP SOLUTIONS, LDA, lffaria@keep.pt</p> <p>Bruno Ferreira, KEEP SOLUTIONS, LDA, bferreira@keep.pt</p> <p>Anders Bo Nielsen, Danish National Archives (Rigsarkivet), abn@sa.dk</p> <p>Krystyna Ohnesorge, Schweizerisches Bundesarchiv, krystyna.ohnesorge@bar.admin.ch</p> <p>Claire Röthlisberger-Jourdan, KOST, claire.roethlisberger@kost.admin.ch</p> <p>Hartwig Thomas, Enter AG, hartwig.thomas@enterag.ch</p> <p>Andreas Voss †, Schweizerisches Bundesarchiv</p>

¹Beitrag zum OC-Besitzdatenbank von Ombac.

info@kost-oe.ch
<https://kost-oe.ch/>

SIARD-2.1.1 – Spécification de format

Page 1 de 75

Titre	SIARD-2.1.1 – Spécification de format
Type	Norme
Stade	Implémenté
Version	2.1.1
Statut	Approuvé
Validation	2019-05-15
Date de publication	2019-05-15
Remplace	eCH-0166 version 2.1
Dépendances	Aucune
Annexes	metadata.xsd, ech-0165_oe.siard ¹
Langues	Allemand (original), français (traduction), anglais (traduction)
Auteurs	<p>Marcel Büchler, Archives fédérales suisses, marcel.buechler@bar.admin.ch</p> <p>Luis Faria, KEEP SOLUTIONS, LDA, lffaria@keep.pt</p> <p>Bruno Ferreira, KEEP SOLUTIONS, LDA, bferreira@keep.pt</p> <p>Anders Bo Nielsen, Danish National Archives (Rigsarkivet), abn@sa.dk</p> <p>Krystyna Ohnesorge, Archives fédérales suisses, krystyna.ohnesorge@bar.admin.ch</p> <p>Claire Röthlisberger-Jourdan, CESCO, claire.roethlisberger@kost.admin.ch</p> <p>Hartwig Thomas, Enter AG, hartwig.thomas@enterag.ch</p> <p>Andreas Voss †, Archives fédérales suisses</p>

¹Basé sur le modèle de base de données OC d'Ombac.

info@kost-oe.ch
<https://kost-oe.ch/>

SIARD Format Specification

Page 1 of 90

Name	SIARD-2.2 Format Specification
Category	Standard
Maturity level	Approved
Version	2.2
Status	Reviewed version
Approval date	2021-08-31
Issue date	2021-08-31
Replaces	SIARD-2.1.1
Prerequisite	None
Attachments	metadata.xsd
Languages	English, German (awaits translation), French (awaits translation), Italian (awaits translation)
Publisher / distributor	DILCIS Board, https://dilcis.eu/ Swiss Federal Archives, https://www.bar.admin.ch/

Summary

This document contains the specification for the SIARD file format version 2.2. SIARD stands for Software Independent Archival of Relational Databases. The format version 1.0 was developed by the Swiss Federal Archives. It is a normative description of a file format for the long-term preservation of relational databases.

The SIARD format is based on standards including the ISO standards Unicode, XML, and SQL:2008, the URI Internet standard, and the industry standard ZIP. The aim of employing internationally recognised standards is to ensure the long-term preservation of, and access to, the widely used relational database model, as well as easy exchange of database content, independent of proprietary "dump" formats.


SIARD 2.2 <https://dilcis.eu/content-types/siard> <https://www.bar.admin.ch>



Development of the SIARD file format


2000 - 2004

ARELDA

 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

2008


SIARD 1.0

 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



2014 - 2017

SIARD 2.0


 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra




E-ARK


2018

SIARD 2.1

 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

2018 - 2019

SIARD 2.1
Review 


 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



E-ARK4ALL

2019-2021

SIARD 2.2 

 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

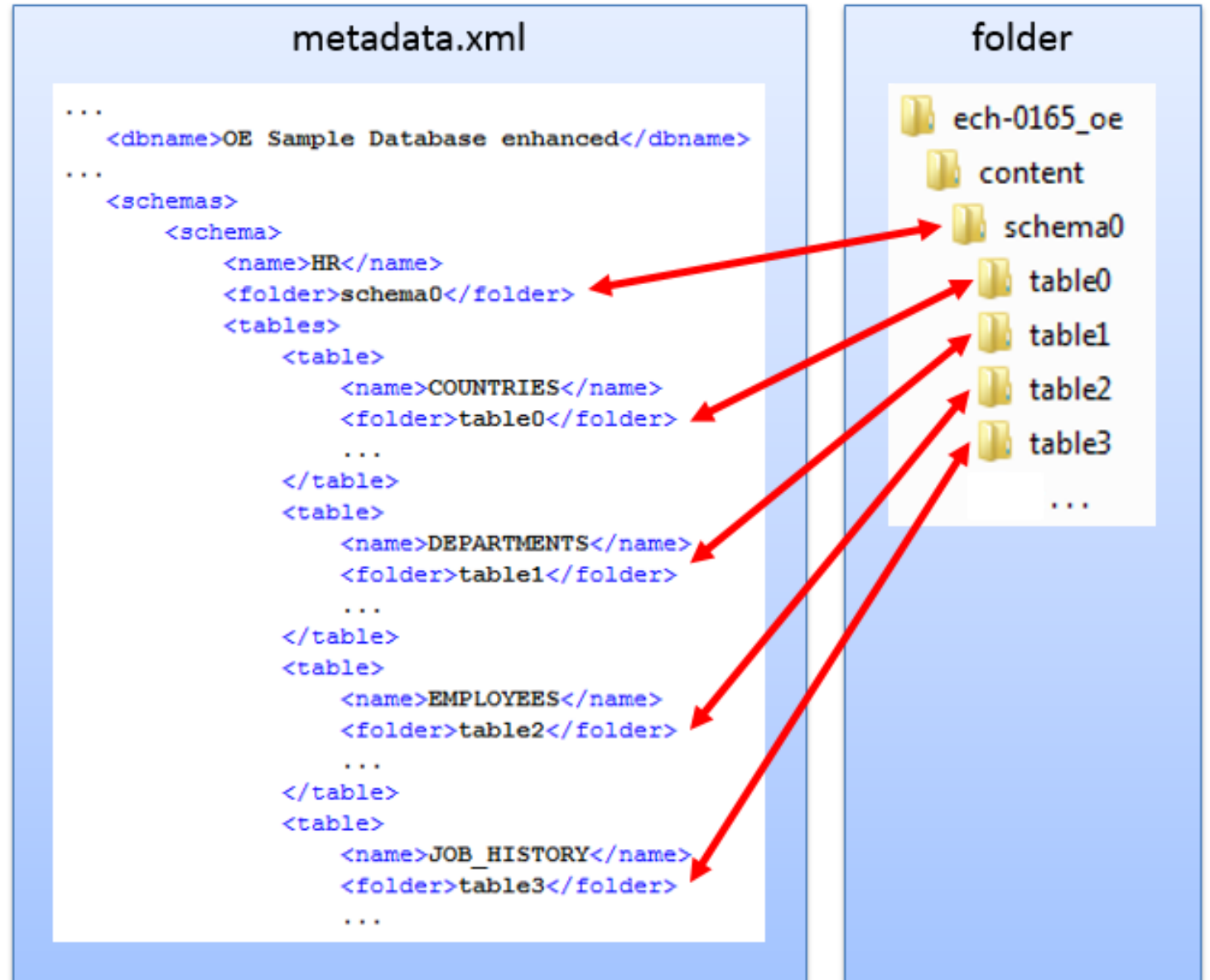


E-ARK3

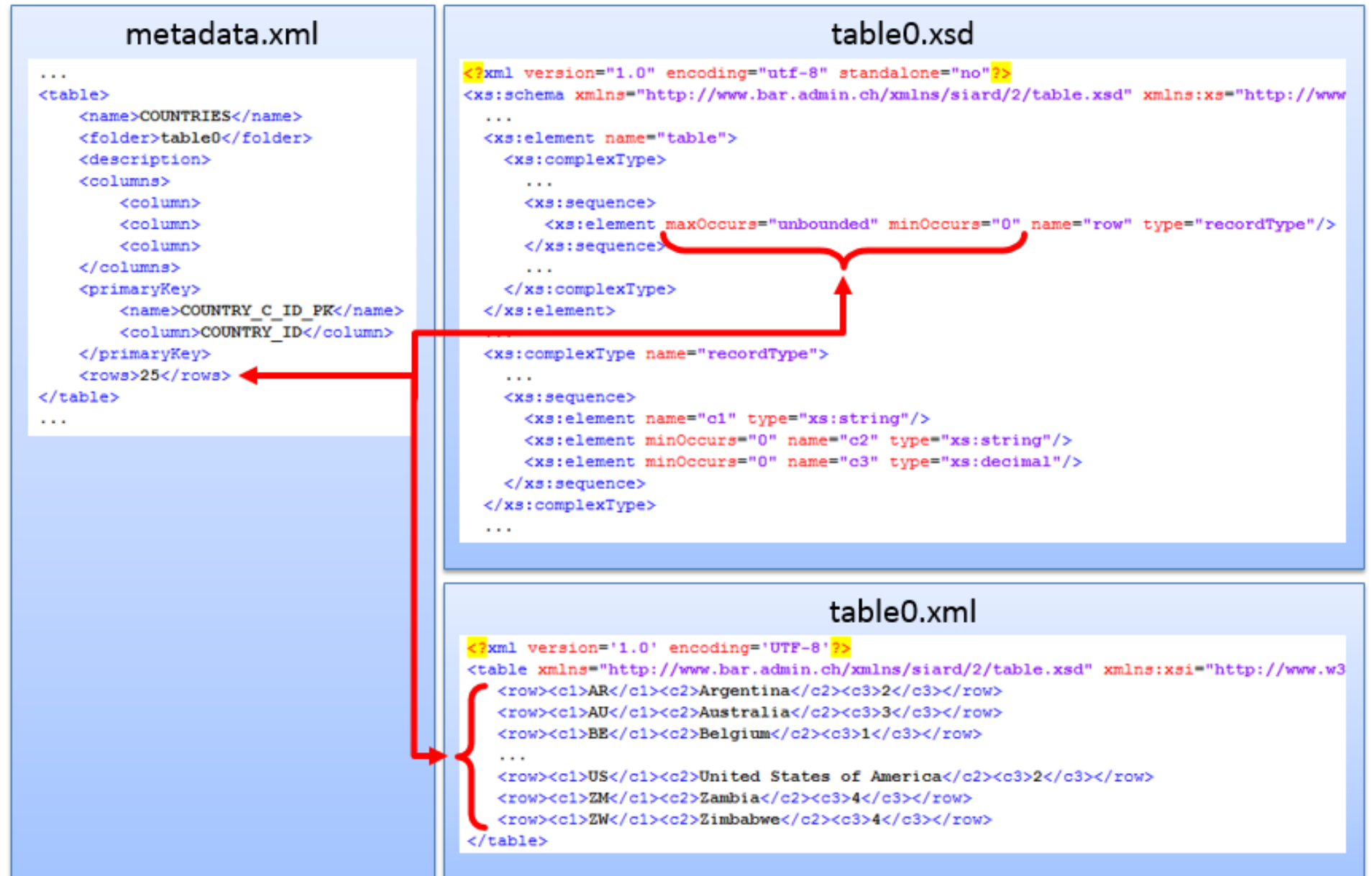
CITS SIARD

SIARD metadata and data

- Metadata.xml refer to folders inside the SIARD file



SIARD database metadata, table spec. and table content



Large objects ISO SQL/MED

- Large object (LOB) is the common name for Binary Large Object (BLOB) and Character Large Object (CLOB). BLOB is content such as video, sound, images, word processing documents etc., and CLOB is text content.

- Support for internal LOBs (ISO/IEC 9075-2:2008 - BLOBS) in SIARD 2.2
 - The SIARD 2.2 format specification supports the SQL:2008 method for using internal LOBS (ISO/IEC 9075-2:2008), as did SIARD 1.0 (SQL:1999).
 - The SIARD 2.2 format supports LOBs stored as files inside the SIARD file and describes this in detail in the SIARD 2.2 format specification (similar to SIARD 1.0).
 - The SIARD 2.2 format supports LOBs stored as files outside the SIARD file (a new feature in SIARD 2.0) and specifies the details in this specification.
- Support for external files (ISO/IEC 9075-9:2008 – SQL/MED) in SIARD 2.2

SIARD 2.2

- Requirements for the folder structure for LOBs stored outside the SIARD file
- Segmenting LOBs and large tables
- An example

```
Northwind.siard <!-- packaged as a ZIP file -->
```

```
  content/  
  header/  
    metadata.xml  
    metadata.xsd  
    siardversion/  
      2.2/
```

```
Northwind_lobs/
```

```
  s0_t2_c4/  
    seg_0/  
      t2_c4_r1.bin  
      t2_c4_r2.bin  
      t2_c4_r3.bin  
      t2_c4_r4.bin
```

```
  seg_1/ <!-- folder file number limit -->
```

```
    t2_c4_r5.bin  
    t2_c4_r6.bin  
    t2_c4_r7.bin
```

```
  seg_2/ <!-- folder file size limit -->
```

```
    t2_c4_r8.bin
```

```
  s0_t2_c8/  
    seg_0/  
      t2_c8_r3.bin
```

```
  s0_t11_c6/  
    seg_0/  
      t11_c6_r7.bin
```

SIARD 2.2

- The previous example for metadata.xml and table2.xml

metadata.xml

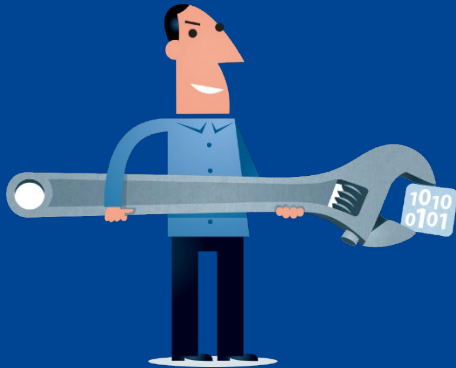
```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>...  
<siardArchive>...<lobFolder>./Northwind_lob/</lobFolder>  
..  
<column>...<lobFolder>s0_t2_c4/</lobFolder>...</column>  
<column>...<lobFolder>s0_t2_c8/</lobFolder>...</column>  
..  
<column>...<lobFolder>s0_t11_c6/</lobFolder>...</column>
```

table2.xml

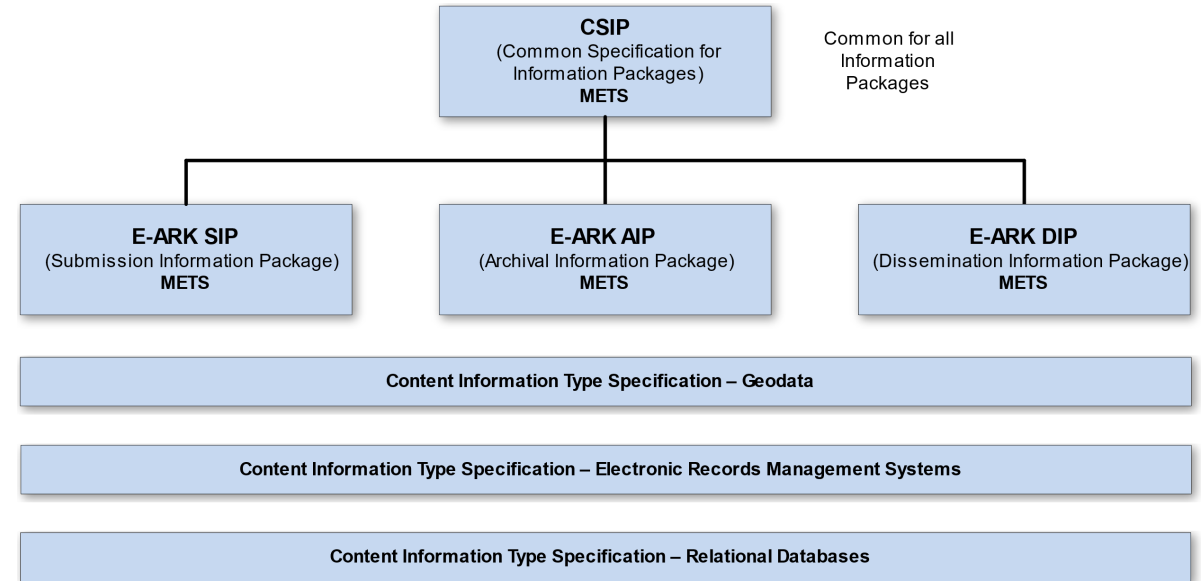
```
<row><c1>1</c1><c2>Beverages</c2><c3>Soft drinks, coffees, teas, beers, and ales</c3>  
<c4 file="seg_0/t2_c4_r1.bin" ... /></row>  
..  
<row><c1>5</c1><c2>Seafood</c2><c3></c3>  
<c4 file="seg_1/t2_c4_r5.bin" ... /></row>  
..  
<row><c1>8</c1><c2>Candy</c2><c3></c3>  
<c4 file="seg_2/t2_c4_r8.bin" ... /></row>
```

CITS SIARD

... is a brand new specification



www.digitalbevaring.dk



<https://earkcsip.dilcis.eu/>

eArchiving Webinar Series 2020:

<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving+webinar+Series+2020>

Special interest: "CSIP", "DBPTK"

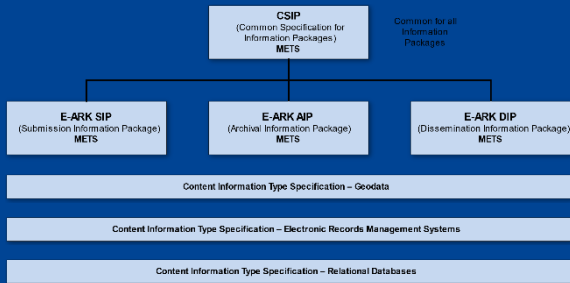
eArchiving Webinar Series 2021:

<https://ec.europa.eu/eusurvey/runner/eArchivingwebinars2021>

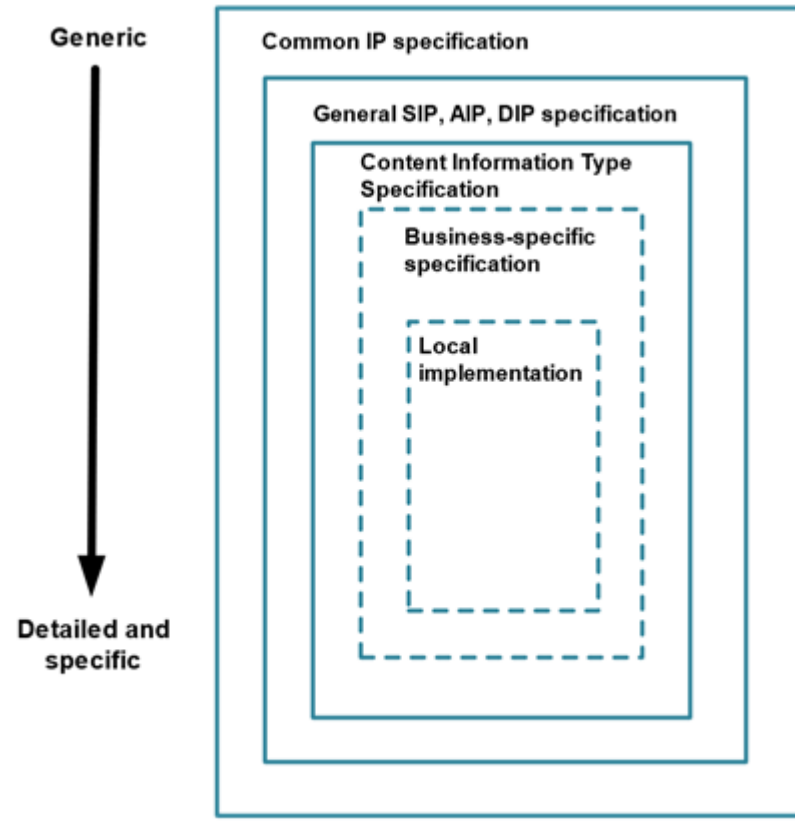
Special interest: "Putting the CITS content in the CSIP package", "Submission Agreements", "E-ARK validation: What's inside the Package?"

CITS SIARD

... in the specification landscape



<https://earksip.dilcis.eu/>



<https://earksip.dilcis.eu/>

CITS SIARD

... can be found at:



“The Digital Information LifeCycle Interoperability Standards Board (DILCIS Board) is an international group of experts committed to maintain and sustain maintain 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.”

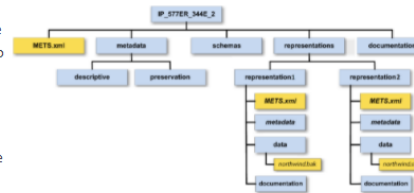
<https://dilcis.eu/content-types/cs-siard>



CITS SIARD

The CITS SIARD (Content Information Type Specification for Relational Databases using SIARD) is a specification that describes how to package and preserve relational database content. This is primarily done by packaging SIARD files into information packages that conform to the [Common Specification for Information Packages](#).

The specification helps you to apply a common way of storing multiple representations of a database (for example a proprietary backup and a SIARD snapshot) in a single package along with appropriate metadata and binary documentation of the dataset.



We welcome all feedback in regard to the SIARD CITS specification - if you have any comments or proposals, please contact us per [e-mail](#) or leave your comments on [GitHub](#)!

Relational Database Archiving Interest Group

The DILCIS Board and eArchiving Building Block also maintain a “Relational Database Archiving Interest Group” which documents and shares best practices on database archiving, the application of the SIARD and SIARD CITS specifications and related tools. If you are interested in joining the interest group please register at <https://forms.gle/b5Qny521G1QmPEK26>

For now the Interest Group has published two international case studies:

- [Case Study 1](#)
- [Case Study 2](#)

Download CITS SIARD

Latest version (v1.0.0, August 31, 2021)

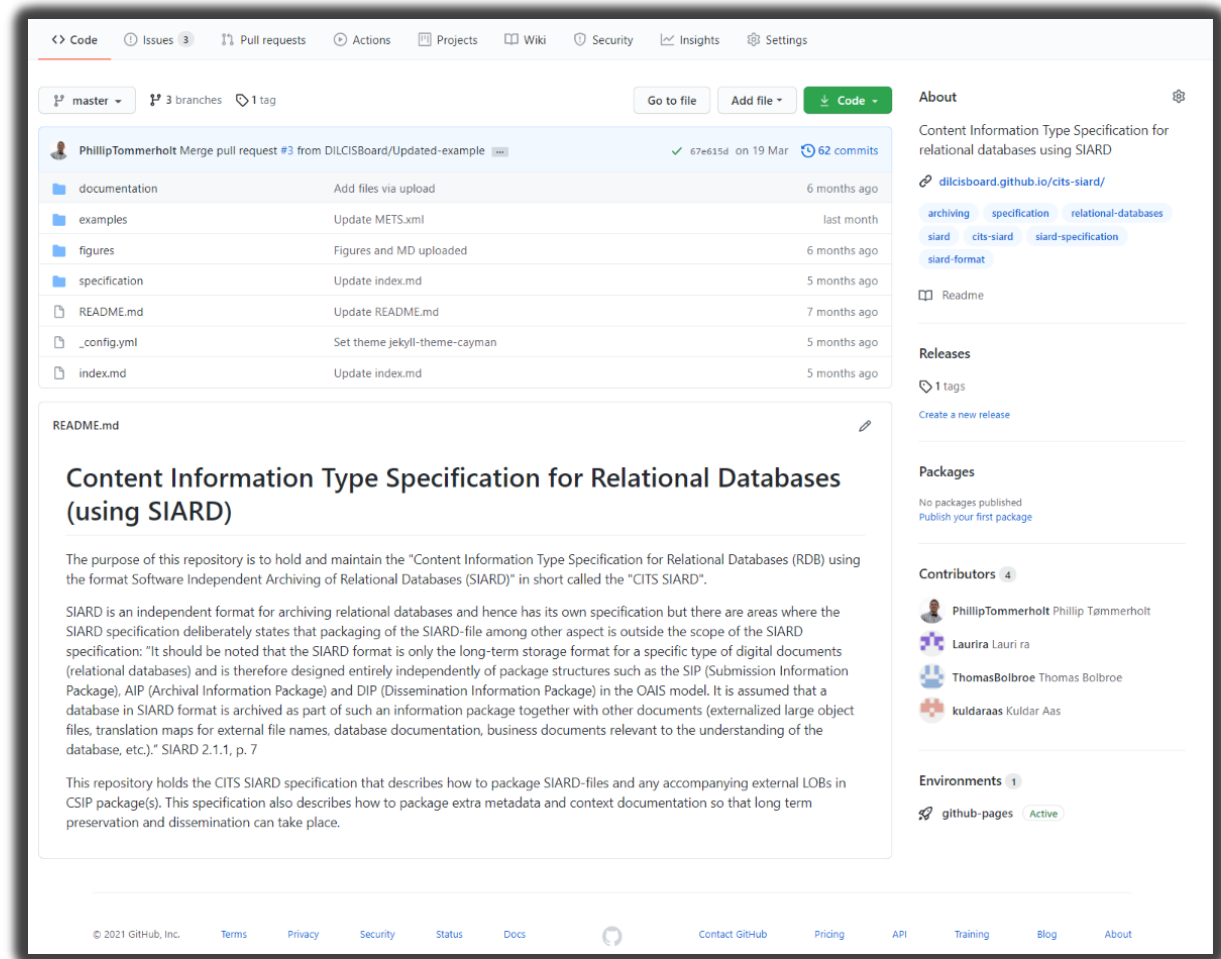
- [CITS SIARD v1.0.0](#)
- [E-ARK-SIARD-ROOT.xml](#)
- [E-ARK-SIARD-REPRESENTATION.xml](#)
- [Guideline_CITS_SIARD_1_0_0.pdf](#)

CITS SIARD

... can be found at:



<https://github.com/DILCISBoard/CITS-SIARD>



The screenshot shows the GitHub repository page for 'CITS SIARD' by the 'DILCISBoard' organization. The repository is on the 'master' branch and has 3 branches and 1 tag. A recent pull request by PhillipTommerholt is visible. The file list includes 'documentation', 'examples', 'figures', 'specification', 'README.md', '_config.yml', and 'index.md'. The 'README.md' file is open, showing the title 'Content Information Type Specification for Relational Databases (using SIARD)'. The README text describes the purpose of the repository and the SIARD format. The right sidebar shows repository statistics, including 62 commits, 1 tag, and 4 contributors.

File	Commit Message	Time Ago
documentation	Add files via upload	6 months ago
examples	Update METS.xml	last month
figures	Figures and MD uploaded	6 months ago
specification	Update index.md	5 months ago
README.md	Update README.md	7 months ago
_config.yml	Set theme jekyll-theme-cayman	5 months ago
index.md	Update index.md	5 months ago

Content Information Type Specification for Relational Databases (using SIARD)

The purpose of this repository is to hold and maintain the "Content Information Type Specification for Relational Databases (RDB) using the format Software Independent Archiving of Relational Databases (SIARD)" in short called the "CITS SIARD".

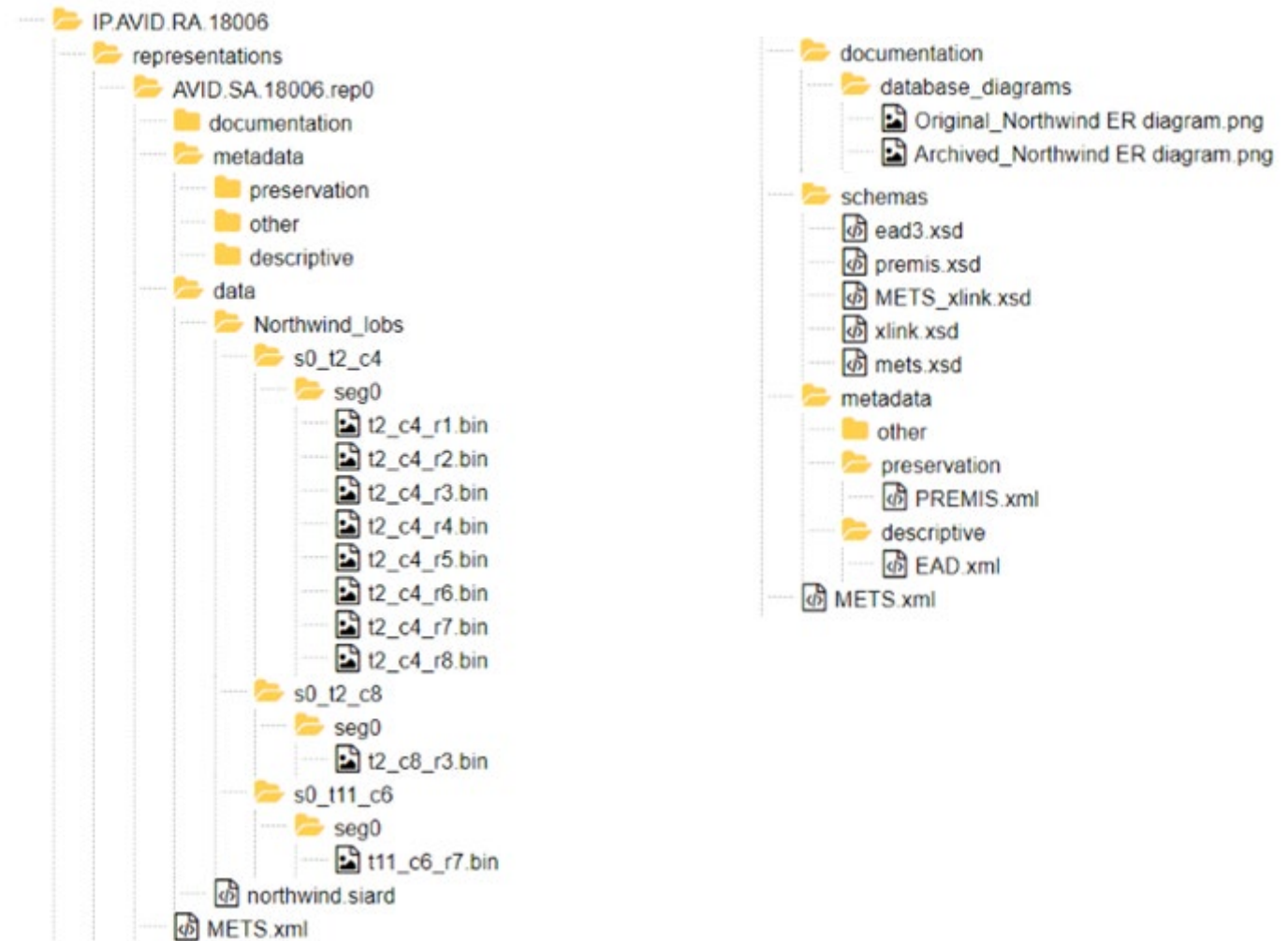
SIARD is an independent format for archiving relational databases and hence has its own specification but there are areas where the SIARD specification deliberately states that packaging of the SIARD-file among other aspect is outside the scope of the SIARD specification: "It should be noted that the 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 (externalized 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 repository holds the CITS SIARD specification that 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.

CITS SIARD

3.1 Folder structure and example

Folder Structure of Northwind Sample Database



CITS SIARD

3.2 Package METS and Representation METS

3.3 Package METS

3.4 Representation METS

3.5 METS requirements between Package and Representation

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

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

Relational Database Archiving Interest Group



Digitalbevaring.dk

- The DILCIS Board and eArchiving Building Block have initiated a “Relational Database Archiving Interest Group” which is expected to document and share best practices on database archiving, the use of the SIARD and SIARD CITS specifications, and related software.
- If you are interested in joining the interest group, please register at:
<https://listserv.dilcis.eu/review/rdb-aig:>
- See also
- <https://dilcis.eu/content-types/siard>
- <https://github.com/DILCISBoard/SIARD>



<https://www.london.gov.uk/what-we-do/mayors-office-policing-and-crime-mopac/data-and-statistics/london-landscape>

E-ARK Validation: Compliance with specifications



Carl Wilson, Open Preservation Foundation



Why E-ARK Validation?

OAIS defines **WHAT COULD** be in an Information Package, but says little about **WHAT SHOULD** be in there or **HOW** the package is arranged:

- There is little guidance for implementers;
- There is no basis for interoperability, information packages from different systems may have very little common ground.

E-ARK sets out to address these issues by defining a formal logical and physical structure for information packages.



The E-ARK validation model

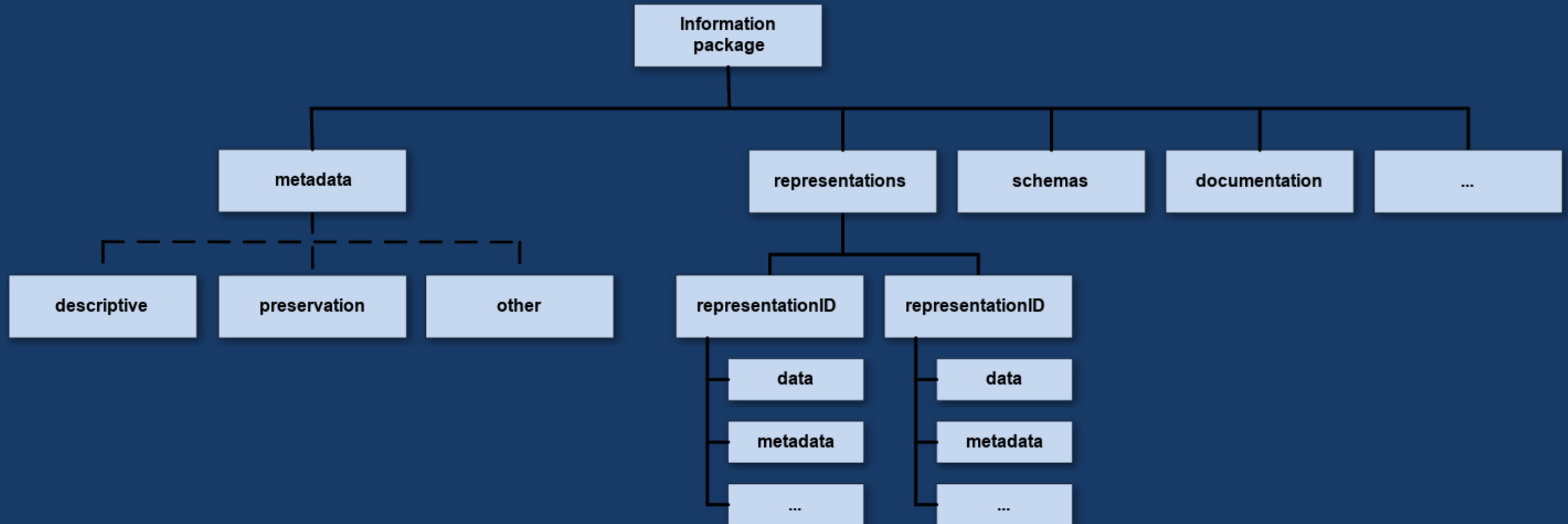
E-ARK separates IP validation into three logical phases:

- **Structure**
Sixteen rules that ensure that the package structure is valid, this covers the presence of specific directories and a main metadata file.
- **Metadata**
Schema validation of metadata files followed by an extended set of metadata checks, currently about 150 rules.
- **Integrity**
Checks the package manifest, ensures that all files are present, verifies any checksums and ensures that no “orphaned” files are present.

E-ARK IP validation

Structure

The E-ARK IP specifications define a standard physical structure:





E-ARK IP validation

Metadata

Metadata validation covers:

- Validation of XML METS files against the METS schema.
- Additional validation against the E-ARK extension schema for additional attributes defined by the project.
- Validation against an extended rule set that is considerably more prescriptive than the METS schema:
 - Published as a METS profile available from the specification sites;
 - Enforced through the use of XML Schematron.

E-ARK IP validation

Integrity

Integrity checking:

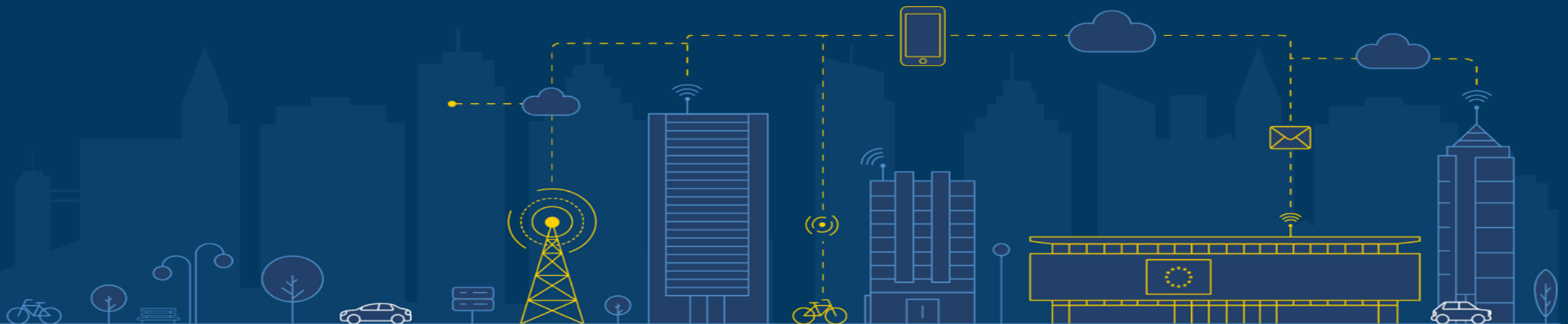
- Ensures all files listed in the metadata documents are present in the package;
 - Verifies the checksums of these files against those in the metadata, if present:
 - This can be time consuming for large packages;
 - Checksum validation is an optional step.
-
- Ensures that there are no orphaned files in the package:
 - An orphan is any file that is not referenced in the metadata somewhere;
 - Possible that the file does not belong in the package;
 - Alternatively, the metadata for the file has been omitted.





Why Schematron? Extensibility at runtime

- METS validation by the official schema is quite limited, a valid document can be VERY sparse;
- METS Profiles go beyond this BUT they aren't machine enforceable;
- Schematron uses XQuery and XPath to test for patterns and conditions in XML documents;
- The ruleset is extensible and can be switched at runtime, you could even write your own rules.



Support for developers of archival software

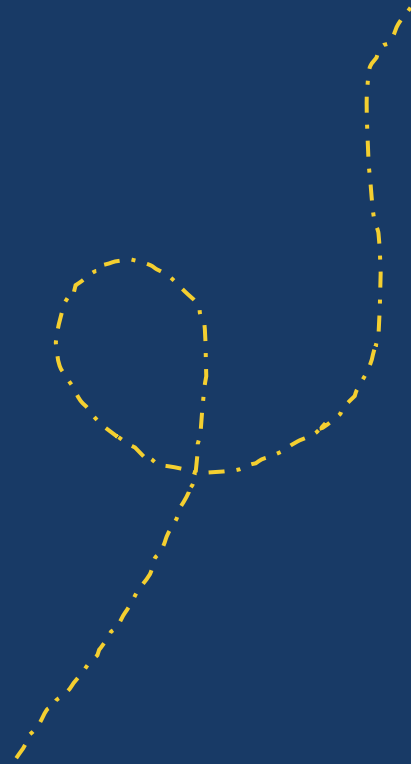


- Validation service for testing packages during development;
- Python library for local validation, e.g. pip install eark-ip-validation;
- Java library for similar, commons-ip;
- Docker images for deploying validation servers locally;
- REST API for validating information packages;
- Swagger definitions for generation of client SDK for REST service integration.

Getting started

There are a few easy-to-use online resources that will help you try out the validation software with sample packages:

- The quickest way to try the E-ARK validator is via the online web demonstrator: <https://pyip.openpreservation.org/>.
- The CEF [ISA2 Interoperability Test Bed](#), an online and self-service platform offered by DIGIT for the conformance testing of software against technical and semantic specifications;
- Sample packages that demonstrate individual errors for testing are available from: <https://dilcisboard.github.io/eark-ip-test-corpus/>.





Questions

Final Words



CEF Digital
Connecting Europe

