eArchiving Webinar #17

eArchiving geospatial records: Approaches and benefits

Start 10:00 (CET)

21 October 2021

Agenda

10:00 – 10:05
Welcome
Pawel Stech, CEF Stakeholder Management Office – DIGIT

10:05 – 10:15
eArchiving Building Block welcome
Jaime Kaminski, eArchiving Building Block training lead

10:15 – 11:00
eArchiving geospatial records: How to benefit from CITS Geospatial
Gregor Završnik, Geoarh

11:00 – 11:40
Impressions from early adopters
Ann-Kristin Egeland, The Danish National Archives

11:40
Q&A

Welcome
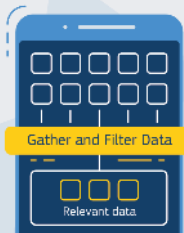
Pawel Stech
CEF Stakeholder Management Office – DIGIT

Building Blocks supported under the CEF Digital programme 2014–2021

Big Data Test Infrastructure

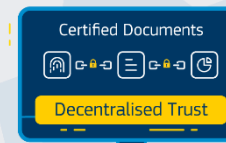Explore and experiment with big data for improved performance and decision making

Context Broker

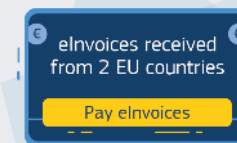Analyze, manage and share data, in real time, at the right time, throughout Europe

eArchiving

Facilitates the preservation, migration, reuse and trust of your data
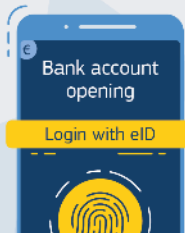
European Blockchain Services Infrastructure

Harness the power of a European-wide network of blockchain services, increasing trust through data security, privacy and transparency

eInvoicing

Promote the implementation of the European standard for electronic invoicing across borders
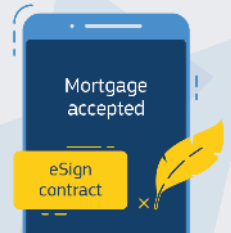
eID

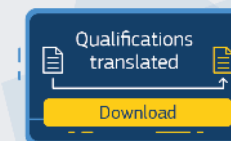Allow citizens to prove who they are across borders, making it easier to access online services in another EU Member State

eDelivery

Exchange online data and documents reliably and securely

eSignature

Create and verify electronic signatures between businesses and EU citizens

eTranslation

Offers machine translation to translate your documents and web content into any official EU language, Norwegian or Icelandic

https://ec.europa.eu/cefdigital/

How to use a Building Block?

Build, buy or reuse the Building Blocks on your own.

Co-develop the solution or partner with other parties.

Co-develop and partner
with other parties

Build
The solution from scratch based on a European standard
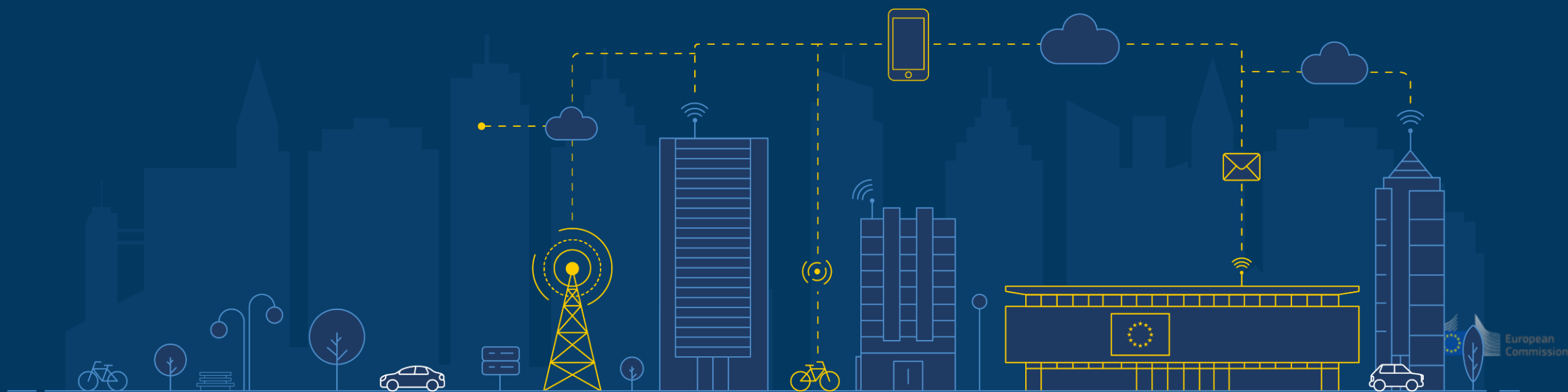
Buy
A compliant solution from the market

Reuse
Sample software available on CEF website

European Standards

Welcome to the eArchiving Building Block

Dr Jaime Kaminski
CEF eArchiving activity lead training

eArchiving Building Block website

https://ec.europa.eu/cefdigital/earchiving

eArchiving Building Block

eArchiving services:

- Technical specifications
- Sample software
- Compliance/validation
- Service Desk
- Outreach/community engagement
- **Training**
  - Webinars
  - Videos
  - Moodle LMS training modules

eArchiving

Preserve and reuse data

WEBINARS : AGENDA & RECORDINGS

Session	What you'll learn	Date & Time*	Webinar presentation & recording	Link to Q&A
Webinar #1: Introduction to CSIP	• CEF eArchiving welcome • Why have a common standard? • Core principles for an information package • Elements and attributes used for describing a package • Extending CSIP to meet more needs • METS in E-ARK CSIP	• 27th February 2020: 10:00 - 11:00	 Recording: CEF webinar CSIP	CEF Webinar #1: Q&A
Webinar #2: Introduction to ESSArch – an open source-based solution for long-term preservation of digital information	• CEF eArchiving welcome • Introduction to ESS and ESSArch • Pre-Ingest and Ingest • Archival and Data Management • Access and Portal • Reports, Statistics, Monitoring and API • Configuration and Administration • ESSArch Installation procedures	• 26th March 2020: 10:00 - 13:00	 Recording: Part 1 - Part 2 - Part 3 - Part 4- Part 5	CEF Webinar #2: Q&A
Webinar #3: Preserving digital geospatial records	• CEF eArchiving welcome • Geospatial data and its role in organisations • How could you benefit form E-ARK specifications for geospatial data preservation? • Strategies for implementing an accessible geospatial records archive • Proactive preservation in new and existing systems	• 23rd April 2020: 10:00 - 11:15		CEF Webinar #3: Q&A

https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving+webinar+Series+2020
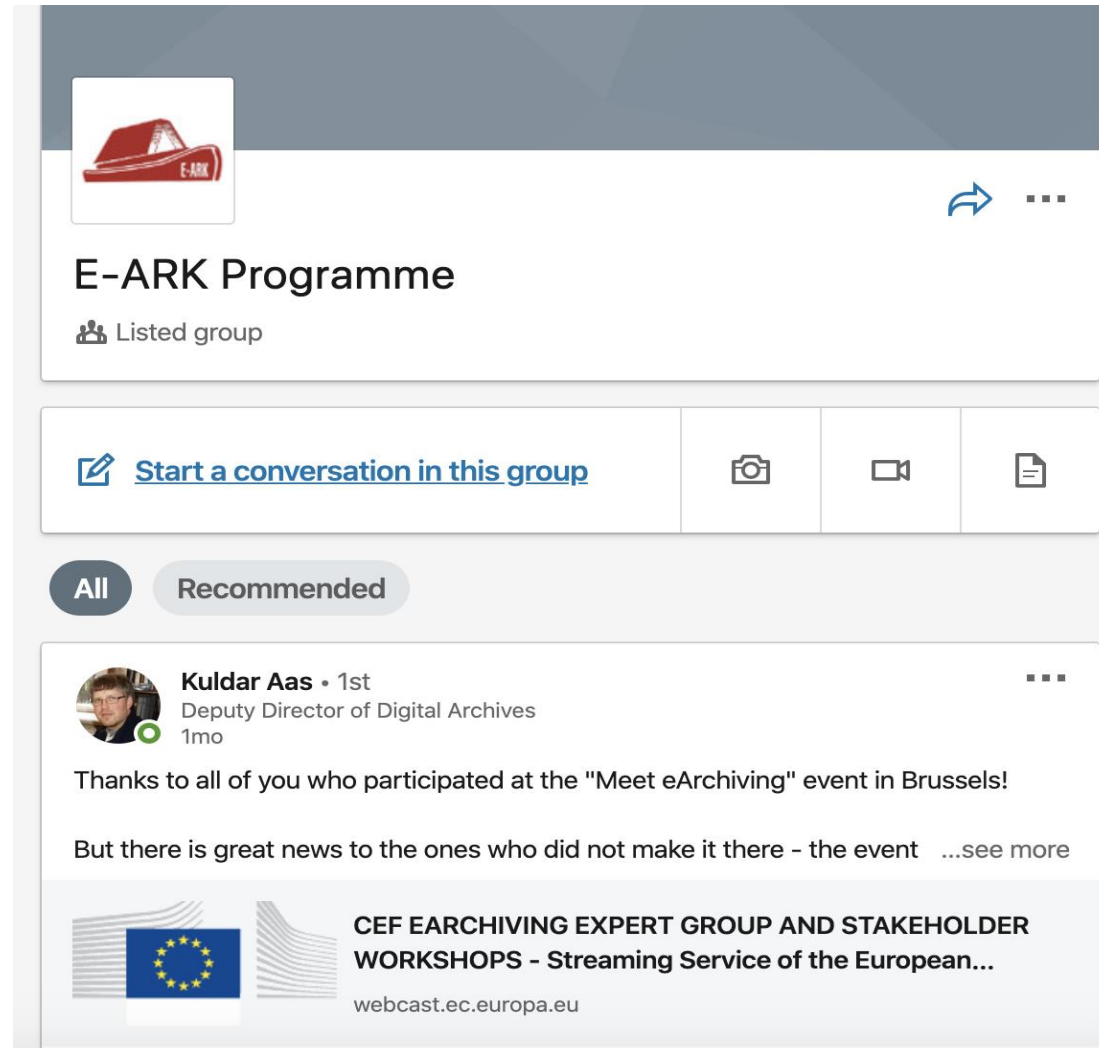
https://www.youtube.com/channel/UCaPOT_MBdE-kL5AJQzrCBDw/videos?view=0&sort=dd&flow=grid

eArchiving outreach

- LinkedIn group
- Twitter #EARKProject

LinkedIn Group: E-ARK Programme
https://www.linkedin.com/groups/8343650/

eArchiving geospatial records:
How to benefit from CITS Geospatial

Gregor Završnik
Geoarh, Slovenia
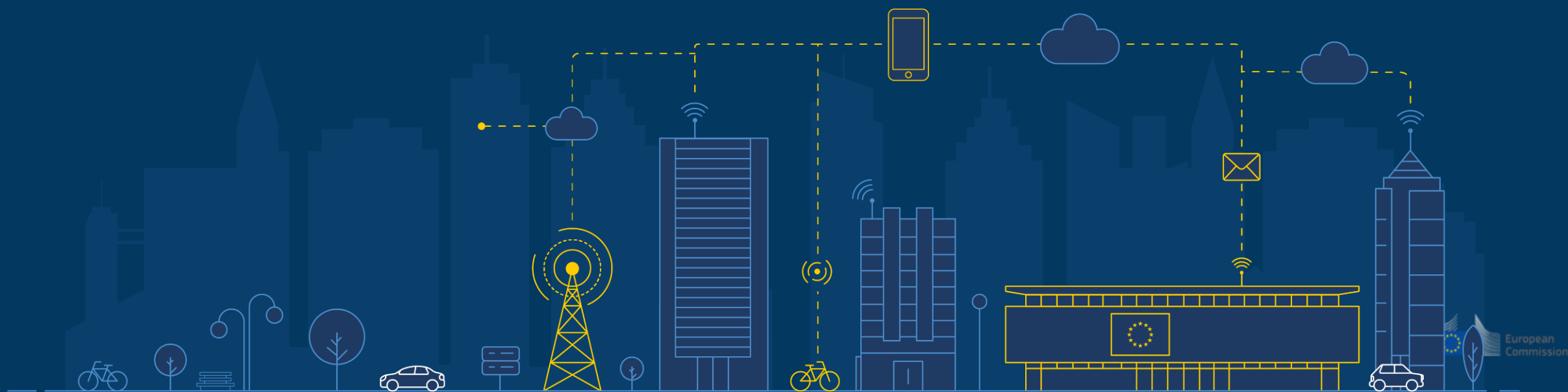
Let's talk about CITS Geospatial

- Why CITS Geospatial?
- What is CITS Geospatial?
  - Categories
  - Guidelines for Geospatial Records
  - Guidelines for GIS
  - What is new since v 2.0
- How you can use it?
  - Getting started
  - Local implementations
- What is next?
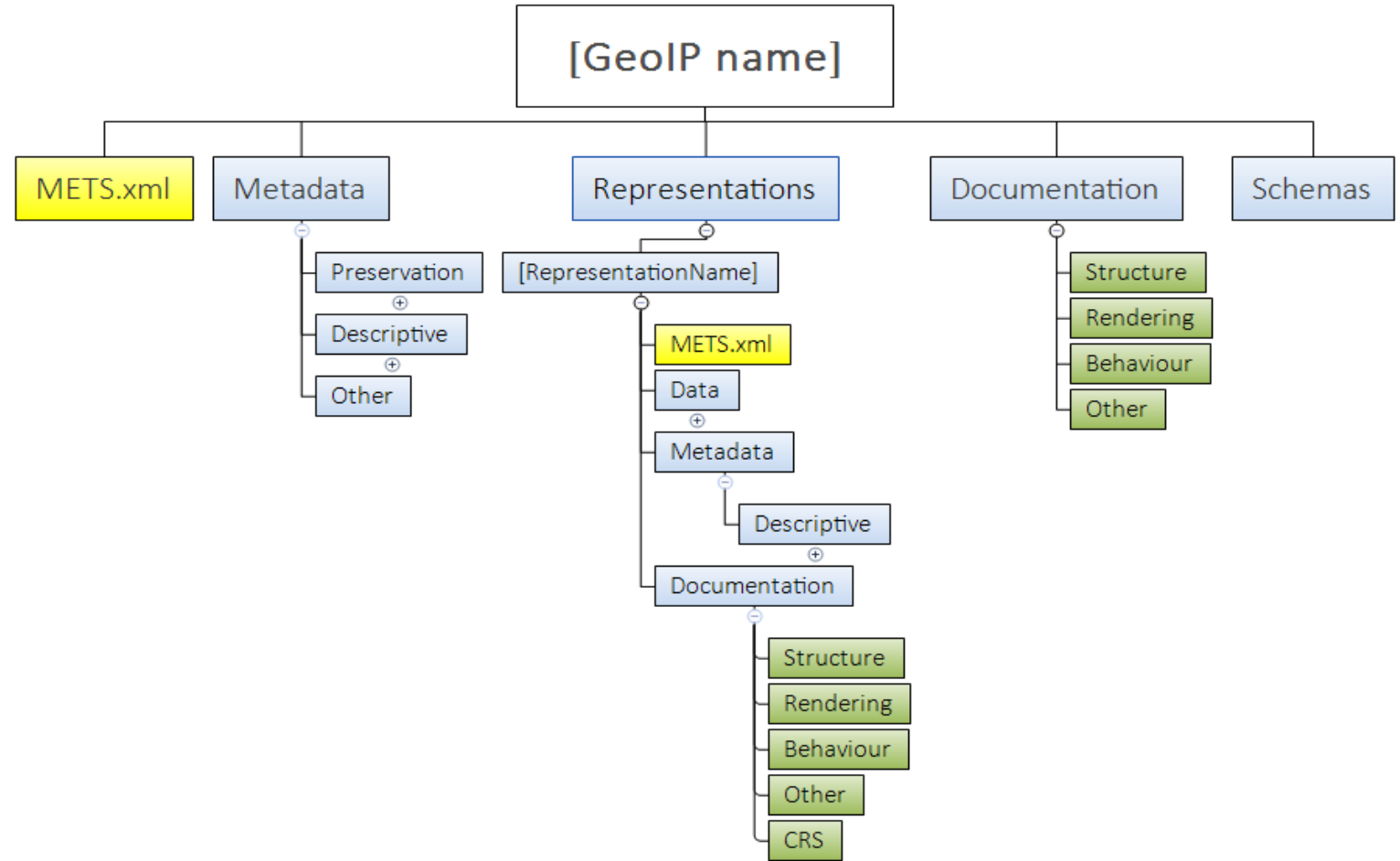
CITS Geospatial is part of eArchiving

The aim is to provide:

- Interoperability by default

- Openness and transparency

- Sustainability and legal compliance

… For Geospatial records
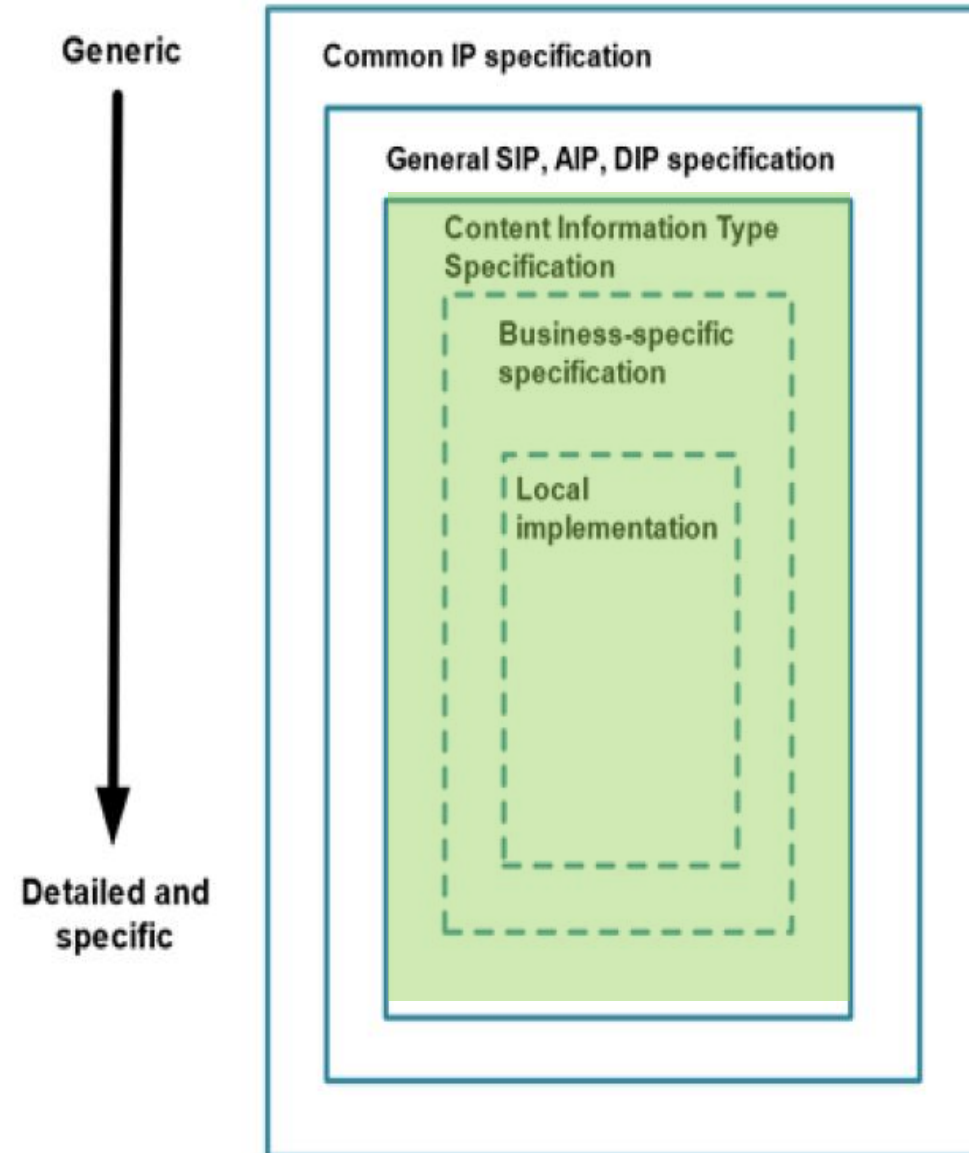
Quick intro:
Why preserve Geospatial records?

- Everything happens somewhere

- Form of official records

- Common denominator for all data

- Cornerstone of Digital Europe Data Economy
  - Analysis across time
  - Innovative solutions
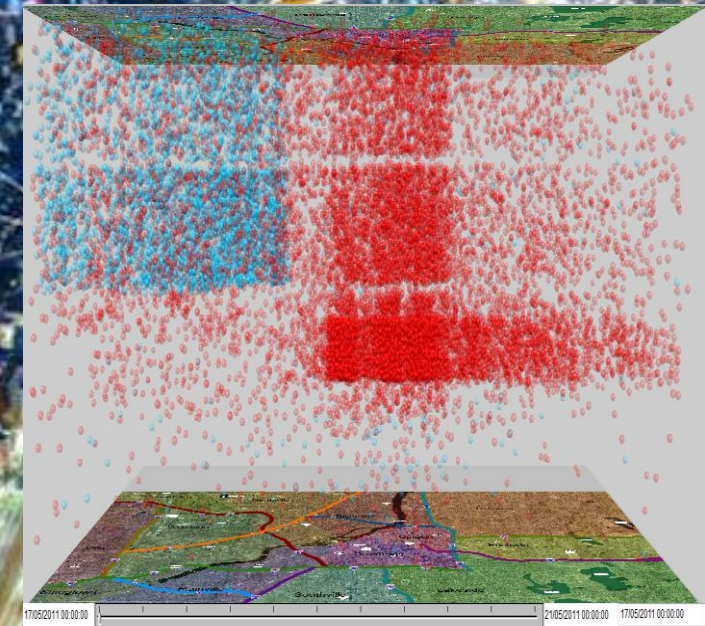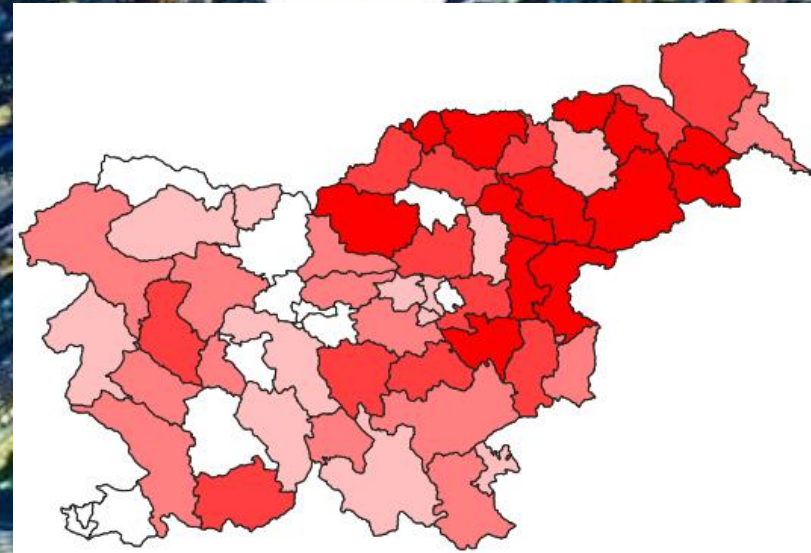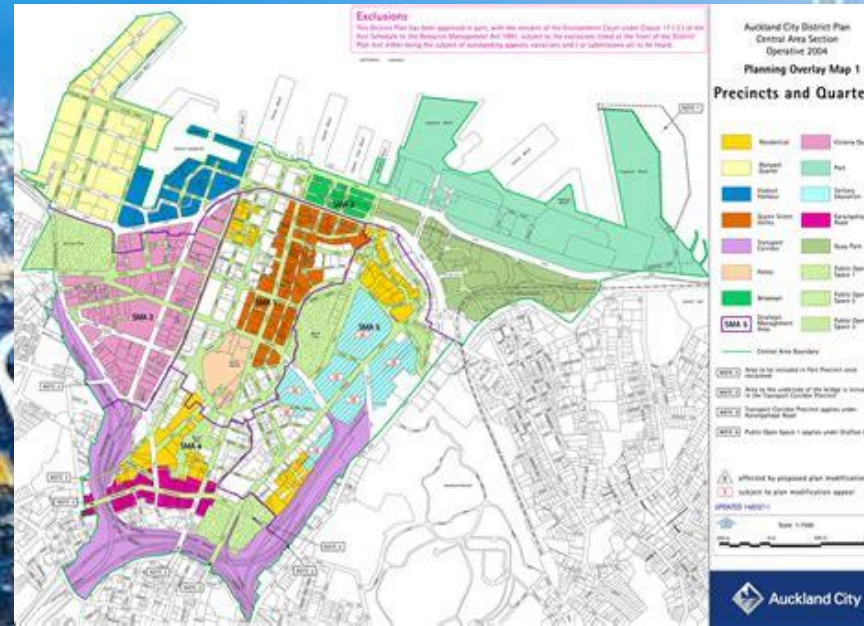
Stakeholders and benefits

Data Producers

- Anyone with information that needs to be preserved and reused

Archives

- Organisations responsible for digital archiving activities

Solution Providers

- Software and service providers

Regulatory agencies, consumers

Data Producers

- Better understanding of the concepts and approaches of long-term preservation of Geospatial records
- Better understanding of the need for archiving
- Better communication with archives

Archives

- Better understanding of geospatial records ensures proper preservation and prevents data-loss
- Interoperability and data reuse potential
- Better communication with all stakeholders

Solution Providers

- Cheaper implementation and maintenance
- Better mapping of product scope to archival activities
- Pan-European market potential

USERS

- Research community
- Innovative digital economy businesses
- Machine applications (AI, ML, DL…)

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/
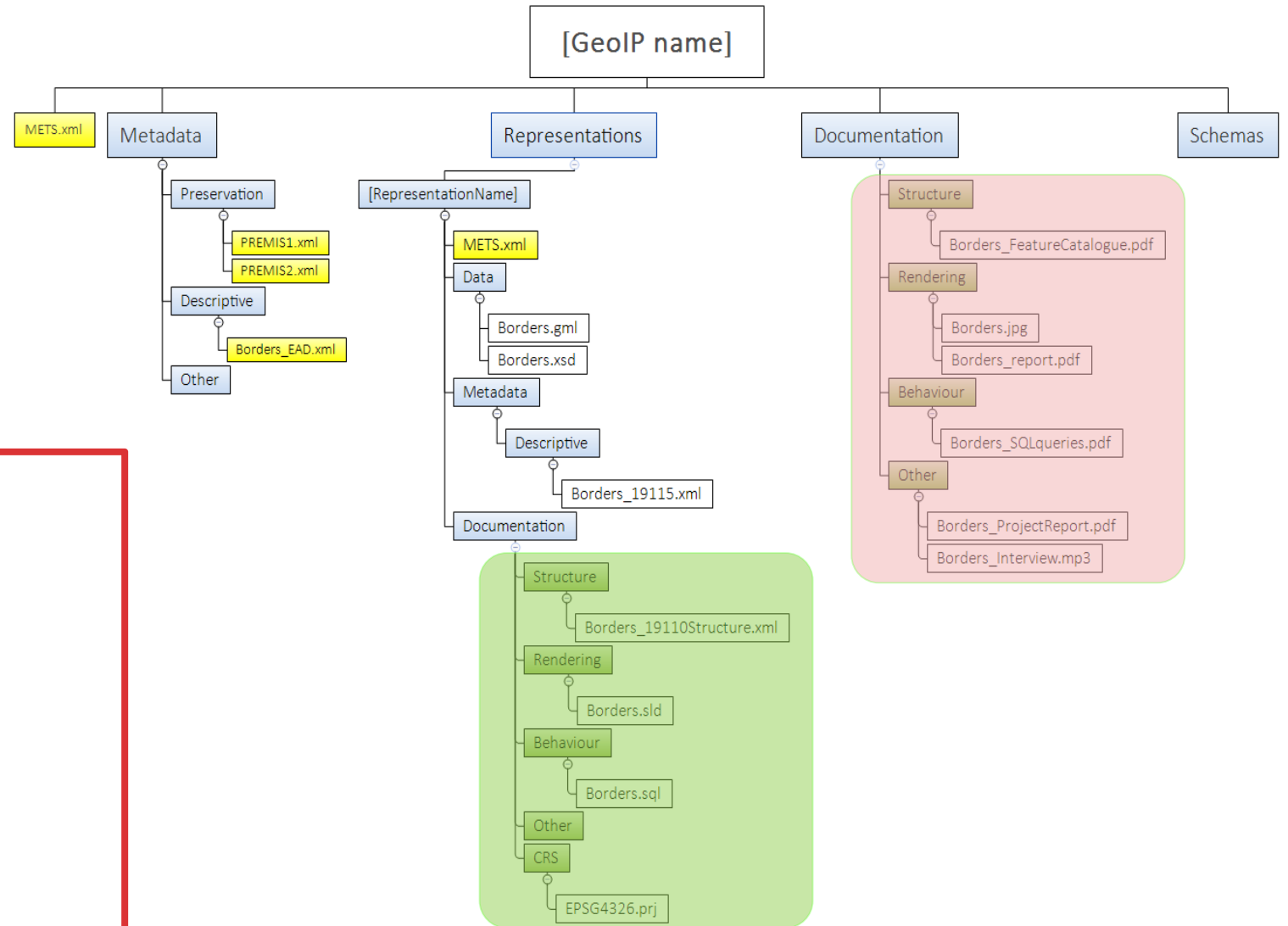
2.2.
METS file requirements

- 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 carring 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 -->
```


GEO_3	Content Information Type Specification	1..1
Ref CSIP4	mets/@csip:CONTENTINFORMATIONTYPE	MUST
	For information packages that primarily contain geospatial data, the value in Package mets/@csip:CONTENTINFORMATIONTYPE MUST be "citsgeospatial_v3_0 " as taken from the CSIP Vocabulary for Detailed Content Type.	
	See also: Content information type specification	
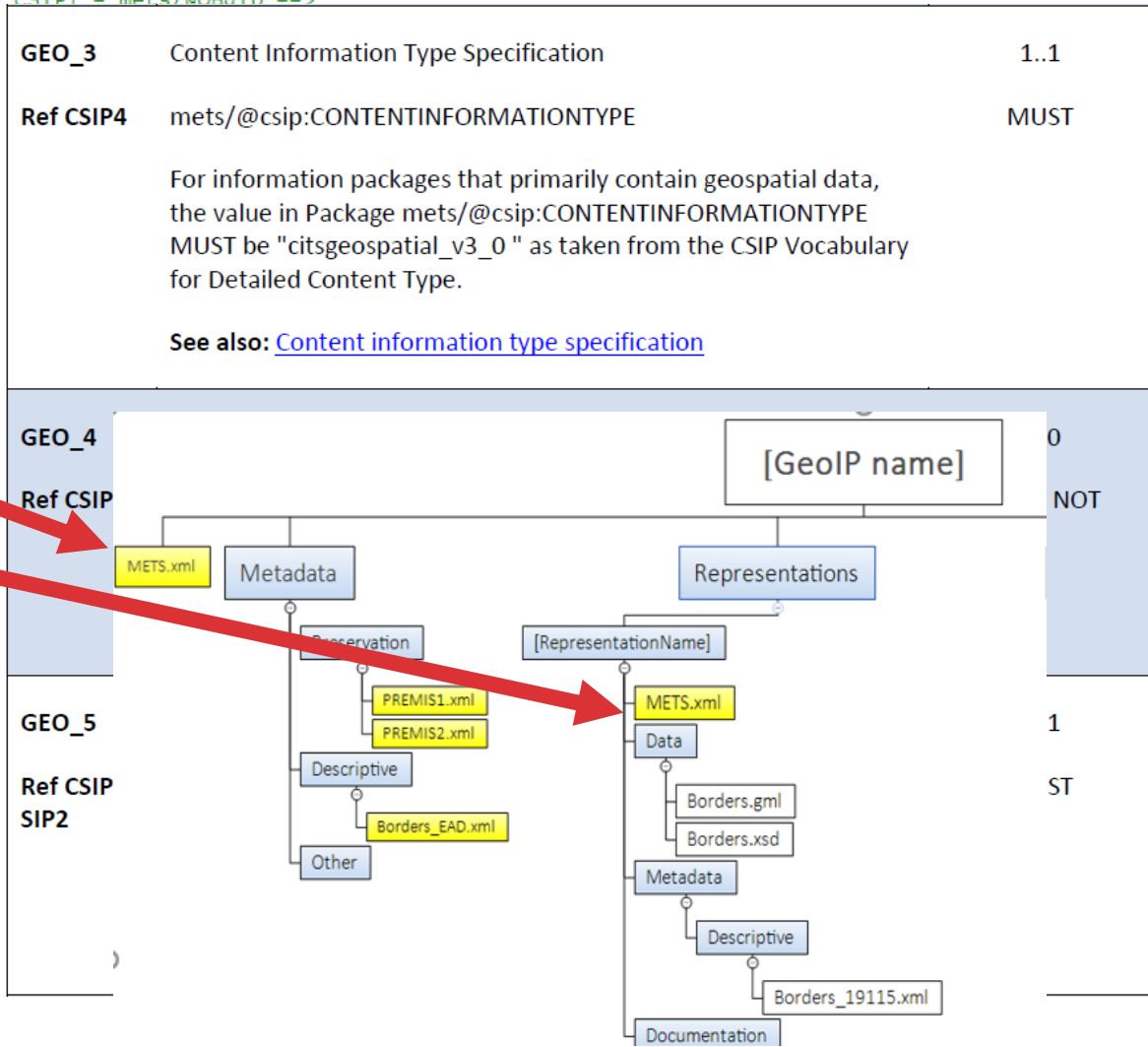
2.3. Data Requirements

- General data requirements

  - Vector requirements
  - Raster requirements

- Long term Preservation format profiles
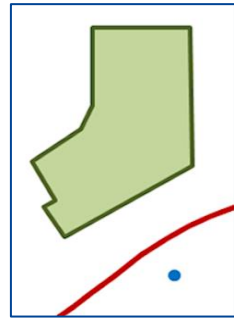
Vector
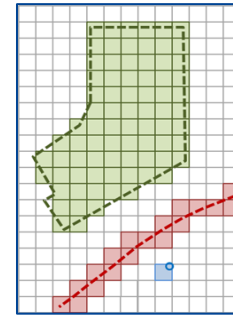
Raster

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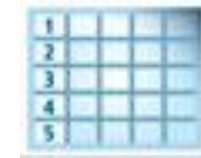
???? GML, SHP, KML, TIFF, JPEG2000, .gpkg, csv …. ????

Title

– 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	GML files larger than 1 GB **MUST** be subdivided into smaller GML files **Recommendation** *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.*	M
D_5.2-2	Geometries and attributes from the same geospatial vector dataset **SHOULD** be kept together within the same GML file	O
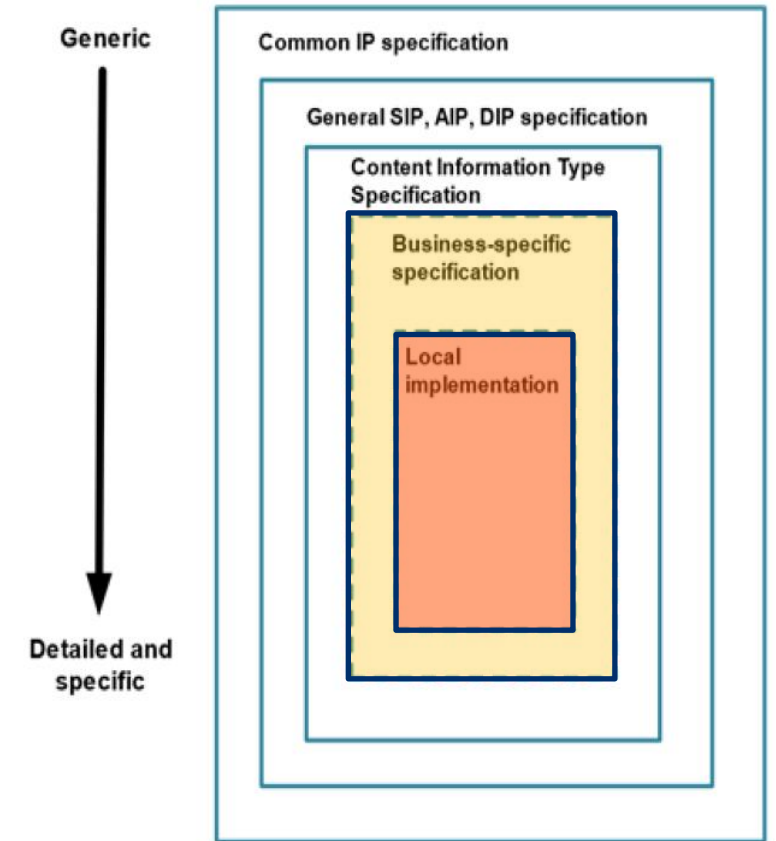
2.4. Documentation requirements

- Requirements on how to document the geodata in the IP
  - Structure of geospatial records
  - Rendering and visualisation
  - Behaviour
  - Coordinate reference systems
  - Other
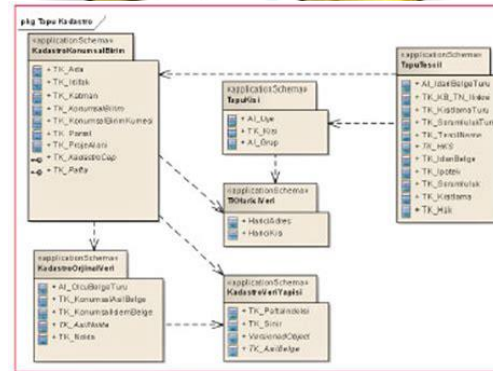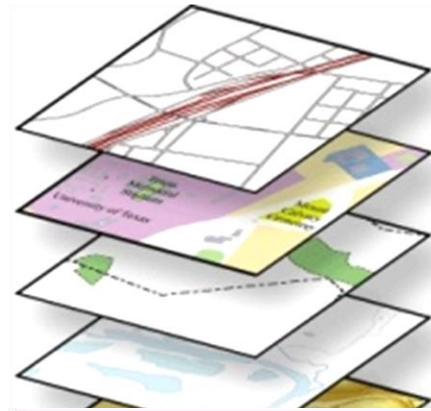
Figure 3: Data groups (packages) of the application schema

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

Source: http://ayresriverblog.com
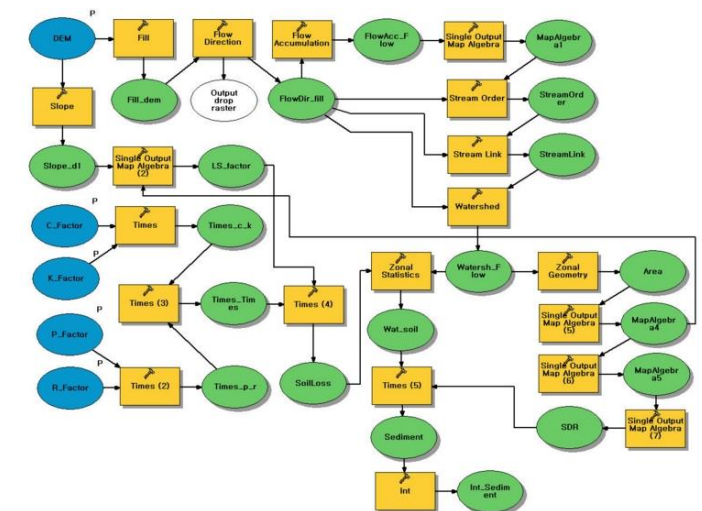
2.5. Geospatial Metadata

- Geospatial metadata enhances findability and reuse

- List of mandatory elements

- Support for standardised machine readable data

  - ISO 19115-1

  - ISO 19115-2

  - ISO 19165-2

- Support for proprietary machine-readable metadata

Guidelines for CITS Geospatial

- Introduction to Geospatial records

- Introduction to "Significant Properties" concepts for Geodata

- Rationalisations of all CITS Requirements

- Examples for many requirements
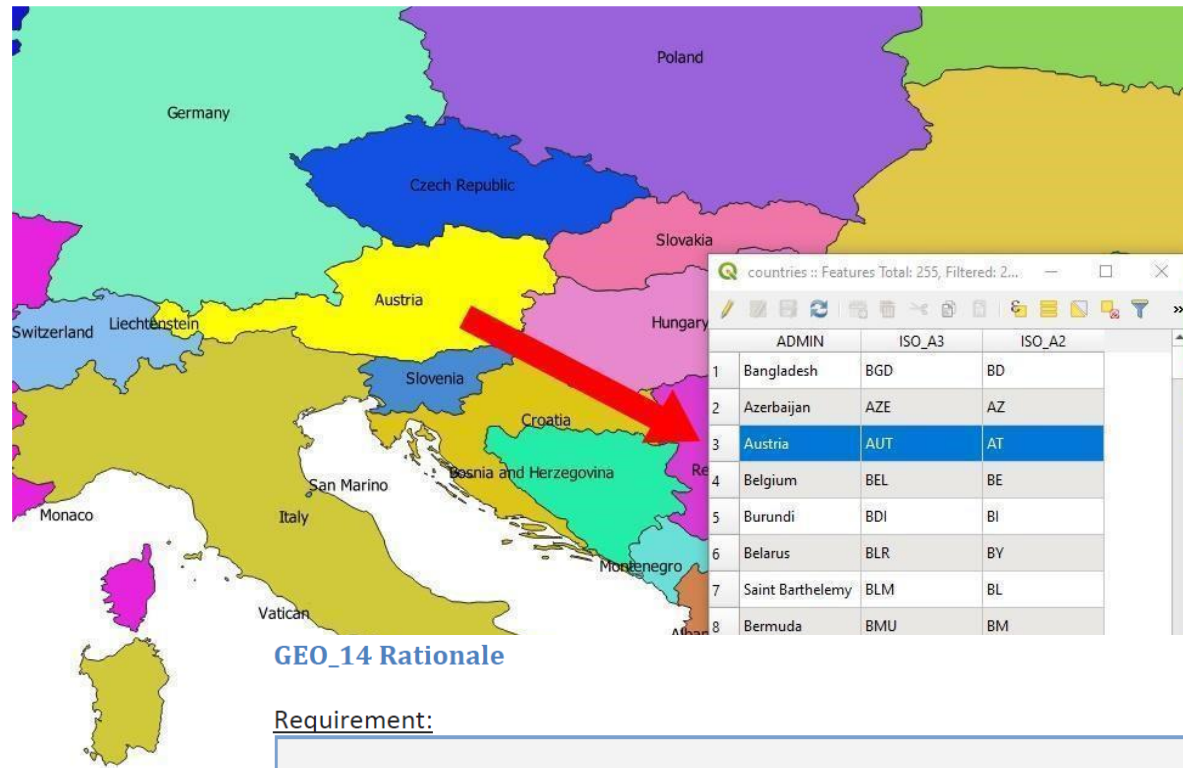
GEO_14 Rationale

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

- Rationalisations of all CITS requirements

- With examples based on standards

  - OWS context
  - WKT2 for CRS
  - SLD
  - etc.
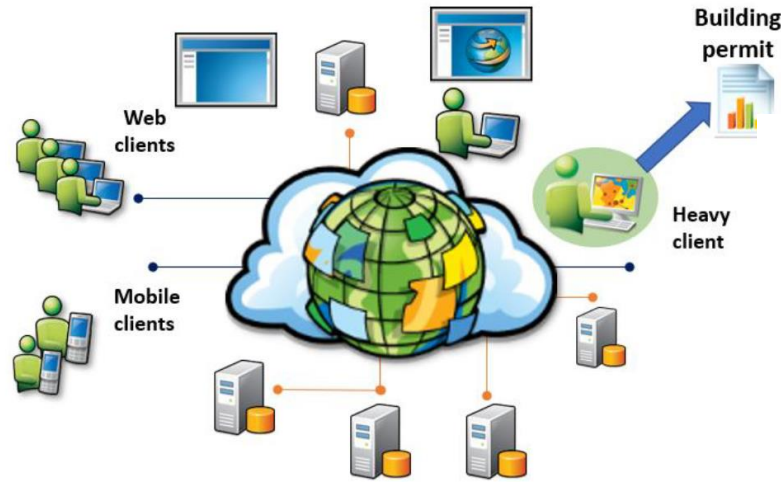
GEO_33 Rationale

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 automatisation.

Example:
An example of Standardised machine-readable formats for the rendering of geospatial records are SLD[20] files. KML[21] files also have some of that capability:

SLD files example
SLD is an OGC[22] (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 QGIS[23]. 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>
```
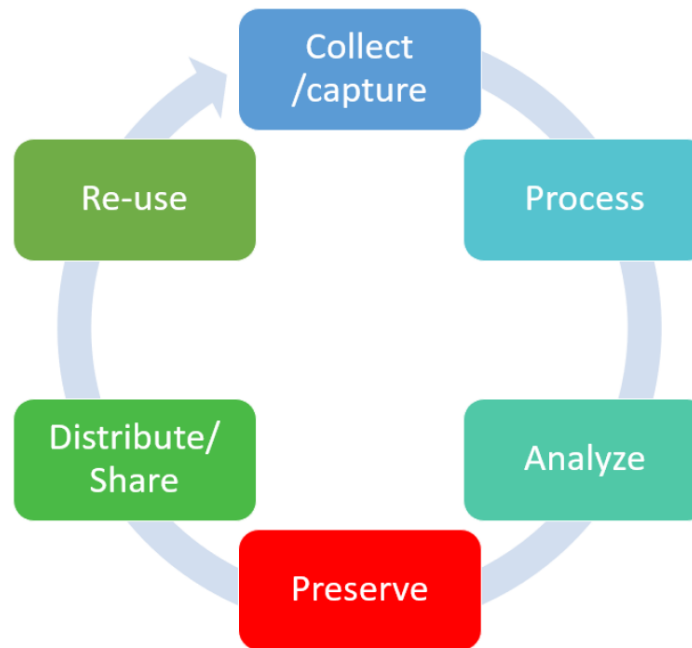
What is new in 3.0

- Alignment with the CSIP package structure

- Requirements redesigned according to "Significant Properties"

- Introducing "Long term preservation Profiles"

- Support for Geospatial metadata Standards

- Guidelines

CITS Geospatial 2.0.

CITS Geospatial 3.0.

Guidelines CITS Geospatial 1.0

Guidelines CITS Geospatial 1.0

Validation criteria

+ METS requirements
+ Updated structure
+ Support for various metadata
+ Long term preservation profiles

Explanations and examples moved into guidelines

NEW in 3.0

+ Intro to GIS Systems
+ Examples for GIS structures
+ Examples for CRS transformations
+ Extendable with Long-term preservation profiles

How to get started with Geospatial data preservation

Learn about Geospatial
Geospatial records are very complex. Learn about geospatial data and the formats. This will help you understand the specification requirements.

Allocate resources
Geospatial records preservation requires:
- A lot of disk space
- Geospatial tools
- Qualified personnel

Revise your workflows
A good and optimised workflow saves a lot of time.
Geospatial data requires adjustment in all the steps of the OAIS Process

Geospatial Records Preservation
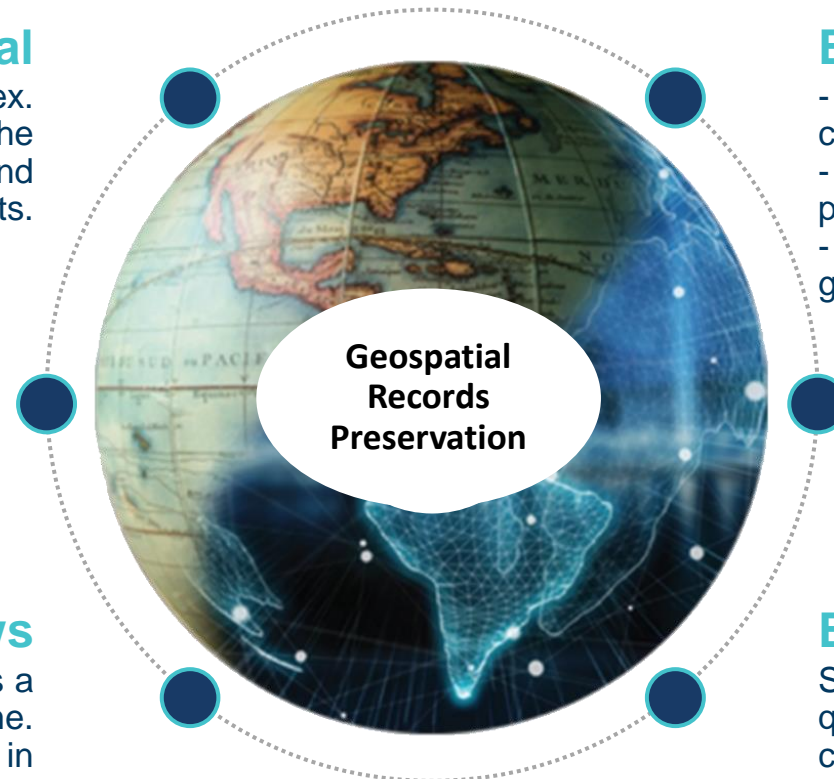
Engage the Community
- Identify their needs through cooperation.
- Educate them on the importance of preservation.
- Cooperate in creation of local guidelines

Acquire authoritative basemaps
Authoritative base maps give your data authentic context.
Prioritise acquisition of base maps first.

Exchange Experiences
Share your experiences, post your questions in DLM and other Preservation communities. Help eArchiving help you.

Geospatial

European Commission

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

Impressions from early adopters

Ann-Kristin Egeland
The Danish National Archives
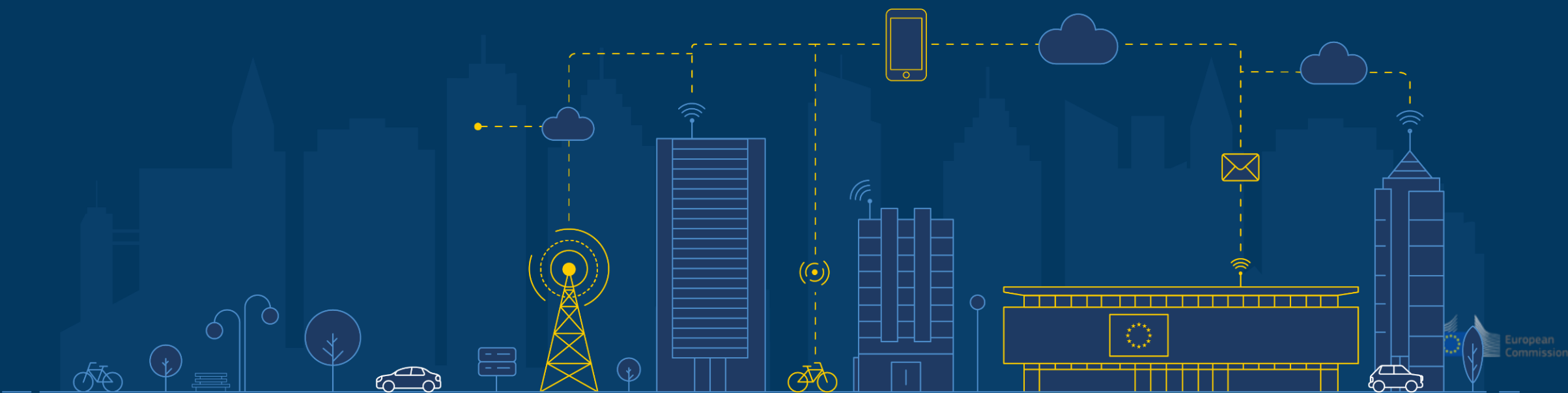
Scope of presentation

Impressions from early adopters of the CITS Geospatial specification

Questions from adopters of the CITS Geospatial specification will be used to explain and illustrate selected core requirements of the specification in more detail

Questions

1. Why is my geodata of Finland displayed next to Great Britain?
2. How do I identify a relevant geodata file in the IP?
3. How far is work on the GeoSIP validator?

CITS Geospatial

Specification for the E-ARK Content Information Type
Specification for digital geospatial data records archiving

DILCIS BOARD

https://dilcis.eu/content-types/cs-geospatial-data

WHY this presentation?

The aim of this presentation is to:

- invite you to help develop the validation tools for a SIP with geodata compliant with the CITS Geospatial specification – let's develop a tool we can all use!

- encourage you to implement the CITS Geospatial specification in your archive, library, company …

- make it easier for you to understand and read the CITS Geospatial specification

- explain to you why these requirements are necessary when preserving geospatial data

WHAT will you learn?

You will:

1. Get a better understanding of selected and central requirements from the specification

2. See what troubles adaptors when preservation geodata using this specification

3. Know what geodata validation tools are available so far and what tools will be developed

Digitalbevaring.dk

Question 1

Why is my geodata of Finland displayed next to Great Britain?

Requirement GEO_15 an GEO_16

"When I open my Finnish dataset and a background map i Q-GIS, I can see that the dataset is in the middle of the ocean – and not in Finland. Why is that?"

Markus Merenmies
National Archives of Finland
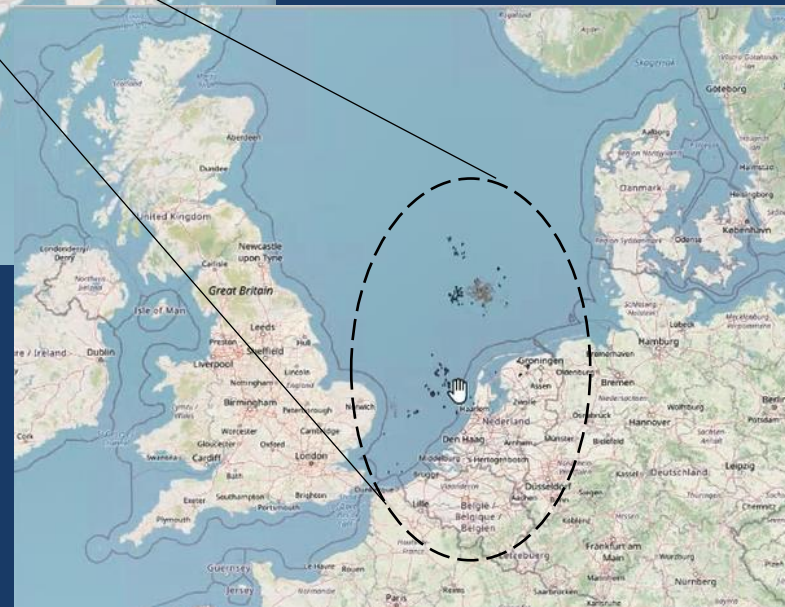
Requirement GEO_15

CRS definition

The answer to the question on previous slide is related to this requirement:

GEO_15	CRS definition	
	Every geospatial dataset **MUST** be accompanied with information about its underlying Coordinate Reference System (CRS) in one of two ways:	1..1 MUST
	• Full description of the CRS together with the archived data (within the geospatial file itself or in an accompanying file) • The geospatial file contains a reference to a CRS registry	

Why is the CRS information important to preserve?

Because you have to display your geodata in a Coordinate Reference System (CRS) that corresponds to the CRS of your geodata.
If not – the geodata will not be placed correctly on the 2D representation of the earth.

Methods/projections
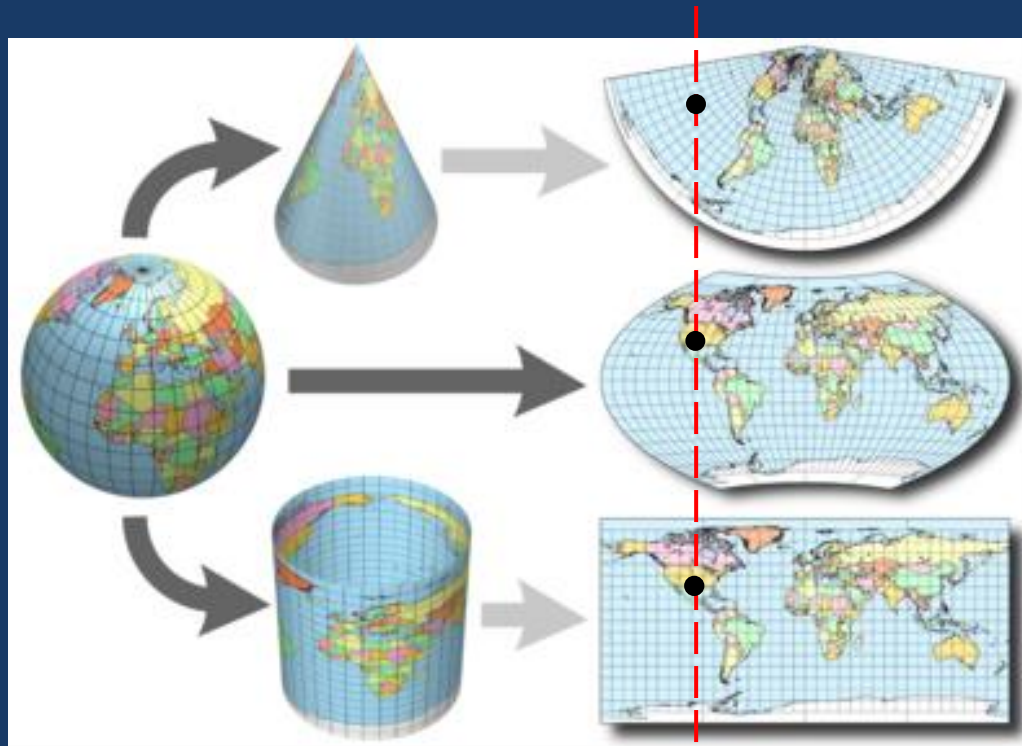
Cone

Cylindrical/curves

Cylindrical/lines

A projected coordinate system is a flat, 2D representation of the 3D earth.

Depending on the method used for the 2D > 3D transformation (also called a projection) the coordinate system will look differently – the countries will be stretched or pressed in the 2D representation of the earth.
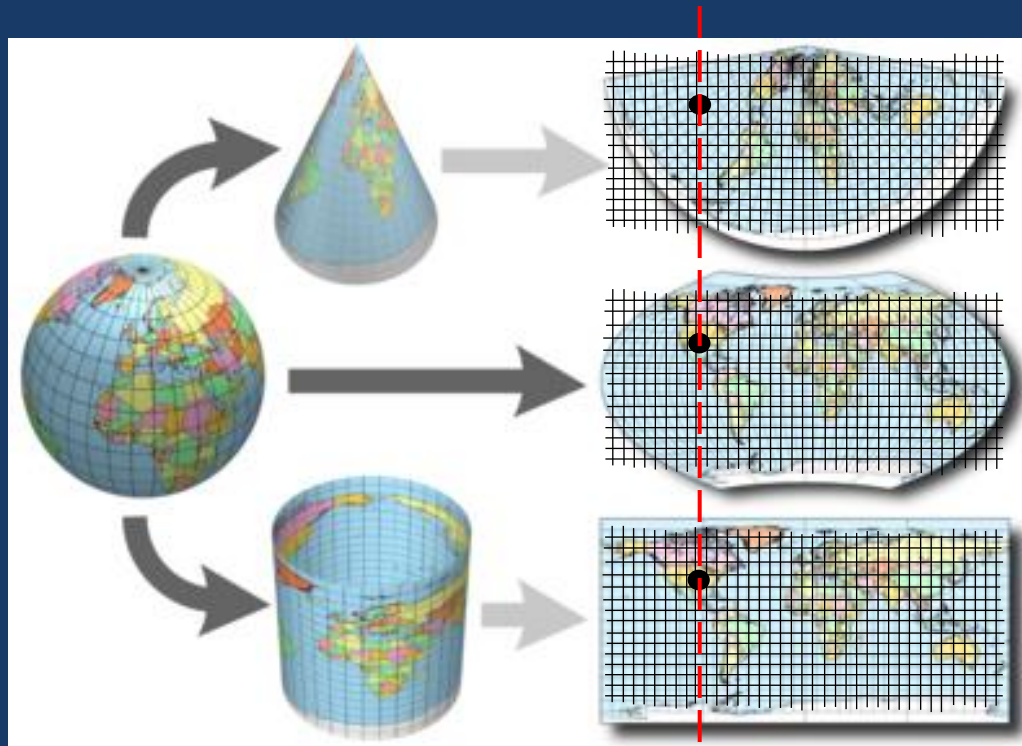
Example: The same geodata point (in a cylindrical/lines CRS) is placed differently in different projections

Why is the CRS information important to preserve?

Methods/projections

Cone

Cylindrical/curves

Cylindrical/lines

Here the CRS of the geodata (coordinate/point/black dot/●) is in a "**cylindrical lines**" projection.

Placed in other CRS/projections ("**cone**" or "**cylindrical curves**") the geodata is not placed correctly on the 2D version of the earth.

In the cone projection the point is in the middle of the ocean and not in Central America

Source: CA Furuti, progonos.com/furuti

Example: The same geospatial dataset (in a cylindrical CRS) is placed differently in different projections

Why is my geodata of Finland displayed next to Great Britain?
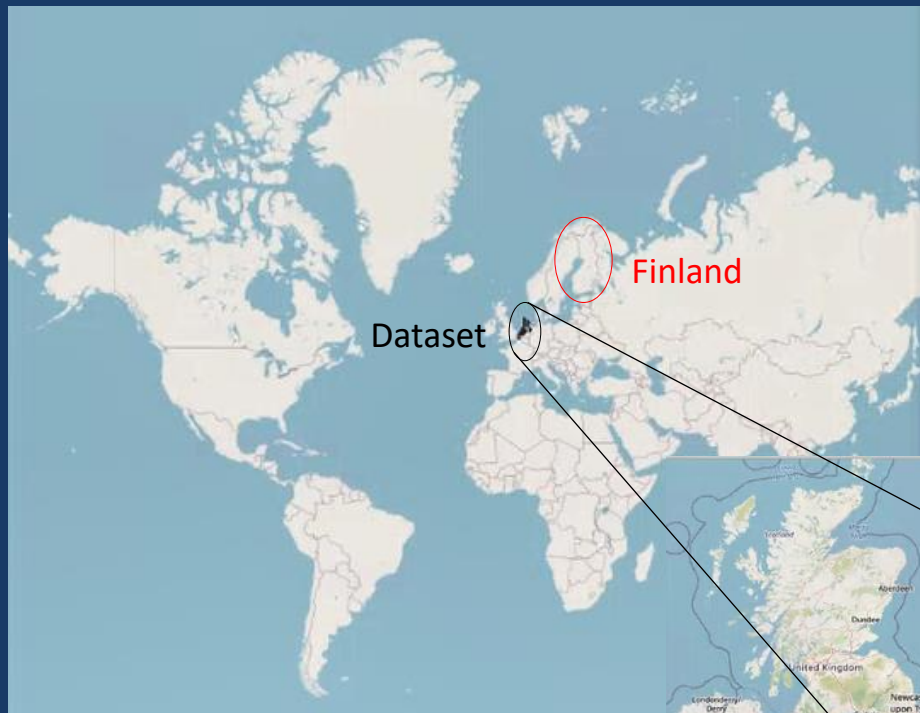
Finland

Dataset

The answer
Because the CRS of the Finnish dataset (EPSG:3067) is different from the CRS of the background map (EPSG:3857)
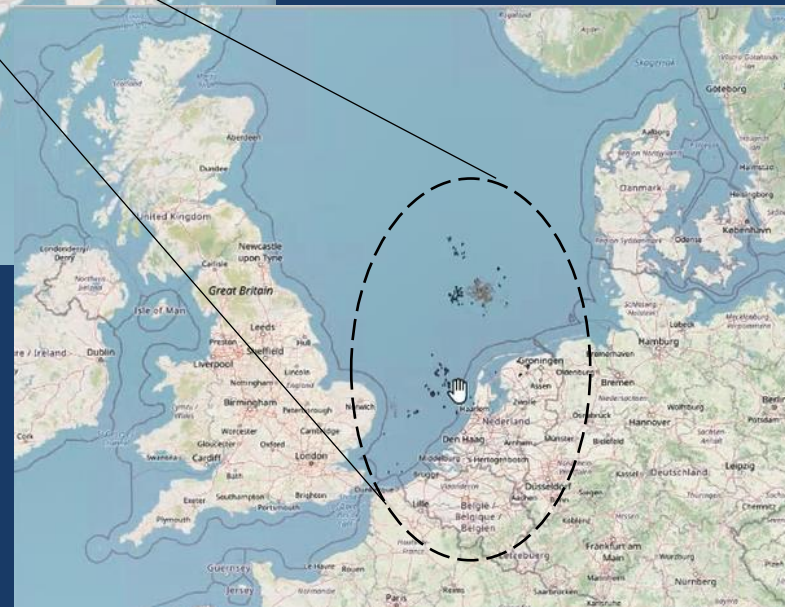
Dataset:
EPSG:3067
ETRS89 / TM35FIN(E,N)
Finland

Background map:
EPSG:3857
WGS84
OpenStreetMap
OSM Standard

Solution to the problem

View your geospatial dataset on a background map in the same CRS as your geodata

To view your Finnish dataset in CRS EPSG:3067, you should use a background map in the same CRS EPSG:3067.

Download a suitable background map from the map agency in the right CRS.

This question raises other important questions:

- Must a geospatial dataset always be viewed on top of the original background map to reflect the original data which a legally binding decision is based on?

- Must the archive preserve historical background maps too?

- Which background maps are suitable for purposes of future users of the archived geospatial dataset? (background map with lakes, streets, city names …?)

Question 1

Another requirement regarding CRS…

Requirement GEO_16

Geografic location validation

This requirement regarding *geographic location validation* reveals when a transformation from one Coordinate Reference System (CRS) to another go wrong during submission or migration in the archive

GEO_16	Geographic location validation	0..1
	The geographies in the geospatial records **SHOULD** be located within a fixed bounding box defined in the submission agreement between the producer and the archive according to the expected location and extent of the dataset	SHOULD

Geographic location validation

Example of a validation of a Danish geospatial dataset

This image shows a validation where some of the coordinates (orange points) in the geospatial dataset is outside the (light blue) bounding box
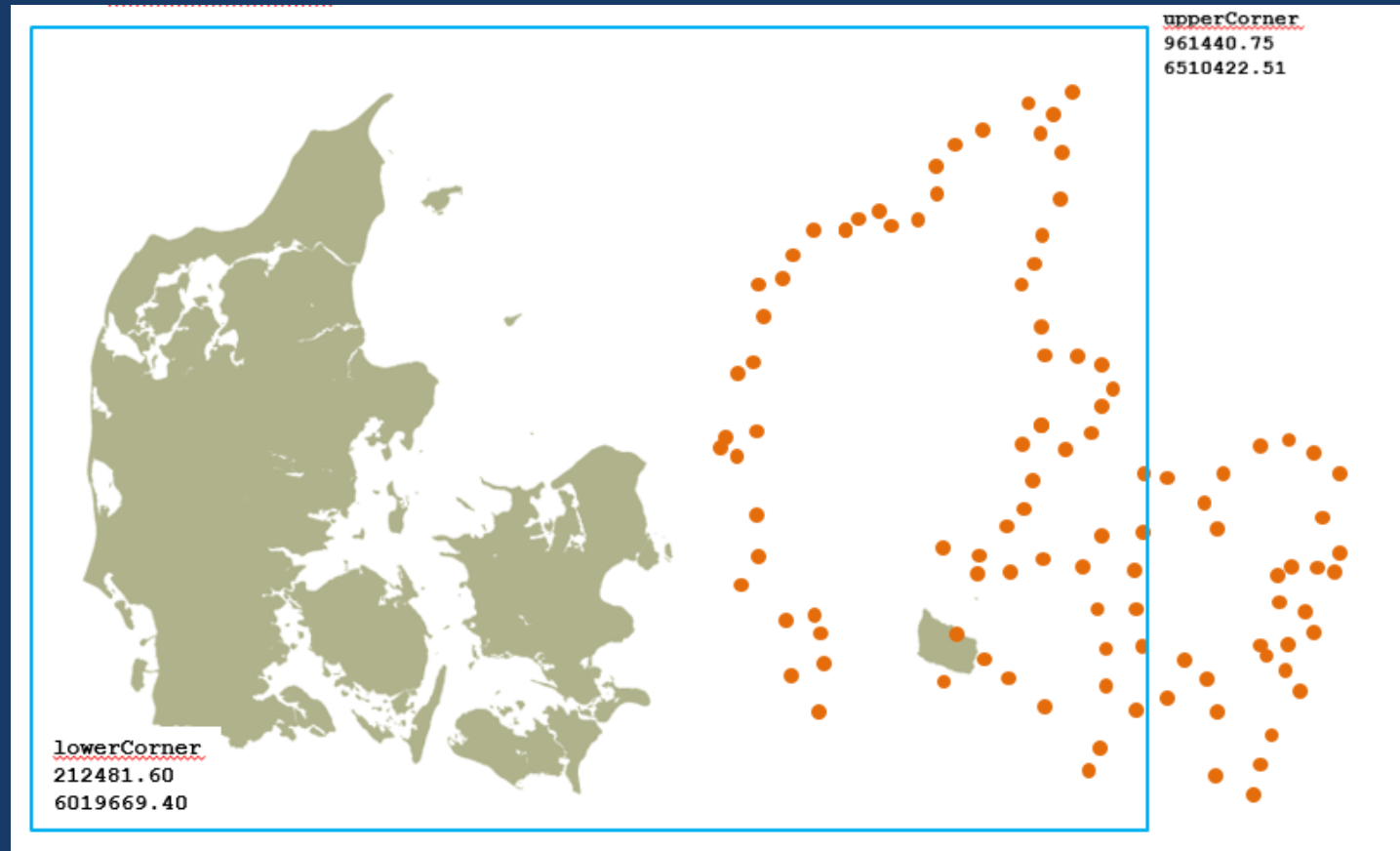
Wrong transformation of geodata

The orange coordinates are not in EPSG:25832

A background map of Denmark is added

In this case the problem is the other way around.

The background map of Denmark (green map) is in correct CRS EPSG:25832.

But the geodata displaying locations of shipwrecks is transformed wrongly and does not correspond to the CRS EPSG:25832 allowed by The Danish National Archives.

The validation of whether the coordinates are located within the light blue bounding box caught this error in transformation of geodata on submission

Correct transformation of geodata

The orange coordinates transformed correctly to EPSG:25832

When the orange coordinates are transformed correctly to CRS EPSG:25832 the shipwrecks are located correctly along the cost line of Denmark
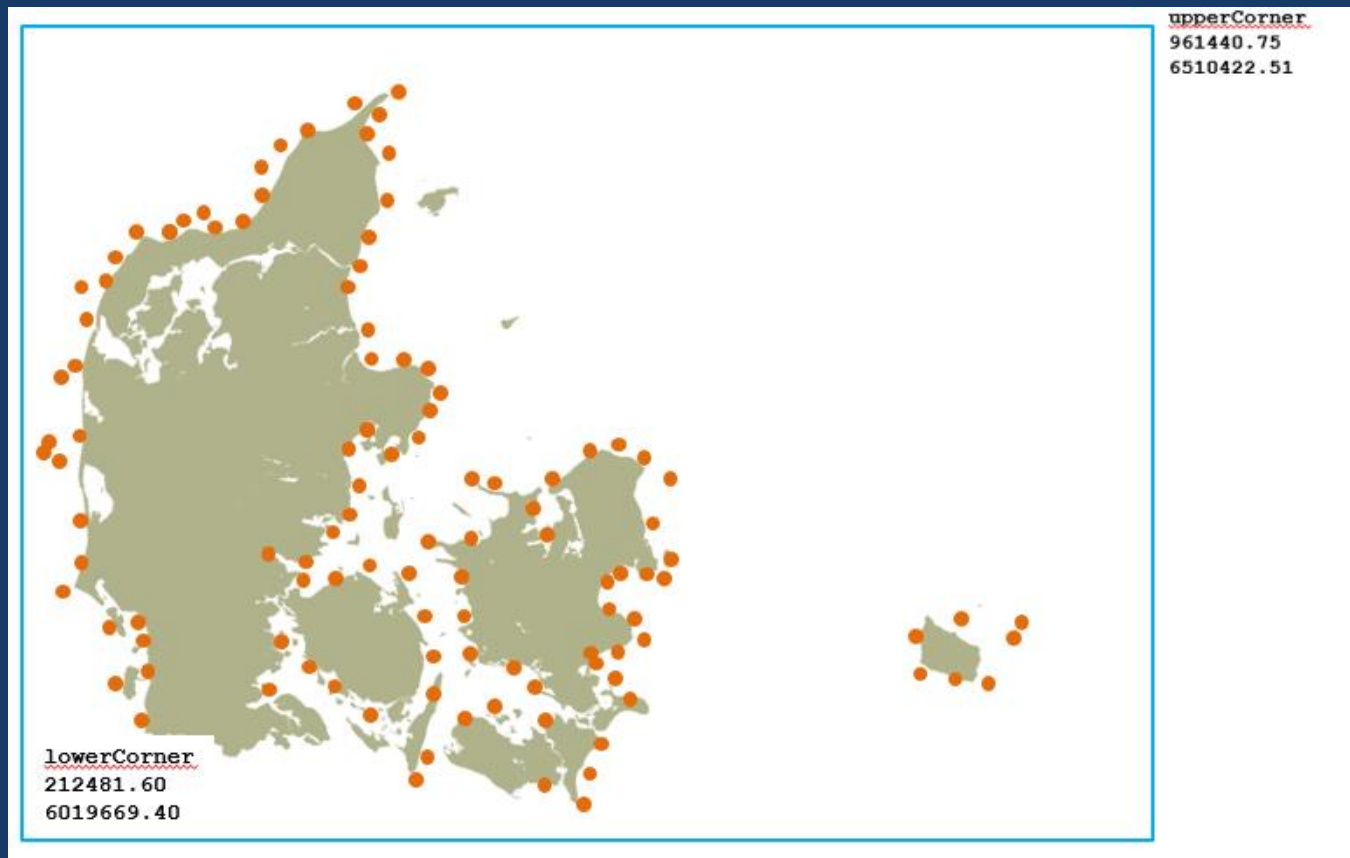
To sum up…

Preserving descriptions in the Information Package about the actual CRS of your geospatial dataset in the IP is very important!

These requirements in the CITS Geospatial specification ensure that:
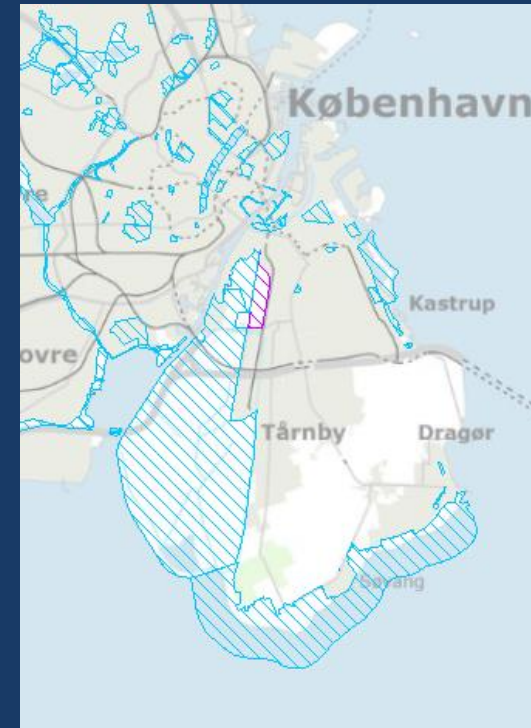
GEO_15 CRS definition
GEO_16 Geographic location validation

Question 2

How do I identify a relevant geodata file in the IP?
Requirement GEO_17, GEO_42 and GEO_43

Often a user of the archive asks for very specific geospatial information, like *"Could you give me a map of protected sites from 2000-2003 from this specific region in Denmark?"*

Ann-Kristin Egeland
The Danish National Archives

Requirement GEO_17

Metadata

This requirement provides a description of each geospatial dataset in the Information Package:

GEO_17	Metadata	
		1..n
	Every geospatial dataset **MUST** be accompanied by a metadata file, that describes the dataset with the basic required information	MUST

Requirement GEO_42

Standardised machine-readable geospatial metadata

And the CITS Geospatial specification recommends that these descriptions of geospatial datasets are provided in standardised machine readable formats, like INSPIRE metadata rules or ISO metadata standards

GEO_42	Standardised machine-readable geospatial metadata	0..n
Ref GEO_17	Descriptive geospatial metadata in the long-term preservation format representation of the Information Package **SHOULD** be provided in the form of standardised machine-readable format compliant with geospatial metadata standards	SHOULD

Metadata standards
Recommend use of standards

- INSPIRE
- ISO 19115-1
- ISO 19115-2
- ISO 19165-1
- …

Example of INSPIRE metadata

Example of IP with 3 GML files and metadata in INSPIRE standard

An Information Package with one geospatial dataset divided into 3 GML files < 1 GB covering the main parts of Denmark
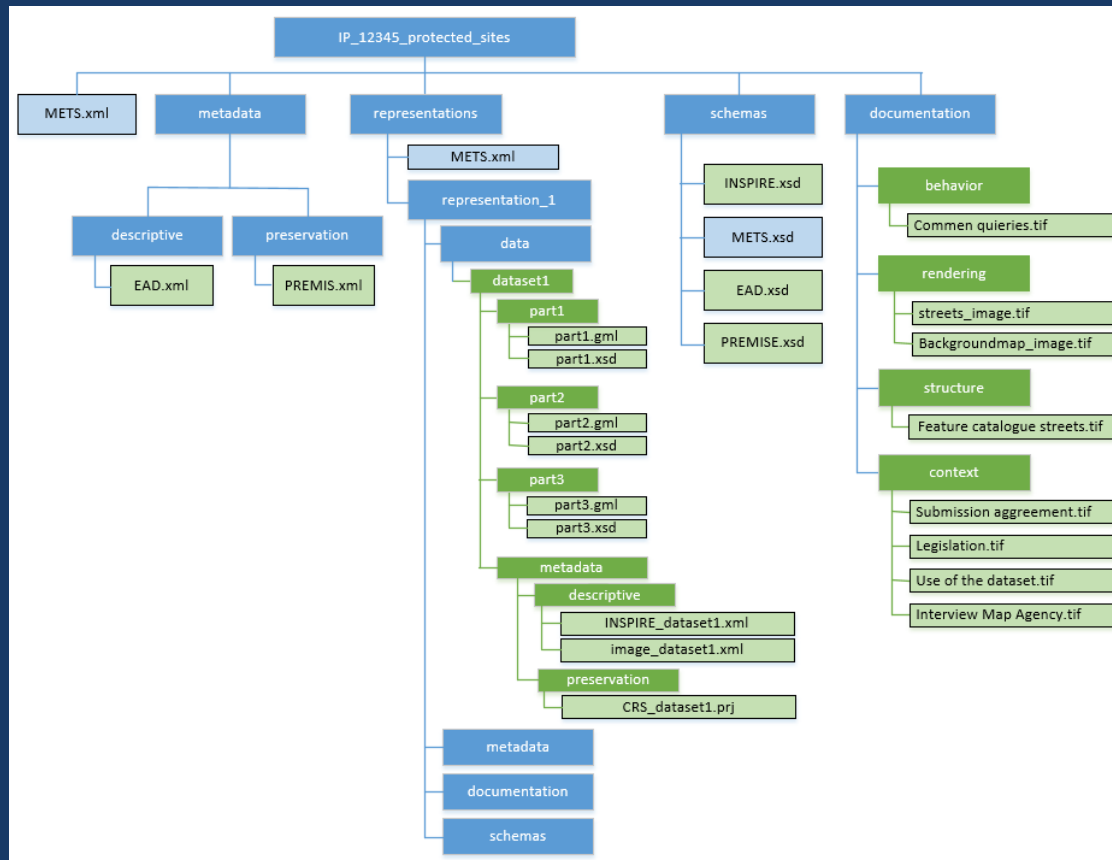
- Which GML file contains geodata from which part of Denmark?

- Should each GML file have an INSPIRE metadata file with regions named to enable search of a specific file with geodata from a specific region?

- In which GML files are geodata of the small islands stored?
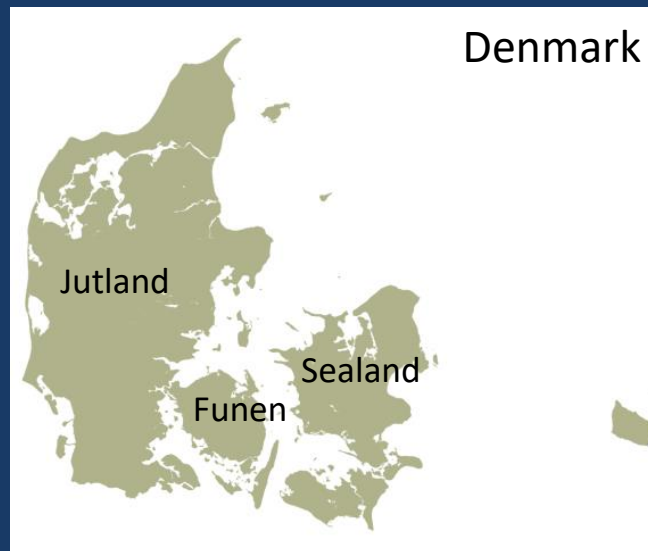
Example of IP with 16 GML files and metadata in INSPIRE standard

An Information Package with one geospatial dataset divided into 16 GML files < 1 GB covering
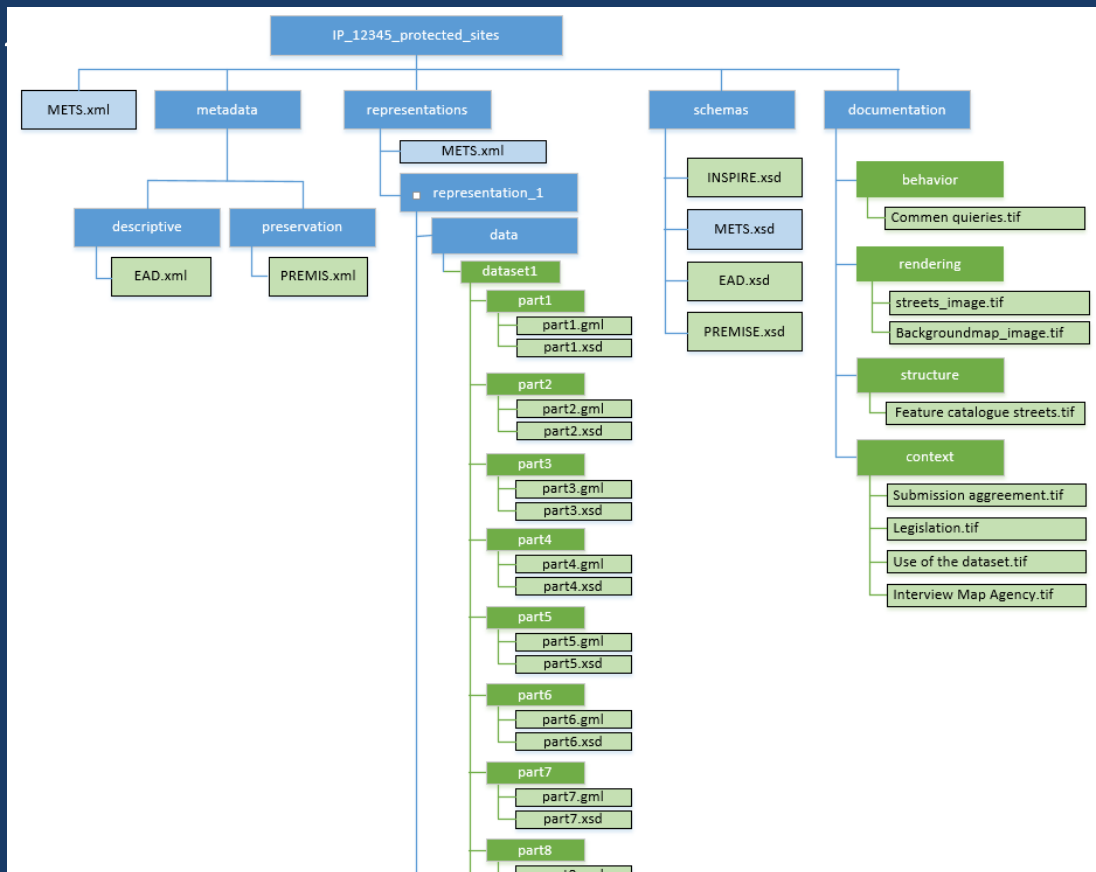
- Which GML file contains geodata from which tile?

- Should each GML have an INSPIRE metadata file to enable search of specific file?

- How would each of the 16 INSPIRE metadata files be described to enable search of which file has geodata from a specific region?
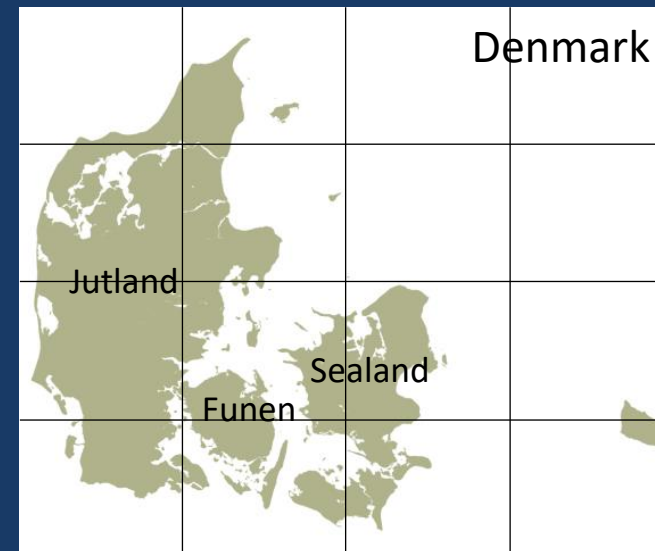
Requirement GEO_43
Non-standardised machine-readable geospatial metadata

To enable search of geodata in an IP from a specific region metadata could be stored in the information package in a database format acoording to this requirement:

GEO_43	Non-standardised machine-readable Geospatial metadata	0..n
Ref GEO_17	A copy of Geospatial metadata in non-long-term preservation representations **MAY** be stored in its original form as databases or documentation. However, if this data is stored in a long-term preservation representation, the formats need to comply with the archival guidelines (stored in approved long-term preservation formats).	MAY

Metadata in database format

An Information Package with one geospatial dataset divided into 16 GML files < 1 GB covering 16 tiles
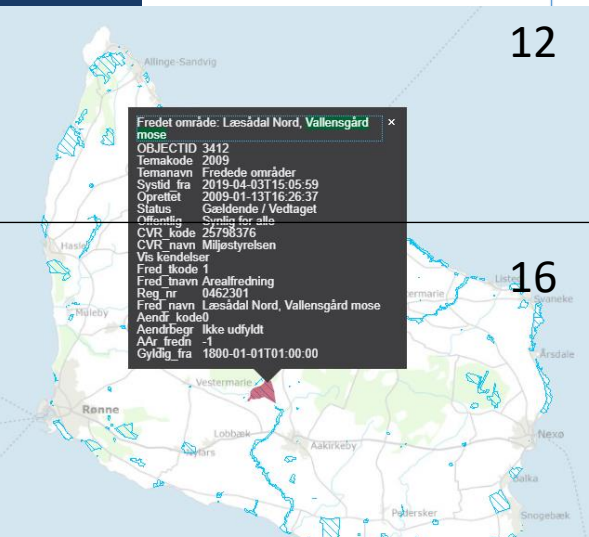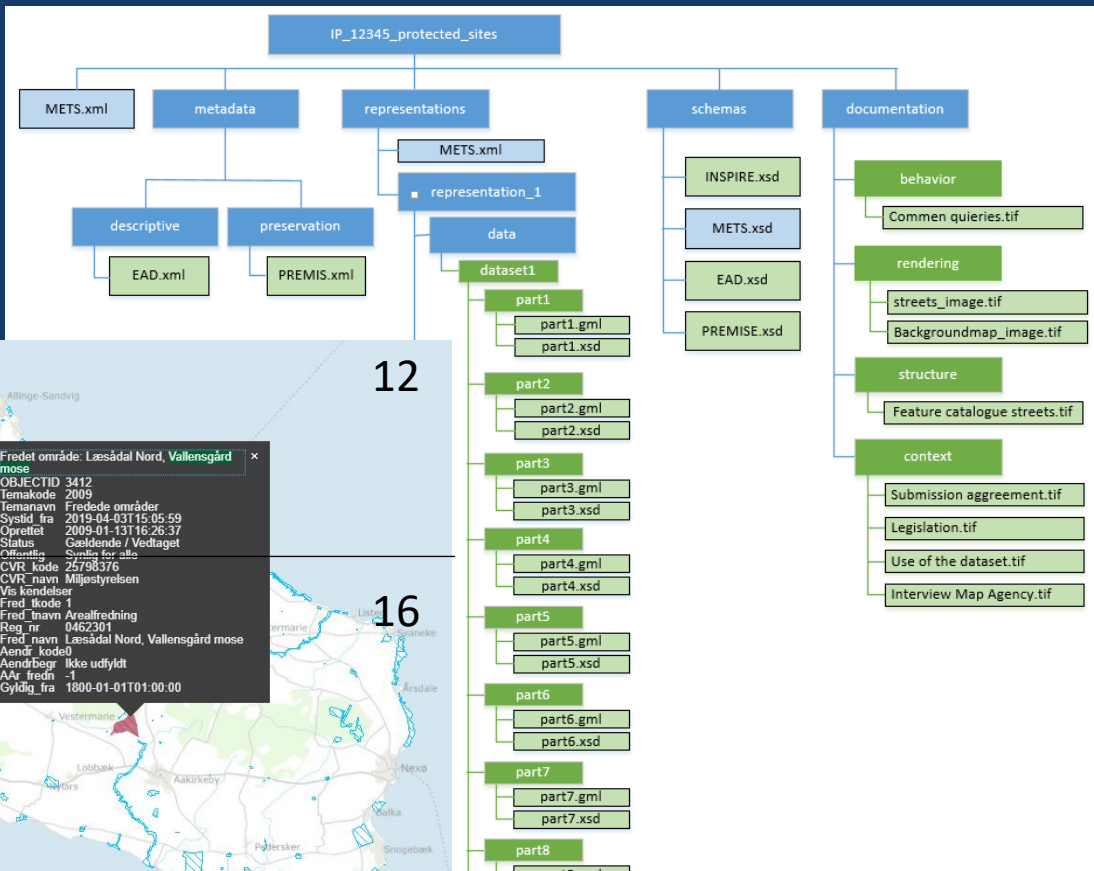
Metadata in database format (eg. SIARD file) in the IP

OBJECT ID	Theme	Name of protected site	Municipal	Region	GML file path
698	Protected Site	Gammelmosen	Gentofte	Hovedstaden	…/part11.gml
3412	Protected Site	Vallensgård mose	Bornholm	Hovedstaden	…/part16.gml
4443	Protected Site	Kongens have	Odense	Syddanmark	…/part10.gml
103	Protected Site	Tranekær park og sø	Langeland	Syddanmark	…/part14.gml
…					…/part2.gml
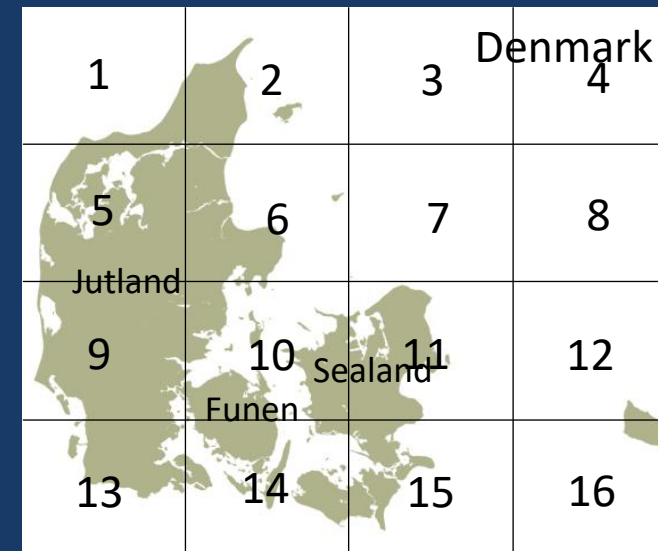
A user of the IP can search for the region "Hovedstaden" in the database file of the IP and find the files part11.gml and part16.gml which both contain protected sites of this region

54

Combination of SIARD and GML in IP

Do we need a new CITS SIARD_Geospatial specification?

To be able to store both geodata files (like GML or GeoTIFF) and a database (SIARD 2.2) we need a new specification (CITS?) for combining these two preservation formats (profiles)

The Danish National Archives are working on this…

CITS Geospatial

+

SIARD 2.2

= ?
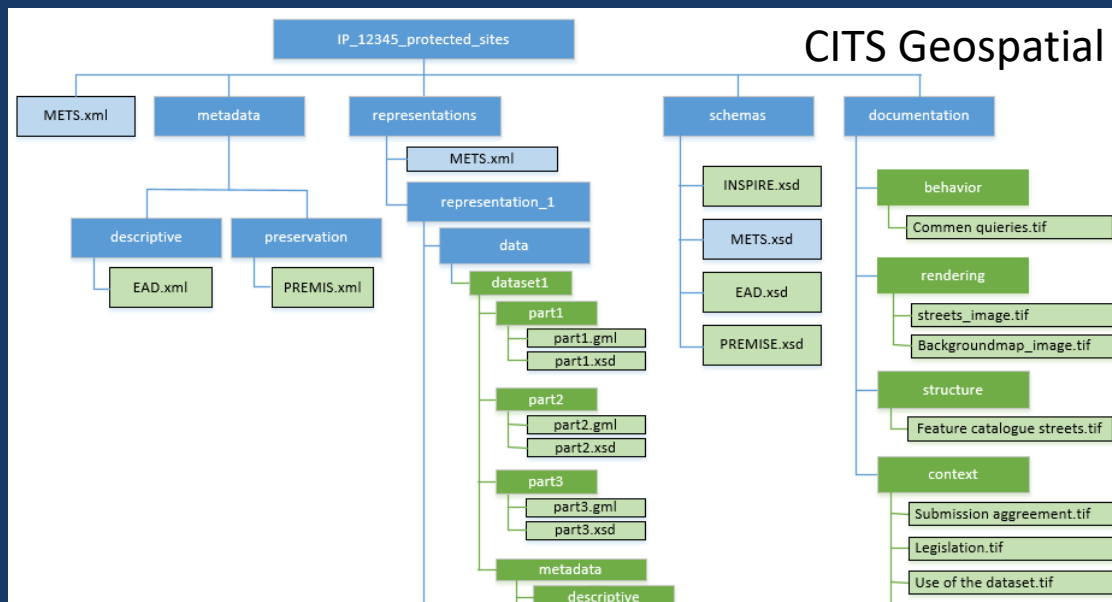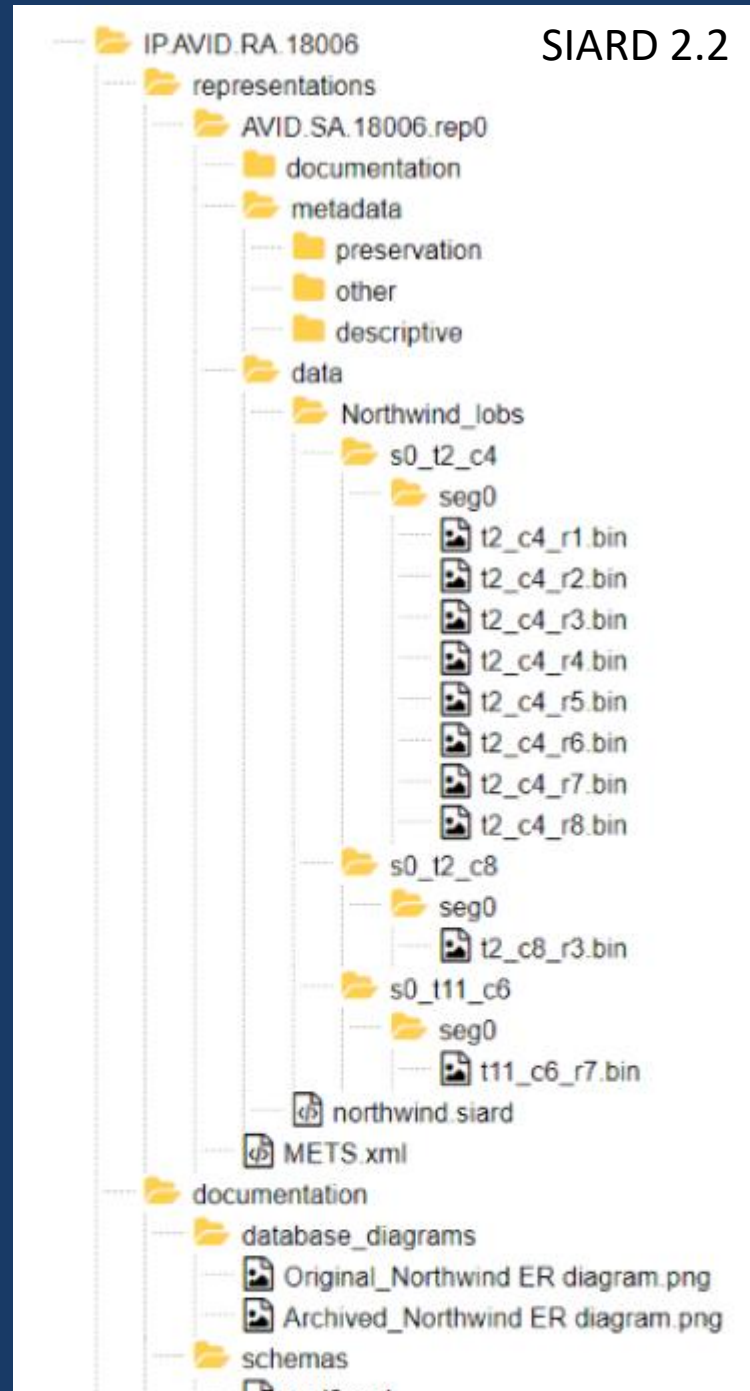
To sum up...

Consider how you describe your geodata files in the IP to enable search of a specific geodata file in the IP

These requirements in CITS Geospatial specification cover that user need:

GEO_17 Metadata
GEO_42 Standardised machine-readable geospatial metadata
GEO_43 Non-standardised machine-readable geospatial metadata

Question 3

How far has work on the GeoSIP validator progressed?

Requirements GEOSTR2 – GEOSTR6, GEO_2 and GEO_3

"How far has the work on the GeoSIP validator progressed? We would like to use it but we would like to know whether we can allow use of Czech names for folders. Not for mandatory E-ARK folders but only for subfolders."

Martin Rechtorik
National Archives, Czech Republic

Digitalbevaring.dk

Question 3a

Can we allow use of Czech names for subfolders?

Answer

Yes, you can allow use of Czech names of subfolders according to the CITS Geospatial specification
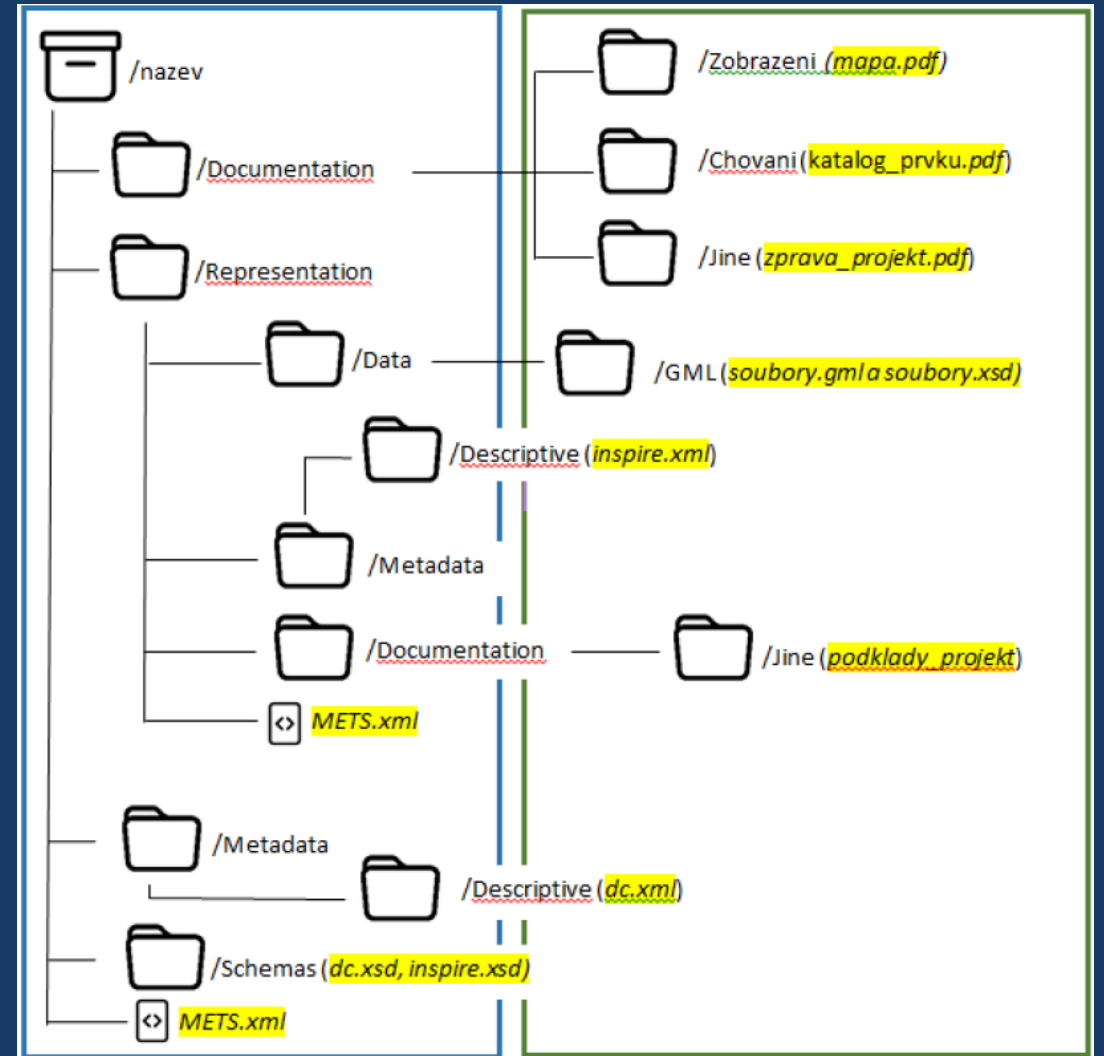

The green box shows subfolders with Czech names
The blue box shows mandatory E-ARK folders

Requirement GEOSTR2 – GEOSTR6

Naming of subfolders

No mandatory requirements regarding naming of subfolders in the *documentation* folder

– only recommendations (SHOULD – optional)

GEOSTR2: A documentation folder on package or representation level **SHOULD** include a subfolder named `structure`. This requirement is an extension of CSIPSTR16.

GEOSTR3: A documentation folder on package or representation level **SHOULD** include a subfolder named `rendering`. This requirement is an extension of CSIPSTR16.

GEOSTR4: A documentation folder on package or representation level **SHOULD** include a subfolder named `behaviour`. This requirement is an extension of CSIPSTR16.

GEOSTR5: A documentation folder on package or representation level **SHOULD** include a subfolder named `CRS`. This requirement is an extension of CSIPSTR16.

GEOSTR6: A documentation folder on package or representation level **SHOULD** include a subfolder named `other`. This requirement is an extension of CSIPSTR16.

Question 3b

How far has work progressed on the GeoSIP validation tool?

Answer

The E-ARK GeoSIP validator is not done yet

It should validate these requirements:
- Whether the IP is compliant with the CSIP specification (folder structure and METS file)
- GEO_1 to GEO_10 in CITS Geospatial specification: METS file requirements for IPs with geodata

Requirement GEO_2

Type

GEO_2	Type	1..1
Ref CSIP2	mets/@TYPE [Description of the element]	MUST
	For information packages that primarily contain geospatial data, the value in Package mets/@TYPE MUST be "Geospatial Data" as taken from the CSIP Vocabulary for Content Category.	
	See also: Content Category	

The purpose of the TYPE value in the METS file is to enable search of IPs with the content category mentioned here

Note that only one value is allowed (1…1)

Guideline to the CITS Geospatial

With an introduction to geodata and answers to "Why this requirement?"

Rationales in 3.2.2 Package METS requirements

GEO_2 Rationale

Requirement:

| GEO_2 | Type | For information packages that primarily contain geospatial data, the value in Package mets/@TYPE **MUST** be "Geospatial Data" as taken from the CSIP Vocabulary for Content Category. | 1..1 |
| Ref CSIP2 | mets/@TYPE | | MUST |

Description:
This requirement ensures that the IP is compliant with the requirement CSIP2 in the CSIP specification which states that there MUST be a TYPE-attribute with a value taken from the provided vocabulary for Content Category (https://earkcsip.dilcis.eu/schema/CSIPVocabularyContentCategory.xml).

"Geospatial Data" is the most relevant value from the Content Category vocabulary found in CSIP.

Example:
`TYPE="Geospatial Data"`

Rationale:
This information in the "Package METS.xml" can be used in a Finding Aid to group the IP as an IP that primarily contains "Geospatial Data".

Digitalbevaring.dk

Requirement GEO_3

Content Information Type Specification

GEO_3	Content Information Type Specification	1..1
Ref CSIP4	mets/@csip:CONTENTINFORMATIONTYPE	MUST
	For information packages that primarily contain geospatial data, the value in Package mets/@csip:CONTENTINFORMATIONTYPE MUST be "citsgeospatial_v3_0 " as taken from the CSIP Vocabulary for Detailed Content Type.	
	See also: Content information type specification	

The purpose of the CONTENTINFORMATIONTYPE value in the METS file is to enable validation of the IP according to the CITS specification mentioned here

```xml
<?xml-model href="DILCISVocabularies.rng" type="application/xml" schematypens="http://relaxng.o
<Vocabularies xmlns="https://DILCIS.eu/XML/Vocabularies/IP">
  <Vocabulary Name="VocabularyDetailedContentType">
    <Information>
      <p>Vocabulary derived from E-ARK</p>
    </Information>
    <Entry>
      <Term lang="en">ERMS</Term>
      <Definition>The specification for Electronic records management is used</Definition>
    </Entry>
    <Entry>
      <Term lang="en">SIARD1</Term>
      <Definition>A relational database is described with SIARD version 1</Definition>
    </Entry>
    <Entry>
      <Term lang="en">SIARD2</Term>
      <Definition>A relational database is described with SIARD version 2</Definition>
    </Entry>
    <Entry>
      <Term lang="eb">SIARDDK</Term>
      <Definition>A relational database is described with SIARD version DK</Definition>
    </Entry>
    <Entry>
      <Term lang="en">GeoData</Term>
      <Definition>The specification for Geodata is used</Definition>
    </Entry>
    <Entry>
      <Term lang="en">MIXED</Term>
      <Definition>A mixed set is described</Definition>
    </Entry>
    <Entry>
      <Term lang="en">OTHER</Term>
      <Definition>A term other than present in the vocabulary is used</Definition>
    </Entry>
  </Vocabulary>
</Vocabularies>
```

Contentinformationtype Vocabulary
The value citsgeospatial_v3_0 will be added to the vocabulary

How is the content of the IP validated automatically?

Do we need a new value in the METS file to enable validation of data, metadata and subfolder etc. in the IP?

GeoSIP validation tool does not validate:

- Data in the IP
- Metadata in the IP
- Subfolders in the IP

How does the E-ARK GeoSIP tool know which other validation module to execute to validate this?

Other requirements in the METS file are needed to enable this validation. Values in the METS file referencing the Format specifications/Profiles used could enable automated validation of the actual content of the IP.

Still to come…

Requirement GEO_6a?

GEO_6a	fileSec Other Content Information Type Specification	0..1
Ref CSIP63	mets/fileSec/fileGrp[@csip:CONTENTINFORMATIONTYPE ='GeoData']/@csip:OTHERCONTENTINFORMATIONTYPE	SHOULD
	For any mets/fileSec/fileGrp[@csip:CONTENTINFORMATIONTYPE with the value "citsgeospatial_v3_0" there SHOULD be a @csip:OTHERCONTENTINFORMATIONTYPE attribute with a value taken from a local vocabulary referencing the long-term preservation profile used for the files in the file group	

Suggestion of new requirement…

65

To sum up…

Work on the E-ARK GeoSIP validator is still going on

This requirement enables automated validation of a GeoSIP against the CITS Geospatial specification:

GEO_3 Content Information Type Specification

WHAT next?

What happens with the CITS Geospatial specification now?

- The CITS Geospatial v.3 specification is available on the DILCIS Board webpage: https://dilcis.eu/content-types/cs-geospatial-data

- Please try the CITS Geospatial and give us feedback!

- Tool development to come:
  - E-ARK are working on completing the validation tool for the CITS Geospatial requirements (GeoSIP validator).

  - The Danish National Archives plan to develop validation tools for vector and raster preservation format specifications (profiles) (an open source module to the GeoSIP validator).

  - The National Archives of Czech Republic develop a simple validator for GeoSIP compliant with CITS Geospatial, but with Czech names for E-ARK root folders also. This tool will validate the METS files, checksums and structure of a GeoSIP.

Digitalbevaring.dk

What do you plan to do?

DIGITAL PRESERVATION

Digitalbevaring.dk

Questions?

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!