AS ISO/IEC 14496.11—2006
ISO/IEC 14966-11:2005

AS ISO/IEC 14496.11—2006

Australian Standard™

Information technology—Coding of audio-visual objects

Part 11: Scene description and application engine

This Australian Standard was prepared by Committee IT-029, Coded Representation of Picture, Audio and Multimedia/Hypermedia Information. It was approved on behalf of the Council of Standards Australia on 14 February 2006.
This Standard was published on 13 March 2006.

The following are represented on Committee IT-029:

Australian Broadcasting Authority (ABA)
Australian Broadcasting Corporation (ABC)
Australian Consumers Association
Australian Subscription Television
CSIRO Mathematical & Information Services
Department of Defence (Australia)
Free TV Australia
School of Computer Science and Mathematics
Victoria University of Technology
Special Broadcasting Service (SBS)
The University of New South Wales
University of Wollongong

This Standard was issued in draft form for comment as DR 05579.

Australian Standard™

Information technology—Coding of audio-visual objects

Part 11: Scene description and application engine

First published as AS ISO/IEC 14496.11—2006.

Published by Standards Australia GPO Box 476, Sydney, NSW 2001, Australia

ISBN 0 7337 7305 2

PREFACE

This Standard was prepared by the Standards Australia Committee IT-029, Coded Representation of Picture, Audio and Multimedia/Hypermedia Information.

This Standard is identical with, and has been reproduced from ISO/IEC 14966-11:2005, *Information technology—Coding of audio-visual objects,* Part 11: *Scene description and application engine*.

The objective of this Standard is to provide the Australian multimedia industry with scene description and application engine, also called BIFS, a binary format for two- or three-dimensional audiovisual content.

The terms 'normative' and 'informative' are used to define the application of the annex to which they apply. A normative annex is an integral part of a standard, whereas an informative annex is only for information and guidance.

As this Standard is reproduced from an international standard, the following applies:

(a) Its number appears on the cover and title page while the international standard number appears only on the cover.

(b) In the source text 'this part of ISO/IEC 14496' should read 'this Australian Standard'.

(c) A full point substitutes for a comma when referring to a decimal marker.

References to International Standards should be replaced by references to Australian Standards, as follows:

Reference to International Standard		*Australian Standards*	
ISO		AS	
3166	Codes for the representation of names of countries and their subdivisions	2632	Codes for the representation of names of countries and their subdivisions
3166-1	Part 1: Country codes	2632.1	Part 1: Country codes
ISO/IEC		AS/NZS	
13818	Information technology — Generic coding of moving pictures and associated audio information	13818	Information technology — Generic coding of moving pictures and associated audio information
13818-2	Part 2: Video	13818.2	Part 2: Video
13818-3	Part 3: Audio	13818.3	Part 3: Audio
11172	Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s	4230	Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s
11172-2	Part 2: Video	4230.2	Part 2: Video
11172-3	Part 3: Audio	4230.3	Part 3: Audio
10918	Information technology — Digital compression and coding of continuous-tone still images	4473	Information technology — Digital compression and coding of continuous-tone still images
10918-1	Part 1: Requirements and guidelines	4473.1	Part 1: Requirements and guidelines

Only International references that had been adopted as Australian or Australian/New Zealand Standards have been listed.

CONTENTS

Page

Page

INTRODUCTION

0.1 Scene Description

0.1.1 Overview

ISO/IEC 14496 addresses the coding of audio-visual objects of various types: natural video and audio objects as well as textures, text, 2- and 3-dimensional graphics, and also synthetic music and sound effects. To reconstruct a multimedia scene at the terminal, it is hence not sufficient to transmit the raw audio-visual data to a receiving terminal. Additional information is needed in order to combine this audio-visual data at the terminal and construct and present to the end user a meaningful multimedia scene. This information, called scene description, determines the placement of audio-visual objects in space and time and is transmitted together with the coded objects as illustrated in Figure 1. Note that the scene description only describes the structure of the scene. The action of assembling these objects in the same representation space is called composition. The action of transforming these audio-visual objects from a common representation space to a specific presentation device (i.e. speakers and a viewing window) is called rendering.

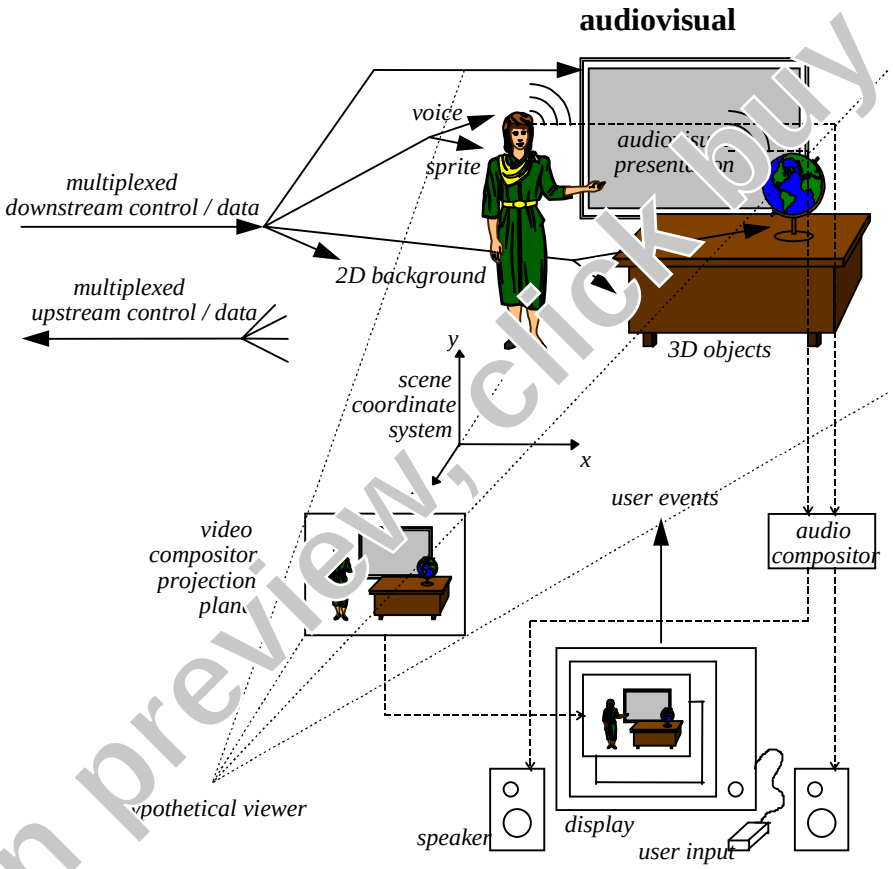

Figure 1 — An example of an object-based multimedia scene

Independent coding of different objects may achieve higher compression, and also brings the ability to manipulate content at the terminal. The behaviors of objects and their response to user inputs can thus also be represented in the scene description.

The scene description framework used in this part of ISO/IEC 14496 is based largely on ISO/IEC 14772-1:1998 (Virtual Reality Modeling Language – VRML).

0.1.2 Composition and Rendering

ISO/IEC 14496-11 defines the syntax and semantics of bitstreams that describe the spatio-temporal relationships of audio-visual objects. For visual data, particular composition algorithms are not mandated since they are implementation-

dependent; for audio data, subclause 7.1.1.2.13 and the semantics of the AudioBIFS nodes normatively define the composition process. The manner in which the composed scene is presented to the user is not specified for audio or visual data. The scene description representation is termed “BInary Format for Scenes” (BIFS).

0.1.3 Scene Description

In order to facilitate the development of authoring, editing and interaction tools, scene descriptions are coded independently from the audio-visual media that form part of the scene. This permits modification of the scene without having to decode or process in any way the audio-visual media. The following clauses detail the scene description capabilities that are provided by ISO/IEC 14496-11.

0.1.3.1 Grouping of audio-visual objects

A scene description follows a hierarchical structure that can be represented as a graph. Nodes of the graph form audio-visual objects, as illustrated in Figure 2. The structure is not necessarily static; nodes may be added, deleted or be modified.

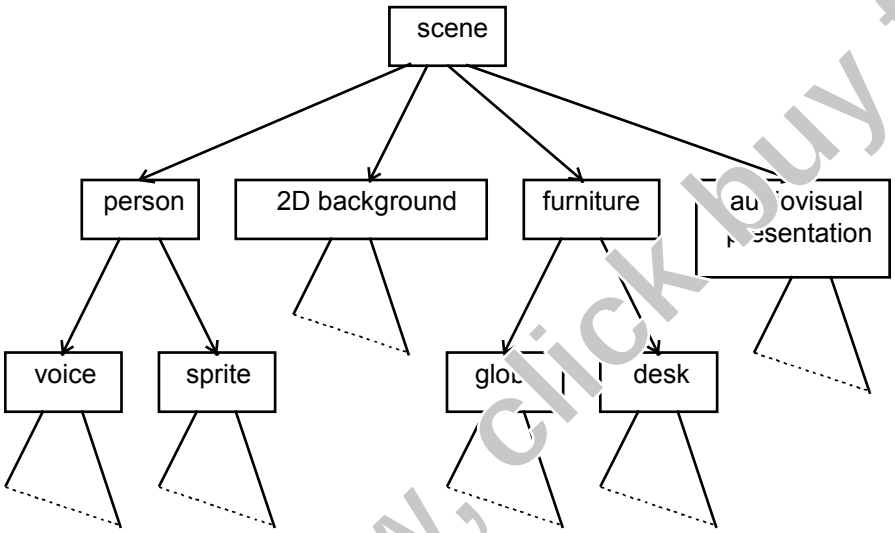

Figure 2 — Logical structure of example scene

0.1.3.2 Spatio-Temporal positioning of objects

Audio-visual objects have both a spatial and a temporal extent. Complex audio-visual objects are constructed by combining appropriate scene description nodes to build up the scene graph. Audio-visual objects may be located in 2D or 3D space. Each audio-visual object has a local co-ordinate system. A local co-ordinate system is one in which the audio-visual object has a pre-defined (but possibly varying) spatio-temporal location and scale (size and orientation). Audio-visual objects are positioned in the scene by specifying a co-ordinate transformation from the object’s local co-ordinate system into another co-ordinate system defined by a parent node in the scene graph.

0.1.3.3 Attributes of audio-visual objects

Scene description nodes expose a set of parameters through which aspects of their appearance and behavior can be controlled.

EXAMPLE – The volume of a sound; the color of a synthetic visual object; the source of a streaming video.

0.1.4 Behavior of audio-visual objects

ISO/IEC 14496-11 provides tools for enabling dynamic scene behavior and user interaction with the presented content. User interaction can be separated into two major categories: client-side and server-side. Client-side interaction is an integral part of the scene description described herein. Server-side interaction is not dealt with.

Client-side interaction involves content manipulation that is handled locally at the end-user’s terminal. It consists of the modification of attributes of scene objects according to specified user actions.

EXAMPLE — A user can click on a scene to start an animation or video sequence. The facilities for describing such interactive behavior are part of the scene description, thus ensuring the same behavior in all terminals conforming to ISO/IEC 14496-11.

0.2 Extensible MPEG-4 Textual Format

0.2.1 Overview

The Extensible MPEG-4 Textual format (XMT) is a framework (illustrated in Figure 3) for representing MPEG-4 scene description using a textual syntax. The XMT allows the content authors to exchange their content with other authors, tools or service providers, and facilitates interoperability with both the Extensible 3D (X3D) being developed by the Web3D and the Synchronized Multimedia Integration Language (SMIL) from the W3C.

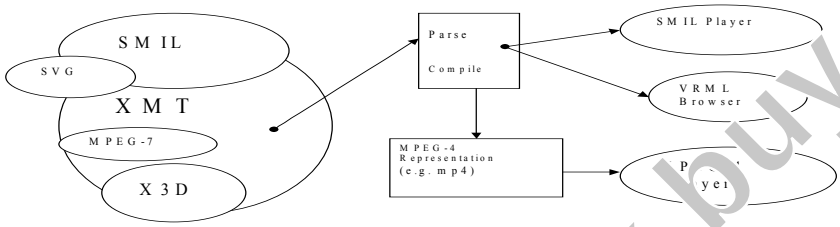

Figure 3 — Overview of the XMT Framework

0.2.2 Interoperability of XMT

The XMT format can be interchangeable between SMIL players, VRML players, and MPEG-4 players. The format can be parsed and played directly by a W3C SMIL player, preprocessed to Web3D X3D and played back by a VRML player, or compiled to an MPEG-4 representation such as mp4, which can then be played by an MPEG-4 player. See below for a graphical description of interoperability of the XMT.

0.2.3 Two-tier Architecture: XMT-A and XMT-Ω Formats

The XMT framework consists of two levels of textual syntax and semantics: the XMT-A format and the XMT-Ω format, which we will abbreviate by A and Ω, respectively, and use them interchangeably where there is no confusion.

The XMT-A is an XML-based version of MPEG-4 content, which contains a subset of the X3D. Also contained in XMT-A is an MPEG-4 extension to the X3D to represent MPEG-4 specific features. The XMT-A provides a straightforward, one-to-one mapping between the textual and binary formats.

The XMT-Ω is a high level abstraction of MPEG-4 features designed based on the W3C SMIL. The XMT provides a default mapping from Ω to A, for there is no deterministic mapping between the two, and it also provides content authors with an escape mechanism from Ω to A.

In addition, an XMT-C (Common) section contains the definition of elements and attributes that may be used within either XMT-A or XMT-Ω.

0.3 MPEG-J

0.3.1 Overview

MPEG-J is a flexible programmatic control system that represents an audio-visual session in a manner that allows the session to adapt to the operating characteristics when presented at the terminal. Two important characteristics are supported: first, the capability to allow graceful degradation under limited or time varying resources, and second, the ability to respond to user interaction and provide enhanced multimedia functionality.

More specifically, 9.7 normatively defines:

The format and delivery of Java byte code by specifying the MPEG-J stream format and the delivery mechanism of such a stream (Java byte code and associated data);

The MPEG-J Session and the MPEG-J application lifecycle; and

The interactions and behavior of byte code through the specification of Java APIs.

0.3.2 Organization MPEG-J specification

10.1 gives an overall architecture of the MPEG-J system. MPEG-J Session start-up is walked through in 10.2. The Delivery of MPEG-J data to the terminal is specified in 10.3. 10.4 specifies the different categories of APIs that a program in the form of Java bytecode would use. 10.5 is an informative annex on starting the Java Virtual Machine. The electronic annex attached to this document lists the normative MPEG-J APIs in the HTML format. 10.6 illustrates the usage of MPEG-J APIs through a few examples.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of patents.

The ISO and IEC take no position concerning the evidence, validity and scope of these patent rights.

The holder of these patent rights have assured the ISO and IEC that they are willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent rights is registered with the ISO and IEC. Information may be obtained from the companies listed in Annex D.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified in Annex D. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

AUSTRALIAN STANDARD

Information technology — Coding of audio-visual objects —

Part 11:
Scene description and application engine

1 Scope

This part of ISO/IEC 14496 specifies:

1. the coded representation of the spatio-temporal positioning of audio-visual objects as well as their behavior in response to interaction (scene description);
2. the Extensible MPEG-4 Textual (XMT) format, a textual representation of the multimedia content described in ISO/IEC 14496 using the Extensible Markup Language (XML); and
3. a system level description of an application engine (format, delivery, lifecycle, and behavior of downloadable Java byte code applications).

2 Normative references

The following referenced documents are indispensable for the application 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 639-2:1998*, Codes for the representation of names of languages — Part 2: Alpha-3 code*

ISO 3166-1:1997, *Codes for the representation of names of countries and their subdivisions — Part 1: Country codes*

ISO 9613-1:1993, *Acoustics — Attenuation of sound during propagation outdoors — Part 1: Calculation of the absorption of sound by the atmosphere*

ISO/IEC 11172-2:1993, *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 2: Video*

ISO/IEC 11172-3:1993, *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 3: Audio*

ISO/IEC 13818-3:1998, *Information technology — Generic coding of moving pictures and associated audio information — Part 3: Audio*

ISO/IEC 13818-7: 2004, *Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC)*

ISO/IEC 14496-2:2004, *Information technology — Coding of audio-visual objects — Part 2: Visual*

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 14772-1:1997/Amd.1:2003, *Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding — Amendment 1: Enhanced interoperability*

ISO/IEC 16262:2002, *Information technology — ECMAScript language specification*

ISO/IEC 13818-2:2000, *Information technology — Generic coding of moving pictures and associated audio information — Part 2: Video*

ISO/IEC 10918-1:1994, *Information technology — Digital compression and coding of continuous-tone still images: Requirements and guidelines*