DRAFT INTERNATIONAL STANDARD
ISO/DIS 24617-11

ISO/TC **37**/SC **4** Secretariat: **KATS**

Voting begins on: **2020-03-16** Voting terminates on: **2020-06-08**

Language resource management — Semantic annotation framework (SemAF) —

Part 11: **Measurable Quantitative information (MQI)**

Gestion des ressources linguistiques — Cadre d'annotation sémantique —

Partie 11: Mesurer l'information quantitative (MQI)

ICS: 01.020

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 24617-11:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
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 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 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: **Error! Hyperlink reference not valid.**.

The committee responsible for this document is ISO/TC 37, *Language and Terminology,* Subcommittee SC 4, *Language resource management*

ISO 24617 consists of the following parts under the general title Language resource management — Semantic annotation framework (SemAF):

— Part 1: Time and events (TimeML)

— Part 2: Dialogue acts (DA)

— Part 3: Named entity

— Part 4: Semantic roles (SR)

— Part 5: Discourse structures (DS)

— Part 6: Principles of semantic annotation (SemAF Principles)

— Part 7: Spatial information

— Part 8: Semantic relations in discourse, core annotation schema (DR-core)

— Part 9: Reference annotation framework (RAF)

— Part 10: Visual information (VoxML)

— Part 11: Measurable quantitative information (MQI)

— Part 12: Quantification

— Part 13: Gestures

Introduction

Measurable quantitative information (MQI) such as '165 cm' or '60 kg' of 'John' that applies to the height or weight of the person is very common in ordinary language. MQI describes one of basic properties which is associated with the magnitude aspect of quantity. Such information is much more abundant in scientific publications or technical reports to the extent that it constitutes an essential part of communicative segments of language in general. The processing of such information is thus required for any successful language resource management.

This document, named 'SemAF-MQI', thus aims to focus on specifying a general annotation scheme with following the principles of semantic annotation laid down in ISO 24617-6 in general and the basic requirements of ISO 24611 Linguistic annotation framework (LAF), that facilitates the processing of MQI in scientific and technical language and to make it interoperable with other semantic annotation schemes, such as ISO 24617 etc.

NOTE 1 ISO 24617-1:2012 (E) TimeML and ISO 24617-7: 2014 (E) Spatial information, for instance, have proposed a way of annotating measures on time (durations or time amounts) and space (distances), respectively. The serious disucssion of annotating measures as part of ISO 24617 was initiated at the 11th joint ACL-ISO/TC 37/SC 4/WG 2 Workshop on Interoperable Semantic Annotation (ISA-11)[1] and was continued at the ISA-13[2], ISA-14[3], and ISA-15[4] workshops. ISO 24612: 2012 (E) LAF provides a pivotal form (GrAF, graphic annotation framework) that makes all the annotation of temporal or spatial measures in these two annotation schemes interchangeable with those measure annotations in the new document SemAF-MQI.

Focusing on measurements in scientifico-technological language, SemAF-MQI as an ISO standard is expected to contribute to information extraction (IR)[5], question answering (QA), text summarization (TS), and other natural language processing (NLP) applications[6].

NOTE 2 To enhance the readability of this document and to correct some obvisous editorial errors, some editorial changes were made on the earlier version of CD 24617-11 MQI that had been submitted to the successful CD ballot (2019-09-11 ~ 2019-11-06) with a 100% approval but with no comments.

- Each item in Bibliography as well as in Clause 2 Normative references was made to be referred to in the main part of the current version of the docment.
- Three of the illustrative examples in clause 7.6 Illustrations of QML_csx were moved to a newly created Annex A (informative) without any change of content change in order to lighten the burden of reading that clause 7.6.
- Incorrect wordings or obvious typos were corrected.
- The white and black coloing of Figure 1 — Metamodel of QML was changed to the multiple coloring to bring out each of the different components of the metamodel.

Language resource management — Semantic annotation framework (SemAF) —

Part 11: Measurable Quantitative information (MQI)

1 Scope

As one of the basic physical properties, quantity is associated with multitude (how many) and magnitude (how much). Focusing on the magnitudinal aspect of quantity, this document, which is named "SemAF-MQI" henceforth, aims at formulating a specification language for the construction of an annotation scheme for measurable quantitative information (MQI) in scientifico-technological language. The main characteristics of SemAF-MQI is that quantitative information is presented as measures expressed in terms of a pair <*n*, *u*>, consisting of a numerically expressed quantity *n* and a unit *u*, which is either basic or derived, or either normalized or conventionally used.

NOTE 1 MQI stands for "measurable quantitative information", whereas SemAF-MQI refers to the part 11 of ISO 24617-11. [See 3.4 for the definition of MQI.]

The scope of SemAF-MQI is restricted to the measurable or magnitudinal aspect of quantity so that it can focus on the technical or practical use of measurements in IR (information retrieval), QA (question answering), TS (text summarization), and other NLP (natural language processing) applications. The scope is restricted to the domains of technology that carry more applicational relevance than some theoretical issues found in the ordinary use of language. The subsequent part of ISO 24617 (Part 12) deals with more general and theoretical issues of quantification and quantitative information.

NOTE 2 The scope of this document is intentionally restricted to the measurable or magnitudinal aspect of quantity so that SemAF-MQI focuses on the technical or practical use of measurements in IR, QA, TS, and other NLP applications. The scope is restricted to domains of technology that carry more applicational relevance than theoretical issues found in the ordinary use of language. Fruit as well as meat is, for instance, sold at markets in terms of weight but not of pieces. Furthermore, the subsequent part of ISO 24617 (Part 12) deals with more general and theoretical issues of quantification and plurals (e.g., "three apples) including quantitative information that includes multitudinal aspects.

The scope of SemAF-MQI also treats temporal durations that are discussed in Part 1 of ISO 24617 SemAF-Time (ISO-TimeML) and spatial measures such as distances that are treated in Part 7 of ISO 24617 Spatial information (ISO-Space), while making them interoperable with other measure types. It also accommodates the treatment of measures or amounts that are introduced in ISO 24617-6 SemAF Principles (Clause 8.3).

NOTE 3 The scope of this document (Part 11) also treats temporal durations that are discussed in Part 1 of ISO 24617 SemAF-Time (TimeML) and spatial measures such as distances that are treated in Part 7 of ISO 24617 Spatial information, while making them interoperable with other measure types. It also accommodates the treatment of measures or amounts that are introduced in ISO 24617-6 SemAF Principles. Its scope thus covers temporal durations treated in XSchema and the TEI Guidelines.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 24612:2012, *Language resource management — Linguistic annotation framework (LAF)*

ISO 24617-1:2012, *Language resource management — Semantic annotation framework (SemAF) — Part 1: Time and events (SemAF-Time, ISO-TimeML)*

ISO 24617-6:2016, *Language resource management — Semantic annotation framework — Part 6: Principles of semantic annotation (SemAF Principles)*

ISO 24617-7:2014, *Language resource management — Semantic annotation framework — Part 7: Spatial information (ISOspace)*

ISO/IEC 14977:1996, *Information technology - Syntactic metalanguage - Extended BNF*

ISO 80000-1:2009, *Quantities and units — Part 1: General*

NOTE 1 The following two documents are de-facto standards to be followed by SemAF-MQI:

TEI P5: Guidelines for Electronic Text Encoding and Interchange, The TEI Consortium, 2019[7].

XML Schema, Part 2: Datatypes, 2nd edition, W3C Recommendation, 28 October 2004[8].

3 Terms and definitions

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

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

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

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

3.1
quantity
property of a measureable object referring to its magnitude (how much) or multitude (how many).

Note 1 to entry: Compare with ISO 80000-1:2009, 3 Terms and Definitions, 3.1: property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed by means of a number and a reference.

3.2
base quantity
quantity in a conventionally chosen subset of a given system of quantities, where no quantity in the subset can be expressed in terms of the other quantities within that subset

Note 1 to entry: Kinds of quantities include seven base quantities defined by the International System of Quantities (ISQ), as listed in Table 1

Table 1 — ISQ base quantities

base quantities	base quantity symbols
length	L
mass	M
time	T
electric current	I
thermodynamic temperature	Θ
amount of substance	N
luminous intensity	J

Note 2 to entry: In ISO 80000-1:2009, 3 Terms and Definition, the symbols such as L and M, which are called base quantity symbols in this document, are called as dimension symbols of quantity

3.3
derived quantity
quantity, in a system of quantities, defined in terms of the base quantities of that system

EXAMPLE Speed is a derived quantity defined by length (distance) over time (LT^{-1}), where length (L) and time (T) are base quantities.

[SOURCE: ISO 80000-1:2009, 3 Terms and Definition, 3.5 derived quantity]

3.4
quantitative information
QI
measure associated with the quantity (3.1) of a measurable object

3.5
measurable quantitative information
MQI
quantitative information (3.3) that can be expressed in unitized numeric terms

3.6
measurable quantitative information markup language
markup language of measurable quantitative information
QML
specification language for the annotation of measurable quantitative information (3.5) extractable from text or other medium types of language

3.7
unit
unit of measurement
measurement unit
scalar basis, defined and adopted by convention, of measuring objects by multiplying their quantitative values expressed in real numbers

Note 1 to entry: The expressions that are used in measurement such as "meter", "liter", and "μmol/kg" are units by the definition given above. The multitude expressions such as "bottles", "boxes", or "two" as in "two bottles of milk", "a box of apples", and "two coffees" sometimes fail to be regarded as units, but they can also be if they are accepted as units by convention or agreement in some communities. ISO 24617 SemAF Part 12: Quantification treats such multitude expressions as genuine units.

Note 2 to entry: There are two major types of units, base and derived

[Refer to ISO 80000-1:2009, 3 Terms and Definitions, 3.9 Unit, 3.10 Base unit, and 3.11 Derived unit.]

[SOURCE: Refer to: ISO 80000-1:2009, 3 Terms and Definitions, 3.9, real scalar quantity, defined and adopted by convention, with which any other quantity of the same kind can be compared to express the ratio of the second quantity to the first one as a number.]

3.8
base unit
measurement unit that is adopted by convention for a base quantity (3.2)

Note 1 to entry: There are seven base units chosen by the International System of Units (SI) associated with seven ISQ base quantities to measure quantities, as shown in Table 2.

Table 2 — base units

SI base unit (unit symbol)	**Associated ISQ base quantity** (base quantity dimension symbol)
meter (m)	length (L)
kilogram (kg)	mass (M)

Table 2 *(continued)*

SI base unit (unit symbol)	**Associated ISQ base quantity** (base quantity dimension symbol)
second (s)	time (T)
ampere (A)	electric current (I)
kelvin (K)	thermodynamic temperature (Θ)
mole (mol)	amount of substance (N)
candela (cd)	luminous intensity (J)

[SOURCE: ISO 80000-1:2009, 3 Terms and Definitions, 3.9 Unit, 3.10 Base unit, and 3.11 Derived unit.]

3.9
derived unit
measurement unit for a derived quantity

EXAMPLE The unit “newton” (N) is a derived unit for a derived quantity “force” (F), which is defined to be “mass times acceleration” (MLT^{-2}), where the quantity “acceleration” is a derived quantity defined by “velocity divided by time” (VT^{-1}) and “velocity” defined by “length (distance) divided by time” (LT^{-1}).

Note 1 to entry: Table 3 illustrates some of the derived units.

[Refer to ISO 80000-1:2009, 3 Terms and Definitions, 3.9 Unit, 3.10 Base unit, and 3.11 Derived unit.]

Table 3 — derived units

Derived unit (unit symbol)	**Associated derived quantity**
kilo-meter per minute(km/min)	speed= length(L)/ time(T)
gram per cubic meter (gram/m^3)	density=mass(M)/volume(L^3)
kilo- gram, meter per square second (kg x m/s^2)	force = mass (M) x length(L)/time(T^2)
lumen per square meter (lm/m^2)	Illuminance = luminous intensity (J)/ area(M^2)

4 Background and Motivations

Quantity exists as a multitude (e.g., “two watermelons”) or magnitude (“one kilogram of watermelon”). The two basic divisions of quantity imply the principal distinction between continuity (continuum) and discontinuity, which are two ways of determining quantity. SemAF-MQI only focuses on the measurement information in scientific and technical texts. Therefore, quantity is regarded as a magnitude property in the document, which is consistent with ISO 80000 - 1:2009 Quantities and units. As in ISO 80000-1:2009, the term “unit” is defined in relation to quantity and is used for real scalar quantity, defined and adopted by convention, with which any other quantity of the same kind can be compared to express the ratio of the second quantity to the first one as a number. There are two types of units: base unit and derived unit.

This document treats complex derived units as unanalyzed wholes. It does not annotate their internal structures and components, unless it is required by some special use cases. Neither does the standard require to specify ways of converting one unit to another. Here are some reasons:

1) Complex derived units such as speed “km/h” (LT-1) or acceleration “m/s2” (LT-2) are understood as they are in ordinary situations.

2) Certain domain specific units cannot be decomposed during their conversion to other equivalent units. For example, Estimated Glomerular Filtration Rate (eGFR) frequently uses the unit “mL/min/1.73m^2” in a medical domain. Thus, a kidney function can be classified into various stages

depending on eGFR, where the stage 1 defines "normal eGFR greater than or equal to 90 mL/min/1.73m^2". In some cases, the unit can be written as "mL/min/((173/100).m^2)". In all these cases, "1.43" or "173/100" in the units cannot be annotated separately for automatic conversion since they are combined with other parts together to be a complete unit.

3) Units can be converted automatically in an effective way such as with the use of a conversion table. For example, by using directly "1 mmol/l" that equals to "18 mg/dl", the computer can more effectively convert the unit into another with one single computation rather than convert each part of unit and then compute the total value.

4) Incomplete units exist. During language processing, there are incomplete units which need to be detected by using different methods such as by formulating some specific rules or guidelines. Such rules could be designed to extend a unit into a more complete representation or to complete missing parts of a derived unit according to some clues such as contextual information or variable-specific default unit information.

With the recent advent of artificial intelligence technologies, many applications in IR and NLP have been developed to acquire meta information from unstructured texts as a core module, such as question answering systems, automatic speech translation systems, and intelligent assistant systems. In the process of running such systems, texts are usually found containing a large amount of measurable quantitative information, constituting an essential portion of meta information for information extraction, text understanding, and data analysis.

Particularly, in such a big data era, demands from industry and academic communities for a precise acquisition of measurable quantitative information have increased. For example, business investment companies frequently need to aggregate various sorts of information covering net sales, gross profit, operating expenses, operating profit, interest expense, net profit before taxes, net income, etc., of the target companies from their annual reports. The fast-growing medical informatics research also needs to process a large amount of medical texts to analyze the dose of medicine, the eligibility criteria of clinical trial, the phenotype characters of patients, the lab tests in clinical records, etc[9]. All these demands either in industry or in medical research require the accurate and consistent extraction and representation of measurable quantitative information for automated processing, computation, and exchange.

However, in the IR and NLP areas, there is no standardized way of extracting and representing measurable quantitative information currently available. Each application system developed in industrial sectors has hitherto used its own format to annotate measurable quantitative information. A flexible, interoperable, and standardized measurable quantitative information representation format for IR and NLP tasks to work with many different application systems is called for. The standard SemAF-MQI specifies an annotation scheme with an XML-based representation format for the annotation of quantitative information, which consists of numeric quantities, units, associated with various types of entities that include eventualities.

It represents measurable quantitative information based on commonly used XML structures. The representation standard aims to provide an easy-to-use and universal specification of the annotation format of measurable quantitative information required to unify the data representation, assist computer computation, and reduce human cost. The annotation format proposed here as an ISO International Standard aims at satisfying this need.

5 Purposes and Requirements

This standard describes QML (Quantitative Markup Language) as an annotation scheme, which is designed to be interoperable with other parts of ISO 24617 Language resource management – Semantic annotation framework (SemAF). It also utilizes various ISO standards on lexical resources and morpho-syntactic annotation frameworks. It aims at being compatible with other existing relevant annotation schemata for the processing or exchange of measurable quantitative information.

QML is normalized at the abstract level that allows various serialization formats representing annotated measurable quantitative information such as an XML-based representation. The normalization of QI (quantitative information) annotation is to be stated at the abstract level of annotation, and the standoff annotation format is to be adopted at the concrete level of serialization.

6 Abstract Specification of SemAF-MQI

6.1 Overview

QML is specified at two levels, abstract and concrete. Some characteristics of QML are listed in clause 6.2. The overall structure of QML is represented by a metamodel, as introduced in 6.3. The abstract syntax of QML as QML_as is a set-theoretic specification of QML in conceptual terms that are independent of ways of representing the annotation (content) of measurable quantitative information. The concrete syntax of QML as QML_cs is a specification of a set of representation formats, based on QML_as, for the annotation of measurable quantitative information in a computationally tractable way. The QML_as is introduced in 6.4, while QML_cs is presented in 7. Equivalent concrete syntaxes, including an XML-based concrete syntax QML_csx and a TEI-based oncrete syntax QML_cst, are described in 7 and 8, respectively.

NOTE 1 There can be many equivalent concrete syntaxes defined on a single abstract syntax.

6.2 Characteristics of SemAF-MQI

SemAF-MQI have many characteristics as follows:

1) SemAF-MQI focuses on the annotation of the measurable attributes of entities. For example, “BMI between 10-20 kg/ m^2”

2) SemAF-MQI provides a way to annotate the relations of measures. For example, “age 40 or older” and “fpg>=100 mg/dl or a1c not less than 5.8%”

3) SemAF-MQI covers the complex uses of unitized numeric quantities. For example, “14.0 x 109”, “glycosylated hemoglobin (hba1c) <1.15 times the upper limit of normal”.

4) Sem-MQI can dramatically facility the identification of normalized numeric, units, as the measurable attribute of an associated entity.

NOTE 1 SemAF-MQI does not specify ways of annotating the normalization (e.g., “millimoles per liter” is normalized to “mmol/L”) or complete specification (e.g., “kg/m” should be “kg/m2” for BMI) of units, which is to be dealt with in another part of ISO 24617 that treats automated implementation of MQI.

6.3 Metamodel

The overall structure of QML is represented by a metamodel below:

