Annotating Measurable Quantitative Information in Language: for an ISO Standard

Tianyong Hao
South China
Normal University
Guangzhou, China
haoty@126.com

Haitao Wang
China National Institute
of Standardization
Beijing, China
wanght@cnis.gov.cn

Xinyu Cao
China National Institute
of Standardization
Beijing, China
caoxy@cnis.gov.cn

Kiyong Lee
Korea University
Seoul, Korea
ikiyong@gmail.com

Abstract

Measurable quantitative information (MQI) such as *165 cm* or *60 kg* that applies to the height or weight of an entity is quite common in ordinary language. Such information is abundant in scientific journals, technical reports or medical records that it constitutes an essential part of communicative segments of language in general. The processing of such information is thus required for successful language resource management. To facilitate such processing in a normalized way, this paper aims to specify a general markup language QML for the semantic annotation of MQI in language, while making it interoperable with other semantic annotation schemes that have been developed as ISO standards.

1 Introduction

Demands from industry and academic communities for a precise acquisition of measurable quantitative information from unstructured texts have increased. For example, as addressed in Hao (2016), 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. The demands require the reliable and consistent extraction and representation of measurable quantitative information.

This paper is aimed at developing a markup language QML as specifying a semantic annotation scheme for measurable quantitative information in language. As first proposed by Bunt (2010) and Bunt (2011), and then laid down as part of the principles of semantic annotation by ISO (2016), QML consists of an abstract syntax QML$_{as}$, a concrete syntax QML$_{cs}$, and a semantics. This paper focuses on developing an XML-based QML$_{cs}$ that provides a format for representing annotation structures that consist of entity structures and link structures.

QML is being developed in the working environment of ISO/TC 37/SC 4/WG 2 Semantic Annotation. It thus aims at following the principles of semantic annotation laid down in ISO (2016) in general. It also follows the basic requirements of the linguistic annotation framework (LAF) (ISO, 2012a): (1) the normalization of MQI annotation is to be stated at the abstract level of annotation and (2) the standoff annotation format is to be adopted at the concrete level of serialization.

QML, as developed in this paper, 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 standards, especially ISO (2009) on quantities and units, while being applicable to TimeML (ISO, 2012b) for event-based temporal measures and to ISO-Space (ISO, 2014) for motion-related spatio-temporal measures.

2 Scope

The proposed markup language QML for quantitative information focuses on *measurable* quantitative information. If quantitative information is measurable, then it is expressible in unitized numeric terms such as *165 cm tall*, *sixty-five kilograms*, *60 miles per hour*, *five books*, or *HbA1c between 5-7 %*. The extent of measurability is wider than that of countability; for example, the measures associated with objects that are referred to by mass nouns such as *whisky* or *gasoline* themselves are not countable measures, although bottles of whisky or gallons of gasoline are countable and measurable. Hence, the scope of QML includes countable measures as well as other types that are considered measurable.

The following three cases are, however, excluded from the scope of QML, although they may be considered as involving quantitative information in one way or another;

1. Pure mathematical equations or formulas for scientific calculations like *1+1=2* and *1 plus 1 equals 2*, pure numerical relation such as *2 is greater than 1*. They express (truth-functional) propositions involving numerical relations only.
2. Quantitative terms such as *a few miles* or *very far* are excluded either because the quantity is not expressed in explicitly numerical terms or because no unit is mentioned.
3. QML does not define what units are. Units are conventionally decided on, depending on their application domains. If needed, QML simply refers to ISO (2009) or more than a dozen of its subsequent parts on *quantities and units*, in which units are rigorously defined with respect to each application domain.

3 Specification of QML

3.1 Metamodel

The overall structure of QML is represented by a metamodel in Figure 1. This metamodel introduces: (1) communicative segments as input to the annotation of MQI, (2) from which markable expressions are extracted, and (3) four basic entity types: **entity, dimension, quantitative relation**, and **measure** with its two subtypes, **numeric** and **unit**, which are all anchored to markables. This metamodel also introduces (4) two relation types, measure link (<mLink>) and comparison link (<cLink>). Triggered by a dimension, each **measure link**, tagged <mLink>, relates a measure to an entity. Triggered by a quantitative relation, each **comparison link**, tagged <cLink> relates a measure to another or other more measures.

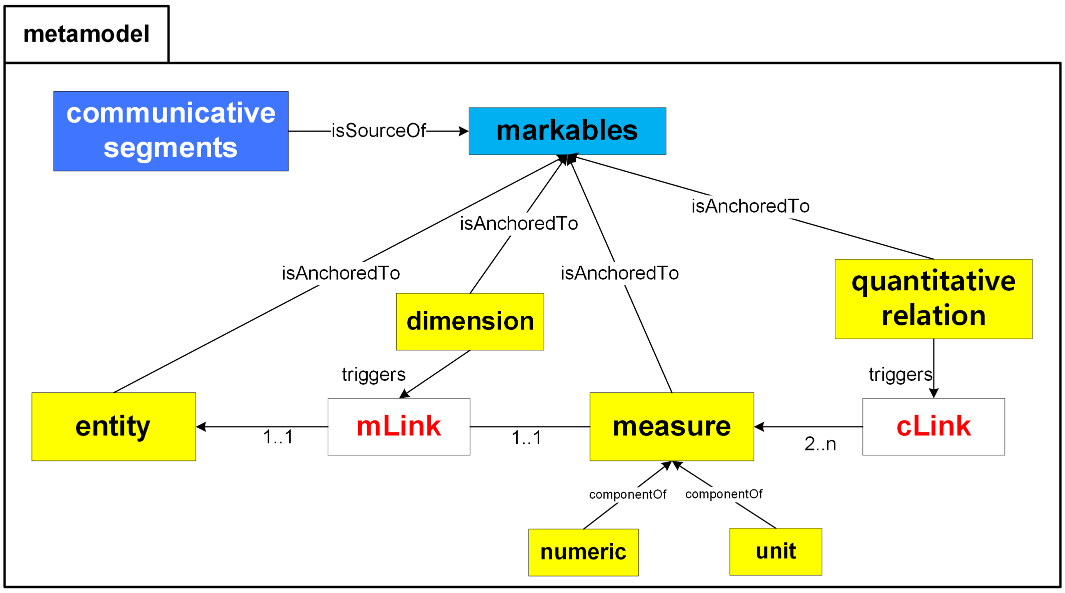

Figure 1: Metamodel of QML

3.2 Abstract syntax of QML (QML$_{as}$)

The abstract syntax specifies an annotation scheme in set-theoretical terms in an abstract way. The abstract syntax QML$_{as}$ for the semantic annotation of measurable quantitative information in communicative segments of language is a quintuple $<C, M, B, R, @>$ such that

1. C is a nonempty set of communicative segments;
2. M is a nonempty set of (possibly null) sequences of segments, called *markables*, from C;

3. B is a nonempty set of basic entity types, anchored to M, consisting of **entity, dimension, quantitative relation**, and **measure** with its two subtypes, **numeric** and **unit**.
4. R is a nonempty set of relation types that consists of **measure link** and **comparison link**.
5. @ is a set of assignments that specify the list of attributes and their value types associated with each of the basic entity types in B and each of the link types in R.

In the abstract syntax QML$_{as}$, @ is expressed in set-theoretic tuples. $@(measure)$ is, for instance, a tuple $<@target, @numeric, @unit>$, consisting of three required attributes such that $@target$ is an anchor to M, the value of $@numeric$ is REALS, and the value of $@unit$ is one of conventionally accepted unit values, provided their values are not complex. The entity types, $@numeric$ and $@unit$, treat complex numerics (e.g. ‘14.0 x 109’) and units (e.g. ‘km/hr’). As specified in ISO (2016), each link r in R is a triplet such that $@r = <\eta, E, \rho>$, where η is an entity structure, E a set of entity structures, and ρ a relation from η to E.

QML$_{as}$ defines well-formed structures, called *annotation structures*, each of which consists of a list of entity structures, each associated with a basic entity type in B, and a list of link structures, each associated with a relation type in R. Semantics then operates on these structures.

3.3 Concrete syntax (QML$_{cs}$)

3.3.1 Overall

Based on an abstract syntax, there can be several equivalent concrete syntaxes. The concrete syntax QML$_{cs}$ is one of such syntaxes based on the abstract syntax QML$_{as}$ as specified above. This concrete syntax QML$_{cs}$ is XML-serialized. Each entity or link structure is represented by a structure, called *element*, consisting of a set of attribute-value pairs with a unique ID that has an attribute name `xml:id`. Each annotation structure is then a list of these structures, enclosed by the element, called *root element*.

Corresponding to each of the basic entity types and the link types in the abstract syntax QML$_{as}$, there is a unique element name, called *tag*, and a unique ID prefix for an XML-based concrete syntax QML$_{cs}$, as shown in Table 1. Each ID prefix is followed by a positive integer. For example, `"x25"` for `<entity>`. In XML, the attribute for ID is prefixed with `xml:id`.

Table 1: List of Tags in QML$_{cs}$

	Tags	**ID prefixes**	**Comment**
root	`<QI>`	`qi`	root tag
Basic entity types			
entity	`<entity>`	`x`	object to which a measure applies
dimension	`<dimension>`	`d`	triggers `<mLink>`
quantitative relation	`<qRelation>`	`qr`	triggers `<cLink>`
measure	`<measure>`	`me`	unitized numeric quantities only
numeric	`<num>`	`nu`	allows complex numeric quantities
unit	`<unit>`	`u`	allows complex units
Link types			
measure relation	`<mLink>`	`mL`	relates a measure to an entity
comparison relation	`<cLink>`	`cL`	relates a measure to another or other more measures

The specification of attribute assignments @ is defined with more specific data types in extended BNF (ISO/IEC, 1996) for the root in 3.3.2, entity types in 3.3.3, and link types in 3.3.4.

3.3.2 Attribute specification of the root `<QI>`

List 1: List of attributes for `<QI>` in extended BNF

```
attributes = identifier, target, [lang], [source], [comment];
target = IDREF | CDATA;
lang = CDATA; {* refer to ISO (2010) on language codes*};
```

```
source = CDATA;
comment = CDATA;
```

3.3.3 Attribute specification of the basic entity types

List 2: List of attributes for `<entity>` in extended BNF

```
attributes = identifier, target, type, [comment];
identifier = x + positive integer;
target = IDREF | CDATA;
type = CDATA; {*ontological types*}
comment = CDATA;
```

List 3: A list of attributes for `<dimension>` in extended BNF

```
attributes = identifier, target, type, [comment];
identifier = d + positive integer;
target = IDREF | CDATA;
type = CDATA; {*For example, "length", "width"*}
comment = CDATA;
```

List 4: A list of attributes for `<qRelation>` in extended BNF

```
attributes = identifier, target, type, [comment];
identifier = qr + positive integer;
target = IDREF | CDATA;
type = CDATA; {* mathematical operators, e.g., "greater than"*}
comment = CDATA;
```

List 5: A list of attributes for `<measure>` in extended BNF

```
attributes = identifier, target, num, unit, [comment];
identifier = me + positive integer;
target = IDREF | CDATA;
num = real numbers;
unit = CDATA | IDREF; {*IDREF refers to a complex unit*}
comment = CDATA;
```

3.3.4 Attribute specification of the link types

By ISO (2016), every link structure is of the form $<\eta, E, \rho>$ such that η is an entity structure, E is a set of entity structures, and ρ is a relation over them.[1] In QML$_{cs}$, we name the entity structure η `"figure"` and the set of entity structure E `"ground"` as in ISO-Space ISO (2014). There are two link types in QML$_{as}$, tagged `<mLink>` and `<cLink>` in QML$_{cs}$. Their attributes are specified as in List 6 and List 7.

List 6: A list of attributes for `<mLink>` in extended BNF

```
attributes = identifier, figure, relType, ground, [trigger], [comment];
identifier = mL + positive integer;
figure = IDREF; {*ID of <measure>*}
ground = IDREFS; {*IDS of <entity>s to which the value of <measure> applies*}
relType = CDATA; {*type of measure that depends on its application domain*}
trigger = IDREF; {*ID of <dimension>*}
comment = CDATA;
```

List 7: A list of attributes for `<cLink>` in extended BNF

```
attributes = identifier, relType, figure, ground, [trigger], [comment];
identifier = cL + positive integer;
relType = CDATA;
figure = IDREF; {*ID of <measure>*}
```

[1] An entity structure is a pair $<m, s>$, where m is a markable and s is semantic information annotated on m.

```
ground = IDREFS; {*IDS of <measure>s*}
trigger = IDREF; {*ID of <qRelation>*}
comment = CDATA;
```

3.4 Illustration of Concrete syntax (QML$_{cs}$)

This clause illustrates how QML$_{cs}$) applies to the representation of measure quantitative information in language in general.[2]

Each of the steps of MQI annotation is illustrated with a sample textual data `"Mia is 165 cm tall."`

Step 1: Pre-processing: word-segmentation and anchoring

Before the semantic annotation of QML-cs applies to some specific language data, the data needs to be pre-processed, for instance, through word segmentation. The word segmentation of the sample data is represented inline as illustrated below.

(1)
```
<wordSeg xml:id="ws1" target="#1a" lang="en">
<w xml:id="w1">Mia</w>
<w xml:id="w2">is</w>
<w xml:id="w3">165</w>
<w xml:id="w4">cm</w>
<w xml:id="w5">tall</w>
<punct xml:id="p1">.</punct>
</wordSeg>
```

This can also be represented in a much simpler way:

(2)
```
<wordSeg xml:id="ws1" target="#1a" lang="en">Mia_w1 is_w2 165_w3 cm_w4 tall_w5.
</wordSeg>
```

Based on the word-segmentation data such as (1) or (2), markables are chosen and each assigned an appropriate entity type, as represented below.

(3)
```
Mia_x1 is [165 cm]_me1 tall_d1.
```

Step 2:Annotation with quantitative information

Word-segmented and anchored data is then annotated with measurable quantitative information, as represented below.

(4) a.
```
<wordSeg xml:id="ws1" target="#1a" lang="en" >Mia_w1 is_w2 165_w3 cm_w4
tall_w5.</wordSeg>
```

b.
```
<QI xml:id="qi1" target="#ws1" lang="en">
<entity xml:id="x1" target="#w1" type="person"/>
<measure xml:id="me1" target="#w3,#w4" num="165" unit="cm"/>
<dimension xml:id="d1" target="#w5" type="length" />
<mLink xml:id="mL1" figure="#me1" ground="#x1" relType="length"
trigger="#d1" />
</QI>
```

With anthor sample data `"John is more than two meters tall."`, we show how a quantitative relation, tagged `<qRelation>`, triggers a comparison relation, tagged `<cLink>`, that relates a measure to another or other more measures. This annotation is interpreted as stating that the height of a person, named *John*, is represented

(5) a.
```
<wordSeg xml:id="ws2" target="#1b" lang="en" >John_w1 is_w2 more_w3 than_w4
two_w5 meters_w6 tall_w7.</wordSeg>
```

b.
```
<QI xml:id="qi2" target="#ws2" >
<entity xml:id="x1" target="#w1" type="person" />
<qRelation xml:id="qr1" target ="#w3, #w4" type="greaterThanOrEqual">
<measure xml:id="me1" target=" " num="" unit=" " />
{*The <measure xml:id="me1"> is a non-consuming tag which is understood as
being anchored to an empty string of texual segments*}
```

[2] We intend to adopt the TEI-serialization as specified in TEI (2016) for QML$_{cs}$. As this stage, however, we have adopted a less embedded mode of representing XML-elements for annotation structures that consist of entity and link structures.

```
<measure xml:id="me2" target="#w5, #w6" num="2" unit="meters" />
<cLink xml:id="cL1" figure="#me1" ground="#me2" relType="
greaterThanOrEqual" trigger="#qr1" />
<dimension xml:id="d1" target="#w7" type="length" />
<mLink xml:id="mL1" figure="#me1" ground="#x1" relType="length"
trigger="#d1" />
</QI>
```

4 Concluding Remarks

Focusing on measurable quantitative information, this paper reports the most recent progress of specification of a markup language QML. We defined the extent of markables of QML and presented a metamodel to show its general structure. In addition, its abstract syntax and an XML-based concrete syntax were formulated with the specification of associated attributes expressed in extended BNF, which is considered expressively more powerful than commonly used data-type declarations (DTD) of XML. The proposed concrete syntax was briefly illustrated with two simple datasets. A fully developed QML is expected to be part of the ISO 24617 standards on semantic annotation for language resource management.

5 Acknowledgments

This paper is supported by grants from National Key R&D Program of China (2016YFF0204205, 2018YFF0213901), China National Institute of Standardization (522016Y-4681, 522018Y-5948, 522018Y-5941), and National Natural Science Foundation of China (No.61772146).

References

Bunt, Harry. 2010. A methodology for designing semantic annotation languages exploiting semantic-syntactic ISO-morphisms. In Alex C. Fang, Nancy Ide, and Jonathan Webster (eds.), *Proceedings of the Second International Conference on Global Interoperability for Language Resources* (ICGL2010), pp.29-46. Hong Kong.

Bunt, Harry. 2011. Abstract syntax and semantics in semantic annotation, applied to time and events. Revised version of Introducing abstract syntax + semantics in semantic annotation, and its consequences for the annotation of time and events. In E. Lee and A. Yoon (eds.), *Recent Trends in Language and Knowledge Processing*, pp.157-204. Hankukmunhwasa, Seoul.

ISO. 2009. *ISO 80000:2009(E) Quantities and units - Part 1: General.* The International Organization for Standardization, Geneva.

ISO. 2010. *ISO 639–1:20109(E) Codes for the representation of names of languages – Part 1: Alpha–2 code.* The International Organization for Standardization, Geneva.

ISO. 2012a. *ISO 24612:2012(E) Language resource management – Linguistic annotation framework (LAF).* The International Organization for Standardization, Geneva.

ISO. 2012b. *ISO 24617–1:2012(E) Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events.* The International Organization for Standardization, Geneva.

ISO. 2014. *ISO 24617–7:2014(E) Language resource management – Semantic annotation framework (SemAF) – Part 7: Spatial information.* The International Organization for Standardization, Geneva.

ISO. 2016. *ISO 24617-6:2016(E) Language resource management - Semantic annotation framework (SemAF)- Part 6: Principles of semantic annotation.* The International Organization for Standardization, Geneva.

ISO/IEC. 1996. *ISO/IEC 14977:1996(E) Information technology - Syntactic metalanguage - Extended BNF.* The International Organization for Standardization and the International Electrotechnical Commission, Geneva.

Lee, Kiyong. 2016. An abstract syntax for ISOspace with its `<moveLink>` reformulated. In Harry Bunt (ed.), *Proceedings of the LREC 2016 Workshop, 12th Joint ACL–ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pp. 28-37. May 23–28, 2016, Portorož, Slovenia.

Lee, Kiyong. 2015. The annotation of measure expressions in ISO standards. *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*

Hao, Tianyong, Wei, Yunyan, Qiang, Jiaqi, Wang, Haitao and Lee, Kiyong. 2017. The representation and extraction of quantitative information. *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, pp. 74-83.

Hao, Tianyong, Liu, Hongfang, Weng, Chunhua. 2016. Valx: A system for extracting and structuring numeric lab test comparison statements from text. *Methods of Information in Medicine*. Vol. 55: Issue 3, pp. 266-275.

TEI. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Intiative Consortium.