Specification for digital geospatial data records archiving

DILCIS BOARD

Date: 31.05.2019
Version: 2.0.0

Preface

Aim of the specification

This specification is one of several related specifications. The single most important aim of all of these specifications is the provision of a common set of specifications for packaging digital information for archiving purposes. The specifications are based on common, international standards for transmitting, describing and preserving digital data. They have been produced to help data creators, software developers and digital archives to tackle the challenge of short-, medium- and long-term data management and reuse in a sustainable, authentic, cost-efficient, manageable and interoperable way.

The foundation upon which the specifications are built is the Reference model for an Open Archival Information System (OAIS) (OAIS Reference model) which has Information Packages as its basis. Familiarity with the core functional entities of OAIS is a prerequisite for understanding the specifications. A visualisation of the current specification network can be seen here:

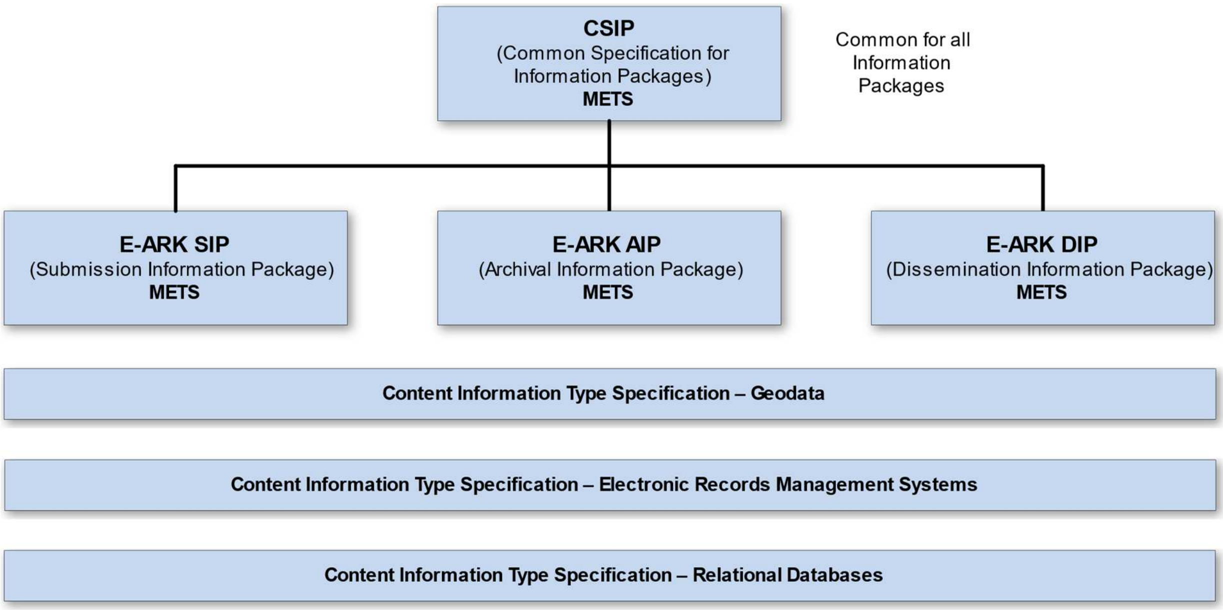

Diagram showing E-ARK specification dependency hierarchy.

Specification	Aim and Goals
Common Specification for Information Packages	This document introduces the concept of a Common Specification for Information Packages (CSIP). Its three main purposes are to: • Establish a common understanding of the requirements which need to be met in order to achieve interoperability of Information Packages.

Specification	Aim and Goals
	• Establish a common base for the development of more specific Information Package definitions and tools within the digital preservation community. • Propose the details of an XML-based implementation of the requirements using, to the largest possible extent, standards which are widely used in international digital preservation. Ultimately the goal of the Common Specification is to reach a level of interoperability between all Information Packages so that tools implementing the Common Specification can be adopted by institutions without the need for further modifications or adaptations.
E-ARK SIP	The main aims of this specification are to: • Define a general structure for a Submission Information Package format suitable for a wide variety of archival scenarios, e.g. document and image collections, databases or geographical data. • Enhance interoperability between Producers and Archives. • Recommend best practices regarding metadata, content and structure of Submission Information Packages.
E-ARK AIP	The main aims of this specification are to: • Define a generic structure of the AIP format suitable for a wide variety of data types, such as document and image collections, archival records, databases or geographical data. • Recommend a set of metadata related to the structural and the preservation aspects of the AIP as implemented by the reference implementation (earkweb). • Ensure the format is suitable to store large quantities of data.
E-ARK DIP	The main aims of this specification are to: • Define a generic structure of the DIP format suitable for a wide variety of archival records, such as document and image collections, databases or geographical data. • Recommend a set of metadata related to the structural and access aspects of the DIP.
Content Information Type Specifications	The main aim and goal of a Content Information Type Specification is to: • Define, in technical terms, how data and metadata must be formatted and placed within a CSIP Information Package in order to achieve interoperability in exchanging specific Content Information. The number of possible Content Information Type Specifications is unlimited. For at list of existing Content Information Type Specifications see, and read more about Content Information Type Specifications in the Common Specification for Information Packages

Organisational support

This specification is maintained by the Digital Information LifeCycle Interoperability Standards Board (DILCIS Board, http://dilcis.eu/). The DILCIS Board was created to enhance and maintain the draft specifications developed in the European Archival Records and Knowledge Preservation Project (E-ARK project, http://eark-project.com/) which concluded in January 2017. The Board consists of eight members, but there is no restriction on the number of participants in the work. All Board documents and specifications are stored in GitHub (https://github.com/DILCISBoard) while published versions are made available on the Board webpage. Since 2018 the DILCIS Board has been responsible for the core specifications in the Connecting Europe Facility eArchiving Building Block https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving.

Authors

A full list of contributors to this specification, as well as the revision history can be found in Appendix 1.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

Term	Description
Archival information package (AIP)	*An Archival Information Package (AIP) is used to transmit archival objects, based on the Submission information package (SIP), into a digital archival system, store the objects within the system, and transmit objects from the system. An IP contains both metadata that describes the structure and content of an archived essence and the actual essence itself. It consists of multiple data files that hold either a logically or physically packaged entity.*
Coordinate Reference System (CRS)	*is a coordinate-based local, regional or global system used to locate geographical entities. A coordinate reference system defines a specific map projection, as well as transformations between different spatial reference systems. Spatial reference systems are defined by the OGC's Simple feature access using well-known text representation of coordinate reference systems, and support has been implemented by several standards-based geographic information systems.*
digital geodata records	*Digital geodata records are records containing a spatial graphical component, describing objects in space. They can be created digitally, or digitized from an analogue source (paper maps)*
DIP	*Distribution Information Package*
Geodata layer	*A Geodata layer is a representation of one or many geospatial datasets within a GIS System. It can contain additional representation information such as visualisation, labelling of the dataset, visibility under certain conditions based on scale, SQL query, etc.*
Geoprocessing workflows	*Geoprocessing workflows are usually defined as a set of Geoprocessing tasks organized into a process. Geoprocessing tasks are functions of a GIS system used to manipulate, transform or manage geospatial datasets, maps and tables.*
Geospatial dataset	*Geospatial datasets are sets of records defining a type of object in space. For example a vector dataset of roads will contain graphical representations of roads and a respective table containing attributes describing each object within the data model of the dataset.*
Geospatial reference	*Geospatial reference is the definition of the coordinate reference system for the geospatial dataset*

GIS	*Abbreviation for Geographical Information System, which is a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data.*
GIS Project	*A GIS project is a document that organizes geospatial datasets into layers, defines the map representations, then reports and stores information on Geoprocessing workflows.*
Information Package	*An Information Package (IP) is used to transmit archival objects into a digital archival system, store the objects within the system, and transmit objects from the system. An IP contains both metadata that describes the structure and content of an archived essence and the actual essence itself. It consists of multiple data files that hold either a logically or physically packaged entity.*
Lossless compression	*Lossless compression is a class of data compression algorithms that allows the original data to be perfectly reconstructed from the compressed data. By contrast, lossy compression permits reconstruction only of an approximation of the original data, though usually with improved compression rates.*
LZW	*Lempel–Ziv–Welch is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch.*
machine-readable documentation	*A machine-readable document is a document whose content can be readily processed by computers. Such documents are distinguished from machine-readable data by virtue of having sufficient structure to provide the necessary context to support the business processes for which they are created.*
Projected coordinate systems	*Geospatial data can have a geographic coordinate system or a projected coordinate system or a geographic projection. Projections are used to define coordinates in distance units as opposed to angular units in geographic coordinate systems.*
Representation Information	*The Representation Information must enable or allow the recreation of the significant properties of the original data object. Essentially, that means Representation Information should be able to recreate a copy of the original. In terms of geospatial data we need all information to reconstruct the usage of the original system to the extent needed to reproduce official records*
SIP	*Submission Information Package*
Well-defined geospatial position	*As defined with a coordinate reference system*
Well-documented graphical component	*Defined using a standardized or proprietary format for geospatial information, that stores coordinates of objects or their representations in space*

1 Introduction

1.1 Purpose and scope

The purpose of this document is to describe the basic required content information for preserving digital geospatial records and provides understanding of the recommended elements needed to archive Geographic Information Systems (GIS) within an Information Package, which would enable the reuse of the geospatial records in an environment similar to the original system.
The definition of required content information for archiving GIS is not the scope of this document. However we provide some basic guidelines on which documentation is to be acquired, but the exact standards and definitions need to be agreed upon or even created by the community.
The document defines which content information is required and what information it needs to contain. It also provides some possible examples and a structure within the Information Package.

1.2 What are geodata records and Geographic Information Systems?

A Geographic Information System (GIS) is a framework for gathering, managing and analyzing data. Outputs from a GIS are a combination of digital **geospatial records (geodata)** and a set of **processes** for transforming basic records into outputs (information products). Outputs can be in the form of maps, lists, new sets of geodata, etc.

In order to archive a GIS in such a way that the outputs from the original system can be reproduced, it is necessary to archive the processes used to manipulate data into outputs.
Geodata are any digital records, that describe an object in space using coordinates based on a geographic coordinate system and a set of descriptive elements called attributes. They mostly come in two forms, vector data (points, lines, polygons) and raster data (sets of one or multiple arrays of pixels).

Increasingly, different geospatial formats include geospatially focused datasets or databases that contain primary information about a geographic location. In addition, ancillary and supplemental data that either are included or can be derived using spatial analysis are considered necessary for the full and effective functioning, interpretation and re-use of the data.

Geodata has many properties that define its accuracy and usability. These properties are commonly described in metadata and should also be part of the archival package. **Processes** are used to organize and transform Geodata in GISs in order to derive value from the dataset itself as well as from the combination with other datasets and databases within the system. Hence in order to transfer data from a GIS, it is necessary to transfer the relations between geodata and other databases and the logic with which the data was used (Representation Information).

2 Elements of a geodata archival information package

In order to transfer data from the original system to the open data repository or the archives, four basic groups of elements have been identified, that compose the archival information package:

- **Geodata** (data object) – This contains the actual geodata records (datasets), exported from the existing GIS system. The same geodata can be stored in one or multiple representations[1]. It is recommended that at least one representation is stored in a long-term preservation format. This is explained in greater detail in the Common Information Specification.

- **Technical documentation** (Representation Information) – This describes the geodata in a way that enables unequivocal interpretation of data and reconstruction of the original data products. Technical documentation should include the: attribute definitions tables, logical structure of the geodatabase, structure of the GIS project, visualisation information for geodata layers (cartography, labelling, etc.), logic of conducting analysis, common queries, etc. The terms are described in the following chapters.

- **Context documentation** (knowledge base) – This may contain any other information that provides contextual information for the geodata such as project reports, user manuals, interviews with the data producer, etc.

- **Geodata specific Metadata** – This includes machine-readable XML files based on geodata metadata standards (ISO 19115:2003[2], EC INSPIRE directive[3]). Any other form of metadata (txt file descriptions, pdf…) is classified as documentation. Metadata specific to the Information Package is described in their respective specifications (SIP, AIP, DIP).

2.1 Geodata

According to OAIS[4], geodata presents the data object in the Information Package. Geodata are records representing and describing objects and phenomena in space. Geodata is always composed of the graphical part in ether vector or raster form, which has a defined location in space in the form of geographical coordinates. The descriptive part connects the graphical element with the descriptions in the form of one or more tables. In the case of geodata in raster format, additional tables may not be always present. When archiving geodata it is necessary to ensure that there is a well-documented graphical component, well defined descriptive attributes and a geographic coordinate system.

[1] As described in the chapter 2 of the SIP Pilot Specification (https://eark-project.com/resources/project-deliverables/51-d33pilotspec/file)

[2] https://www.iso.org/standard/26020.html

[3] https://inspire.ec.europa.eu/

[4] http://www.oais.info/

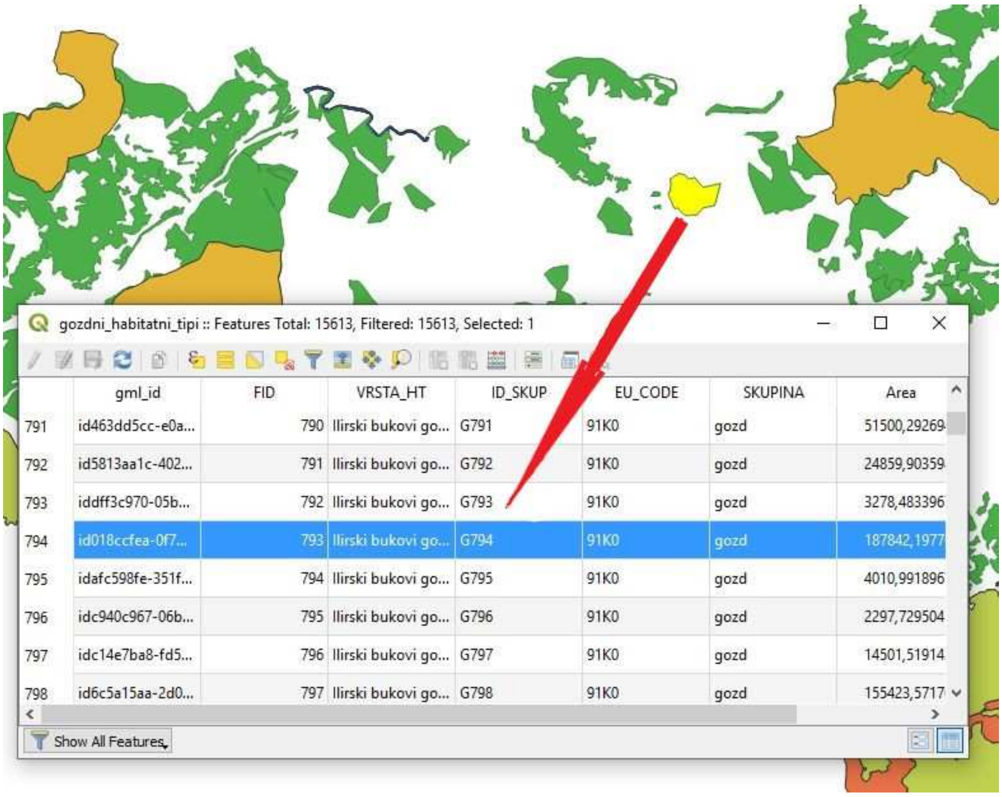

Figure 1: Each Graphical component is connected to the descriptive component in a table

In order to ensure long-term preservation of geodata, proprietary and undocumented formats must be converted into a long-term preservation format that is well described and defines all the necessary elements of the geospatial dataset (unique geospatial reference, usable independent of the system and well documented). Even if some formats are a *de facto* standard today, they may become unreadable in the distant future.

2.1.1 Vector geodata

Vector geodata are sets of coordinates, representing objects in space. Depending on the geometry of the object, it can be represented in the form of points, lines or polygons. Vector geodata comes in many formats, depending on the system of origin. Generally, it takes the form of tables, where one of the columns contains the graphical information and the other columns represent descriptive information which enables the visualisation, querying and analysis of the data.

EXAMPLE 2.1.1.1:
An appropriate long-term preservation format for vector geodata is the Geography Markup Language (GML) format (version 3.2.1. / ISO 19136:2007)[5]. This format is well-documented, because it is an ISO standard. It contains a definition of the geospatial coordinate system and enables descriptions of both simple elements (point, line, polygon) and complex vector elements (arcs, topology definitions, etc.). The GML format is not as common as some *de*

[5] https://www.iso.org/standard/32554.html

facto standards, such as the ESRI shapefile[6], however this is less important when considering long-term preservation where the ability to read it after a long period of time is critical.

2.1.2 Raster geodata

Geodata in the form of raster datasets are actually array-oriented data structures, where the information is stored in the value of the pixels (or array cells) that compose the raster. In this way the raster covers a larger area and is mainly used as continuity surfaces that present remote sensing imagery, digital elevation, scanned maps, etc. This often calls for special requirements for raster formats that enables the storage of multiple types of values per location (text, number), or a combination of up to 400 colour bands (hyperspectral imagery) in one image or a multidimensional representation (netCDF[7]) and others. Another property determining the choice of a long-term preservation format is its size.

When choosing a long-term preservation format it is necessary to ensure that all information that the original format provides is retained and that there is a well-defined geospatial position of the raster and the coordinate system.

EXAMPLE 2.1.2.1 – TIFF:

A good long-term preservation format for most raster datasets is TIFF[8], however in its basic form it lacks the geospatial position definition. TIFF also supports LZW[9] as a lossless compression. A plain TIFF file should be augmented with additional files describing the geospatial reference, which depend on the GIS Tool. (an example would be a tfw world file[10] describing coordinates and prj[11] file defining the geographic coordinate reference system and its projection)

EXAMPLE 2.1.2.2 – GeoTIFF:

GeoTIFF[12] is a public domain geospatial metadata standard which allows geo-referencing information to be embedded within a TIFF file. However, there are some uncommon projected coordinate systems which are not supported. In such cases it is necessary to augment the GeoTIFF with the additional definition of the coordinate system, in a similar manner as is the case with TIFF.

2.1.3 Additional tabular information

Geodata is often a part of a complex data structure, stored in a database along with ordinary tables. In order to reproduce the information products from a GIS, it is often necessary to store additional tables with the geodata. These tables do not have their own geospatial component. In this case, it is important to store the relationships and logic of the data structure, so that it can be reconstructed in the future. For long-term preservation of additional tabular information (attribute tables, code lists, etc.) along with geodata formats proposed for RDBMS archiving are used.[13]

---

[6] https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml
[7] https://www.loc.gov/preservation/digital/formats/fdd/fdd000330.shtml
[8] TIFF - https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml
[9] LZW Compression - https://www.loc.gov/preservation/digital/formats/fdd/fdd000074.shtml
[10] TFW World file: See EXAMPLE 2.1.4.3 and for more info see: https://en.wikipedia.org/wiki/World_file
[11] PRJ Example – See EXAMPLE 2.1.4.2 in this document
[12] GeoTIFF - https://www.loc.gov/preservation/digital/formats/fdd/fdd000279.shtml
[13] https://github.com/DILCISBoard/SIARD/blob/master/specification/index.md

EXAMPLE 2.1.3.1 – CSV[14]
In simple cases *.csv can be used as a text-based format for storing tabular information. It is important that the structure of the table and its nominal code page is defined in the representation information part of the archival information package.

2.1.4 Geospatial reference

In order to properly render the geodata in any future GIS, it is necessary to specify the geospatial reference. The Coordinate Reference System (CRS) provides information about how to locate geodata objects anywhere on the earth's surface. Elements of the spatial reference system are **projection**, **geodetic datum**, and **unit of measurement.**

The CRS definition for geodata within an archival information package can be accomplished in many different ways, including:

- embedded within the data itself (e.g. GML 3.2.1. and GeoTIFF)
- written in the accompanying files (e.g. GML 3.1.1. and earlier, ESRI Shapefile, TIFF, JPEG2000 and GMLJP2[15])
- recorded in the accompanying documentation which needs to be recreated for every subsequent dataset.

Ease of use of geodata in a DIP diminishes with the number of steps needed to define the CRS for every geodata layer. If geodata was used using a standard and documented CRS, it can be referenced by linking to a well-documented list of CRS (example is EPSG.[16])

EXAMPLE 2.1.4.1 – CRS in GML 3.2.1
Georeferencing information in GML is a mandatory part of the file itself and it is embedded in the geodata file itself:

```
<gml:boundedBy>
   <gml:Envelope srsName="urn:x-ogc:def:crs:EPSG:4326">
       <gml:lowerCorner>50.23 9.23</gml:lowerCorner>
       <gml:upperCorner>50.31 9.27</gml:upperCorner>
   </gml:Envelope>
</gml:boundedBy>
```

The attribute "srsName" holds the value of the coordinate reference system code according to EPSG. In this example the code is 4326.

[14] CSV - https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml
[15] http://www.opengeospatial.org/standards/gmljp2
[16] http://www.epsg.org

EXAMPLE 2.1.4.2 – CRS in ESRI Shapefile (shp)

Although the ESRI shapefile is not the best choice of formats for a long-term preservation format, it could be used as a current DIP format. A SHP needs a <shapefilename>.prj file in order to be properly georeferenced. A .prj file is a .txt file, containing a definition of the coordinate reference system and all of its elements. Here is an example:

```
PROJCS["NAD_1983_UTM_Zone_10N",GEOGCS["GCS_North_American_1983",
DATUM["D_North_American_1983",SPHEROID["GRS_1980",6378137,298.257222101]],
PRIMEM["Greenwich",0],UNIT["Degree",0.0174532925199433]],
PROJECTION["Transverse_Mercator"],PARAMETER["False_Easting",500000.0],PARAMETER["False_N
orthing",0.0],PARAMETER["Central_Meridian",-123.0],PARAMETER["Scale_Factor",0.9996],
PARAMETER["Latitude_of_Origin",0.0],UNIT["Meter",1.0]]
```

EXAMPLE 2.1.4.3 – CRS in a TIFF

A TIFF raster geodata file does not contain any CRS information. A TIFF file must therefore be accompanied by the tfw file that contains its initial coordinates and pixel size and a prj file that defines the geospatial coordinate system. For example, a D240143.tif file would be accompanied by a D240143.tfw file. That is a txt file, containing information about the coordinates and the size of the first top left pixel. Here is an example:

```
0.42333
0.0
0.0
-0.42333
394250.00
```

2.1.5 Other geodata types

Geodata can be presented in different formats and as a part of different data structures (such as geometric networks, coverages, structures combining raster and vector data, etc.), as web services, automated maps, etc.

This document focuses on the fundamental structures of geodata, which are used as basic input components for complex structures. In order to replicate the complex data structures or services in the future, their organizational logic and the way they are used within applications needs to be documented.

2.2 Technical documentation

Geodata is rarely in a form that is sufficiently self-explanatory to be used and properly interpreted by itself. Consequently additional information is required in order to enable the user to properly understand, interpret and use geodata. This chapter describes the required technical documentation for geodata datasets (where it is applicable). Ideally a machine-readable format of documentation is preferred, however any other form of documenting the system is welcome.

In this document technical documentation and general contextual documentation of geodata is differentiated. The purpose of the technical documentation is to provide Representation Information that enables the reuse of geodata in a way that is similar enough to its use in the initial system. Technical documentation can be stored with the rest of the documentation, as long as it contains all the required information.

Such documentation spans from basic information (how to cartographically render a simple layer or how to set up a basic query using a simple combination of datasets) up to a detailed documentation of the initial system, that would allow the reconstruction of a technical environment in the future that could produce similar information products as the initial system. When creating an SIP, it is required that the producer ensures the following elements of technical documentation (if applicable):

2.2.1 Attribute definition

Attribute definition describes the value types for each column in the table of a geodata dataset or an accompanying table.

EXAMPLE 2.2.1.1 – Attribute definition in GML with XSD schema

The attribute definition of table elements in a GML is expressed in an accompanied XSD schema. That is why a XSD schema should be present for every GML file in the Information Package, if this format is used. It also provides means for GML validation and at the same time it documents the dataset structure.

2.2.2 Feature catalogue

The feature catalogue represents a logical structure of attributes. It provides an understanding of the meaning, use and structure of the spatial data and provides a unified classification of spatial data in feature types (classes). Feature types are distinguished by their attributes (properties), by importance and by the relationships between them. ISO 19110:2005 describes this in greater detail.[17]

2.2.3 Visualisation information

Data visualisation provides an illustration and representation of spatial data. The catalogue of cartographic symbol is a collection of agreed cartographic symbols, which are used via visualisation of spatial data sets to display objects in space. Cartographic symbols are shown in the legend, which explains their meaning.

For certain geodata the visualisation is already made by the producer in the form of (geo-located) raster images or paper maps. In these cases, it is reasonable to archive that kind of visualisation. For each spatial data set it is possible to produce any number of different visualisations with the appropriate software. It is proposed, that:

- Every dataset is described with at least a screenshot image of the geodata dataset in full extent, to enable easy discovery and identification in the archival catalogue.

[17] https://www.iso.org/standard/39965.html

- If a cartographic key exists, it should be documented in a way that it can be reproduced in a future system to a satisfying extent.
- If geodata was used for the production of complex maps, the logic is preserved in such a way that a similar representation is possible in the future.
- If a visualisation was created using well-documented machine-readable files, they should be preserved.

EXAMPLE 2.2.3.1 – SLD files

SLD[18] is an OGC[19] 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 which is provided in the documentation in QGIS. 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.

```
<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>
               <Rule>
                   <PointSymbolizer>
                       <Graphic>
                           <Mark>
                               <WellKnownName>circle</WellKnownName>
                               <Fill>
                                   <CssParameter name="fill">#FF0000</CssParameter>
                               </Fill>
                           </Mark>
                           <Size>6</Size>
                       </Graphic>
                   </PointSymbolizer>
               </Rule>
           </FeatureTypeStyle>
       </UserStyle>
   </NamedLayer>
</StyledLayerDescriptor>
```

Listing 26: Example of an SLD file

2.2.4 Table relationships

If the IP contains two or more geospatial datasets, or additional tables, and if it is applicable, the technical documentation should describe the relationships between tables in a database or within a

[18] SLD – Styled Layer Description: https://portal.opengeospatial.org/files/?artifact_id=22364
[19] http://www.opengeospatial.org/

GIS project, in order to enable the reconstruction of queries and provide greater understanding of the usage of complex data structures (example Topologies, Geometric networks, etc.).

2.2.5 GIS Project (logical structure of layers)

When a larger GIS project (containing many geodata datasets and additional tables) is documented, in order to be able to reproduce an information product (a web service), or repeat a workflow, the following should be documented:

- Logical structure of geodata layers and tables in a GIS Projects
- Geodata layer properties (definition query, scale dependent display, visualisation parameters, etc.)
- Labelling (which layers were used for labelling, scale dependent display of labels, text rendering properties, etc.)

EXAMPLE 2.2.5.1 – QGS file - GIS Project definition file on QGIS

QGIS application stores its Project structure of geodata layers and its rendering properties (Projection, symbology, labelling..) in a *.qgs file, which has an XML structure. So it can be opened with any text reader and layer hierarchy and its parameters can be read from the file. However if used for archiving, additional documentation on the qgs file structure should be stored in the Information Package or available in the archive.

2.2.6 Common queries and geoprocessing workflows

In order to produce information products in the initial GIS, there is often a requirement to run certain database queries or geo-specific processes (geoprocessing workflows). Common information products from a GIS are maps, lists, charts, new geodata derived from existing data, web services, etc. In order to reproduce this type of GIS information product in the future, documentation of the queries and geoprocessing workflows is needed. Ideally these queries should be described using machine-readable formats, however any other form of documentation is welcome.

2.3 General (contextual) documentation

This part of the AIP describes all remaining information about the geodata. Included here are links to relevant documentation describing the lineage and provenance of the spatial data set. The documentation includes: user manuals, related practices in the EU and worldwide, methodological rules, scientific articles, publications, etc.

2.4 Geospatial Metadata folder

When Geospatial data is described with a machine readable xml file within a GIS, and the xml schema is based on a standardized metadata format, we can use this content to harvest descriptive elements into Archival metadata (EAD, ISADG, etc.). That is why we need to define a special location for such files, so that an automatic harvesting process knows where to look. A proposal for translation tables between the INSPIRE metadata schema and ISAD(G) and for the EAD3 schemas is presented in chapter 4.

In general, GIS systems offer a possibility to describe the geospatial dataset with additional technical and descriptive metadata. In some systems, the metadata structure is proprietary and in some metadata adheres to local or global standards. Technical metadata can be derived from the dataset (geometry type, number of records, etc.), however the descriptive data must be entered by the data creator or its manager. And this is the part that we should consider as part of the Information Package. In reality we might encounter systems, where geodata is not accompanied by adequate metadata.

Depending on the age of the GIS systems, we might encounter different descriptive information of each individual dataset:

- None or very limited (Metadata can be derived from separate documentation)
- Metadata is available in separate files that are not in a machine-readable format (pdf, unstructured txt…)
- Metadata is available in machine-readable format (xml or other).

Geodata specific metadata commonly contains descriptions that are specific to geodata (CRS info, bounding coordinates, scale, etc.) and would not be easily added to archival contextual metadata formats like ISADG or EAD without extending the schema. So we recommend that we store this geospatial metadata as content of the data. It can be used to harvest elements from geospatial metadata, into archival descriptive metadata of the Information Package (GEO IP), during the ingest process.

We recommend, that the producer provides metadata based on existing metadata standards for Geodata (ISO 19115:2003 or later)[20] or its adoption by the European Directive INSPIRE.

[20] https://www.iso.org/standard/26020.html

According to the mandatory elements, that are used in describing datasets by the INSPIRE directive, we propose that the following metadata elements are present within the descriptive metadata structure for each geospatial dataset, which is present in the data folder:

1	**Resource title**	*Name by which the cited resource is known*
2	**Resource abstract**	*Brief narrative summary of the content of the resource(s)*
3	**Resource type**	*Scope to which metadata applies.* *This is the type of resource being described by the metadata and it is filled in with a value from a classification of the resource based on its scope. The choice of Resource Type will be probably the first decision made by the user and it will define the metadata elements that should be filled.* *(Example: dataset)*
4	**Unique resource identifier**	*Value uniquely identifying an object within a namespace*
5	**Resource language**	*Language(s) used within the datasets*
6	**Topic category**	*Main theme(s) of the dataset. (Example: Hydrography, administrative areas, transportation, etc.)*
7	**Keyword value**	*Commonly used word(s) or formalized word(s) or phrase(s) used to describe the subject*
8	**Originating controlled vocabulary**	*Name of the formally registered thesaurus or a similar authoritative source of keywords*
9	**Geographic bounding box**	*Western-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees (positive east). Eastern-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees (positive east) Northern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees (positive north) Southern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees (positive south).*
10	**Temporal extent**	*Time period covered by the content of the dataset*
11	**Date of publication**	*Reference date for the cited resource - publication*
12	**Date of last revision**	*Reference date for the cited resource - revision*
13	**Date of creation**	*Reference date for the cited resource - creation*
14	**Lineage**	*General explanation of the data producer’s knowledge about the lineage of a dataset. This information can also reference other documents, that cover this description in*

		greater detail explaining more about how and from what source datasets were creates, methodology of the process, etc.
15	**Spatial resolution**	*• Equivalent scale: level of detail expressed as the scale denominator of a comparable hardcopy map or chart* *• Distance: ground sample distance*
16	**Specification**	*Citation of the product specification or user requirement against which data is being evaluated*
17	**Limitations on public access [and use]**	*Access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the resource* *Limitations on public access:* *Access constraints - Example: Other Restrictions (limitation not listed).* *Other constraints - Example: No limitations* *Classification - Example: unclassified*
18	**Conditions applying to access and use**	*Restrictions on the access and use of a resource or metadata*
19	**Metadata date**	*Date that the metadata was created*
20	**Metadata language**	*Language used for documenting metadata*

3 Example structures of a Geodata Archival Information Package (GEO IP)

In this chapter, we provide possible information package structures containing geodata (GEO IP), following the Common Specification for Information Packages[21] at least one representation should be put in the representation directory (Figure 2).

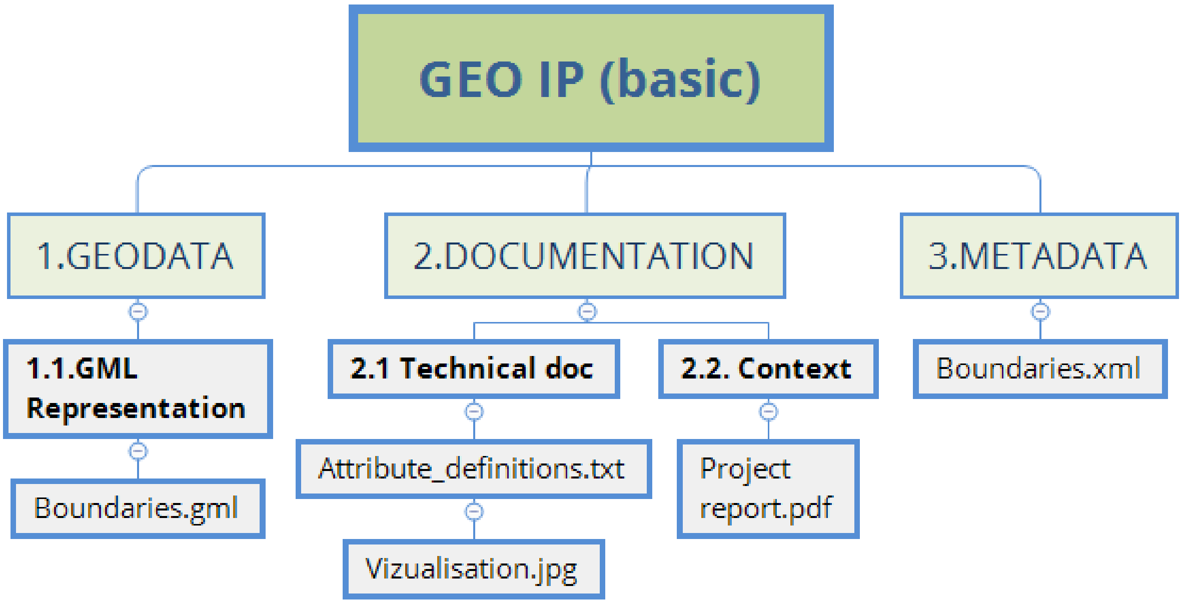

Figure 2: Basic Geodata IP

Figure 2 shows a very basic example, containing only one representation of vector geodata in GML format, which could be used as a long-term preservation format. Additional documentation is stored in other folders (Technical documentation and Context).

Technical documentation describes the meaning of attribute values in the table of the Boundaries.gml geodata layer and shows an overview image of how the data was visualised. This folder should contain the representation information needed to render the data in Geodata folder.

Context folder contains all other information we want to store about the record.

The IP Package contains only one representation folder (in this case named 'geodata'- although the use of other names is allowed) and the additional documentation is stored within that representation.

The geospatial metadata are to be stored within the "Metadata" folder under the "Data" folder (In this case GEO IP (Basic).

Other scenarios are also possible, such as multiple representations of the same geodata in the same IP and different possibilities of storing additional documentation:

[21] https://github.com/DILCISBoard/E-ARK-CSIP/tree/master/specification

3.1 GeoSIP containing multiple vector representations

In this case, a GeoSIP package contains one representation in GML format and an original representation of the same data in ESRI Shapefile format. All other documentation, required to properly interpret both representations is put in the documentation folder (Figure 3). It is also possible to include a logical link to point to additional documentation being stored in a different SIP (in case of a larger time series of the same data or similar records but from different organizational units).

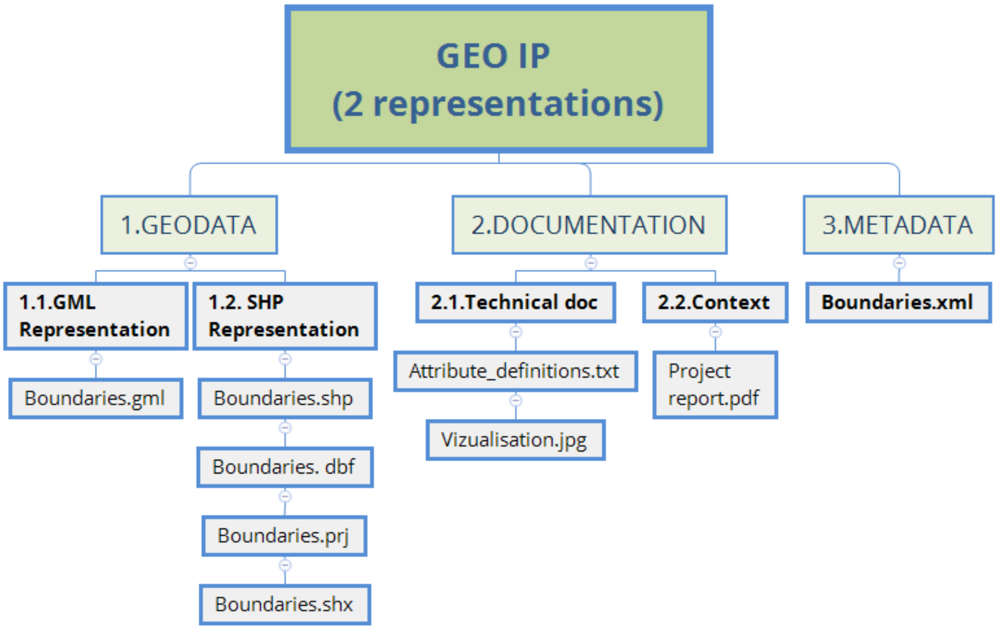

Figure 3: Folder structure of the GeoSIP containing multiple vector representations and documentation on the top

3.2 GeoSIP containing one representation of multiple raster datasets

In this case, the IP contains one representation of multiple raster images covering an area with an accompanying vector file – containing positions of the raster images (Figure 4). Documentation for the raster datasets is located in the top Documentation folder. In the case of a large volume of data we could split the data into multiple SIPs and record the organisation of the split within the documentation folder.

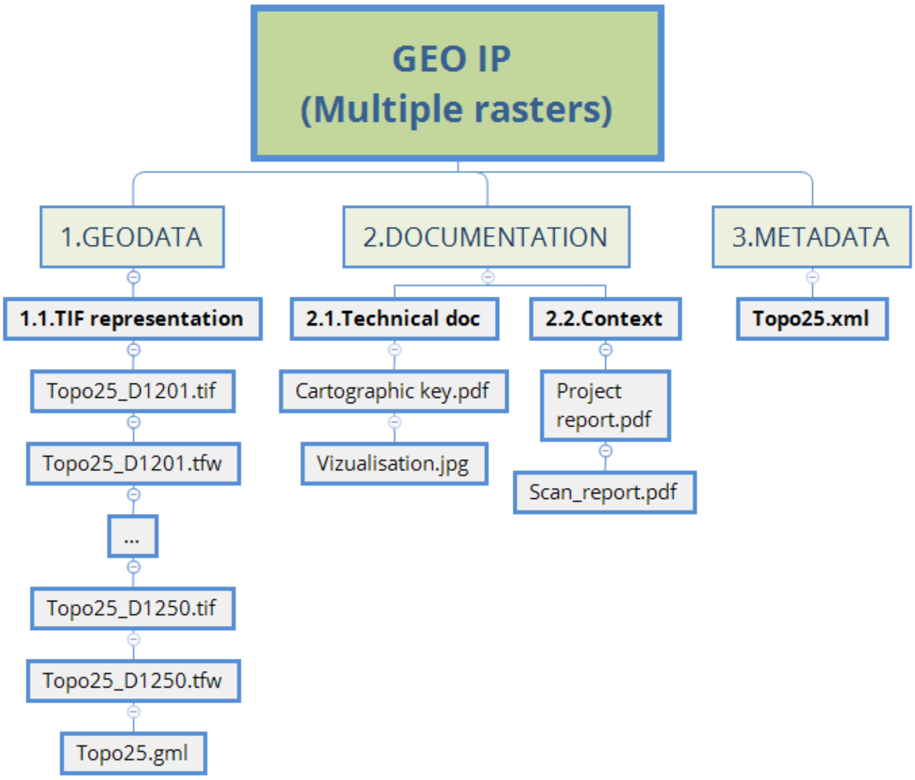

Figure 4: Folder structure of the GeoSIP containing one representation of multiple rasters

3.3 Proposed cardinality for elements in a GEO IP

In this section we describe the proposed cardinality of elements within the GEO IP according to the types of elements it holds. If we only store records in the form of geospatial datasets, see the following table:

Geodata elements

	Information	Cardinality
	Graphical information	1..n
	Data attributes	1..n
	Geo referencing information	1..n
	Visualisation information	0..n
	GIS Metadata[22]	0..n
	Graphical information	1..n
	Data attributes	0..n
	Geo referencing information	1..n
	Visualisation information	0..n
	GIS Metadata	
	Attribute definition	1..n
	Object catalogue	0..n
	Relationship to geodata	1..n

Table 1: Proposed cardinality for elements in a Geo IP

GIS System elements

If we are preserving a more complex GIS, we propose that the Geo IP also contains the following elements:

	Information	Cardinality
	List of elements in a project	1..n
	Object relations (geodata layers	1..n
	Geo referencing transformations	1..n
	Data layer properties	0..n
	Labeling	0..n
	Map visualisation	0..n
	Common queries	1..n
	Geoprocessing workflows	0..n
	Common reports	0..n

Table 2: GIS System element

[22] GIS Metadata refers to the information covered in section 2.4 of this document

4 Metadata translations

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 applies	CodeList	0..1	3.1.5 Extent and medium of the unit of description	CodeList (see annex B.5.25 of ISO 19115)
2.2.5	Unique resource identifier	Value uniquely identifying an object within a namespace	text	0..1	3.1.1 Reference code	
2.2.6	Coupled resource	Provides information about the datasets that the service operates on	URI	0..*	3.5.3 Related units of description	
2.2.7	Resource language	Language(s) used within the datasets	CodeList	0..*	3.4.3 Language/scripts of material	LanguageCode (ISO/TS 19139)
2.3.1	Topic category (INSPIRE specific)	Main theme(s) of the dataset	CodeList	1..1		Separate descriptor? List of values. See Part D 2 of the INSPIRE Metadata Regulation 1205/2008/EC) ISO19115:B.5.27 MD_TopicCategoryCode
2.3.2	Spatial data service type	A service type name from a registry of services	CodeList	0..1		Separate descriptor? List of values. See section 1.3.1 in INSPIRE Metadata Implementing Rules

2.4.1	Keyword value	Commonly used word(s) or formalised word(s) or phrase(s) used to describe the subject	text	0..*	?	descriptors
2.6.1	Temporal extent	Time period covered by the content of the dataset	Date	1..*	3.1.3 Date(s)	
2.6.2	Date of publication	Reference date for the cited resource – publication	Date	0..1	3.1.3 Date(s)	
2.6.3	Date of last revision	Reference date for the cited resource – revision	Date	0..1	3.1.3 Date(s)	
2.6.4	Date of creation	Reference date for the cited resource – creation	Date	0..1	3.1.3 Date(s)	
2.7.1	Lineage	General explanation of the data producer's knowledge about the lineage of a dataset	text	0..1	/	This element is Geodata specific, so we propose that searching using this criteria is done by using Geospatial metadata catalogues and not Archival catalogues or that data is added into "Scope and content" element (3.3.1. ISAD(G))
2.7.2	Spatial resolution	• Equivalent scale: level of detail expressed as the scale denominator of a comparable hardcopy map or chart • Distance: ground sample distance	text	0..1	/	Same as with Lineage
2.9.1	Limitations on public access [and use]	Access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the resource	CodeList	1..1	3.4.1 Conditions governing access	
2.9.2	Conditions applying to access and use	Restrictions on the access and use of a resource or metadata	text	1..1	**3.4.1 Conditions governing access;** 3.4.2 Conditions governing reproduction; 3.4.4 Physical characteristics and technical requirements	

Table3: Proposed translation schema the INSPIRE metadata descriptions for geospatial resources in ISAD(G)

4.2 Proposed translation between INSPIRE metadata elements and EAD3 metadata elements

The following table displays the identified counterparts of the mandatory INSPIRE metadata elements used in the EAD3 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.

Further information on EAD elements can be found in the TagLibrary-VersionEAD3.pdf available at:

http://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf

INSPIRE el. Nr.	INSPIRE element name	Explanation	Proposed Cardinality	EAD3 mapping	EAD description	Comments
2.2.1	Resource title	Name by which the cited resource is known	1..1	<unittitle>	**<unittitle>** is for recording the title statement, either formal or supplied, of the described materials. The title statement may consist of a word or phrase. <unittitle> is used at both the highest unit or <archdesc> level (e.g., collection, record group, or fonds) and at all the subordinate <c> levels (e.g., subseries, files, items, or other intervening stages within a hierarchical description).	

2.2.2	Resource abstract	Brief narrative summary of the content of the resource(s)	0..1	<scopecontent>	**<scopecontent>** contains a narrative statement that summarizes the range and topical coverage of the materials. It provides the researcher with the information necessary to evaluate the potential relevance of the materials being described. <scopecontent> may include information about the form and arrangement of the materials; dates covered by the materials; significant organizations, individuals, events, places, and subjects represented in the materials; and functions and activities that generated the materials being described. It may also identify strengths of or gaps in the materials.	
2.2.3	Resource type	Scope to which metadata applies. This is the type of resource being described by the metadata and it is filled in with a value from a classification of the resource based on its scope. The choice of Resource Type will be probably the first decision made by the user and it will define the metadata elements that should be filled. (Example: dataset)	0..1	<physdesc>	**<physdesc>** is a wrapper element for bundling information about the appearance or construction of the described materials, such as their dimensions, a count of their quantity or statement about the space they occupy, and terms describing their genre, form, or function, as well as any other aspects of their appearance, such as colour, substance, style, and technique or method of creation. The information may be presented as plain text, or it may be divided into the <dimension>, <extent>, <genreform>, and <physfacet> sub-elements.	**CodeList (see annex B.5.25 of ISO 19115)**
2.2.5	Unique resource identifier	Value uniquely identifying an object within a namespace	0..1	<originalsloc>	**<originalsloc>** may be used to provide information about the location, availability, and/or destruction of originals.	If applicable

2.2.6	Coupled resource	Provides information about the datasets that the service operates on	0..*	<relatedmaterial>; <arrangement>	**<relatedmaterial>** is used to identify associated materials in the same repository or elsewhere. These materials may be related by sphere of activity, or subject matter; Use **<arrangement>** to record the logical or physical groupings within a hierarchical structure and their relationships. This includes how the described materials have been subdivided into smaller units, e.g., record groups into series. May also indicate the filing sequence of the described materials, for example chronological or alphabetical arrangement.	
2.2.7	Resource language	Language(s) used within the datasets	0..*	<langmaterial>	**<langmaterial>** records information about languages and scripts represented in the materials being described. <langmaterial> must contain one or more <language> or <languageset> elements, but cannot contain text.	LanguageCode (ISO/TS 19139)
2.3.1	Topic category (Specific to INSPIRE)	Main theme(s) of the dataset	1..1			A separate descriptor. List of values. See Part D 2 of the INSPIRE Metadata Regulation 1205/2008/EC) ISO19115:B.5.27 MD_TopicCategoryCode
2.3.2	Spatial data service type	A service type name from a registry of services	0..1			A separate descriptor List of values. See section 1.3.1 in INSPIRE Metadata Implementing Rules
2.4.1	Keyword value	Commonly used word(s) or formalised word(s) or phrase(s) used to describe the subject	0..*	<controlaccess> <subject> <part>	**<controlaccess>** An element that binds together elements containing access headings for the described materials. **<subject>** An element for encoding topics represented in the materials being described. **<part>** A required and repeatable child of controlled access elements used to encode one or more parts of an access term.	Each keyword is given in its own <part> element.

2.5.1	Geographic bounding box	Western-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees (positive east). Eastern-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees (positive east) Northern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees (positive north) Southern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees (positive north).	0..*	<geographiccoordinates>	Use **<geographiccoordinates>** to express a set of geographic coordinates such as latitude, longitude, and altitude representing a point, line, or area on the surface of the earth.	
2.6.1	Temporal extent	Time period covered by the content of the dataset	1..*	<unitdate>; <unitdatestructured>	**<unitdate>** is for indicating the date or dates the described materials were created, issued, copyrighted, broadcast, etc. <unitdate> may be in the form of text or numbers, and may consist of a single date, a date range, or a combination of single dates and date ranges; **<unitdatestructured>** provides a machine-processable statement of the date or dates the materials described were created, issued, copyrighted, broadcast, etc. <unitdatestructured> must contain one of the following child elements: <datesingle>, <daterange>, or <dateset>. <unitdatestructured> may contain only one child, therefore <dateset> must be used in situations where complex date information needs to be conveyed and requires at least two child elements. A date set may combine two or more <datesingle> and <daterange> elements.	

2.6.2	Date of publication	Reference date for the cited resource - publication	0..1	<unitdate>, <unitdatestructured>	**<unitdate>** is for indicating the date or dates the described materials were created, issued, copyrighted, broadcast, etc. <unitdate> may be in the form of text or numbers, and may consist of a single date, a date range, or a combination of single dates and date ranges; **<unitdatestructured>** provides a machine-processable statement of the date or dates the materials described were created, issued, copyrighted, broadcast, etc. <unitdatestructured> must contain one of the following child elements: <datesingle>, <daterange>, or <dateset>. <unitdatestructured> may contain only one child, therefore <dateset> must be used in situations where complex date information needs to be conveyed and requires at least two child elements. A date set may combine two or more <datesingle> and <daterange> elements.	

2.6.3	Date of last revision	Reference date for the cited resource - revision	0..1	<unitdate>, <unitdatestructured>	**<unitdate>** is for indicating the date or dates the described materials were created, issued, copyrighted, broadcast, etc. <unitdate> may be in the form of text or numbers, and may consist of a single date, a date range, or a combination of single dates and date ranges; **<unitdatestructured>** provides a machine-processable statement of the date or dates the materials described were created, issued, copyrighted, broadcast, etc. <unitdatestructured> must contain one of the following child elements: <datesingle>, <daterange>, or <dateset>. <unitdatestructured> may contain only one child, therefore <dateset> must be used in situations where complex date information needs to be conveyed and requires at least two child elements. A date set may combine two or more <datesingle> and <daterange> elements.	

2.6.4	Date of creation	Reference date for the cited resource - creation	0..1	<unitdate>, <unitdatestructured>	**<unitdate>** is for indicating the date or dates the described materials were created, issued, copyrighted, broadcast, etc. <unitdate> may be in the form of text or numbers, and may consist of a single date, a date range, or a combination of single dates and date ranges; **<unitdatestructured>** provides a machine-processable statement of the date or dates the materials described were created, issued, copyrighted, broadcast, etc. <unitdatestructured> must contain one of the following child elements: <datesingle>, <daterange>, or <dateset>. <unitdatestructured> may contain only one child, therefore <dateset> must be used in situations where complex date information needs to be conveyed and requires at least two child elements. A date set may combine two or more <datesingle> and <daterange> elements.	
2.7.1	Lineage	General explanation of the data producer's knowledge about the lineage of a dataset, covering the history of data creation and methodologies used during the process	0..1		It is a separate element in INSPIRE, however it would best fit as an addition of the **<scopecontent>**. So we recommend appending it to this element.	
2.7.2	Spatial resolution	• Equivalent scale: level of detail expressed as the scale denominator of a comparable hardcopy map or chart • Distance: ground sample distance	0..1		It is a separate element, that is often used as a search criterion within Geospatial Metadata repositories. Since we did not identify the element in EAD, that would unequivocally define it	

2.9.1	Limitations on public access [and use]	Access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the resource Limitations on public access: Access constraints - Example: otherRestrictions (limitation not listed). Other constraints - Example: No limitations Classification - Example: unclassified	1..1	<accessrestrict> <legalstatus>	Record in **<accessrestrict>** information about the availability of the described materials, whether due to the nature of the information in the materials being described, the physical condition of the materials, or the location of the materials. Examples include restrictions imposed by the donor, legal statute, repository, or other agency, as well as the need to make an appointment with repository staff. May also indicate that the materials are not restricted; Use **<legalstatus>** to identify the status of the material being described as defined by law, for example, the Public Records Act of 1958 in the United Kingdom.	This element has 3 separate elements
						General set of limitations (ISO 19115 B.5.24)
						Description of the limitation
						Level of confidentiality (ISO 19115 B.5.11)

2.9.2	Conditions applying to access and use	Restrictions on the access and use of a resource or metadata	1..1	<accessrestrict>; <userestrict>	Record in **<accessrestrict>** information about the availability of the described materials, whether due to the nature of the information in the materials being described, the physical condition of the materials, or the location of the materials. Examples include restrictions imposed by the donor, legal statute, repository, or other agency, as well as the need to make an appointment with repository staff. May also indicate that the materials are not restricted; Use **<userestrict>** for information about any limitations, regulations, or special procedures imposed by a repository, donor, legal statute, or other agency. These conditions may be related to reproduction, publication, or quotation of the described materials after access to the materials has been granted. <userestrict> may also be used to indicate the absence of restrictions, such as when intellectual property rights have been dedicated to the public.	

Table4: Proposed translation schema of INSPIRE metadata descriptions in EAD3

5 Appendix 1

AUTHOR(S)	
Name(s)	Organisation(s)
Gregor Završnik	Geoarh

REVIEWER(S)	
Name(s)	Organisation(s)
Jaime Kaminski	Highbury IVS
Janet Anderson	Highbury IVS

Project co-funded by the European Commission within the ICT Policy Support Programme Dissemination Level		
P	**Public**	X
C	**Confidential, only for members of the Consortium and the Commission Services**	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Submitted Revisions History

Revision No.	Date	Authors(s)	Organisation	Description
0.1	31 October 2018	Gregor Završnik	Geoarh	Draft outline based on SFSB SMURF document.
1.0	20.December 2018	Gregor Završnik,	Geoarh,	
1.1	31.May 2019	Gregor Završnik	Geoarh	Changes introduced based on received comments from the users

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.