Specifications session

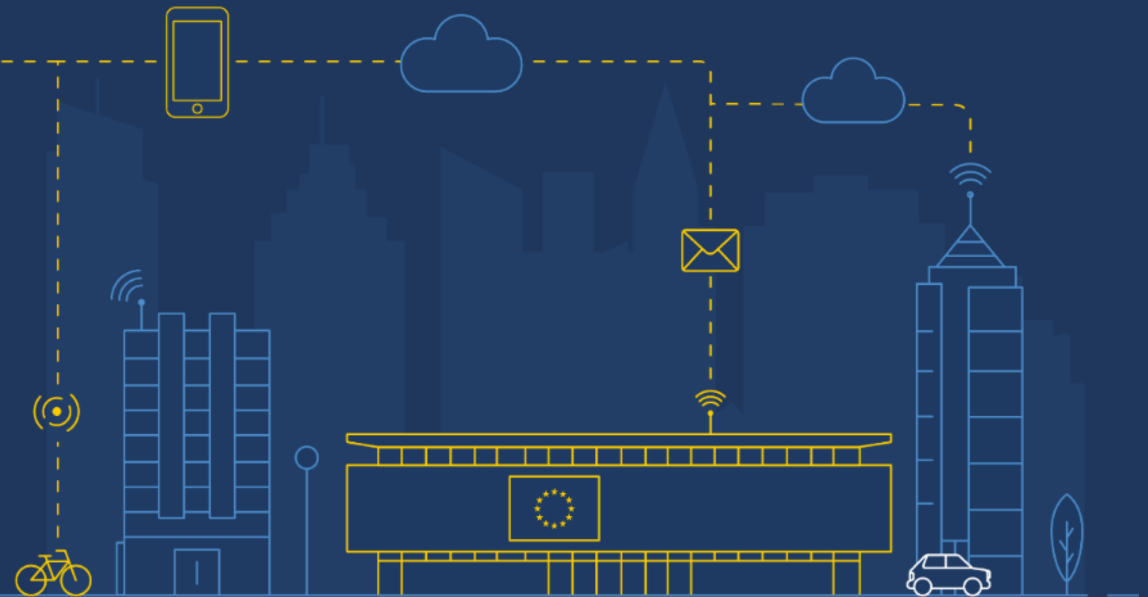
Specifications session

Practical use cases

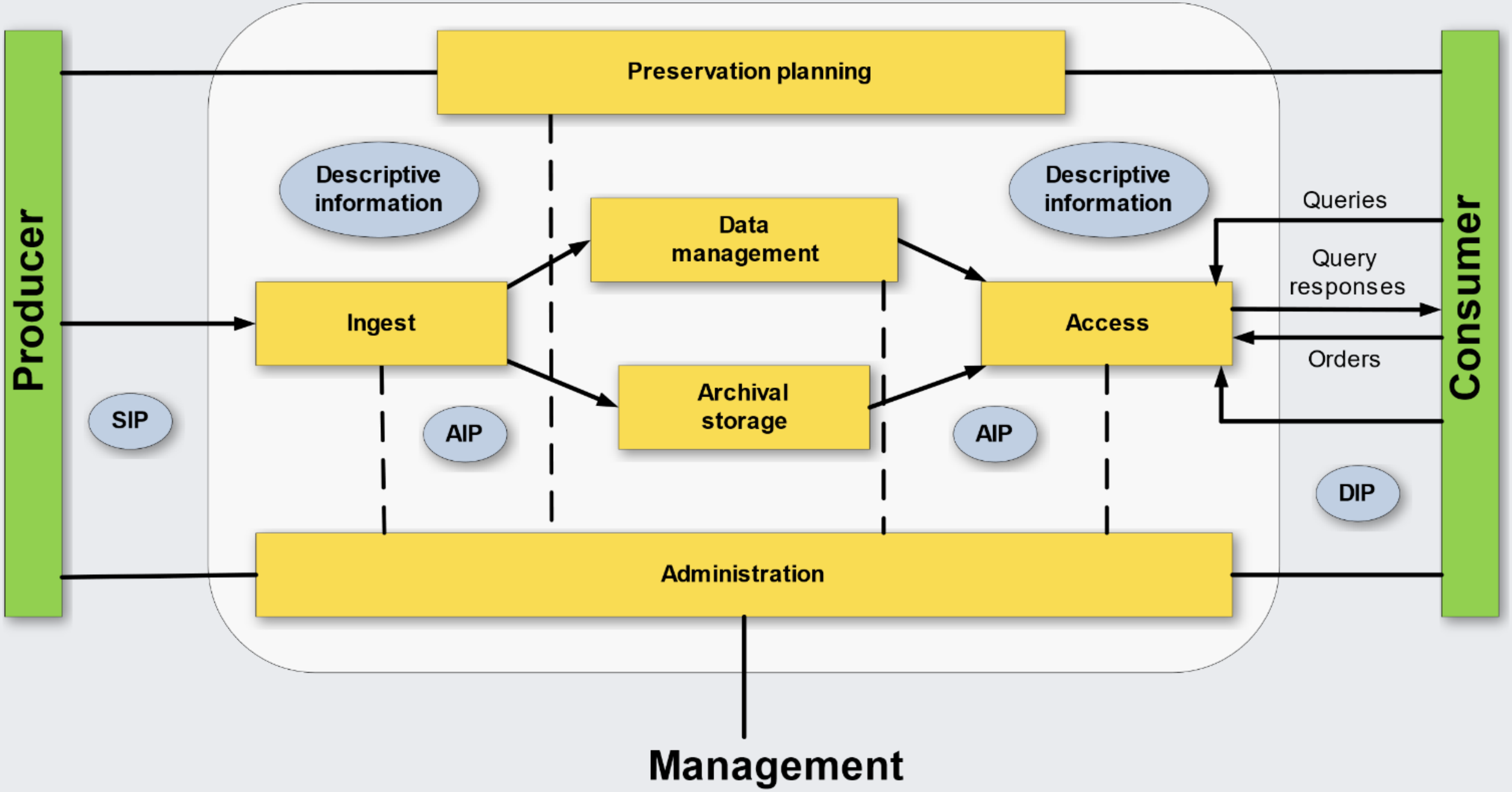
Conformance testing

Karin Bredenberg
Metadata Strategist
Sydarkivera

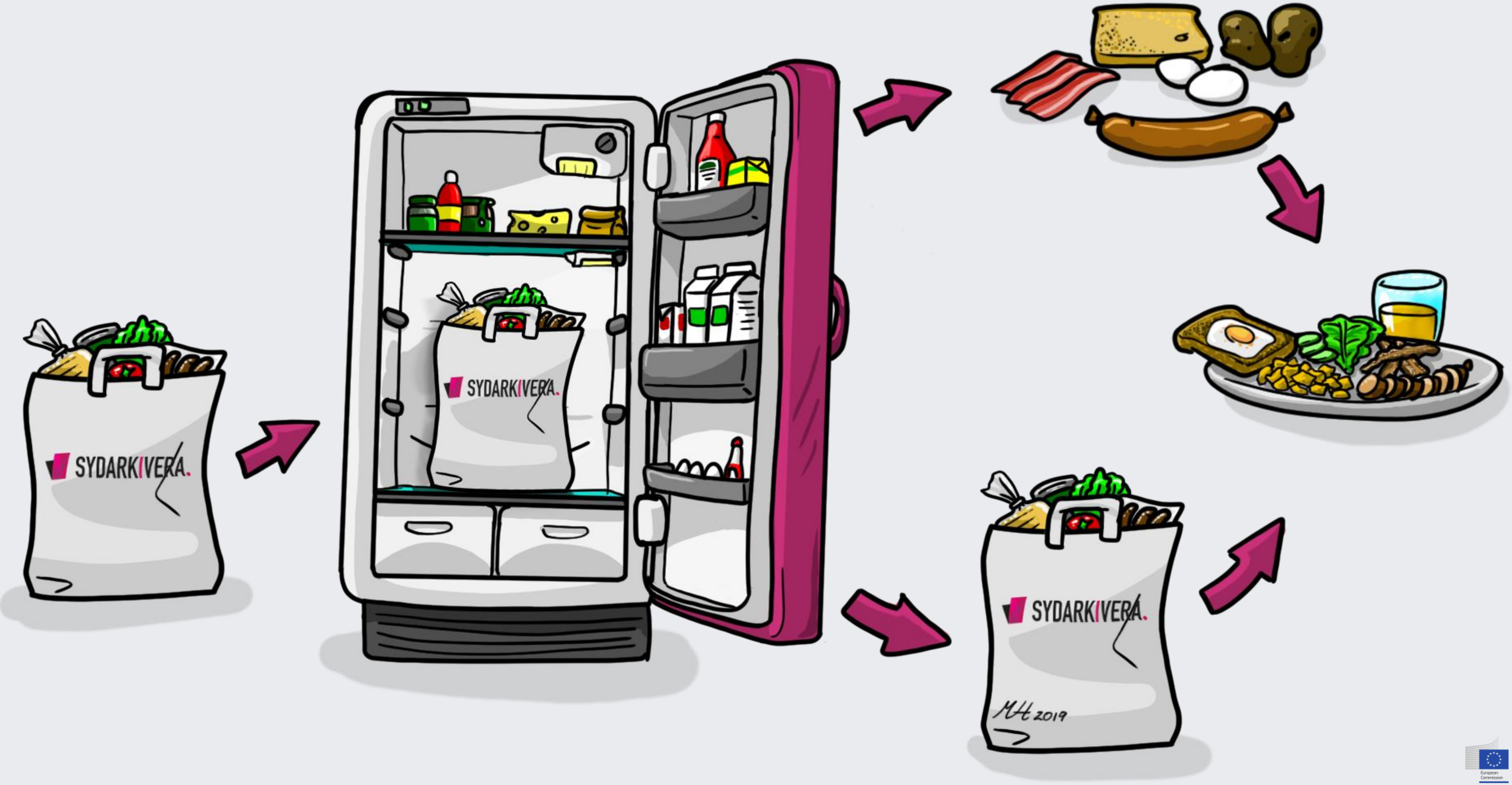
Carl Wilson
Technical lead
Open Preservation Foundation



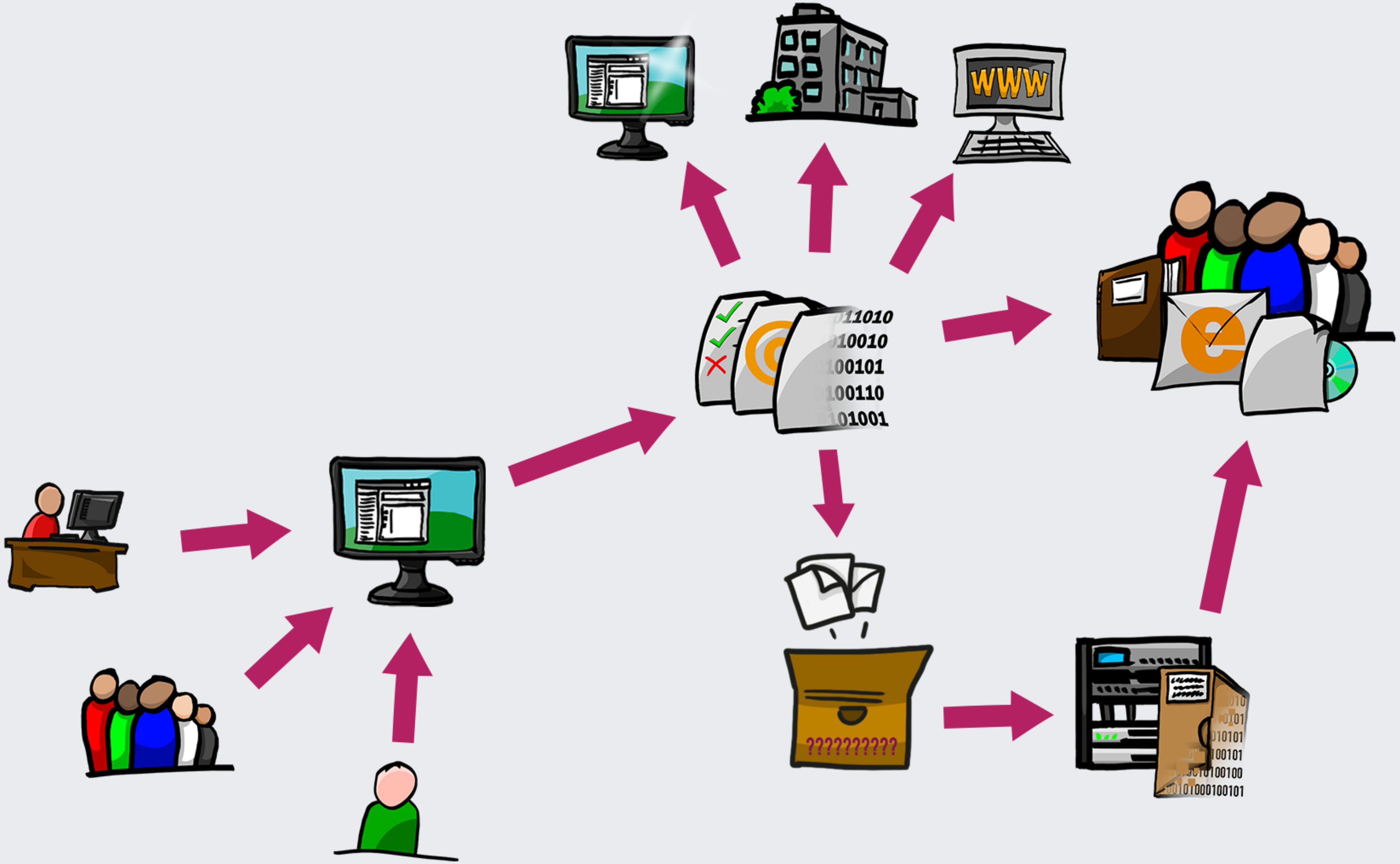
We use the OAIS reference model as the basis for facilitating data transfer and conformance



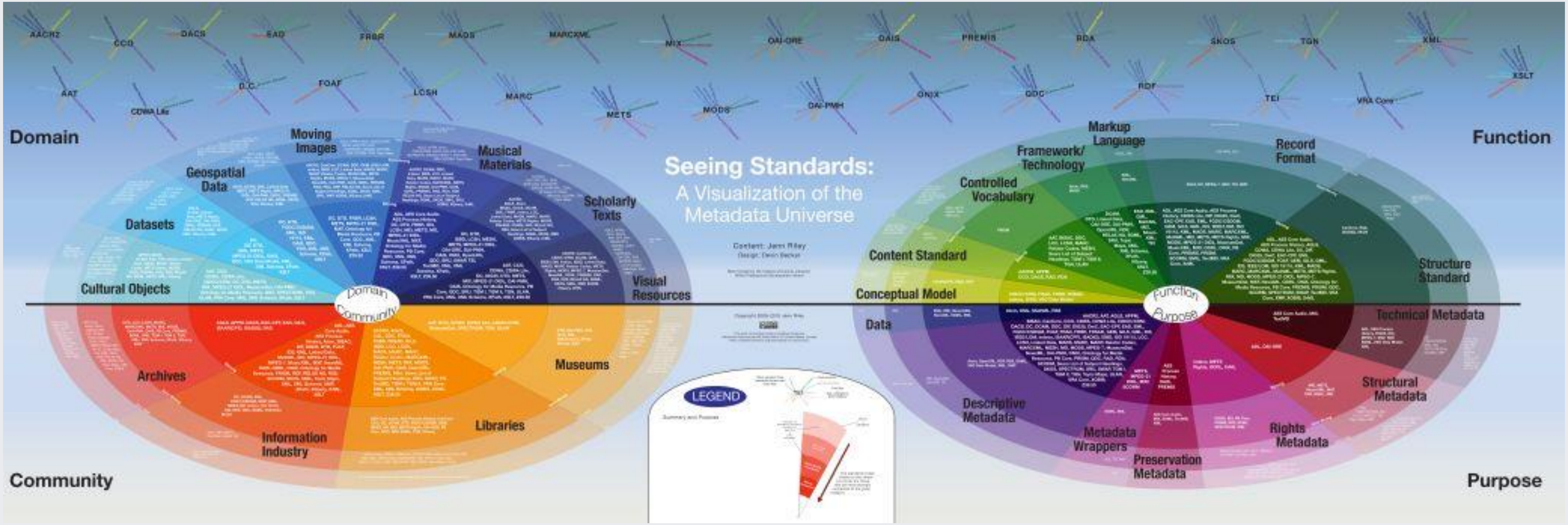
Let's explain it in an easy way: The OAIS Reference Model



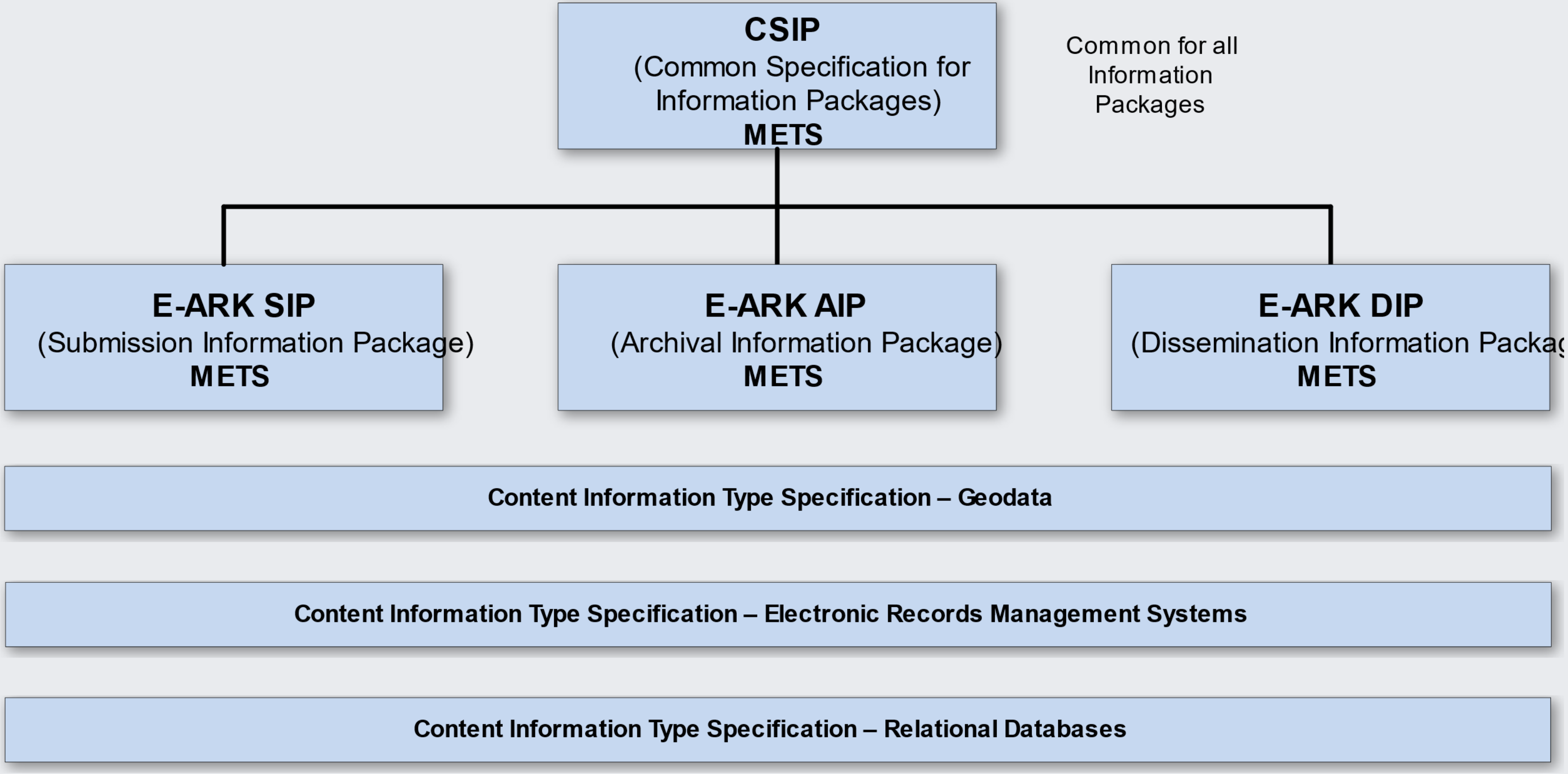
The data can be used in numerous ways by many different users



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



We need to agree to use a defined set of standards and specifications



CSIP
(Common Specification for Information Packages)
METS

Common for all Information Packages

E-ARK SIP
(Submission Information Package)
METS

E-ARK AIP
(Archival Information Package)
METS

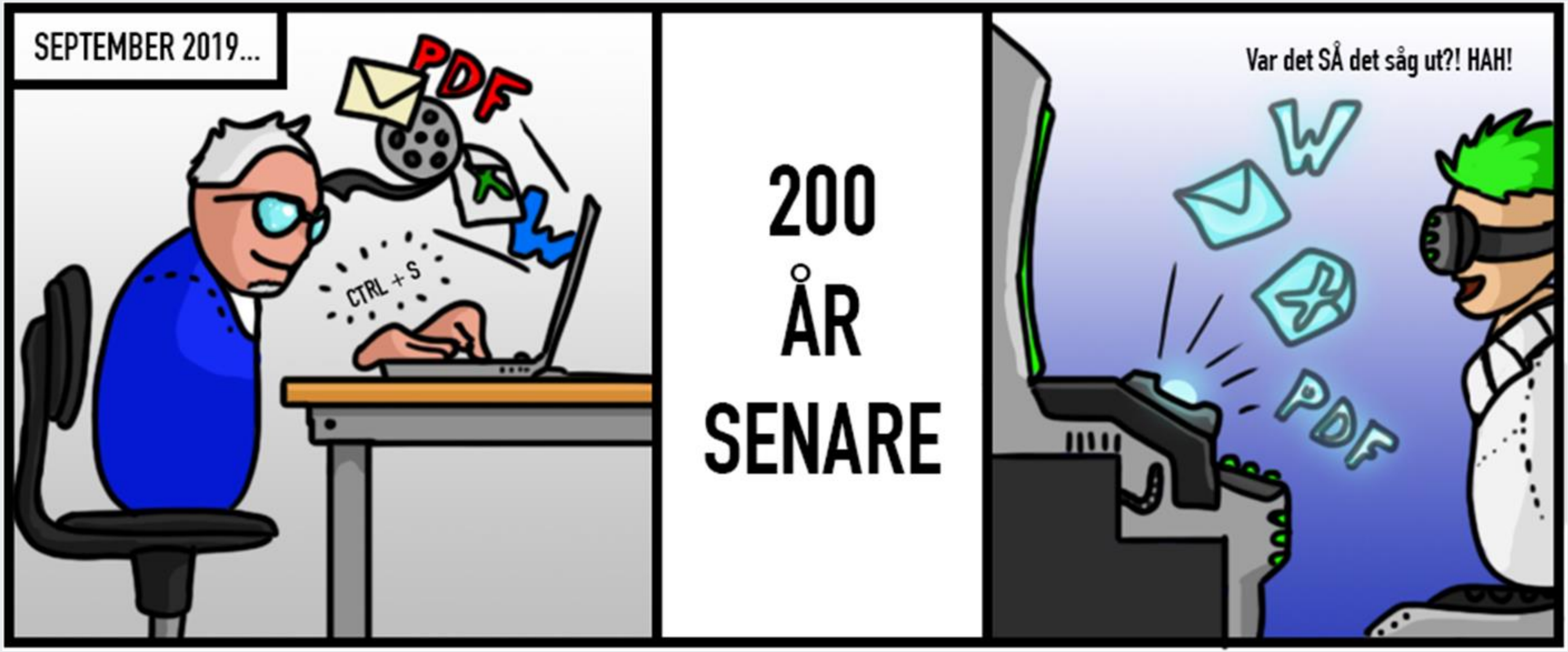
E-ARK DIP
(Dissemination Information Packag
METS

Content Information Type Specification – Geodata

Content Information Type Specification – Electronic Records Management Systems

Content Information Type Specification – Relational Databases

When we use the same specifications, we make preservation, migration, reuse and trust of your data easy



There are two types of eArchiving specifications:

Information Package and

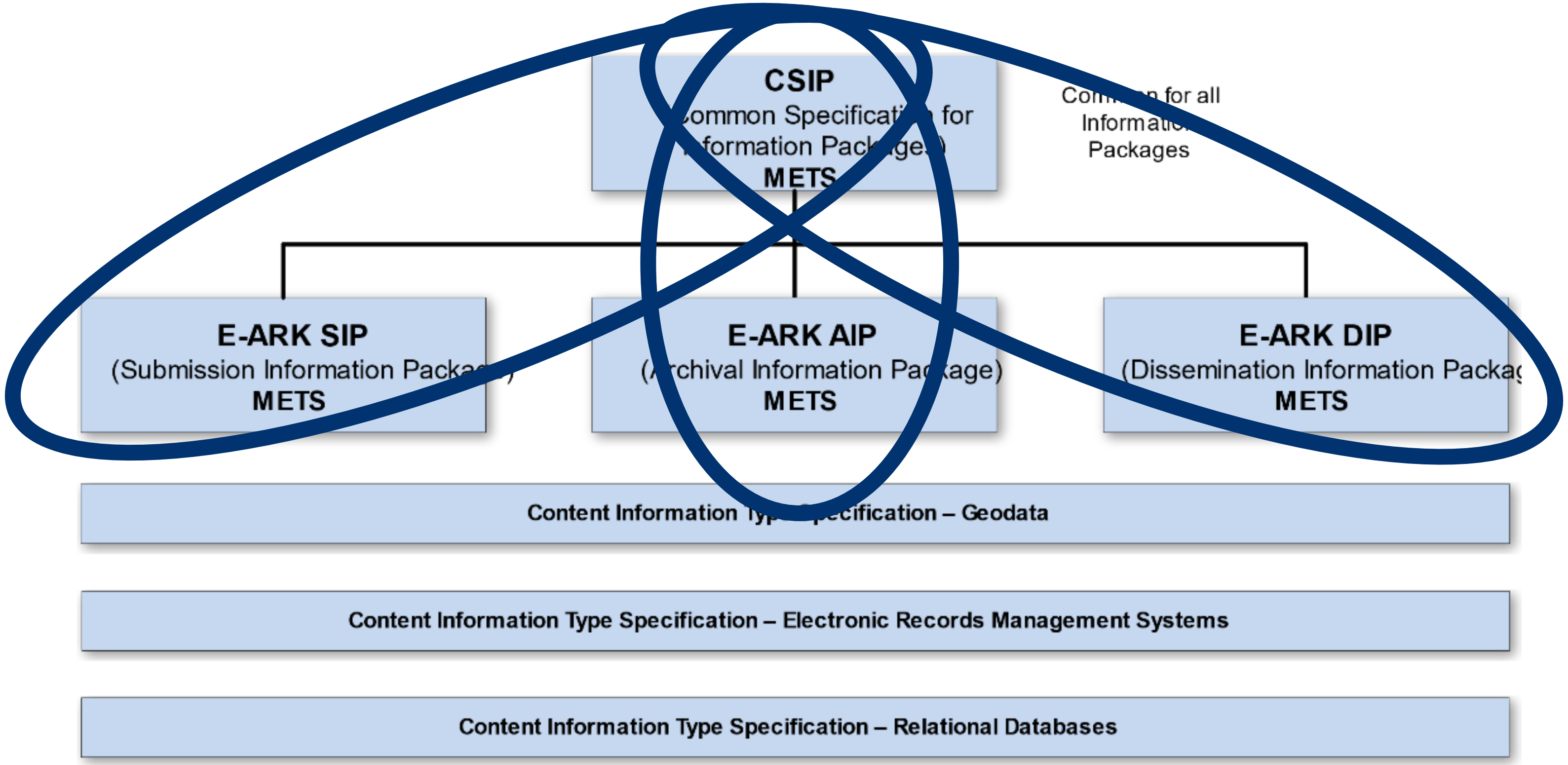
Content Information Type

The package

The box we fill with data is an Information Package

CSIP, SIP, AIP and DIP

- The CSIP provides a common basis for all information package specifications
- That is the SIP, AIP and DIP all build upon the CSIP



CSIP
(Common Specification for Information Packages)
METS

Common for all Information Packages

E-ARK SIP
(Submission Information Package)
METS

E-ARK AIP
(Archival Information Package)
METS

E-ARK DIP
(Dissemination Information Packag
METS

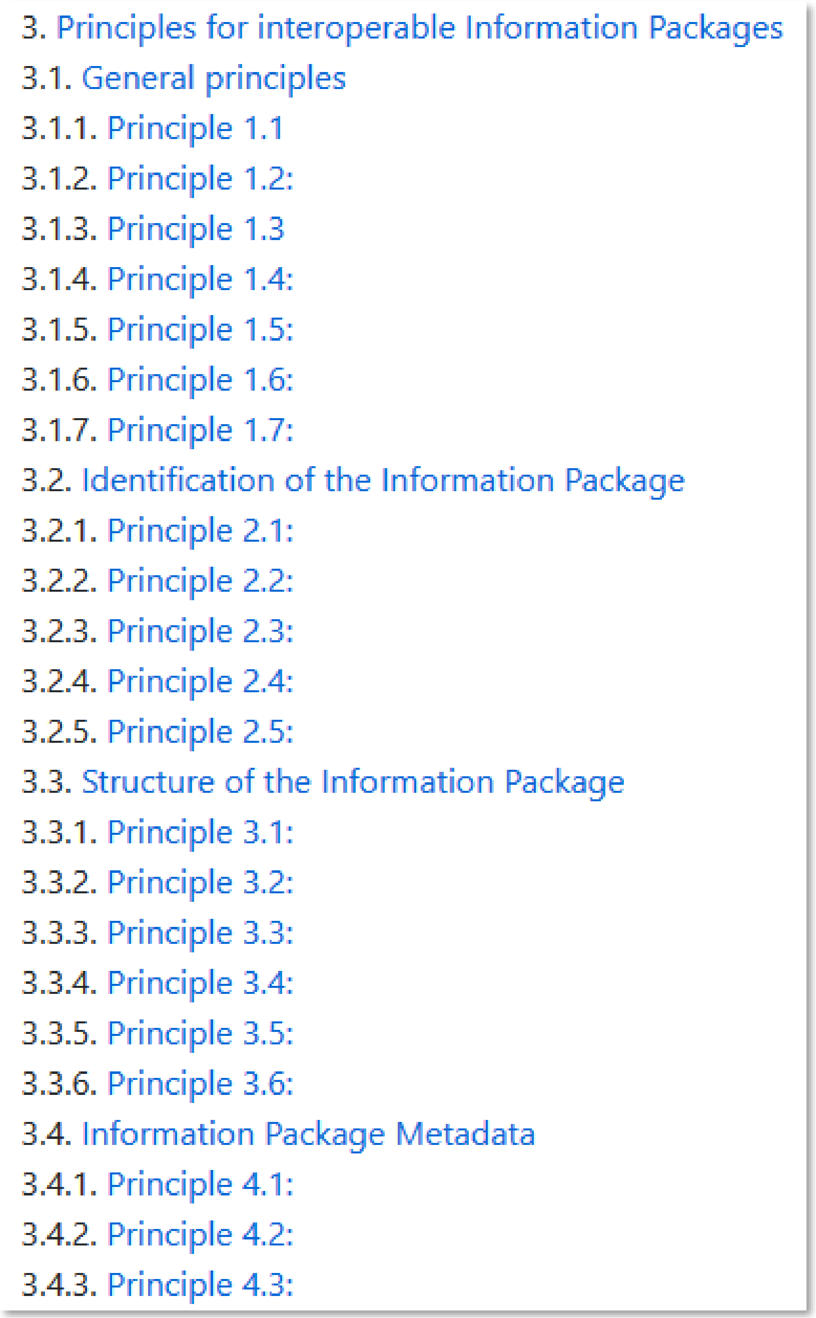
Content Information Type Specification – Geodata

Content Information Type Specification – Electronic Records Management Systems

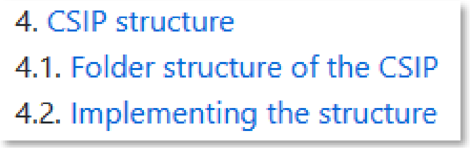
Content Information Type Specification – Relational Databases

The core principles for a package

- For example:
  - What makes a package a package?
  - How is the package identified?
  - How is the package structured?
  - What metadata is needed for a package?



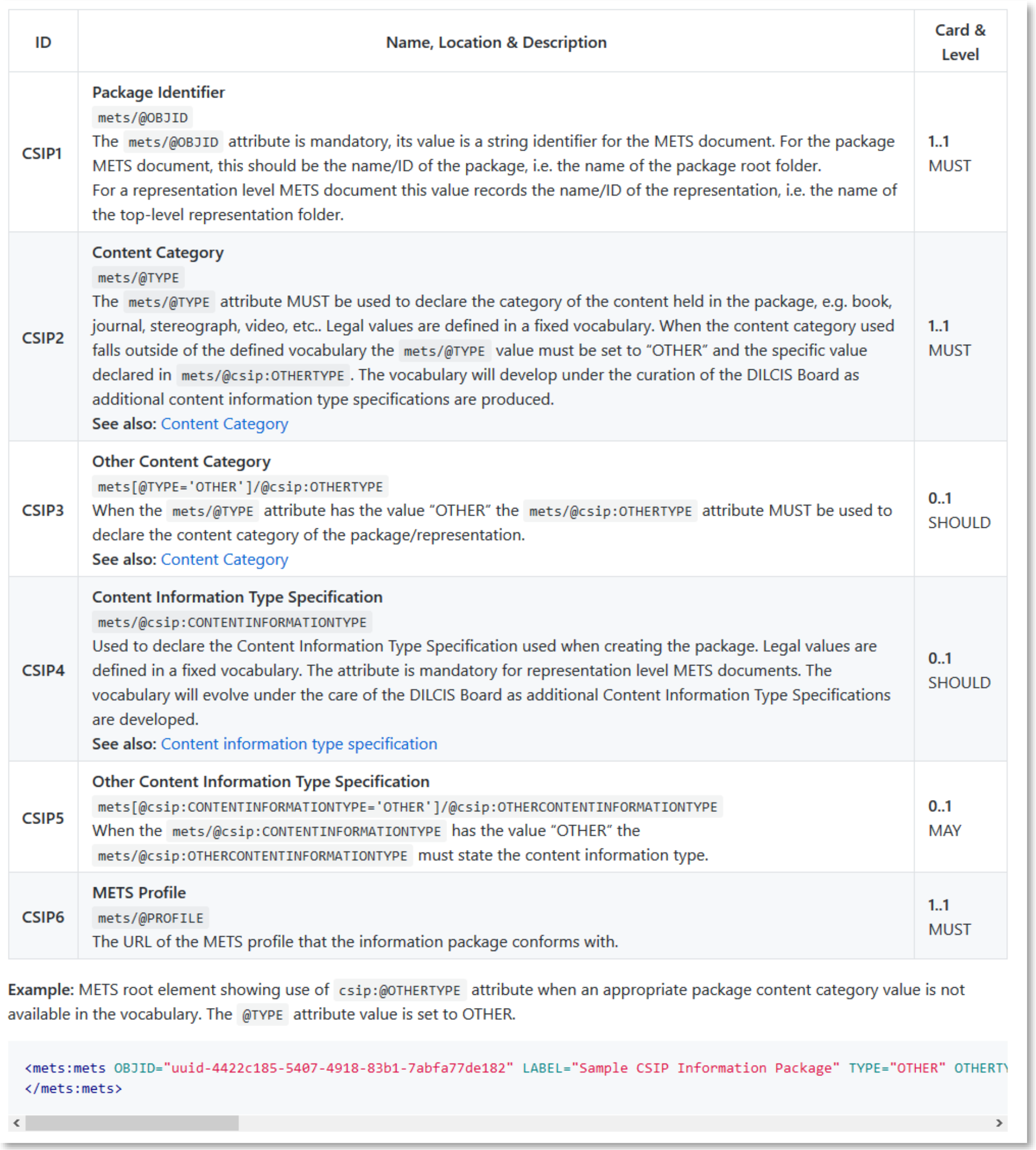
3. Principles for interoperable Information Packages
3.1. General principles
3.1.1. Principle 1.1
3.1.2. Principle 1.2:
3.1.3. Principle 1.3
3.1.4. Principle 1.4:
3.1.5. Principle 1.5:
3.1.6. Principle 1.6:
3.1.7. Principle 1.7:
3.2. Identification of the Information Package
3.2.1. Principle 2.1:
3.2.2. 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:
3.3.3. Principle 3.3:
3.3.4. Principle 3.4:
3.3.5. Principle 3.5:
3.3.6. Principle 3.6:
3.4. Information Package Metadata
3.4.1. Principle 4.1:
3.4.2. Principle 4.2:
3.4.3. Principle 4.3:



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

The CSIP describes the elements and attributes used in the transfer

- We utilise the elements and attributes from the de-facto standard, Metadata Encoding and Transmission Standard (METS)



ID	Name, Location & Description	Card & Level
CSIP1	**Package Identifier** `mets/@OBJID` The `mets/@OBJID` 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** `mets/@TYPE` The `mets/@TYPE` 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 `mets/@TYPE` value must be set to “OTHER” and the specific value declared in `mets/@csip:OTHERTYPE` . 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** `mets[@TYPE='OTHER']/@csip:OTHERTYPE` When the `mets/@TYPE` attribute has the value “OTHER” the `mets/@csip:OTHERTYPE` 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** `mets/@csip:CONTENTINFORMATIONTYPE` 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** `mets[@csip:CONTENTINFORMATIONTYPE='OTHER']/@csip:OTHERCONTENTINFORMATIONTYPE` When the `mets/@csip:CONTENTINFORMATIONTYPE` has the value “OTHER” the `mets/@csip:OTHERCONTENTINFORMATIONTYPE` must state the content information type.	0..1 MAY
CSIP6	**METS Profile** `mets/@PROFILE` 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" OTHERTY
</mets:mets>
```

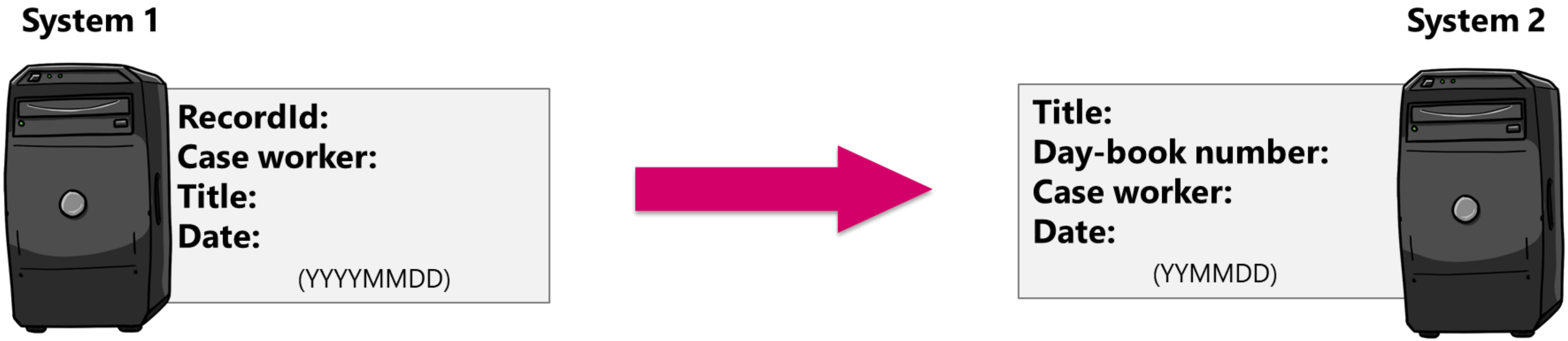
The data

The box is filled with data following a Content Information Type Specification

Why do we need these Content Information Type Specifications (CITS)?

We want to transfer our data from one system to another

- We want to transfer data from system1 to system2
- Observe the different element names!



System 1

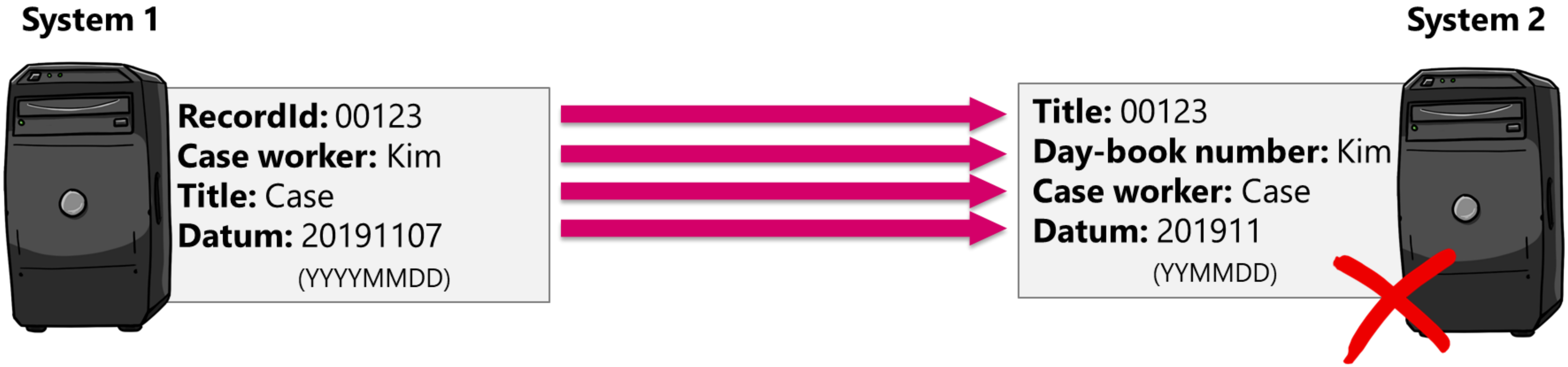
RecordId:
Case worker:
Title:
Date:
(YYYYMMDD)

System 2

Title:
Day-book number:
Case worker:
Date:
(YYMMDD)

Just move the data!

- We just move the data in System 1 to System 2



System 1

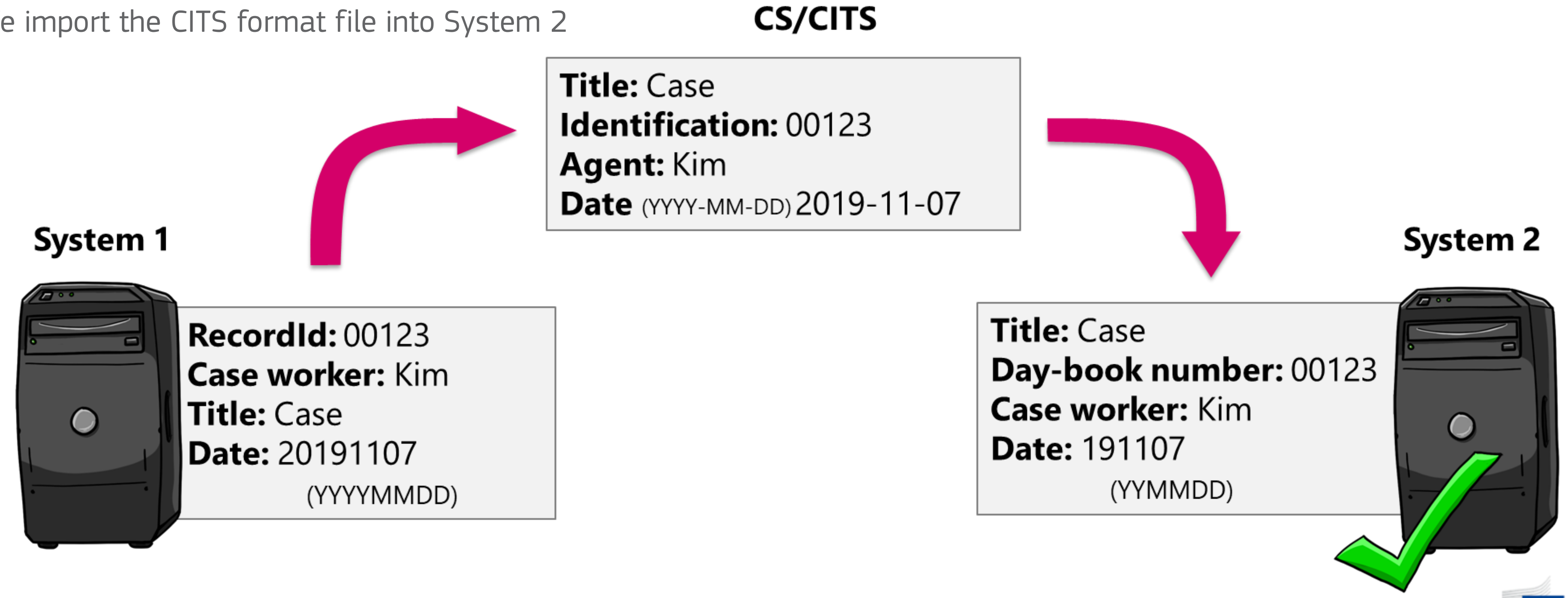
RecordId: 00123
Case worker: Kim
Title: Case
Datum: 20191107
(YYYYMMDD)

System 2

Title: 00123
Day-book number: Kim
Case worker: Case
Datum: 201911
(YYMMDD)

We use a content information type specification for moving data!

- We export to CITS format from System 1
- We obtain a file that conforms to the CITS
- We import the CITS format file into System 2



CS/CITS

Title: Case
Identification: 00123
Agent: Kim
Date (YYYY-MM-DD) 2019-11-07

System 1

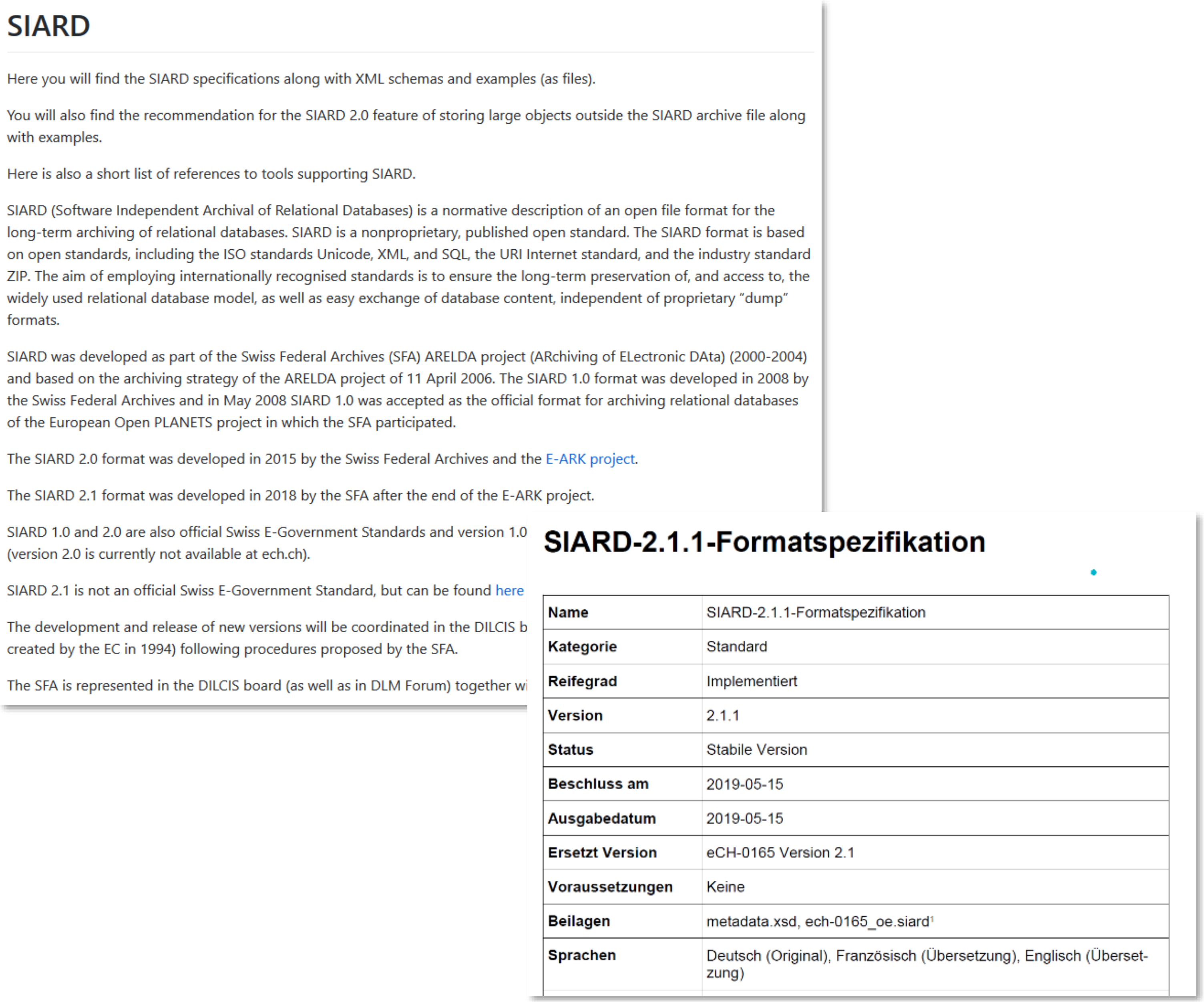
RecordId: 00123
Case worker: Kim
Title: Case
Date: 20191107
(YYYYMMDD)

System 2

Title: Case
Day-book number: 00123
Case worker: Kim
Date: 191107
(YYMMDD)

All CITS are based on existing standards

- No new wheels!
- For example the CITS for Relational Databases is based upon SIARD



SIARD

Here you will find the SIARD specifications along with XML schemas and examples (as files).

You will also find the recommendation for the SIARD 2.0 feature of storing large objects outside the SIARD archive file along with examples.

Here is also a short list of references to tools supporting SIARD.

SIARD (Software Independent Archival of Relational Databases) is a normative description of an open file format for the long-term archiving of relational databases. SIARD is a nonproprietary, published open standard. The SIARD format is based on open standards, including the ISO standards Unicode, XML, and SQL, 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 was developed as part of the Swiss Federal Archives (SFA) ARELDA project (ARchiving of ELectronic DAta) (2000-2004) and based on the archiving strategy of the ARELDA project of 11 April 2006. The SIARD 1.0 format was developed in 2008 by the Swiss Federal Archives and in May 2008 SIARD 1.0 was accepted as the official format for archiving relational databases of the European Open PLANETS project in which the SFA participated.

The SIARD 2.0 format was developed in 2015 by the Swiss Federal Archives and the E-ARK project.

The SIARD 2.1 format was developed in 2018 by the SFA after the end of the E-ARK project.

SIARD 1.0 and 2.0 are also official Swiss E-Government Standards and version 1.0 (version 2.0 is currently not available at ech.ch).

SIARD 2.1 is not an official Swiss E-Government Standard, but can be found here

The development and release of new versions will be coordinated in the DILCIS b created by the EC in 1994) following procedures proposed by the SFA.

The SFA is represented in the DILCIS board (as well as in DLM Forum) together wi

SIARD-2.1.1-Formatspezifikation

Name	SIARD-2.1.1-Formatspezifikation
Kategorie	Standard
Reifegrad	Implementiert
Version	2.1.1
Status	Stabile Version
Beschluss am	2019-05-15
Ausgabedatum	2019-05-15
Ersetzt Version	eCH-0165 Version 2.1
Voraussetzungen	Keine
Beilagen	metadata.xsd, ech-0165_oe.siard[1]
Sprachen	Deutsch (Original), Französisch (Übersetzung), Englisch (Übersetzung)

The CITS tells us where in the box to put the data and how we classify it

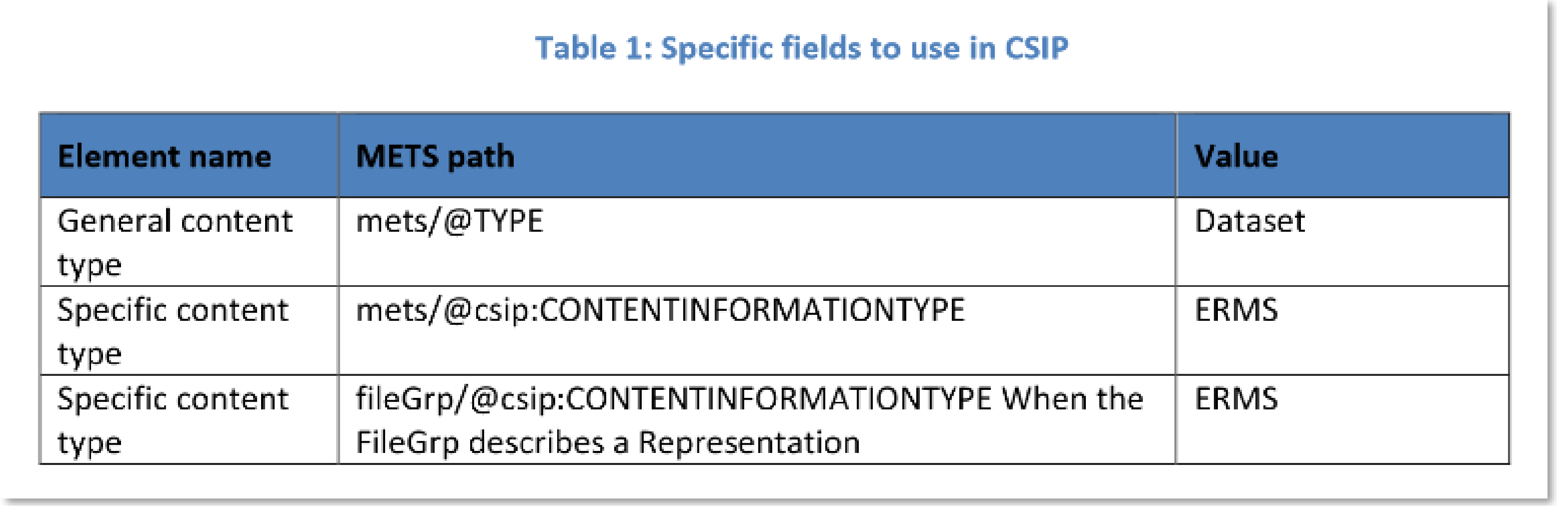
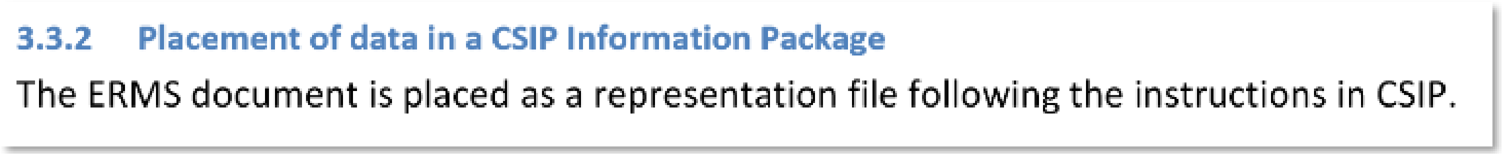


Table 1: Specific fields to use in CSIP

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.

Specification work

The worker bees of the specifications

Currently, all specification work is undertaken by the DILCIS Board

Digital Information LifeCycle Interoperability Standards Board

DILCIS BOARD

New CITS will be created

- CITS Archivial Description
- CITS SIARD
- CITS GEODATA GIS
- CITS PREMIS

How can we increase the number of CITS in the future?

Certification/Endorsment

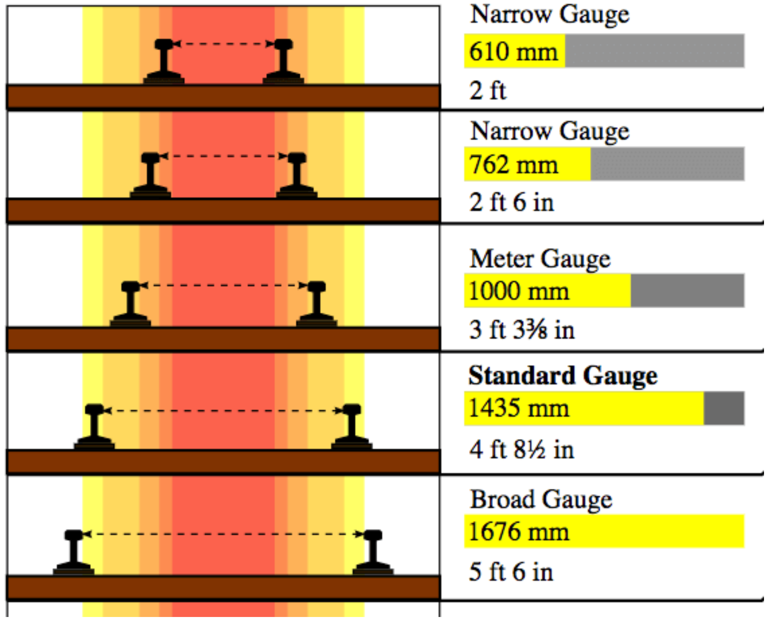
Supporting tools and software for eArchiving implementors

The Components of eArchiving Validation

An introduction to the eArchiving validation process

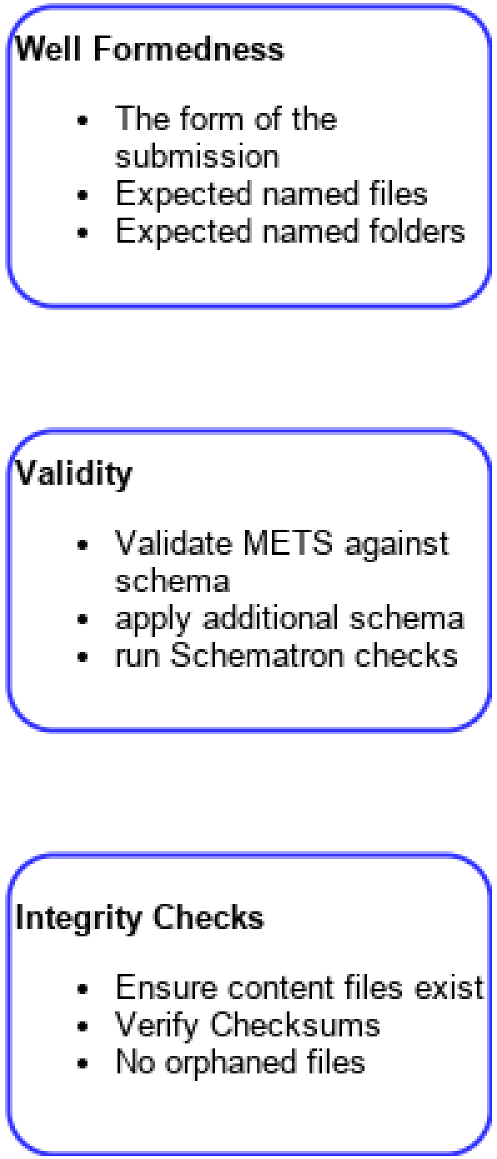
The Case for Automated Support

- Why develop validation tools?
- Validation software for users.
- Validation software libraries for developers.



The eArchiving Validation Process

- Structural Conformance
- Syntactic Conformance
- Package Integrity



Well Formedness

- The form of the submission
- Expected named files
- Expected named folders

Validity

- Validate METS against schema
- apply additional schema
- run Schematron checks

Integrity Checks

- Ensure content files exist
- Verify Checksums
- No orphaned files

What We're Making

- Test packages
- Shareable and reusable validation rules
- Validation software libraries
- Online validation service

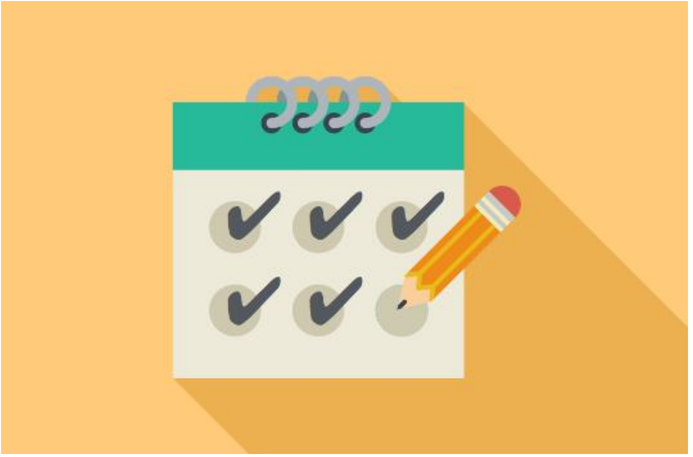
schematron
A language for making assertions about patterns found in XML documents

Building the Validation Components

How we go about producing test corpora, validation rules and software.

The Big Picture

- The importance of test data
- Validation rules
- Putting it all together



Establishing Baselines

- Why test data is needed
- The hidden benefits of producing test data

Validation Rules

- Why we're using Schematron
  - Reusable rules
  - Consistent implementation
- Quality assurance for validation rules

schematron

A language for making assertions about patterns found in XML documents

CONSISTENT QUALITY

Where we're going and how you can join the journey

Software for Specification Users

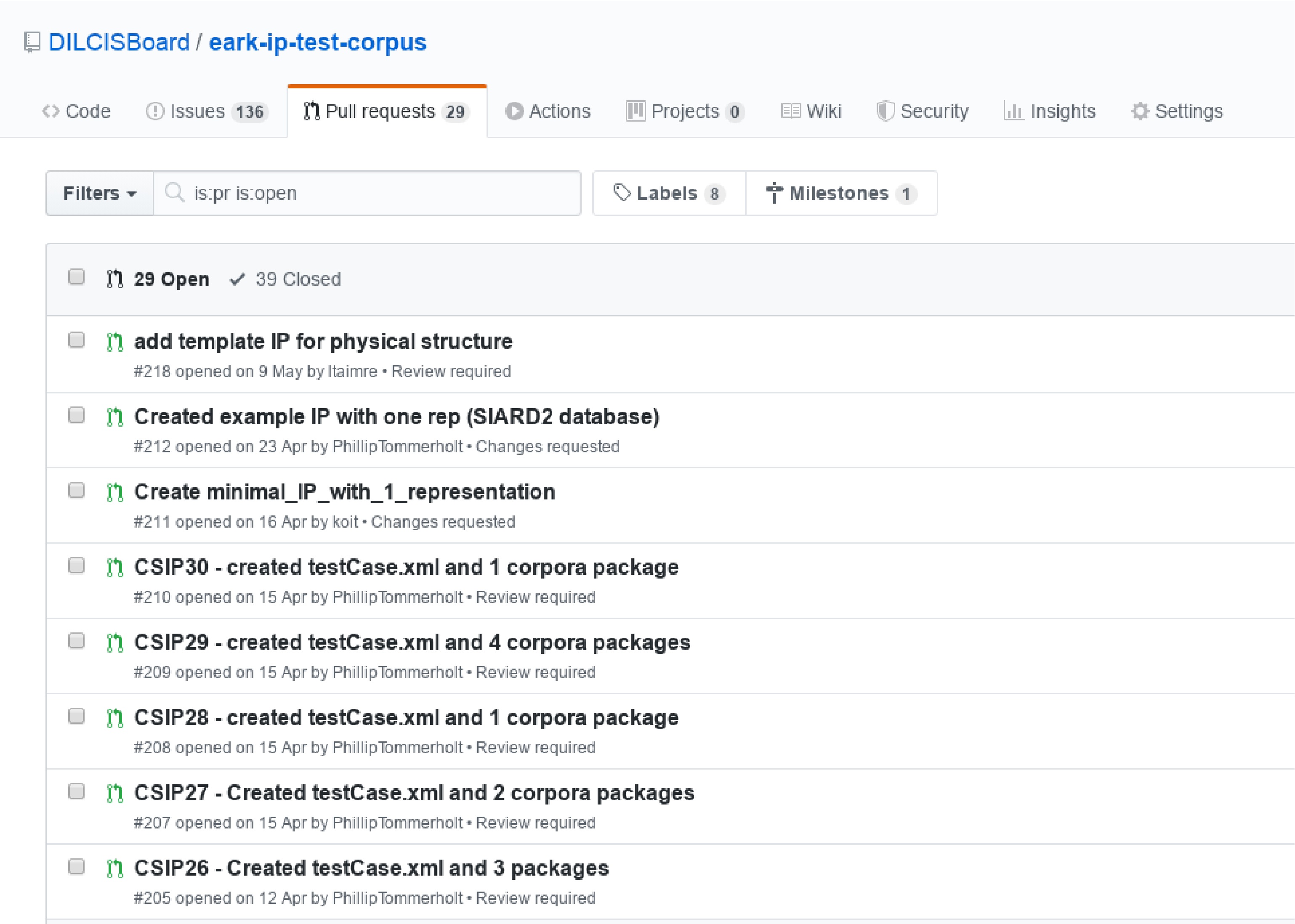
- Getting started with Information Packages
- Online validation service

Support for eArchiving Developers

- Software libraries for:
  - Third party developers wishing to integrate eArchiving support into their products
  - In house development staff at any institution working with eArchiving standards.

Join the eArchiving Community

- Open process with focus on GitHub for:
  - Specifications
  - Test Corpora
  - Validation Rules
  - Software
- Have your say by giving feedback
- Make thing better by contributing.



DILCISBoard / eark-ip-test-corpus

29 Open ✓ 39 Closed

□ **add template IP for physical structure**
#218 opened on 9 May by ltaimre • Review required

□ **Created example IP with one rep (SIARD2 database)**
#212 opened on 23 Apr by PhillipTommerholt • Changes requested

□ **Create minimal_IP_with_1_representation**
#211 opened on 16 Apr by koit • Changes requested

□ **CSIP30 - created testCase.xml and 1 corpora package**
#210 opened on 15 Apr by PhillipTommerholt • Review required

□ **CSIP29 - created testCase.xml and 4 corpora packages**
#209 opened on 15 Apr by PhillipTommerholt • Review required

□ **CSIP28 - created testCase.xml and 1 corpora package**
#208 opened on 15 Apr by PhillipTommerholt • Review required

□ **CSIP27 - Created testCase.xml and 2 corpora packages**
#207 opened on 15 Apr by PhillipTommerholt • Review required

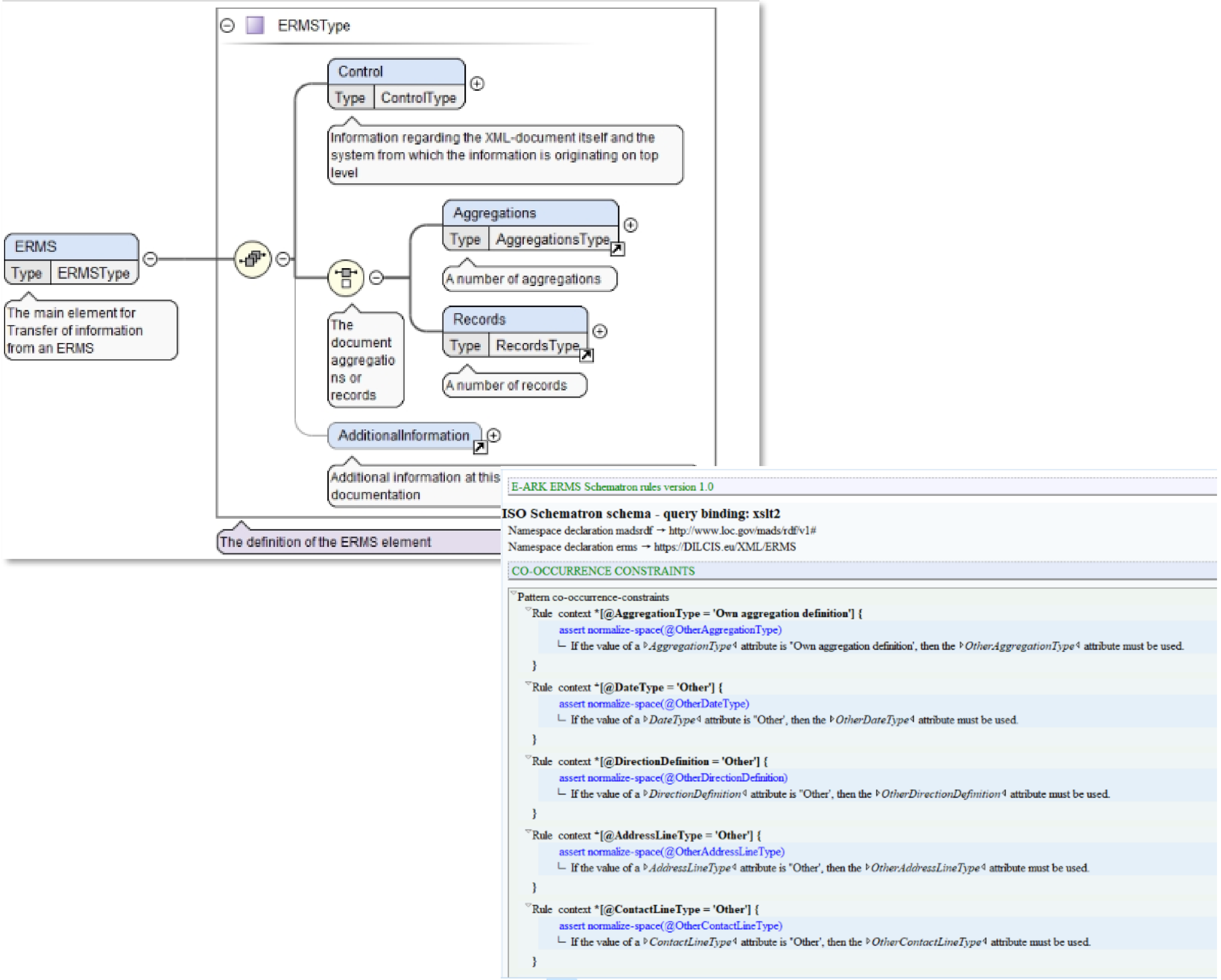
□ **CSIP26 - Created testCase.xml and 3 packages**
#205 opened on 12 Apr by PhillipTommerholt • Review required

With some help from our friends

To use the specifications we also need access to some supporting elements

Schemas facilitate validation rules

- XML-schemas
- Schematron documents
- Draft examples are shown


ERMSType
Control
Type ControlType
Information regarding the XML-document itself and the system from which the information is originating on top level
Aggregations
Type AggregationsType
A number of aggregations
Records
Type RecordsType
A number of records
ERMS
Type ERMSType
The main element for Transfer of information from an ERMS
The document aggregations or records
AdditionalInformation
Additional information at this documentation
The definition of the ERMS element

E-ARK ERMS Schematron rules version 1.0

ISO Schematron schema - query binding: xslt2

Namespace declaration madsrdf → http://www.loc.gov/mads/rdf/v1#
Namespace declaration erms → https://DILCIS.eu/XML/ERMS

CO-OCCURRENCE CONSTRAINTS

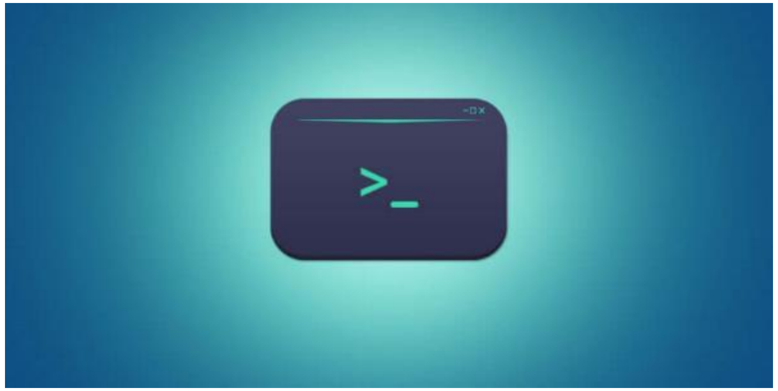
```
Pattern co-occurrence-constraints
  Rule  context *[@AggregationType = 'Own aggregation definition'] {
      assert normalize-space(@OtherAggregationType)
      └ If the value of a AggregationType attribute is 'Own aggregation definition', then the OtherAggregationType attribute must be used.
  }
  Rule  context *[@DateType = 'Other'] {
      assert normalize-space(@OtherDateType)
      └ If the value of a DateType attribute is "Other', then the OtherDateType attribute must be used.
  }
  Rule  context *[@DirectionDefinition = 'Other'] {
      assert normalize-space(@OtherDirectionDefinition)
      └ If the value of a DirectionDefinition attribute is "Other', then the OtherDirectionDefinition attribute must be used.
  }
  Rule  context *[@AddressLineType = 'Other'] {
      assert normalize-space(@OtherAddressLineType)
      └ If the value of a AddressLineType attribute is "Other', then the OtherAddressLineType attribute must be used.
  }
  Rule  context *[@ContactLineType = 'Other'] {
      assert normalize-space(@OtherContactLineType)
      └ If the value of a ContactLineType attribute is "Other', then the OtherContactLineType attribute must be used.
  }
```

Guidelines

- How to use the different specifications
- Examples
- Detailed explanations

Software tools and libraries

- Online services for non-technical users
- Command line software for batch processing and researchers
- Software libraries for in-house and commercial developers



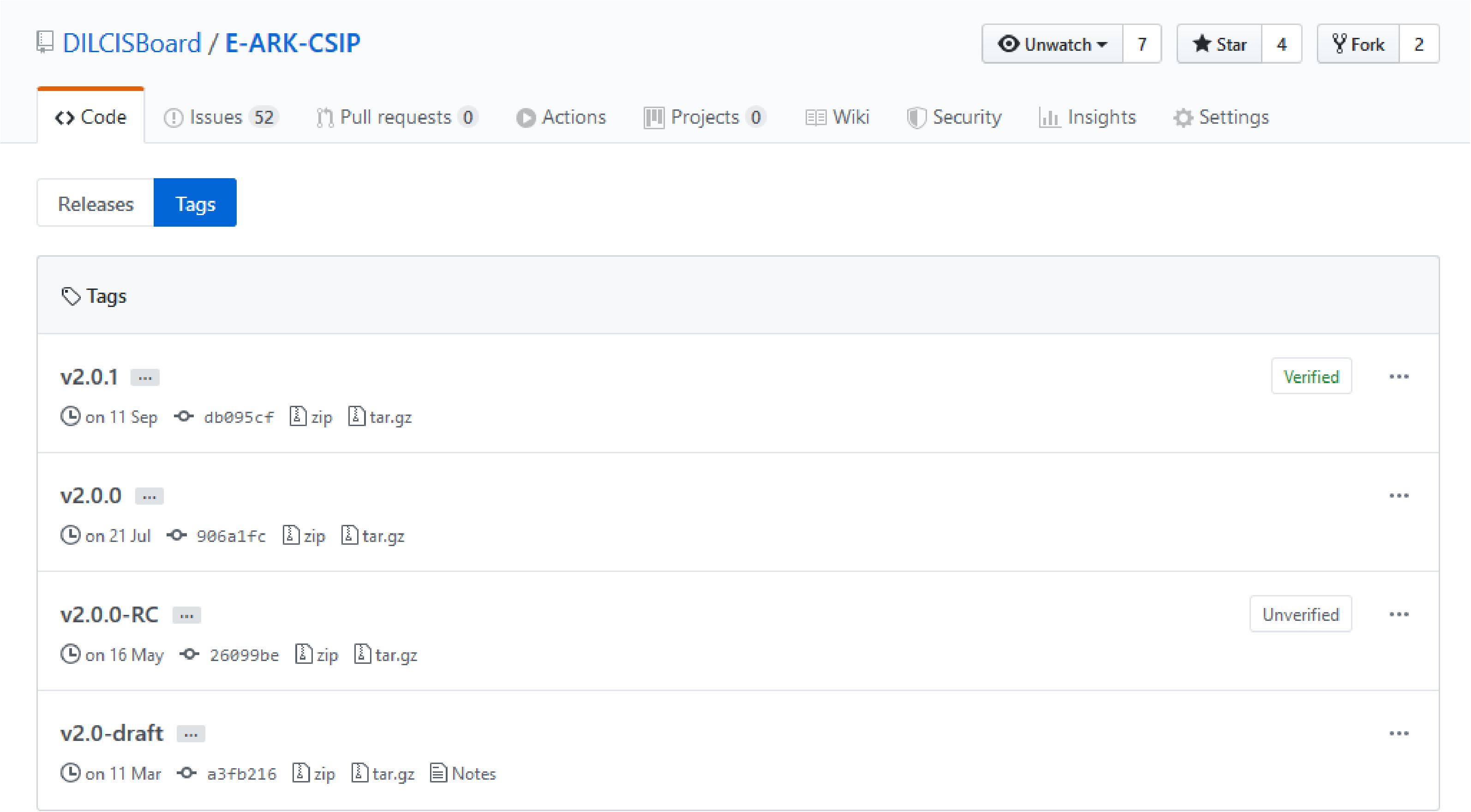
GitHub

Specifications created by the DILCIS Board are hosted on GitHub

GitHub

- The largest host of open source software and specification on the planet
- Provides an infrastructure for hosting, managing and participating in open source development.

All versions of the different specifications are found in the GitHub repositories



Tags

v2.0.1 … Verified
on 11 Sep db095cf zip tar.gz

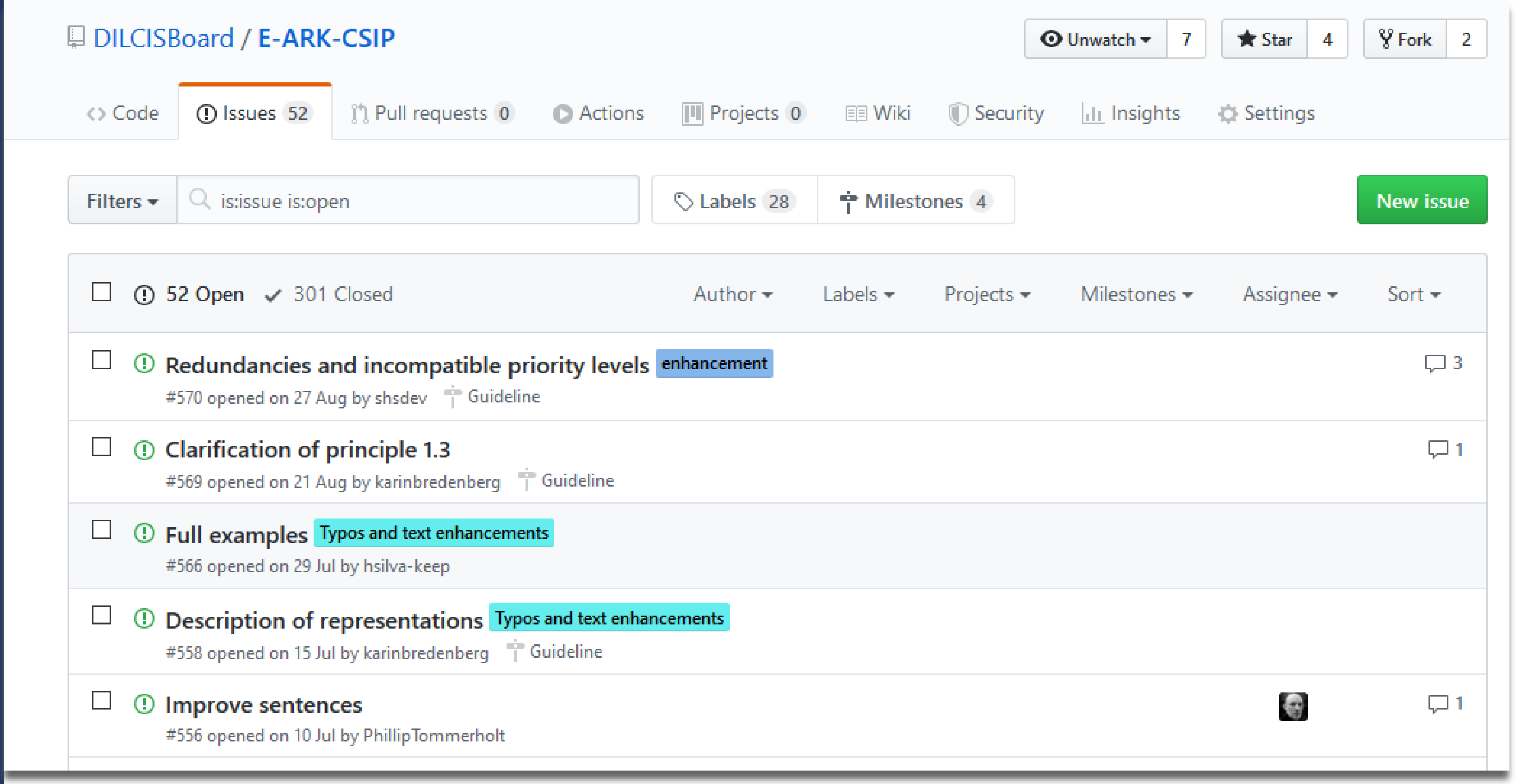
v2.0.0 …
on 21 Jul 906a1fc zip tar.gz

v2.0.0-RC … Unverified
on 16 May 26099be zip tar.gz

v2.0-draft …
on 11 Mar a3fb216 zip tar.gz Notes

We track issues and comments on GitHub's Trackers

- All issues are addressed!
- GitHub users can create and comment on issues!
- Remember the Service Desk!



DILCISBoard / E-ARK-CSIP

Unwatch 7 | Star 4 | Fork 2

Code | Issues 52 | Pull requests 0 | Actions | Projects 0 | Wiki | Security | Insights | Settings

Filters | is:issue is:open | Labels 28 | Milestones 4 | New issue

52 Open ✓ 301 Closed | Author | Labels | Projects | Milestones | Assignee | Sort

□ Redundancies and incompatible priority levels enhancement
#570 opened on 27 Aug by shsdev Guideline 3

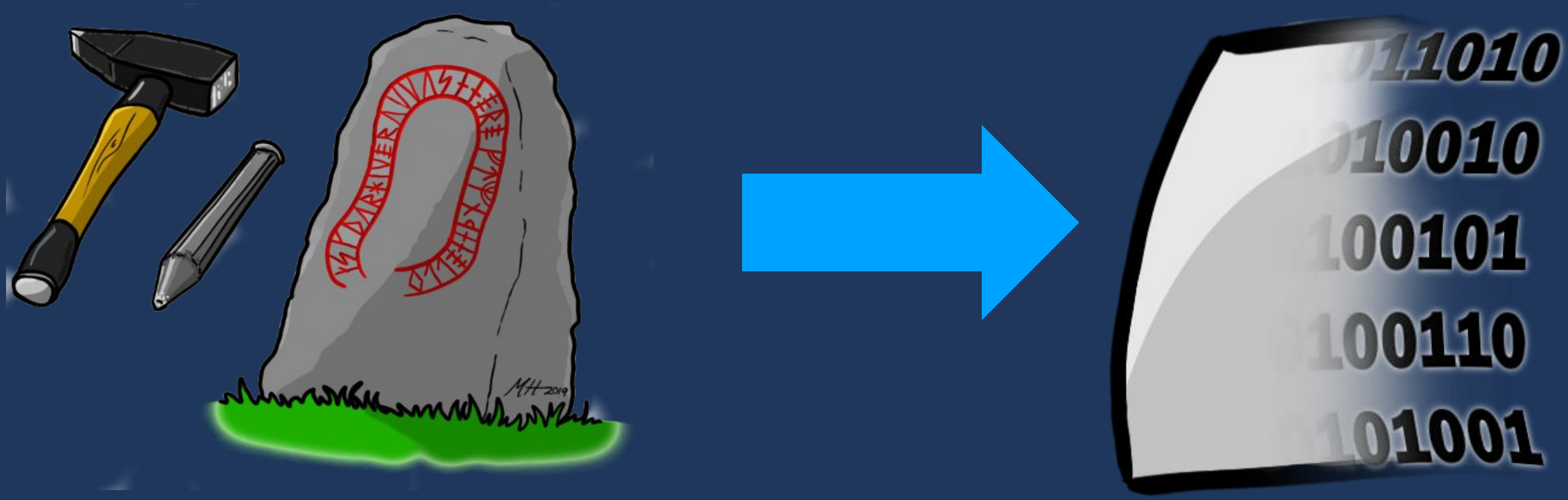
□ Clarification of principle 1.3
#569 opened on 21 Aug by karinbredenberg Guideline 1

□ Full examples Typos and text enhancements
#566 opened on 29 Jul by hsilva-keep

□ Description of representations Typos and text enhancements
#558 opened on 15 Jul by karinbredenberg Guideline

□ Improve sentences
#556 opened on 10 Jul by PhillipTommerholt 1

Data has been saved for a long time and we will continue to preserve, migrate, reuse and trust our data regardless of its form using common specifications and conformance



Specifications and conformance are a community effort!

We use the same standards and specifications to make preservation, migration, reuse and trust of the data easy

We use conformance testing to make preservation, migration, reuse and trust of the data easy

Take the opportunity to ask us questions!

We are here both days!

Links

- http://jennriley.com/metadatamap/
- http://www.loc.gov/standards/mets/
- https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving
- https://dilcis.eu/
- https://dilcis.eu/specifications/common-specification
- https://dilcis.eu/specifications/sip
- https://dilcis.eu/specifications/aip
- https://dilcis.eu/specifications/dip
- https://dilcis.eu/content-types/cserms
- https://dilcis.eu/content-types/cs-geospatial-data
- https://dilcis.eu/content-types/siard

GitHub Links

- https://github.com/DILCISBoard
- https://github.com/DILCISBoard/E-ARK-CSIP
- https://github.com/DILCISBoard/E-ARK-SIP
- https://github.com/DILCISBoard/E-ARK-AIP
- https://github.com/DILCISBoard/E-ARK-DIP
- https://github.com/DILCISBoard/eark-ip-test-corpus

European
Commission

Thank you!

Karin Bredenberg

SYDARKIVERA.

Carl Wilson

Images thanks to Magnus Heimonen, Sydarkivera

Lunch break

We will resume at 14:00