

Appendix 1: Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1

Guideline for the specification for the E-ARK Content Information Type Specification for digital geospatial data records archiving (CITS Geospatial)

A proper front page will be created for the publication occurring after implementation of review comments.

The correct preface will be inserted for the publication occurring after implementation of review comments.

The correct postface will be inserted for the publication occurring after implementation of review comments.

Appendix 1: Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1

Name	Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1
Category	Example of a vector profile for the CITS Geospatial guideline
Maturity level	Draft
Version	0.1.0.00
Status	Not final – Serves as an example of a profile in the review version (Request for comments) of the CITS Geospatial Guideline
Approval date	Not applicable
Issue date	2021-04-09
Replaces	None
Prerequisite	None
Attachments	No sample Information Package illustrates the use of this profile yet
Languages	English
Publisher / distributor	DILCIS Board, https://dilcis.eu/

Summary

This document contains the Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1 and serves as Appendix 1 for the CITS Geospatial specification guideline. It is a normative description of a standard for the long-term preservation of geospatial vector data.

The specification is based on standards including GML 3.2.1 and XML 1.0 and W3C XML Schema Definition Language (XSD) 1.1 Part 1.

The aim of employing internationally recognised standards is to ensure the long-term preservation of, and access to, the widely used geospatial vector data, as well as the easy exchange of geospatial vector data independent of proprietary vector formats.

DRAFT

Version history

Relationship of the present version to previous versions:

Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1	First draft of the profile
--	----------------------------

DRAFT

Table of Contents

1 Introduction	5
1.1 Status	5
1.2 Area of application	5
1.2.1 Addressees / target group	5
1.2.2 Background	5
1.2.3 Distinctions	5
2 Structure of the document	6
2.1 Structure of chapters	6
2.2 ID for requirements	6
2.3 The distinction between mandatory and optional requirements	7
2.4 Notation of folders, files and folder structures	7
3 General requirements	7
3.1 Use of standards	7
3.2 Character sets and characters	8
3.3 File format extensions	8
4 Requirements for package structure	8
5 Requirements for data	11
5.1 XML schema definition requirements	11
5.2 GML file requirements	15
6 Requirements for metadata	19

1 Introduction

1.1 Status

This document was approved by the DILCIS Board.

1.2 Area of application

1.2.1 Addressees / target group

This is a technical document for IT specialists involved in the long-term archiving of vector geodata.

1.2.2 Background

This is a standard for the long-term archiving of geospatial vector data based on the GML format.

Long-term archiving is the preservation, normally without a time limit, of the information stored in the gml files while retaining the bit stream and the ability to interpret and display the data in a way that is human-readable and comprehensible.

If the structure and content of any geospatial vector data format are translated into GML 3.2.1, it will subsequently be possible to access and exchange the data in the gml file at any time, even when the original geospatial vector format software is no longer available or can no longer be run. This has been achieved by the use of suitable standards for geospatial vector data exchange that are widely supported internationally.

This long-term interpretability of the geospatial vector data content is essentially based on two standards GML 3.2.1 and XML 1.0.

1.2.3 Distinctions

It should be noted that this profile is only the long-term storage format for a specific type of geodata: geospatial vector data. It is assumed that geospatial vector data preserved using this profile is archived as part of an Information Package (CSIP) together with other documents (e.g. metadata files, raster objects, SIARD file, documentation, etc.). Requirements for a geospatial Information Package is specified in the CITS Geospatial specification. This Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1 provides additional requirements to the CITS Geospatial specification.

Just as an XML-based Word or e-mail file contains an internal file structure consisting of metadata, primary data and various auxiliary data, geospatial vector data preserved using this profile contains its own metadata describing the data more precisely in addition to the actual data – regardless of the metadata catalogue that an archive records in its OASIS packages.

2 Structure of the document

2.1 Structure of chapters

Each chapter in this profile is constructed according to the same pattern. After a brief introduction, the requirements are listed in a table.

ID	Description of requirement	M/O
contains the ID of the requirement	contains the text of the requirement	stipulates whether mandatory or optional

A requirement is frequently further explained by means of recommendations, notes and examples, each of which is specifically indicated as such.

ID	Description of requirement	M/O
A_3.1-1	Text of requirement Example Text of example Note Text of note Recommendation The text of recommendations is in italics.	M

2.2 ID for requirements

The requirements are unambiguously identifiable by means of an ID.

ID
G_2.1-1

This ID is constructed according to the following pattern:

- G_ Letter + _ identifies main chapters
- G_ = General requirements / principles
- P_ = Requirements for package structure
- D_ = Requirements for data
- M_ = Requirements for metadata

- 2.1-1 The number begins with the number of the chapter (which groups together requirements on the same topic), and the number after the dash is consecutive, thus designating all the requirements in the chapter.

2.3 The distinction between mandatory and optional requirements

Each requirement is either mandatory or optional. This is indicated by a letter:

Abbreviation	Meaning
M	Mandatory requirement This requirement must be met.
O	Optional requirement This requirement should be met. It simplifies handling and constitutes best practice.

2.4 Notation of folders, files and folder structures

The following symbols and parameters are used for the notation of folders, files, etc.

Symbol	Meaning
/	Folder
header/	A folder with the name "header"
xy.txt	File (with file extension "txt")
dir1/	Example folders (in red)
abc.pdf	Example files (in red)
...	Placeholder for files, folders or elements that are not relevant to the explanation
<i>element</i>	An element, attribute, class or namespace used in XML or GML syntax

3 General requirements

3.1 Use of standards

To ensure that the contents of geospatial vector data remain interpretable over a long period, this Long-Term preservation format Profile for Geospatial Vector data using GML 3.2.1 is essentially based on two ISO standards: GML and XML.

ID	Description of requirement	M/O
G_3.1-1 GEO_18 and GEO_20	A geospatial vector dataset is stored as a file in GML in accordance with specifications from OGC (Open Geospatial Consortium) in ISO 19136 GML version 3.2.1 ¹ and must comply with the syntax rules for XML version 1.0	M
G_3.1-2	Geospatial metadata is stored in an XML file compliant with the INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119 ³	M

3.2 Character sets and characters

ID	Description of requirement	M/O
G_3.2-1	GML files and schemas are encoded in the UTF-8 character set, which must be specified in the XML declaration of the files. The use of characters is limited as described in ISO/IEC 10646:2003 Annex D and in The Unicode Standard 5.1, Chapter 3.	M

3.3 File format extensions

ID	Description of requirement	M/O
G_3.3-1	The GML data file has the file extension <code>.gml</code>	M
G_3.3-2	The XML schema file has the file extension <code>.xsd</code>	M

4 Requirements for package structure

This section lists requirements on how to place different files of a geospatial vector dataset an Information Package.

Each geospatial vector dataset in an Information Package preserved using this profile consists of several files:

Data files

- a *GML data file* with geometries and attributes. Geometry is defined as a point or an aggregate of points representing anything in the world that has a location, e.g., a

¹ OGC (Open Geospatial Consortium) in ISO 19136 GML version 3.2.1, <https://www.ogc.org/standards/gml>

³ INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, <https://inspire.ec.europa.eu/documents/inspire-metadata-implementing-rules-technical-guidelines-based-en-iso-19115-and-en-iso-1>

point, line or polygon (feature types). An attribute is an information describing the geometry.

- a *PRJ file* with a full description of the Coordinate Reference System (CRS) used for data in the gml-file

Metadata files

- a *XML metadata file* with descriptive geospatial metadata about at dataset level compliant with the INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119

Schema files

- a *XML schema definition* with metadata, which defines the XML storage format of the geospatial vector dataset and describes the geometry and attributes used in the gml file
- *OGC schema files for GML 3.2.1* used for validation that gml files in the IP is valid GML 3.2.1

Documentation files

- a *text, image, audio or video file* describing the dataset (contextual documentation)

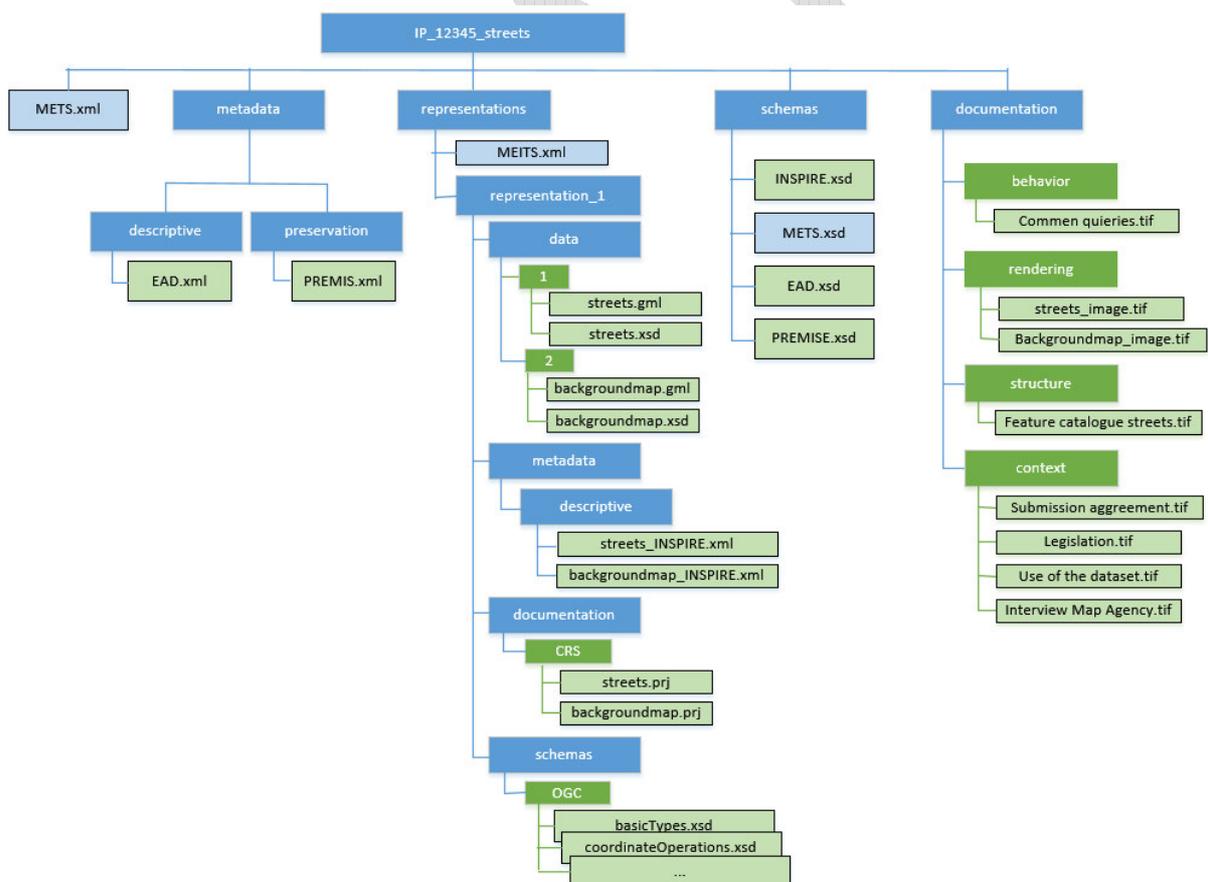


Figure 1: Folder structure of the Long-Term preservation format profile for Geospatial Vector data using GML 3.2.1

ID	Description of requirement	M/O
P_4.0-1 Ref GEO_11 and GEO_12 and GEO_13	If the value in mets/@csip: CONTENTINFORMATIONTYPE is "GeoData" and the IP contains geospatial vector data then there MUST exist at least one file with extension .gml in representations/[RepresentationName]/data or in any of its subfolders	M
P_4.0-2 Ref GEO_12	If there are more gml files in a representation each gml file COULD be placed in subfolders in representations/[RepresentationName]/data named consecutively with a number starting with 1	O
P_4.0-3	If the value in mets/@csip: CONTENTINFORMATIONTYPE is "GeoData" and the IP contains geospatial vector data a folder named OGC containing the schema collection from Open Geospatial Consortium to GML version 3.2.1 in an unchanged form SHOULD be provided in representations/[RepresentationName]/schemas	O
P_4.0-4	There MUST be an XML schema definition xy.xsd for each GML file xy.gml that indicates the XML storage format of the geospatial vector dataset	M
P_4.0-5	An XML schema definition, complementary to a single gml file SHOULD be placed in representations/[RepresentationName]/data in the same folder as the gml file that the schema definition validates	O
P_4.0-6	If an XML schema definition specify the same syntax rules for several gml files in the IP the schema definition SHOULD be placed in representations/[RepresentationName]/schema or in the schema folder on package level	O
P_4.0-7 Ref GEO_15 and GEO_38	A full description of the Coordinate Reference System (CRS) for all geospatial records is provided in one of two ways: <ul style="list-style-type: none"> - in an accompanying projection file (TFW) in representations/[RepresentationName]/documentation/data - in an accompanying projection file (TFW) in representations/[RepresentationName]/documentation/CRS 	O
P_4.0-8 Ref GEO_31	An image displaying the overall view of any geospatial dataset in the IP MUST be provided in a documentation/rendering	M

5 Requirements for data

A geospatial vector dataset is always stored in a GML file. An XML schema definition is generated for each GML datafile.

ID	Description of requirement	M/O
D_5.0-1	<p>The schema definition <code>xy.xsd</code> must be complied with for the <code>xy.gml</code> file. This means that <code>xy.gml</code> must pass validation against <code>xy.xsd</code></p> <p>Example The schema definition <code>streets.xsd</code> must be complied with for the <code>streets.gml</code> file. This means that <code>streets.gml</code> must pass validation against <code>streets.xsd</code></p>	M

5.1 XML schema definition requirements

The XML schema definition defines the XML storage format of the geospatial vector dataset and describes the data contained in the GML datafile. It is used for automatic validation and dissemination of the GML datafile. Requirements for the content and structure of the XML schema definition is defined below.

ID	Description of requirement	M/O
D_5.1-1 Ref GEO_18 and G_3.1-1	A XML schema definition MUST comply with the syntax for schemas according to W3C XML Schema Definition Language (XSD) 1.1 Part 1.	M
D_5.1-2	<p>In the root element of the XML schema definition, the following namespaces MUST be defined: <code>gml</code>, <code>xlink</code>, <code>xsi</code> and the namespace for the submission (in the example below <code>streets</code>).</p> <p>Example</p> <pre>xmlns="http://www.w3.org/2001/XMLSchema" xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:streets="http://www.streets.dk/gml/streets" targetNamespace="http://www.streets.dk/gml/streets" elementFormDefault="qualified"></pre> <p>WHY? To enable automated validation of the GML file.</p>	M
D_5.1-3	<p>The XML schema definition MUST contain an import element for the namespace for GML version 3.2.1</p> <p>Example</p>	M

	<pre><import namespace="http://schemas.opengis.net/gml" schemaLocation="http://schemas.opengis.net/gml http://schemas.opengis.net/gml/3.2.1/gml.xsd"/></pre>	
D_5.1-4	<p>The XML schema definition SHOULD contain a general description of the content of the GML file using the XML documentation elements <i>annotation</i> and <i>documentation</i></p> <p>Example</p> <pre><annotation> <documentation>Dataset with streets in Denmark </documentation> </annotation></pre>	O
D_5.1-5 Ref GEO_19 and GEO_16 and GEO_27	<p>The XML schema definition MUST contain at least one local feature with associated GML geometry.</p> <p>Example</p> <pre><element name="Street" type="integer" minOccurs="1" maxOccurs="1"> ... <element ref="gml:LineString" minOccurs="1"></pre> <p>Recommendation</p> <p><i>A local feature can also hold an ID referencing a feature catalogue or table data in the IP with more attributes describing the content of the GML-file.</i></p>	M
D-5.1-6	<p>All local features and geometries in a XML schema definition SHOULD be described using XML documentation elements <i>annotation</i> and <i>documentation</i></p> <p>Example</p> <pre><sequence> <element name="Street" type="string" minOccurs="1" maxOccurs="1"> <annotation> <documentation>Name of the street </documentation> </annotation> </element> <element ref="gml:LineString" minOccurs="1"> <annotation> <documentation>Line placing the geographical position of the street </documentation> </annotation> </element> </sequence></pre>	O
D_5.1-7	<p>Local features and geometries in a XML schema definition SHOULD be annotated as a sequence of features with the <i>sequence</i> element</p> <p>Example</p> <pre><sequence> <element name="Street" type="string" minOccurs="1" maxOccurs="1"> <annotation> <documentation>Name of the street</pre>	O

	<pre> </documentation> </annotation> </element> <element ref="gml:LineString" minOccurs="1"> <annotation> <documentation>Line placing the geographical position of the street </documentation> </annotation> </element> </sequence> </pre>	
D_5.1-8	<p>A collection element MUST be defined in an XML schema definition and inherits from the class <i>gml:Feature</i> annotated using the XML attribute <i>substitutionGroup</i> and annotated with the elements <i>complexType</i> and <i>complexContent</i></p> <p>Example In this example the collection element is named GEOMETRY:</p> <pre> <element name="GEOMETRY" type="streets:GEOMETRYType" substitutionGroup="gml:_Feature"/> ... <complexType name="GEOMETRYType"> <complexContent> </pre>	M
D_5.1-9	<p>A XML schema definition MUST inherit the properties of the GML class <i>gml:AbstractFeatureType</i>.</p> <p>Example</p> <pre> <extension base="gml:AbstractFeatureType"> </pre>	M
D_5.1-10	<p>A local feature in a XML schema definition COULD establish a link to other information in the Information Package. The link between other information and a GML instance can be documented in the XML schema file for the local feature which is a part of the reference.</p> <p>Example</p> <pre> <element name="FileID" type="string" minOccurs="1" maxOccurs="1"> <annotation> <documentation>Unique number identifying a case file in an electronic record management system (ERDMS). This local feature references the column "CaseFileNumber" in the table "Files" stored in the siard-file with data from an ERDMS in this SIP. </documentation> </annotation> </pre>	O

Figure 2: Example of an XML schema definition (streets.xsd)

```

<?xml version="1.0" encoding="UTF-8"?>
<schema
xmlns="http://www.w3.org/2001/XMLSchema"
xmlns:gml="http://www.opengis.net/gml"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:streets="http://www.streets.dk/gml/streets"
targetNamespace="http://www.streets.dk/gml/streets" elementFormDefault="qualified">

  <import namespace="http://www.opengis.net/gml"
  schemaLocation="http://schemas.opengis.net/gml http://schemas.opengis.net/gml/3.2.1/gml.xsd"/>

  <element name="GEOMETRY" type="sa:GEOMETRYType" substitutionGroup="gml:_Feature"/>

  <annotation>
  <documentation>Dataset with streets in Denmark.
  </documentation>
  </annotation>

  <complexType name="GEOMETRYType">
  <complexContent>

  <extension base="gml:AbstractFeatureType">

  <sequence>

  <element name="FileID" type="string" minOccurs="1" maxOccurs="1">
  <annotation>
  <documentation> Unique number identifying a case file in an electronic record management system
  (ERDMS). This local feature references the column "CaseFileNumber" in the table "Files" stored in the
  siard-file with data from an ERDMS in this SIP</documentation>
  </annotation>
  </element>

  <element name="Street" type="string" minOccurs="1" maxOccurs="1">
  <annotation>
  <documentation>Name of the street</documentation>
  </annotation>
  </element>

  <element ref="gml:LineString" minOccurs="1">
  <annotation>
  <documentation>Line placing the geographical position of the street</documentation>
  </annotation>
  </element>

  </sequence>
  </extension>
  </complexContent>
  </complexType>
</schema>

```

5.2 GML file requirements

ID	Description of requirement	M/O
D_5.2-1	<p>GML files larger than 1 GB MUST be subdivided into smaller GML files</p> <p>Recommendation <i>It is recommended that GML files larger than 1 GB is subdivided into smaller GML files because GML files larger than 1-2 GB are impossible to produce, test, correct or visualize in a GIS.</i></p>	M
D_5.2-2	<p>Geometries and attributes from the same geospatial vector dataset SHOULD be kept together within the same GML file</p>	O
D_5.2-3	<p>A <code>gml:FeatureCollection</code> element MUST be used as root element in the GML file.</p> <p>Example</p> <pre data-bbox="363 913 1294 1167"> <gml:FeatureCollection xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:streets="http://www.streets.dk/gml/streets" xsi:schemaLocation="http://www.streets.dk/gml/streets ./streets.xsd"> ... </gml:FeatureCollection> </pre>	M
D_5.2-4	<p>In the root element of a GML file, the following namespaces MUST be defined: <code>gml</code>, <code>xlink</code>, <code>xsi</code> and the namespace for the submission</p> <p>Example</p> <p>In this example the namespace for the specific submission contained in this GML-file is streets:</p> <pre data-bbox="363 1435 1294 1675"> <gml:FeatureCollection xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:streets ="http://www.streets.dk/gml/streets" ... "> </pre>	M
D_5.2-5	<p>The root element of a GML file MUST specify a schema reference <code>xsi:schemaLocation</code> with the namespace of the submission and a relative path for the location of the XML schema definition in the Information Package</p> <p>Example</p>	M

	<p>In this example the path <code>./streets.xsd</code> references an XML schema definition placed in the same folder as the GML file in the IP:</p> <pre><gml:FeatureCollection ... xsi:schemaLocation="http://www.streets.dk/gml/streets ./streets.xsd"></pre>	
D_5.2-6 Ref GEO_15	<p>A GML file MUST reference the underlying Coordinate Reference System (CRS) using the GML attribute <code>srsName</code> in the GML element <code>gml:boundedBy</code> with the sub elements <code>Envelope</code></p> <p>Example</p> <p>In this example an EPSG code is used as reference:</p> <pre><gml:boundedBy> <gml:Envelope srsName="EPSG:25832" ... </gml:Envelope> </gml:boundedBy></pre>	M
D_5.2-7	<p>All geometries (<code>gml:featureMember</code> elements) in the GML file SHOULD be located entirely within a bounding box defined in the GML elements <code>lowerCorner</code> and <code>upperCorner</code> in the <code>gml:boundedBy</code> element</p> <p>Example</p> <pre><gml:boundedBy> <gml:Envelope srsName="EPSG:25832" srsDimension="2"> <gml:lowerCorner>212481.60 6019669.40</gml:lowerCorner> <gml:upperCorner>961440.75 6510422.51</gml:upperCorner> </gml:Envelope> </gml:boundedBy></pre> <p>Recommendation</p> <p><i>The bounding box should be kept as small as possible</i></p>	O
D_5.2-8 Ref GEO_16	<p>The geometries in a GML file MUST 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</p>	M
D_5.2-9	<p>The dimension of the geometry in the GML file MUST be defined using the GML attribute <code>srsDimension</code> in the GML element <code>Envelope</code></p> <p>Example</p> <pre><gml:Envelope ... srsDimension="2"></pre>	M
D_5.2-10	<p>A GML file MUST contain at least one <code>gml:featureMember</code> element</p> <p>Example</p> <pre><gml:featureMember></pre>	M

	<pre> <streets:GEOMETRY> <streets:Street>Kalvebod brygge</Street> <gml:LineString> <gml:posList>724944.42 6175249.02 724881.34 6175219.90 723779.41 6174147.20 </gml:posList> </gml:LineString> </streets:GEOMETRY> </gml:featureMember> </pre>	
<p>D_5.2-11</p>	<p>A collection element containing the local features and geometries of the GML file MUST be placed in a <i>gml:featureMember</i> element</p> <p>Example</p> <p>In this example the collection element is named GEOMETRY</p> <pre> <gml:featureMember> <streets:GEOMETRY> <streets:Street>Kalvebod brygge</Street> <gml:LineString> <gml:posList>724944.42 6175249.02 724881.34 6175219.90 723779.41 6174147.20 </gml:posList> </gml:LineString> </streets:GEOMETRY> </gml:featureMember> </pre>	<p>M</p>
<p>D_5.2-12</p>	<p>The collection element MUST be included in the namespace of the submission</p> <p>Example</p> <p>In this example the namespace of the submission is streets:</p> <pre> <gml:featureMember> <streets:GEOMETRY> <streets:Street>Kalvebod brygge</Street> <gml:LineString> <gml:posList>724944.42 6175249.02 724881.34 6175219.90 723779.41 6174147.20 </gml:posList> </gml:LineString> </streets:GEOMETRY> </gml:featureMember> </pre>	<p>M</p>

Figure 3: Example of a GML-file (streets.gml)

```

<?xml version="1.0" encoding="UTF-8"?>

<gml:FeatureCollection
  xmlns:gml="http://www.opengis.net/gml"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:streets="http://www.streets.dk/gml/streets"
  xsi:schemaLocation="http://www.streets.dk/gml/streets ./streets.xsd">

  <gml:boundedBy>
    <gml:Envelope srsName="EPSG:25832" srsDimension="2">
      <gml:lowerCorner>212481.60 6019669.40</gml:lowerCorner>
      <gml:upperCorner>961440.75 6510422.51</gml:upperCorner>
    </gml:Envelope>
  </gml:boundedBy>

  <gml:featureMember>

    <streets:GEOMETRY>

      <streets:FileID>2001-01000</streets:FileID>

      <streets:Street>Kristian Erslevs Gade</streets:Street>

      <gml:LineString>
        <gml:posList>724297.34 6174822.74 724372.67 6174738.59
      </gml:posList>

    </gml:LineString>
  </streets:GEOMETRY>
</gml:featureMember>

  <gml:featureMember>

    <streets:GEOMETRY>

      <streets:FileID >2001-07501</street:FileID >

      <streets:Street >Kalvebod Brygge</streets:Street >

      <gml:LineString>
        <gml:posList>724944.42 6175249.02 724881.34 6175219.90 723779.41 6174147.20
      </gml:posList>

    </gml:LineString>
  </sa:GEOMETRY>
</gml:featureMember>

</gml:FeatureCollection>

```

6 Requirements for metadata

ID	Description of requirement	M/O
M_6.0-1 Ref GEO_42 Ref G_3.1-2	Geospatial metadata in the long-term preservation format representation of the Information package must comply with the INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Metadata is placed in <i>Representations/[RepresentationName]/metadata/descriptive</i>	M

DRAFT