INTERNATIONAL STANDARD

ISO/IEC 23000-13

Second edition
2017-11

Information technology - Multimedia application format (MPEG-A) —

Part 13: Augmented reality application format

Technologies de l'information - Format des applications multimedias —

Partie 13: Format pour les Applications de Realité Augmentée

Reference number
ISO/IEC 23000-13:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

This second edition cancels and replaces the first edition (ISO/IEC 23000-13:2014), which has been technically revised.

It also incorporates the Amendment ISO/IEC 23000-13:2014/Amd. 1:2015.

A list of all parts in the ISO/IEC 23000 series can be found on the ISO website.

Introduction

Augmented Reality (AR) applications refer to a view of a real-world environment (RWE), whose elements are augmented by content, such as graphics or sound, in a computer driven process. Augmented Reality Application Format (ARAF) is a collection of a subset of the ISO/IEC 14496-11 Scene Description and Application Engine standard, combined with other relevant MPEG standards (e.g. ISO/IEC 23005, MPEG-V), designed to enable the consumption of 2D/3D multimedia content. Consequently, this document focuses not on client or server procedures, but on the data formats used to provide an augmented reality presentation.

Information technology - Multimedia application format (MPEG-A) —

Part 13:
Augmented reality application format

1 Scope

This document specifies the following:

— scene description elements for representing AR content;

— mechanisms to connect to local and remote sensors and actuators;

— mechanisms to integrated compressed media (image, audio, video, graphics);

— mechanisms to connect to remote resources such as maps and compressed media.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 10646-1:2012, *Information technology — Universal multiple-octet coded character set (UCS) — Part 1: Architecture and basic multilingual plane*

ISO/IEC 14496-1:2010 + Amd. 2:2014, *Information technology — Coding of audio-visual objects — Part 1: Systems*

ISO/IEC 14496-3:2009, *Information technology — Coding of audio-visual objects — Part 3: Audio*

ISO/IEC 14496-11:2015, *Information technology — Coding of audio-visual objects — Part 11: Scene description and application engine*

ISO/IEC 14496-16:2011, *Information technology — Coding of audio-visual objects — Part 16: Animation Framework eXtension (AFX)*

ISO/IEC 14772-1:1997, *Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding*

ISO/IEC 23005-5, *Information technology — Media context and control — Part 5: Data formats for interaction devices*

3 Terms, definitions, and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 23000-1 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1.1
ARAF browser
augmented reality application format compliant browser

3.1.2
MAR scene
textual result of the MAREC creation, played by an *ARAF browser* (3.1.1)

Note 1 to entry: The result is a MAR experience.

3.1.3
MAR experience
act of playing the ARAF scene using an *ARAF browser* (3.1.1)

Note 1 to entry: The ARAF browser interprets the ARAF scene and presents the result on the end-user's device.

3.1.4
content creator
creator of the media files that are being used within the *MAR experience* (3.1.3)

Note 1 to entry: The media files can be 2D and/or 3D graphics, images, videos and/or sounds.

3.1.5
end-user device
smartphone or mobile device used by an end-user to play a *MAR scene* (3.1.2)

Note 1 to entry: The device shall have an ARAF browser installed.

3.1.6
processing server
server that offers at least one required processing functionality for a *MAR experience* (3.1.3) and it is capable of communicating with an *ARAF browser* (3.1.1)

3.1.7
target resource
target image or target image descriptor

Note 1 to entry: The target image represents the image that shall be detected and recognized by a recognition library. The target image descriptor is represented by the visual descriptors extracted from a target image. The target resources may be specified by the MAREC or they can be already stored in databases on remote servers.

3.1.8
prerecorded video
prerecorded 2D video whose location is specified by MAREC

Note 1 to entry: The video file can be stored locally (on the device where the MAR experience is played) or remotely (anywhere else on the web). The recognition process shall be performed on the frames (still images) composing the video.

3.1.9
live video camera (stream)
live 2D video camera feed

Note 1 to entry: The URL of the camera providing the real time capture is specified by the MAREC. The URL can point to one of the cameras of the device where the MAR experience is played or to any other camera that can provide a live video stream and the ARAF browser can connect to.

3.1.10
image recognition library
library that is able to recognize *target resources* (3.1.7) in a video

Note 1 to entry: The library can run locally (implemented in the ARAF browser) or remotely (on a processing server). The result of an image recognition library is an array of indexes of the recognized target resources.

3.1.11
image recognition and tracking library
library that is able to recognize and track *target resources* (3.1.7) in a video

Note 1 to entry: The library can run locally (implemented in the ARAF browser) or remotely (on a processing server). The result of a recognition and tracking library is an array of indexes of the recognized target resources and their pose matrixes. Each recognized target resource shall have a pose matrix associated or a default value if the corresponding pose matrix could not be computed.

3.1.12
augmentation resource
media objects that are used in the augmentation of the *MAR experience* (3.1.3)

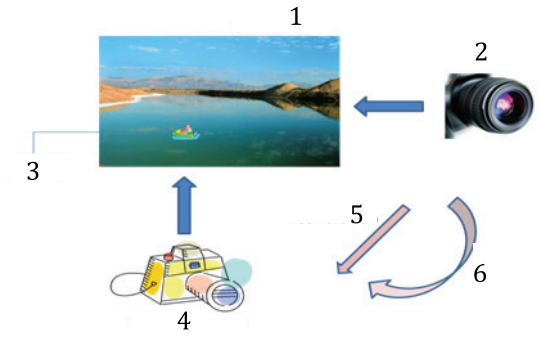
Note 1 to entry: A valid augmentation resource can be a 2D/3D grapic element, an image, a video, a sound or a BIFS scene. The augmentation resources can be stored locally in the MAR Scene or remotely anywhere on the Web, as long as the ARAF browser is capable of accesing their locations. In this case, a URL pointing to the augmentation resource is stored in the MAR scene.

3.2 Abbreviated terms

AR	Augmented Reality
ARAF	Augmented Reality Application Format
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
MAR	Mixed and Augmented Reality
MARE	Mixed and Augmented Reality Experience
MAREC	Mixed and Augmented Reality Experience Creator
PROTO	A PROTOtype is a mechanism used to group together scene graph elements in order to implement one or several specific functionalities.
RTR	Recognized Target Resource

4 ARAF principle and context

Augmented Reality (AR) applications refer to a view of a real-world environment whose elements are augmented by content, such as graphics or sound, in a computer driven process. Figure 1 illustrates two real and virtual cameras and the composition of a real image and graphics objects. Annex C describes several application scenarios for augmented reality.



Key

1 real picture
2 real camera
3 graphic object
4 virtual camera
5 calibration
6 position and orientation

Figure 1 — Simplified illustration of the AR principle

The Augmented Reality Application Format (ARAF) is an extension of a subset of the MPEG-4 part 11 Scene Description and Application Engine standard, combined with other relevant MPEG standards (MPEG-4, MPEG-V), designed to enable the consumption of 2D/3D multimedia content as depicted in Figure 2.

An ARAF, available as a file or stream, is interpreted by a device, called **ARAF device**. The nodes of the ARAF scene point to different sources of multimedia content such as 2D/3D image, 2D/3D audio, 2D/3D video, 2D/3D graphics and sensor/sensory information sources/sinks that are either remote or/and local.

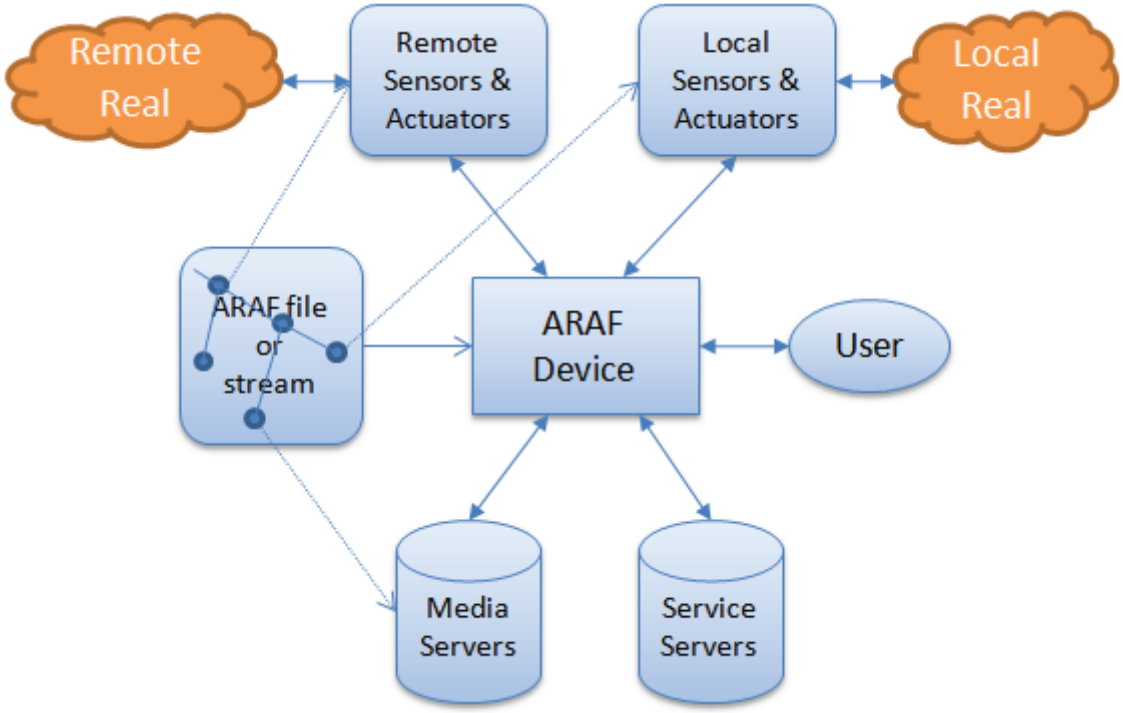


Figure 2 — The ARAF context

5 ARAF scene description

5.1 General

To describe the multimedia scene, ARAF is based on ISO/IEC 14496-11 which at its turn is based on ISO/IEC 14772-1 (VRML97). About two hundred nodes are standardized in MPEG-4 BIFS and VRML, allowing various kinds of scenes to be constructed. ARAF is referring to a subset of MPEG-4 BIFS nodes and external prototypes as defined in ISO/IEC 14496-11:2015, Annex E for scene description as presented in Table 1.

Table 1 — ARAF nodes and prototypes

Category	Sub-category	Node, prototypes/elements name in MPEG-4 BIFS/XMT	Type
Elementary media	Audio	AudioSource	Node
		Sound	Node
		Sound2D	Node
	Image and video	ImageTexture	Node
		MovieTexture	Node
	Textual information	FontStyle	Node
		Text	Node
	Graphics	Appearance	Node
		Color	Node
		LineProperties	Node
		LinearGradient	Node
		Material	Node
		Material2D	Node
		Rectangle	Node
		Shape	Node
		SBVCAnimationV2	Node
		SBBone	Node
		SBSegment	Node
		SBSite	Node
		SBSkinnedModel	Node
		MorphShape	Node
		Coordinate	Node
		TextureCoordinate	Node
		Normal	Node
		IndexedFaceSet	Node
		IndexedLineSet	Node
Programming information		Script	Node

Table 1 *(continued)*

Category	Sub-category	Node, prototypes/elements name in MPEG-4 BIFS/XMT	Type
User interactivity		InputSensor	Node
		OutputActuator	Node
		SphereSensor	Node
		TimeSensor	Node
		TouchSensor	Node
		MediaSensor	Node
		PlaneSensor	Node
		Background	Node
Scene related) information (spatial and temporal relationships		Background2D	Node
		CameraCalibration	PROTO
		Group	Node
		Inline	Node
		Layer2D	Node
		Layer3D	Node
		Layout	Node
		NavigationInfo	Node
		OrderedGroup	Node
		LocImg	PROTO
		RemImgProxy	PROTO
		RemImgServer	PROTO
		RemImgComp	PROTO
		LocAud	PROTO
		RemAud	PROTO
		Switch	Node
		Transform	Node
		Transform2D	Node
		Viewpoint	Node
		Viewport	Node
		Form	Node
Dynamic and animated scene		OrientationInterpolator	Node
		ScalarInterpolator	Node
		CoordinateInterpolator	Node
		ColorInterpolator	Node
		PositionInterpolator	Node
		Valuator	Node
Communication and compression		BitWrapper	Node
		MediaControl	Node
	Maps	Map	PROTO
		MapOverlay	PROTO
		MapMarker	PROTO
		MapPlayer	PROTO
Terminal		TermCap	Node

All the above listed elements are specified in MPEG-4 Part 11. However, to facilitate the implementation of ARAF content, the current document contains their XML syntax as well as the semantics and functionality.

MPEG-4 Part 11 describes a scene with a hierarchical structure that can be represented as a graph. Nodes of the graph build up various types of objects, such as audio video, image, graphic, text, etc. Furthermore, to ensure the flexibility, a new, user-defined type of node derived from a parent one can also be defined on demand by using the *Proto* method.

In general, nodes expose a set of parameters, through which aspects of their appearance and behavior can be controlled. By setting these values, scene designers have a tool to force a scene-reconstruction at clients' terminals to adhere to their intention in a predefined manner. In more complicated scenario, the structure of BIFS nodes is not necessarily static; nodes can be added or removed from the scene graph arbitrarily.

Certain types of nodes called *sensors*, such as TimeSensor, TouchSensor, can interact with users and generate appropriate triggers, which are transmitted to others nodes by routing mechanism, causing changes in state of these receiving nodes. They are bases for the dynamic behavior of a multimedia content supported by MPEG-4.

The maximum flexibility in the programmable feature of MPEG-4 scene is carried out with the *Script* node. By routing mechanism to Event In *valueIn* attribute of Script node, the associated function (defined in its URL attribute) with the same name Event In *valueIn* () will be triggered. The behavior of this function is user-defined, i.e. scene-designer can freely process some computations, and then sets the values for every Event Out *valueOut* attribute, which consecutively affect the states of other nodes linked to them.

Direct manipulation of nodes' states is also possible in MPEG-4 Part 11: the Field *field* attribute can refer to any node in the scene; through this link, all attributes of the contacted node will be exposed to direct setting and modifying operators within the *Script* node. The syntax of the language used to implement the function of Script node is ECMAScript (see ISO/IEC 16262).

ARAF supports the definition and reusability of complex objects by using the MPEG-4 PROTO mechanism. The PROTO statement creates its own nodes by defining a configurable object prototype; it can integrate any other node from the scene graph.

ARAF makes extensive use of EXTERNPROTO mechanism which are nodes whose syntax is identified by URNs as given in ISO/IEC 14496-11:2915, Annex E, while their names are only informative and for convenience can be changed by the content creator in the declaration step.

Table 1 indicates the MPEG-4 Part 11 nodes that are included in ARAF. For each node, it is specified in the version of this document when it was published. Further, the XML syntax as well as the semantics and functionality of these elements are described.

5.1.1 Elementary media

5.1.1.1 Audio

The following audio related nodes are used in ARAF: AudioSource, Sound, Sound2D.

5.1.1.1.1 AudioSource

5.1.1.1.1.1 XSD description

```
<complexType name="AudioSourceType">
 <all>
   <element ref="xmta:IS" minOccurs="0"/>
   <element name="children" form="qualified" minOccurs="0">
     <complexType>
       <group ref="xmta:SFAudioNodeType" minOccurs="0" maxOccurs="unbounded"/>
     </complexType>
```

```
      </element>
    </all>
    <attribute name="url" type="xmta:MFUrl" use="optional"/>
    <attribute name="pitch" type="xmta:SFFloat" use="optional" default="1"/>
    <attribute name="speed" type="xmta:SFFloat" use="optional" default="1"/>
    <attribute name="startTime" type="xmta:SFTime" use="optional" default="0"/>
    <attribute name="stopTime" type="xmta:SFTime" use="optional" default="0"/>
    <attribute name="numChan" type="xmta:SFInt32" use="optional" default="1"/>
    <attribute name="phaseGroup" type="xmta:MFInt32" use="optional"/>
    <attributeGroup ref="xmta:DefUseGroup"/>
   </complexType>
   <element name="AudioSource" type="xmta:AudioSourceType"/>
```

5.1.1.1.1.2 Functionality and semantics

As defined in ISO/IEC 14496-11:2015, 7.2.2.15.

This node is used to add sound to a BIFS scene. See ISO/IEC 14496-3 for information on the various audio tools available for coding sound.

The addChildren eventIn specifies a list of nodes that shall be added to the children field. The removeChildren eventIn specifies a list of nodes that shall be removed from the children field.

The children field allows buffered AudioBuffer or AdvancedAudioBuffer data to be used as sound samples within a structured audio decoding process. Only AudioBuffer and AdvancedAudioBuffer nodes shall be children to an AudioSource node, and only in the case where url indicates a structured audio bitstream. The pitch field controls the playback pitch for the structured audio, the parametric speech (HVXC) and the parametric audio (HILN) decoder. It is specified as a ratio, where 1 indicates the original bitstream pitch, values other than 1 indicate pitch-shifting by the given ratio. This field is available through the getttune() core opcode in the structured audio decoder (see ISO/IEC 14496-3:2009, Clause 5). To adjust the pitch of other decoder types, use the AudioFX node with an appropriate effects orchestra.

The speed field controls the playback speed for the structured audio decoder (see ISO/IEC 14496-3:2009, Clause 5), the parametric speech (HVXC) and the parametric audio (HILN) decoder. It is specified as a ratio, where 1 indicates the original speed; values other than 1 indicate multiplicative time-scaling by the given ratio (i.e. 0,5 specifies twice as fast). The value of this field shall be made available to the structured audio decoder indicated by the url field. ISO/IEC 14496-3:2009, 5.7.3.3.6, list item 8 describes the use of this field to control the structured audio decoder. To adjust the speed of other decoder types, use the AudioFX node with an appropriate effects orchestra (see ISO/IEC 14496-3:2009, 5.9.14.4).

The startTime and stopTime exposedFields and their effects on the AudioSource node are described in ISO/IEC 14496-11:2015, 7.1.1.1.6.2. The numChan field describes how many channels of audio are in the decoded bitstream.

5.1.1.1.2 Sound

5.1.1.1.2.1 XSD description

```
<complexType name="SoundType">
 <all>
    <element ref="xmta:IS" minOccurs="0"/>
    <element name="source" form="qualified" minOccurs="0">
        <complexType>
          <group ref="xmta:SFAudioNodeType" minOccurs="0"/>
        </complexType>
    </element>
 </all>
 <attribute name="direction" type="xmta:SFVec3f" use="optional" default="0 0 1"/>
 <attribute name="intensity" type="xmta:SFFloat" use="optional" default="1"/>
 <attribute name="location" type="xmta:SFVec3f" use="optional" default="0 0 0"/>
 <attribute name="maxBack" type="xmta:SFFloat" use="optional" default="10"/>
 <attribute name="maxFront" type="xmta:SFFloat" use="optional" default="10"/>
 <attribute name="minBack" type="xmta:SFFloat" use="optional" default="1"/>
```



```
    <attribute name="minFront" type="xmta:SFFloat" use="optional" default="1"/>
    <attribute name="priority" type="xmta:SFFloat" use="optional" default="0"/>
    <attribute name="spatialize" type="xmta:SFBool" use="optional" default="true"/>
    <attributeGroup ref="xmta:DefUseGroup"/>
  </complexType>
  <element name="Sound" type="xmta:SoundType"/>
```

5.1.1.1.2.2 Functionality and semantics

As defined in ISO/IEC 14496-11:2015, 7.2.2.116.

The Sound node is used to attach sound to a scene, thereby giving it spatial qualities and relating it to the visual content of the scene. The Sound node relates an audio BIFS sub-graph to the rest of an audio-visual scene. By using this node, sound may be attached to a group, and spatialized or moved around as appropriate for the spatial transforms above the node. By using the functionality of the audio BIFS nodes, sounds in an audio scene described using ISO/IEC 14496-11 may be filtered and mixed before being spatially composited into the scene. The semantics of this node are as defined in ISO/IEC 14472-1:1997, 6.42, with the following exceptions and additions.

The source field allows the connection of an audio sub-graph containing the sound. The spatialize field determines whether the Sound shall be spatialized. If this flag is set, the sound shall be presented spatially according to the local coordinate system and current listeningPoint, so that it apparently comes from a source located at the location point, facing in the direction given by direction. The exact manner of spatialization is implementation-dependant, but implementators are encouraged to provide the maximum sophistication possible depending on terminal resources. If there are multiple channels of sound output from the child sound, they may or may not be spatialized, according to the phaseGroup properties of the child, as follows. Any individual channels, that is, channels not phase-related to other channels, are summed linearly and then spatialized. Any phase-grouped channels are not spatialized, but passed through this node unchanged. The sound presented in the scene is thus a single spatialized sound, represented by the sum of the individual channels, plus an “ambient” sound represented by mapping all the remaining channels into the presentation system as described in ISO/IEC 14496-11:2015, 7.1.1.2.13.2.2. If the spatialize field is not set, the audio channels from the child are passed through unchanged, and the sound presented in the scene due to this node is an “ambient” sound represented by mapping all the audio channels output by the child into the presentation system as described in ISO/IEC 14496-11:2015, 7.1.1.2.13.2.2.

As with the visual objects in the scene, the Sound node may be included as a child or descendant of any of the grouping or transform nodes. For each of these nodes, the sound semantics are as follows. Affine transformations presented in the grouping and transform nodes affect the apparant spatialization position of spatialized sound. They have no effect on “ambient” sounds. If a particular grouping or transform node has multiple Sound nodes as descendants, then they are combined for presentation as follows. Each of the Sound nodes may be producing a spatialized sound, a multichannel ambient sound, or both. For all of the spatialized sounds in descendant nodes, the sounds are linearly combined through simple summation from presentation. For multichannel ambient sounds, the sounds are linearly combined channel-by-channel for presentation.

5.1.1.1.3 Sound2D

5.1.1.1.3.1 XSD description

```
<complexType name="Sound2DType">
 <all>
   <element ref="xmta:IS" minOccurs="0"/>
   <element name="source" form="qualified" minOccurs="0">
     <complexType>
        <group ref="xmta:SFAudioNodeType" minOccurs="0"/>
     </complexType>
   </element>
 </all>
 <attribute name="intensity" type="xmta:SFFloat" use="optional" default="1"/>
 <attribute name="location" type="xmta:SFVec2f" use="optional" default="0 0"/>
 <attribute name="spatialize" type="xmta:SFBool" use="optional" default="true"/>
 <attributeGroup ref="xmta:DefUseGroup"/>
```