 ISO/IEC JTC 1/SC 29/WG 6 N0340

**ISO/IEC JTC 1/SC 29/WG 6**

**MPEG Audio Coding Convenorship: DE**

**Document type:** Output Document

**Title:** Metadata format for audio coding for machines

**Status:** Approved

**Date of document:** 2025-04-04

**Source:** ISO/IEC JTC 1/SC 29/WG 6

**Expected action:** None

**Action due date:** None

**No. of pages:** 6 (without cover page)

**Email of Convenor:** Thomas.Sporer@idmt.fraunhofer.de

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg6>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 6**

**MPEG AUDIO CODING**

**ISO/IEC JTC 1/SC 29/WG 6 N0340**

**April 2025, Online**

|  |  |
| --- | --- |
| **Title** | **Metadata format for audio coding for machines** |
| **Source** | **WG 6, MPEG Audio Coding** |
| **Status** | **Approved** |
| **Serial Number** | **25012** |

# Abstract

This document describes the metadata format for ACoM, which adopts a unified structure for machine-readable metadata that describes audio content.

# ACoM Metadata Input Format

## **Overview of ACoM Metadata Group Structure**

A characteristic property of MPEG-ACoM is the ACoM metadata as shown in the introduction. The list of its structured elements is as follows.

Groups of values:

* **General** describe the recording situation
* For each recording channel **AudioStream** specifies the file and the microphone parameters
* **Annotation** describes information that a machine is expected to infer while performing a task
* **Object** describes the object recorded
* **Rights** describes the usage rights of the data

## **Field Specification**

In general, most fields are optional and use case dependent. Whether field is optional or non-optional see attached XSD.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | **Field** | **Description** | **Values (example)** | **Default** |
| General | SamplingRate | Sampling rate in Hz or wav (values from wav file header used) | String | 48000.0 |
| Channels | Number of channels | Positive Int (1) | 1 |
| Duration | Total audio file length [second] | Float |  |
| AudioFormat | Specifies int16/int24/int32/float32 or wav (values from wav file header used) | String | Int16 |
| Humidity | Relative humidity | String | 50% |
| Temperature | Temperature in C | String | 20.0 |
| AirPressure | Air pressure in hPascal | String | 1013.25 |
| NmberOfSensors | Number of AudioStreams | String (16) |  |
| NmberOfAnnotations | Number of Annotations | String |  |
| DateTime | Date and time of recording in UTC | String (2025-01-06 11:00:00) |  |
| Sensor | Id | Unique identifier of sensor channel | String  (Chan1) (Right) |  |
| File | File name of recording channel relative to metadata file | String (sensor1.wav) |  |
| SensorType | Brand name and type of sensor | String (Shure SM58) |  |
| SensorDirectivity | Either Omni, cardioid, super-cardioid, hypercardiod, or filename of balloon data specifying sensor | String | Omni |
| SensorPosition | Comma separated of cartesian coordinates of sensor in meters (1) or string defining medical sensor definition (2) or position label of loudspeaker (e.g. in the case of a stereo mix) (3) | string 1: (“10,15,10”) 2: (“FPz”) 3: (“Left”) |  |
| SensorTolerance | Comma separated of tolerance of cartesian coordinates of sensor in meters | string (“0.1,0.1,0.1”) | 0,0,0 |
| SensorDirection | Cartesian vector of sensor direction | String,string,string (-10,-15,10) | Pointing towards 0,0,0 |
| SensorDirectionTolerance | Tolerance of sensor direction | string  (0.2,0.2,0.2) | 0,0,0 |
| SensorGain | Scaling of sensor signal in dB (positive: data in file is amplified compared to raw output of sensor) | String | 0.0 dB |
| Annotation | Id | Unique identifier of annotation | String |  |
| File\* | File name of annotation. (see more in Section 2.3)  If this field exist, AnnotationValue, Onset and Offset field are not allowed. | String (TimeVaryingOperationMode.csv,  FaultStatus\_with\_onset.csv,  Transcript.csv) |  |
| AudioFile | File name of recording channel associated to annotation file | String  (sensor\_01.wav) |  |
| AnnotationType | Type of annotation, identical to the column header in the annotation file, (see more in Section 2.3) | String  (OperationMode, SoundEvent, FaultStatus, Transcript, Gender, SpakerName,  AudioSceneClass, …) |  |
| AnnotationValue | Actual value of specific annotation, that could be time-varying.  If AnnotationValue field exist, File field is not allowed. | String  (for AnnotionType is SoundEvent, it would be ‘Music’, ‘Speech’, etc) |  |
| Onset | Start time of described annotation in the audio file in seconds.  If Onset field exist, File field is not allowed. | String, 0.0 |  |
| Offset | End time of described annotation in the audio file in seconds.  If Offset field exist, File field is not allowed. | String, 10.0 |  |
| Object | Name | Verbal description of object measured | String (“electric drill, Einhell 4711“) |  |
| OperationMode\*\* | OperationMode like “heavy load”, “broken”, “gear defect”, “motor blocked”, ….  This might also include a specification of material processed and tool used (if machine has different tools) | String (“5mm drill in steel A3”) |  |
| ObjPosPrecision | Comma separated of tolerance of cartesian coordinates of object in meters | String (“0.1,0.1,0.1”) | 0,0,0 |
| ObjOrientation | Cartesian vector of object direction | String (-10,-15,10) | Pointing towards 0,0,0 |
| ObjOrientDescr | Text describing semantics of ObjOrientation | String (“direction where the drill points to, handle pointing down”) |  |
| ObjOrientTolerance | Tolerance of ObjOrientation | String (“0.2,0.2,0.2”) | 0,0,0 |
| Transcript | For content containing speech recordings this might contain a transcript of spoken words (ground truth) | String | OPTIONAL |
| Age | Self-reported age value | Int  (25) |  |
| AgeCategory | Age category | String  (Child) |  |
| Gender | Self-reported gender | String  (male, M) |  |
| RespiratoryCondition | The patient has other respiratory conditions | Boolean  (TRUE) |  |
| FeverMusclePain | The patient has fever or muscle pain | Boolean  (FALSE) |  |
| Height | Height in centimeters | Int  (115) |  |
| Weight | Weight in kilograms | Float  (17.1) |  |
| PregnancyStatus | Did the patient report being pregnant during screening | Boolean  (FALSE) |  |
| Murmur | Indicates if a murmur is present, absent or unidentifiable for the annotator | String  (Present) |  |
| Rights | Type | Define the usage rights of the data.  Valid values are “public\_domain”, “real-time\_processing”, “MPEG\_internal” | String | MPEG\_internal |
| File | link to the rights file relative to directory of audio data | String |  |
| ContentOwner | Ownership of data (name, Institute, department) | String (“Fraunhofer IDMT, Germanx”) |  |
| Further | Category | Kind of content (industry, domestic, tool,..) |  |  |
| Usage | All further relevant information |  |  |

\*Annotation can be specified in the XML file or in external annotation file.

\*\*If the value of OperationMode is static and does not change over time, it should be described within the object group using the operationMode field. If OperationMode has time-varying annotation values, it should be described in the annotation group with OperationMode as the AnnotationType.

## **Structure of Annotation File**

Annotation files are structured in CSV format, with the first row specifying the column headers. These headers vary depending on the type of annotation. For time-varying annotations, the headers typically include *AudioFile*, *onset*, *offset*, and *AnnotationType*, where *AudioFile* and *AnnotationType* correspond to a field defined in the Annotation Group. These fields indicate the start and end times of a given annotation value, along with the actual value associated with the specified *AnnotationType*. Figure 2 provides an example of a CSV file used for time-varying annotations. Each row specifies the audio file associated with the annotation, the start and end times of the sound event, and the label describing the sound event.

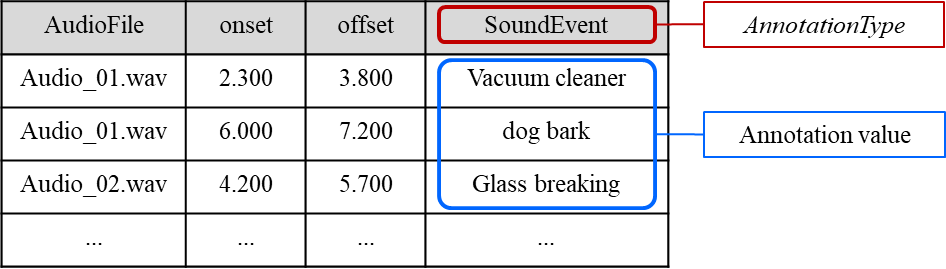


Figure 2. An example of a CSV file used for time-varying annotations

The meaning of each annotation value is determined by its corresponding *AnnotationType*. For instance, if the *AnnotationType* is "OperationMode", the annotation value may describe the machine's state, such as "idle", "heavy load", or "motor blocked". Alternatively, if the *AnnotationType* is "SoundEvent", the annotation value represents the type of detected sound event during the specified time segment — for example, "dog bark", "vacuum cleaner", or "glass breaking".

In contrast, static annotations provide information that remains constant throughout the entire audio recording. In such cases, the annotation consists of two fields, *AudioFile* and *AnnotationType* whose value applies to the full duration of the file without any associated time metadata.

# Schema definition toolchain for ACoM metadata

It is intended to store the metadata in XML format [1]. One advantage of this file format is that it can be read and processed by both humans and machines.

With “Visual Studio Code*”* [2] a free editor is available. Suitable extensions for editing XML files are available with “XML Language support*”* [3] and “XML Tools for Visual Studio Code*”* [4].

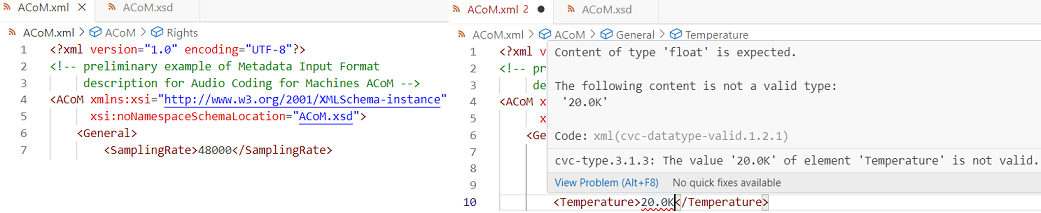


Figure 4: XML Language support in Visual Studio Code in action,  
left: valid XML, right: invalid value for temperature element

The schema of an XML file can be described with an XSD file. One advantage of XSD is that the cardinality of the individual elements can be specified. This allows optional and mandatory elements to be defined. Also the format of the individual elements can be specified in detail.

Examples can be found in the appendix, although these are not to be understood as final specifications for ACoM metadata.

# Reference

1. <https://www.w3.org/TR/xml>
2. https://code.visualstudio.com
3. <https://marketplace.visualstudio.com/items?itemName=redhat.vscode-xml>
4. <https://marketplace.visualstudio.com/items?itemName=DotJoshJohnson.xml>

Appendix

## Exemplary XSD file (XML based scheme description)

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

<!-- preliminary prototype of metadata input format

     scheme description for Audio Coding for Machines ACoM -->

<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

    <xs:simpleType name="cartesian">

        <xs:restriction base="xs:string">

            <xs:pattern value="[0-1000].[0-99],[0-1000].[0-99],[0-1000].[0-99]"/>

        </xs:restriction>

    </xs:simpleType>

    <xs:simpleType name="angular">

        <xs:restriction base="xs:string">

            <xs:pattern value="[0-359],[0-359],[0-359]"/>

        </xs:restriction>

    </xs:simpleType>

    <xs:simpleType name="pint">

        <xs:restriction base="xs:positiveInteger"/>

    </xs:simpleType>

    <xs:element name="ACoM">

        <xs:complexType>

            <xs:sequence>

                <xs:element name="General" minOccurs="1" maxOccurs="1">

                    <xs:complexType>

                        <xs:sequence>

                            <xs:element name="SamplingRate" type="pint" minOccurs="1" maxOccurs="1"/>

<xs:element name="Channels" type="pint" minOccurs="0" maxOccurs="1"/>

<xs:element name="Duration" type="xs:float" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="AudioFormat" type="xs:string" minOccurs="1" maxOccurs="1"/>

                            <xs:element name="Humidity" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="Temperature" type="xs:float" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="AirPressure" type="xs:float" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="NumberOfSensors" type="pint" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="NumberOfAnnotations" type="pint" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="DateTime" type="xs:dateTime" minOccurs="0" maxOccurs="1"/>

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

                <xs:element name="Sensor" minOccurs="0" maxOccurs="unbounded">

                    <xs:complexType>

                        <xs:sequence>

                            <xs:element name="Id" type="xs:string" minOccurs="1" maxOccurs="1"/>

                            <xs:element name="File" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorType" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorDirectivity" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorPosition" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorTolerance" type="cartesian" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorDirection" type="angular" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorDirectionTolerance" type="cartesian" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="SensorGain" type="xs:string" minOccurs="0" maxOccurs="1"/>

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

                <xs:element name="Annotation" minOccurs="0" maxOccurs="unbounded">

                    <xs:complexType>

<!-- mutual exclusion of file vs. AnnotationValue will be added in the future by xsl:choose -->

                        <xs:sequence>

                            <xs:element name="Id" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="File" type="xs:string" minOccurs=“0" maxOccurs="1"/>  
 <xs:element name=“AudioFile" type="xs:string" minOccurs=“0" maxOccurs="1"/>  
 <xs:element name=“AnnotationType" type="xs:string" minOccurs=“0" maxOccurs="1"/>

<xs:element name=“AnnotationValue" type="xs:string" minOccurs=“0" maxOccurs=“1"/>

<xs:element name=“Onset" type="xs:string" minOccurs=“0" maxOccurs=“1"/>

<xs:element name=“Offset" type="xs:string" minOccurs=“0" maxOccurs=“1"/>

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

                <xs:element name="Object" minOccurs="0" maxOccurs="unbounded">

                    <xs:complexType>

                        <xs:sequence>

                            <xs:element name="Name" type="xs:string" minOccurs="1" maxOccurs="1"/>

                            <xs:element name="OperationMode" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="ObjPosPrecision" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="ObjOrientation" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="ObjOrientDescr" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="ObjOrientTolerance" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="Transcript" type="xs:string" minOccurs="0" maxOccurs="1"/>

<xs:element name="Age" type="xs:int" minOccurs="0" maxOccurs="1"/>

<xs:element name="AgeCategory" type="xs:string" minOccurs="0" maxOccurs="1"/>

<xs:element name="Gender" type="xs:string" minOccurs="0" maxOccurs="1"/>

<xs:element name="RespiratoryCondition" type="xs:boolean" minOccurs="0" maxOccurs="1"/>

<xs:element name="FeverMusclePain" type="xs:boolean" minOccurs="0" maxOccurs="1"/>

<xs:element name="Height" type="xs:int" minOccurs="0" maxOccurs="1"/>

<xs:element name="Weight" type="xs:float" minOccurs="0" maxOccurs="1"/>

<xs:element name="PregnancyStatus" type="xs:boolean" minOccurs="0" maxOccurs="1"/>

<xs:element name="Murmur" type xs:string" minOccurs="0" maxOccurs="1"/>

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

                <xs:element name="Rights" minOccurs="1" maxOccurs="1">

                    <xs:complexType>

                        <xs:sequence>

                            <xs:element name="Type" type="xs:string" minOccurs="0" maxOccurs="1"/>

                            <xs:element name="File" type="xs:string" minOccurs="1" maxOccurs="1"/>

                            <xs:element name="ContentOwner" type="xs:string" minOccurs="1" maxOccurs="1"/>

                            <xs:element name="Privacy" minOccurs="0" maxOccurs="1"/>

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

                <xs:element name="Further" minOccurs="0" maxOccurs="unbounded">

                    <xs:complexType>

                        <xs:sequence>

                            <xs:element name="Category" />

                            <xs:element name="Usage" />

                        </xs:sequence>

                    </xs:complexType>

                </xs:element>

            </xs:sequence>

        </xs:complexType>

    </xs:element>

</xs:schema>

## Exemplary XML file

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

<!-- preliminary example of Metadata Input Format

     description for Audio Coding for Machines ACoM -->

<ACoM xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

      xsi:noNamespaceSchemaLocation="ACoM.xsd">

    <General>

        <SamplingRate>48000</SamplingRate>

        <AudioFormat>int16</AudioFormat>

<Channels>1</Channels>

<Duration>1.0</Duration>

        <Humidity>50%</Humidity>

        <Temperature>20.0</Temperature>

        <AirPressure>1013.25</AirPressure>

        <NumberOfSensors>1</NumberOfSensors>

        <NumberOfAnnotations>1</NumberOfAnnotations>

        <DateTime>2025-01-06T11:00:00</DateTime>

    </General>

    <Sensor>

        <Id>1</Id>

<SensorType>microphone</SensorType>

        <SensorDirectivity>Omni</SensorDirectivity>

        <SensorPosition></SensorPosition>

        <SensorTolerance>0.1,0.1,0.1</SensorTolerance>

        <SensorDirection>0,0,0</SensorDirection>

        <SensorDirectionTolerance>0.1,0.1,0.1</SensorDirectionTolerance>

        <SensorGain>0.0 dB</SensorGain>

    </Sensor>

    <Annotation>

        <Id>1</Id>

        <File>annotation.csv</File>

        <AnnotationType>audioScene</AnnotationType>

    </Annotation>

    <Object>

        <Name>electric drill</Name>

        <OperationMode>idle</OperationMode>

        <ObjPosPrecision>0,0,0</ObjPosPrecision>

        <ObjOrientation>0,0,0</ObjOrientation>

        <ObjOrientDescr>drilling down</ObjOrientDescr>

        <ObjOrientTolerance>0,0,0</ObjOrientTolerance>

    </Object>

    <Rights>

        <Type>MPEG\_internal</Type>

        <File>license.txt</File>

        <ContentOwner>The company</ContentOwner>

    </Rights>

    <Further>

        <Category>Private information</Category>

        <Usage>Any useful information</Usage>

    </Further>

</ACoM>