International
Standard

ISO 24617-10

Language resource management — Semantic annotation framework (SemAF) —

Part 10: **Visual information**

Gestion des ressources linguistiques - Cadre d'annotation sémantique —

Partie 10: informations visuelles (VoxML)

First edition
2024-08

Reference number
ISO 24617-10:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 document 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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

For an explanation of 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 www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology,* Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24617 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document standardizes the specification of a semantic annotation scheme for visual information, based on a modelling language for constructing three-dimensional (3D) visualizations of concepts denoted by natural language (NL) expressions. This modelling language serves as a semantic basis of interpreting the semantic forms of annotation structures model-theoretically by constraining the models for interpretation. This document focuses on the introduction of the modelling language as a semantic basis for interpretation, since the syntactic specification of the annotation scheme for visual information is a simplified formulation based on the abstract specification of the spatio-temporal annotation schemes, such as those specified in ISO 24617-1, ISO 24617-7 and ISO 24617-14. These three standards lay a theoretical basis for this document, which specifies ways of annotating visual information involving motions and actions that are spatio-temporally characterized.

The modelling language, named "VoxML" (visual object concept structure modelling language), where "Vox" abbreviates "visual object concept structure" (VOCS), can be used as the platform for creating multimodal semantic simulations in the context of human-computer communication. VoxML encodes semantic knowledge of real-world objects represented as 3D models, and of events and attributes related to and enacted over these objects. VoxML is intended to overcome the limitations of existing 3D visual markup languages by allowing for the encoding of a broad range of semantic knowledge that can be exploited by a variety of systems and platforms, leading to multimodal simulations of real-world scenarios using conceptual objects that represent their semantic values.

NOTE 1 The main content of this document is based on References [1] and [2]. Reference [1] was developed by the Brandeis University Computer Science Department in the context of communicating with computers (CwC), a Defence Advanced Research Projects Agency (DARPA) effort to identify and construct computational semantic elements, for the purpose of carrying out joint plans between a human and computer through NL discourse.

NOTE 2 This document adopts VoxML as a semantic basis for enriching the model for interpreting the descriptions of objects, actions and relations involving dynamic visual information.

This document outlines a specification:

a) to formulate the annotation scheme for visual information;

b) to represent semantic knowledge of real-world objects represented as 3D models.

It uses a combination of parameters that can be determined from the object's geometrical properties as well as lexical information from NL, with methods of correlating the two where applicable. This information allows for visualization and simulation software to fill in information missing from the NL input and allows the software to render a functional visualization of programs being run over objects in a robust and extensible way. Currently, a voxicon, which is the structured repository of visual object concepts, contains 500 object (noun) voxemes, lexemes or entries of the voxicon, and 10 program (verb) voxemes.

NOTE 3 As this library of available voxemes continues to grow, the specification elements will operationalize an increasingly large library of various and more complicated programs. A voxeme library and visualization software where users will be able to conduct visualizations of available behaviours driven by VoxML after parsing and interpretation is available from Reference [25].

Language resource management — Semantic annotation framework (SemAF) —

Part 10: Visual information

1 Scope

This document specifies an annotation language for visual information, based on VoxML (visual object concept structure modelling language), a modelling language for the visualizations of concepts and actions denoted by natural language (NL) expressions in three dimensions (3D).

The specification of the VoxML-based annotation scheme conforms to the requirements given in ISO 24617-1, ISO 24617-7 and ISO 24617-14. The adoption of VoxML, specified in ISO 24617-14 as a semantic basis, is necessary for the 3D simulation and visualization of actions and motions taken by both human and artificial agents in real-life situations.

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 24610-1:2006, *Language resource management — Feature structures — Part 1: Feature structure representation*

ISO 24617-1, *Language resource management — Semantic annotation framework (SemAF) — Part 1: Time and events (SemAF-Time, ISO-TimeML)*

ISO 24617-7, *Language resource management — Semantic annotation framework — Part 7: Spatial information*

ISO 24617-14, *Language resource management — Semantic annotation framework (SemAF) — Part 14: Spatial semantics*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

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

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

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

3.1
affordance
affordance structure
set of specific actions, described along with the requisite conditions, that the object may take part in

3.1.1
Gibsonian affordance
GA
set of specific actions that an agent can perform with an object that is presented to the agent

EXAMPLE Hold, grasp, move.

3.1.2
telic affordance
set of goal-oriented or intentionally situated actions of an agent on an object presented to the agent

EXAMPLE An agent *eating* an apple when it is presented to the agent.

3.2
habitat
representation of an object situated within a partial minimal model

3.3
minimal embedding space
MES
three-dimensional (3D) region within which the state is configured, or the event unfolds

3.4
qualia
qualia structure
QS
relational forces or aspects of a lexical item or concept

3.5
telic
purpose or function *qualia* (3.4) of an object

3.6
voxeme
basic entries in *voxicon* (3.7)

3.7
voxicon
lexicon or list of basic visual object concepts of VoxML (visual object concept structure modelling language)

4 Abbreviated terms

3D	three dimensional
A	agentive role
ARG	argument
AS	atomic structure
$\mathrm{AS}_{\mathrm{visML}}$	annotation scheme for visual information markup language
$\mathrm{ASyn}_{\mathrm{visML}}$	abstract syntax for visual information markup language
$\mathrm{CSyn}_{\mathrm{visML}}$	concrete syntax for visual information markup language
C	constitutive property
F	formal property
GA	Gibsonian affordance

ID	identifier
MES	minimal embedding space
NL	natural language
NLP	natural language processing
QS	qualia structure
T	telic role
Vox	visual object concept structure
VoxML	visual object concept structures modelling language
XML	extensible markup language

5 Basic semantic assumptions — Habitats and affordances

Before introducing the VoxML specification, this document reviews two basic assumptions regarding the semantics underlying the model. Following the Generative Lexicon,[3] lexical entries in the object language are given a feature structure consisting of a word's basic type, its parameter listing, its event typing and its qualia structure. In accordance with ISO 24610-1:2006, each feature structure shall be typed, consisting of pairs of features (attributes) and values, either atomic or complex. If a value is a variable, then it is bound either universally, existentially, or by the lambda operator, as shown in Example 1.

The semantic structure of an object shall be analysed into the following four sub-structures:

a) atomic structure (formal): objects expressed as basic nominal types;

b) subatomic structure (constitutive): mereo-topological structure of objects;

c) event structure (telic) and (agentive): origin and functions associated with an object;

d) macro-object structure: how objects fit together in space and through coordinated activities.

Objects can be partially contextualized through their qualia structure. For example, a food item has an atelic value of “eat”; an instrument for writing has a telic value of “write”; a cup has a telic value of “hold”, etc. As a further example, the lexical semantics for the noun “chair” carries a telic value of “sit_in”:

EXAMPLE 1

$$\lambda x \exists y \begin{bmatrix} \textbf{chair} \\ \text{AS} = [\text{ARG1} = x{:}e] \\ \text{QS} = \begin{bmatrix} \text{F} = \mathit{phys(x)} \\ \text{T} = \lambda z,e[\mathit{sit\_in}(e, z, x)] \end{bmatrix} \end{bmatrix}$$

where

AS	is an atomic structure;
QS	is a qualia structure;
ARG1	is argument 1;
F	is a formal property;
T	is a telic role.

While an artefact is designed for a specific purpose (its telic role), this can only be achieved under specific circumstances. Reference [4] introduces the notion of an object's “habitat”, which encodes these circumstances. References [5] and [6] further define the notion of habitat and how it interacts with affordances. It is assumed that for an artefact, x, given the appropriate context C, performing the action π will result in the intended or desired resulting state, R, i.e. $C \rightarrow [\pi]R$. That is, if a context C (a set of contextual factors) is satisfied, then every time the activity of π is performed, the resulting state *R* will occur. It is necessary to specify the precondition context *C* since this enables the local modality to be satisfied.

Using this notion, a habit is defined as representing an object situated within a partial minimal model; it is a directed enhancement of the qualia structure. Multi-dimensional affordances determine how habitats are deployed and how they modify or augment the context, and compositional operations include procedural (simulation) and operational (selection, specification, refinement) knowledge.

The habitat for an object is built by first placing it within an embedding space and then contextualizing it. For example, to use a table, the top must be oriented upward, the surface must be accessible, etc. A chair also must be oriented up, the seat must be free and accessible, it must be able to support the user, etc. An illustration of how the resulting knowledge structure for the habitat of a chair is shown in Example 2.

EXAMPLE 2

$$\lambda x \left[\begin{array}{l} \textbf{chair}\,\textit{hab} \\ \mathrm{F} = [phys(x), on(x, y_1), in(x, y_2), orient(x,up)] \\ \mathrm{C} = [sit(x_1), back(x_2), leg(x_3), clear(x_1)] \\ \mathrm{T} = [\lambda z \lambda e[C \rightarrow [sit(e, z, x)]\ \boldsymbol{R} sit(x)] \\ \mathrm{A} = [made(e', w, x)] \end{array}\right]$$

where

F is a formal property;

C is a constitutive property;

T is a telic role;

A is an agentive role.

As described in more detail in 6.4, event or action simulations are constructed from the composition of object habitats, along with some constraints imposed by the dynamic event structure inherent in the verb itself, when interpreted as a program.

The final step in contextualizing the semantics of an object is to operationalize the telic value in its habitat. This effectively means identifying the “affordance structure” for the object.[7][8] The affordance structure available to an agent, when presented with an object, is the set of actions that can be performed with it. These are referred to as “Gibsonian affordances” and they include “grasp”, “move”, “hold”, “turn”, etc. This is to distinguish them from more goal-directed, intentionally situated activities, referred to as “telic affordances”.

6 VoxML specification

6.1 Metamodel and VoxML elements

The spatio-temporal annotation schemes given in ISO 24617-1, ISO 24617-7 and ISO 24617-14 shall apply.

The metamodel, graphically depicted by Figure 1, represents a small world of basic elements modelled in VoxML. These elements form a set of categories:

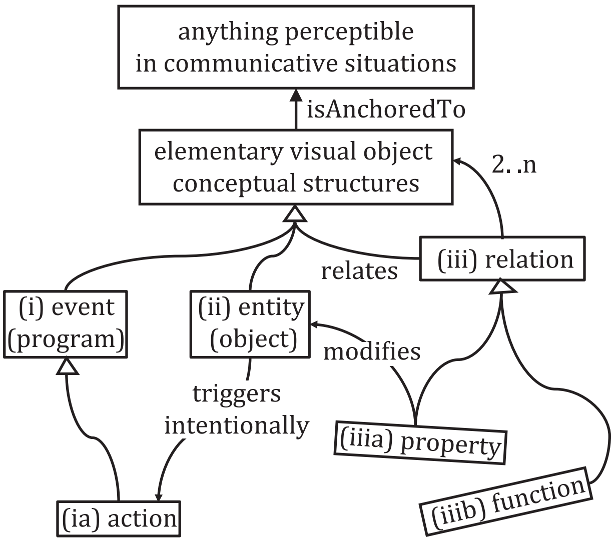
a) event (program);

b) entity (object);

c) relation over them.

Events, especially actions, work as programs while taking simple objects or spatio-temporally localized objects as arguments. Entities as objects are individuals or groups that may behave as agents. Relations can be divided into properties, often referred to as “attributes”, and functions as subcategories. Attributes and relations evaluate to states, and functions evaluate to geometric regions. These elements can then compose into visualizations ns of NL concepts and expressions.

The metamodel of VoxML, presented in Figure 1, has no regions or times. These are introduced by functions such as *loc* and *τ.* The function *loc,* for instance, maps an object *x to* the region *loc(x)* to which it is anchored. Likewise, *τ(x)* maps an event to an event time, the time of its occurrence. Similarly, the function *seq* or the function *vec* maps a set of regions to a path or a vector. Thereby, the ontology of VoxML is enriched with spatio-temporal entities and dynamic paths.



NOTE 1 The empty triangular head of an arrow represents a subcategorization relation. Each directed arrow with a smaller filled-in arrowhead relates one element to one or other more elements while its labelling specifies such a relation. An entity as an agent, for example, triggers intentionally an action, while the action is a subcategory of an event, treated as a program.

NOTE 2 SOURCE: Reference [2], reproduced with the permission of the authors.

Figure 1 — Metamodel

6.2 Representation of VoxML structures

This document follows the convention of the current version of VoxML and Voxicon (see Reference [1]). Basic VoxML structures called “voxemes” are conventionally represented as feature structures, each consisting of a set of attribute-value specifications, conforming to ISO 24610-1. Voxemes are mostly formed by complex feature structures, having at least one of their substructures embedded in them as a feature structure, as illustrated in this clause.

NOTE 1 ISO 24610-1 avoids the use of the term “attribute-value”. Instead, it uses the term “feature-value”, thus defining a feature structure as a function from a set of features to a set of values.

In the concrete syntax, adopted for representing these feature structures of VoxML in this document, the names of its attributes are represented in all uppercase characters, while the names of elements start with their first character in upper case (e.g. the attribute LEX for the element Object as in Figure 2).

NOTE 2 This document follows the convention of the current version of VoxML and Voxicon for representing attribute names in upper case characters.