# ETSI GS LIS 004 V2.0.0 (2012-07)


# Localisation Industry Standards (LIS); Global Information Management Metrics eXchange Volume (GMX-V)

***Disclaimer***

This document has been produced and approved by the Localisation Industry Standards (LIS) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGS/LIS-0004

Keywords
ICT, XML

***ETSI***

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

# Contents

# Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://ipr.etsi.org).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

# Foreword

This Group Specification (GS) has been produced by ETSI Industry Specification Group Localisation Industry Standards (LIS) and represents the views of those members who participated in this ISG. It does not necessarily represent the views of the entire ETSI membership.

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

The present document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at http://portal.etsi.org/tb/status/status.asp.

If you find errors in the present document, please send your comment to one of the following services: http://portal.etsi.org/chaircor/ETSI_support.asp.

# 1 Scope

The present document scope is to recreate, enhance and maintain former LISA OSCAR SIG GMX-V standard. GMX-V stands for Global Information Management Metrics eXchange Volume. The present document is the new GMX-V version V2.0 of January 2012 edited by Andrzej Zydroń <azydron@xml-intl.com> and Arle Lommel <alommel@gala-global.org> and made available on-line at http://www.xtm-intl.com/manuals/gmx-v/GMX-V-2.0.html for the final version in XHTML.

Past LISA/OSCAR standards are available now online at http://www.gala-global.org/lisa-oscar-standards and http://www.ttt.org/oscarStandards.

In March 2011 the Localization Industry Standards Association (LISA) was declared insolvent. As a result LISA's portfolio of standards has been authorized to be posted under a Creative Commons Attribution 3.0 License that allows for reuse and creation of derivative works based on the LISA standards. Note that LISA has designated the European Telecommunications Standards Institute (ETSI) Localization Industry Standards (LIS) Industry Specification Group (ISG) as its successor organization for its standards portfolio.

The present document defines the LISA ETSI (formerly LISA) Global Information Management Metrics eXchange Volume (GMX-V) Version 2.0 specification.

The purpose of this vocabulary is to define the metrics that allow for the unambiguous calculation of the size in terms of word and character counts of a given electronic document (so called verifiable metrics), as well as providing a method of exchanging said counts via an XML document. In addition GMX/V provides a means of exchanging non-verifiable metrics (such as manual page and screen counts) using the same XML vocabulary. Verifiable metrics can be proven using a computer program based on the GMX/V specification.

GMX-V is one of the tripartite planned Global Information Management standards which encompass volume (GMX-V), complexity (GMX-C) and quality (GMX-Q).

GMX/V Version 2.0 is backwards compatible with GMX/V Version 1.0 and introduces the following new features:

1) An overall character count which includes white space and punctuation character counts as well as the actual alpha numeric character count.

2) Word count factors for electronic documents encoded with Chinese, Japanese, Korean and Thai scripts.

# 2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

## 2.1 Normative references

The following referenced documents are necessary for the application of the present document.

[1] Past versiosn of five LISA OSCAR SIG standards: TMX, TBX, SRX, GMX-V and xml:tmp.

NOTE: Available at http://www.ttt.org/oscarStandards or http://www.gala-global.org/lisa-oscar-standards

[2] OASIS XML Localisation Interchange File Format (XLIFF) TC.

NOTE: Available at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff

[3] Unicode 6.1.0.

NOTE: Available at http://www.unicode.org/versions/Unicode6.1.0/

[4] Unicode Standard, Annex #29, Version 4.1.0, Text Boundaries.

NOTE: Available at http://www.unicode.org/reports/tr29/tr29-9.html.

[5] Unicode Standard, Annex #15, Version 4.1.0, Unifcode Normalization Forms.

NOTE: Available at http://www.unicode.org/reports/tr15/

[6] ISO 8601: 2004: "Data elements and interchange formats -- Information interchange -- Representation of dates and times".

NOTE: Available at http://www.iso.org/iso/catalogue_detail?csnumber=40874

[7] IETF RFC 4646 (September 2006): "Tags for Identifying Languages".

NOTE: Available at http://www.rfc-editor.org/rfc/rfc4646.txt

[8] XLIFF 1.2 Specification. OASIS XLIFF Committee Specification, February 2008.

NOTE: Available at http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html

[9] GMX/V Java Reference implementation - Okapi XLIFF extractor.

NOTE: Available at http://okapi.sourceforge.net/Release/Utilities/Help/extraction.htm

[10] IETF RFC 1766: "Tags for the Identification of Languages".

NOTE: Available at http://www.ietf.org/rfc/rfc1766.txt

## 2.2 Informative references

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] The XML schema for GMX-V.

NOTE: Available at http://www.xtm-intl.com/manuals/gmx-v/gmx-v.xsd

[i.2] ISO, International Organization for Standardization Web site.

NOTE: Available at http://www.iso.org/iso/home.html

[i.3] European Telecommunications Standards Institute Web site.

NOTE: Available at http://www.etsi.org/WebSite/homepage.aspx

[i.4] European Telecommunications Standards Institute, Localisation Industry Standards Web site.

NOTE: Available at http://portal.etsi.org/lis

[i.5] OASIS (Organization for the Advancement of Structured Information Standards Web site). .

NOTE: Available at https://www.oasis-open.org/org

[i.6] World Wide Web Consortium Web site (W3C).

NOTE: Available at http://www.w3.org/

# 3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

| | |
|---|---|
| DOM | W3C Document Object Model |
| GMX-V | Global Information Management Metrics eXchange |
| OSCAR | LISA special interest group (Open Standards for Container/Content Allowing Re-use) |
| RTF | Rich Text Format |
| SIG | Special Interest Group |
| Unicode | Unicode is the official way to implement ISO/IEC 10646 - universal character encoding standard |
| UTC | UTC stands for Coordinated Universal Time |
| WSDL | Web Services Definition Language |
| XLIFF | OASIS Standard for XML Localization Interchange File Format |
| XML | eXtensible Markup Language |

# 4 GMX-V V2.0 Specification

## 4.1 Introduction

GMX-V addresses the issue of quantifying the workload for a given localization or translation task. This is often commonly referred to as word counts. Word counts, however, do not convey the true range of possible metrics that can be used to assess the cost of localizing a document such as the number of screen shots for a software localization project, or page counts for a document layout task. GMX-V is a more precise definition of the metrics required for billing and sizing purposes.

In defining GMX-V, care has been taken to provide a definitive and unambiguous definition that also offers the widest possible scope for achieving an adequate description of the task load for a given Global Information Management project.

Metrics fall into two categories: directly verifiable from the file contents (words and characters) and unverifiable (pages, file units, lines, etc.). Some metrics may fall into both categories, depending on the circumstances. For instance, page counts may be verifiable from the file contents under some circumstances; however, under other circumstances, only a printout for a given page format can provide the basis for a count.

GMX-V does not preclude the existence of unverifiable counts, but is concerned with defining precise rules for verifiable counts based on the file unit that is being counted.

To this end, it is proposed that GMX-V cover both word and character counts, as well as allowing for other relevant count categories that cannot be verified electronically. Character counts convey the most precise definition of a translation task, whereas word counts are the most commonly used metric in the translation industry. GMX-V encompasses both measurements, thus affording both translation suppliers and customers with a choice as to which measurement most adequately reflects the translation task in question as well as allowing for other relevant metrics.

From the implementation point of view GMX-V is designed to co-exist as a namespace within other XML documents. The main target will primarily be XLIFF and Translation Web Services WSDL compliant documents, but any XML document that allows namespace extension points could host GMX-V namespace.

GMX-V has therefore the following aims:

1) To provide an unambiguous specification for counting words and characters for translation related tasks.

2) To provide a rich set of qualifiers to help accurately define the actual translation workload for translation related tasks.

3) To provide an XML notation for exchanging Global Information Management metrics for any Global Information Management task whether it entails translation activity or not.

GMX-V defines a precise mechanism for word and character counts irrespective of the context within which the standard is being used. It can therefore be used for the unambiguous definition of word and character counts for any electronic document. Section 2.14 defines aspects of the GMX-V specification that are specific to localization tasks.

GMX-V makes no assumptions about the format of the input document. Although it is envisaged that GMX-V will be applied predominantly to XLIFF documents, it is not limited to XLIFF or XML based documents. Any electronic document can be counted using GMX-V as long as the text is converted into the required canonical form described in Section 2.2. Canonical Form.

## 4.2 Key Concepts

The following concepts are fundamental to the GMX-V specification.

### 4.2.1 Text Unit

A document is made of a number of text units. A text unit is either a stand alone piece of text within a document, or a subdivision of a stand alone piece of text into recognizable segments. Document metrics will be based on an accumulation of the word and character statistics of the individual text units. Any segmentation should be detailed in an SRX (Segmentation Rules eXchange format) compliant document. A separate count of text units can be maintained within the GMX-V specification (see clause 4.3.8.).

Where the metrics are based on an XLIFF 1.2 file the <source> child of a <trans-unit> element constitutes a text unit regardless of any additional segmentation information that may be available as per section '2.9 Segmentation' of the XLIFF 1.2 specification. The <source> element should always form the basis of verifiable metrics such as word and character counts in an XLIFF file. The reason for this is to insure consistency with the XLIFF file structure and is also due to the fact that even where it is used <seg-source> data is not guaranteed to be comprehensive.

### 4.2.2 Canonical Form

A precise canonical (base) form for individual text units is required to provide an accurate and unambiguous basis for conducting metrics. Native forms of the text are often encumbered with extraneous proprietary formatting codes, which make the production of unambiguous statistics difficult.

The XLIFF (XML Localization Interchange File Format) normalized XML form using Unicode encoding of the source language text shall be used as the basis for the canonical form for GMX-V. XLIFF is an OASIS standard. This is the format in which the text appears in the <source> element of an XLIFF file <trans-unit> element. For audit purposes an XLIFF file is required for GMX-V metrics. The GMX-V count is undertaken on the basis of this XLIFF document. The <source> element shall always be the basis of the verifiable word and character based counts.

EXAMPLE:

```
<source>An example of the canonical form of a text unit.</source>
```

The canonical form does not contain any embedded formatting characters, such as those that exist in an XLIFF document extracted from a RTF file. Any such characters must be removed to produce the canonical form. In addition any formatting characters representing a space must be converted to the standard SPACE character (U+0020).

Original XLIFF <source> element with embedded RTF codes:

```
<source>The <bpt i="1" x="1">{\b </bpt>black<ept i="1">}</ept><bpt i="2" x="2">{\i </bpt>
cat<ept i="2">}</ept> eats.</source>
```

Canonical form:

```
<source>The <bpt/>black<ept/><bpt/> cat<ept/> eats.</source>
```

The GMX-V metrics for the above are:

```
 words: 4, characters: 15, inline elements: 4, punctuation characters: 1, white space characters:
3
```

GMX-V does not mandate that the electronic form of the document being counted is in XLIFF. Nevertheless it is much easier to conduct GMX-V counts on an XLIFF file. Independent verification of the GMX-V metrics requires that an XLIFF version of the file is available.

GMX-V assumes that the text that is presented for counting contains only the text that has been deemed as relevant for the task. For a localization task this would be only the text that is required for localization purposes.

Where the metrics are based on an XLIFF 1.2 file and segmentation information is available as per section '2.9 Segmentation' of the XLIFF 1.2 specification then the segmented data should be used rather than the unsegmented contents of the XLIFF 1.2 <source> element.

### 4.2.3 Unicode (ISO 10646)

Unicode Version 4.1.0 forms the fundamental basis for the XLIFF canonical form for character encoding and for establishing word boundaries. Apart from the ISO 10646 two Unicode Technical Reports are used for establishing the canonical form:

- Unicode Standard Annex #29 - Text Boundaries (TR 29-9).
- Unicode Standard Annex #15 - Unicode Normalization Forms (TR 15) - Normalized Form C.

Unicode TR 29 establishes the word boundaries that allow for words and characters to be counted. TR 15 establishes the actual canonical form for Unicode characters themselves. Normalized Form C is the form mandated by W3C for XML documents and can normally be taken for granted during any conversion from non-Unicode encoding form to Unicode using industry standard programming libraries.

### 4.2.4 Word Boundaries

Word and character counts are governed by Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 4 Word Boundaries, which in turn relies on the Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 3 Grapheme Cluster Boundaries rules. The present document unambiguously defines words as opposed to stand alone punctuation, white space or enclosing punctuation characters. All word and character counts will be on the basis of the Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 4 Word Boundaries.

A full definition of the application of Unicode TR 29 Word Boundaries to the GMX-V specification is provided in clause 4.2.8.

### 4.2.5 Verifiable and Non-Verifiable Metrics

Not all GMX-V metrics can be strictly defined or verified. Verifiable metrics can be defined for an electronic document in XLIFF canonical form. Non-verifiable metrics require a mechanism such as manual counting to establish its accuracy.

Non-verifiable metrics are not subordinate in any way to verifiable metrics; it is only that they cannot be proven on the basis of a given electronic document.

### 4.2.6 Inline Element Transparency

For word and character counts, the code for any inline elements (either empty or having content) within the canonical XLIFF representation will be treated as being totally transparent, that is, they are be treated as not being present. Inline elements will be counted separately. This is detailed in the clause 4.2.11.

EXAMPLE:

```
<source>In this <g id="g1">exa<x id="x1"/>mple</g> the in-line codes do not form
part of the word or character counts but are counted separately.</source>
```

would be counted as:

```
<source>In this example the in-line codes do not form
part of the word or character counts but are counted separately.</source>
```

The GMX-V metrics for the above are:

```
 words: 20, characters: 91, inline elements: 3, punctuation characters: 1, white space
characters: 19
```

In the canonical form, any inline XLIFF elements that signify a space or new line character must have the equiv-text attribute set to a single space character if the space character were otherwise not present in the canonical XLIFF form. If an inline element has spatial characteristics, then it is up to the program that is generating the XLIFF file to set the contents of the equiv-text attribute accordingly. If the inline element has content, then the space character must precede and follow the start and end tags of the element if the equiv-text attribute is used to denote spatial characteristics.

GMX-V is totally agnostic regarding how the XLIFF file is created and cannot imply anything regarding inline elements.

```
   <source>The HTML break element<x id="x1" ctype="x-html-br" equiv-text=" "/>represented here by
the in-line "x" element was
    not preceded by a space in the original document.</source>
```

Sub flow text within place holder elements needs to always be preserved. Sub flow text always implies that it is preceded and followed by a white space character. If no white space is present then it must be inserted in the canonical form:

```
    <source>Start<bpt id="2">code<sub>Text</sub></bpt>end<ept id="2">code</ept>.</source>
```

Canonical form:

```
    <source>Start<bpt><sub> Text </sub></bpt><ept>end</ept>.</source>
```

The GMX-V metrics for the above are:

```
    words: 3, characters: 12, inline elements: 6, punctuation characters: 1, white space characters:
2
```

A separate "Inline Element" count is maintained for inline elements.

### 4.2.7 White Space Characters

The following list defines white space characters:

- Unicode space characters (SPACE_SEPARATOR, LINE_SEPARATOR, or PARAGRAPH_SEPARATOR) but not non-breaking space ('\u00A0', '\u2007', '\u202F').
- '\u0009', HORIZONTAL TABULATION.
- '\u000A', LINE FEED.
- '\u000B', VERTICAL TABULATION.
- '\u000C', FORM FEED.
- '\u000D', CARRIAGE RETURN.
- '\u001C', FILE SEPARATOR.
- '\u001D', GROUP SEPARATOR.
- '\u001E', RECORD SEPARATOR.
- '\u001F', UNIT SEPARATOR.
- '\u200B', ZERO WIDTH SPACE.

In the XLIFF canonical form white space characters are trimmed at the start and end of a text unit. Within a text unit multiple white space characters are reduced to a single space. The only exception is where the xml:space="preserve" attribute has been set for an element. In this case no normalization of spaces will occur and all white space characters will be counted.

A separate count 'WhiteSpaceCharacterCount' will be maintained for Unicode white space. White space characters are not included in the main character count.

EXAMPLE:

```
<source>This sentence has a word count of 14 words and 13 white space characters.</source>
```

### 4.2.8 Words

Words form the basic unit for counting for the GMX-V specification. The character count is also based on identified words, with the exception of scripts that do not use space word separation. Word separation is described in this clause.

Words are defined according to Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 4 Word Boundaries, which in turn relies on the Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 3 Grapheme Cluster Boundaries rules. Unicode TR 29 Section 4 defines detailed Boundary Property Values and Boundary Rules which distinguish words from other grapheme clusters such as punctuation characters. These form an integral part of the GMX-V specification.

The following example, taken from Unicode TR 29, shows an example of the identification of grapheme boundaries:

EXAMPLE 1: Word Boundaries.

| The | quick | ( | " | brown | " | ) | fox | can't | jump | 32,3 | feet | , | right | ? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Followed by the extracted words:

EXAMPLE 2: Extracted Words.

| The | quick | brown | fox | can't | jump | 32,3 | feet | right |
|---|---|---|---|---|---|---|---|---|

In addition Unicode TR 29, section 4 provides an optional rule for the apostrophe character which relates to French and Italian usage such as "l'objectif". This rule known as "Break between apostrophe and vowels (French, Italian)" must also be applied for GMX-V. Apostrophe includes U+0027 (') APOSTROPHE and U+2019 (’) RIGHT SINGLE QUOTATION MARK (curly apostrophe).

Thai, Lao, Khmer, Myanmar, Chinese, Japanese and Korean scripts do not use space characters between words. See clause 4.2.13 for details of how these scripts are treated within the GMX-V standard.

Hyphen characters will not be treated as word break characters. Hyphens include U+002D HYPHEN-MINUS, U+2010 HYPHEN, U+058A ARMENIAN HYPHEN and U+30A0 KATAKANA-HIRAGANA DOUBLE HYPHEN, and will form part of the character count if they appear as part of a word as in 'Italian-American'.

No additional tailoring of the Unicode TR 29 Version 4.1.0 - Text Boundaries, Section 4 Word Boundaries rules is permitted in the GMX-V specification.

EXAMPLE 3:

```
<source>This sentence has a word count of 9 words.</source>
<source>This sentence/text unit has a word count of 11 words.</source>
```

### 4.2.9 Characters

The character count is predicated on the word count detailed in clause 4.2.8. For Thai, Lao, Khmer, Myanmar, Chinese, Japanese, Korean and other scripts that do not use spaces between words, the character counts are based on the non-punctuation grapheme boundaries. For all other scripts the character count is based on the identifiable words. Please refer to clause 4.2.8 for a detailed explanation.

Characters are counted based on Unicode encoding according to Unicode TR 15 - using Unicode Normalization Form C.

### 4.2.10 Punctuation Characters

An additional separate count is maintained for punctuation characters.

The following list defines what are to be considered Unicode punctuation characters:

- Basic Latin punctuation characters in the ranges of '\u0021' - '\u002F', '\u003A' - '\u0040', '\u005B' - '\u0060', '\u007B' - '\u007E'.
- !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~.
- The division sign ÷ \u00F7 and multiplication sign × \u00D7.
- The Spanish inverted exclamation and question marks, \u00A1 (¡) and \u00BF (¿).
- The Armenian full stop \u0589.
- The Hebrew colon \u05C3, maqaf \u05BE and paseq \u05C0.
- The Arabic semicolon \u061B.
- General Unicode Punctuation: '\u2000'–'\u+206F'.
- CJK Symbols and Punctuation: '\u3000' – '\u303F'.

The only exceptions are the 'hyphen' and 'apostrophe' characters if they appear within a word as in: can't, or out-of-the-box. These will be counted as normal characters and form part of the 'TotalCharacterCount' as detailed in clause 4.3.4.

The apostrophe character count is qualified for French and Italian as per the Unicode TR 29 Version 4.1.0 - Text Boundaries, section 3 Grapheme Cluster Boundaries rules "Break between apostrophe and vowels (French, Italian)" rule as described in clause 4.2.8. In this rule the 'apostrophe' acts as a word break and is also counted as part of the character count.

Hyphens include U+002D HYPHEN-MINUS, U+2010 HYPHEN, U+058A ARMENIAN HYPHEN and U+30A0 KATAKANA-HIRAGANA DOUBLE HYPHEN, and will form part of the character count if they appear as part of a word as in 'Italian-American'. Please refer to Section 2.8 above for a detailed explanation.

A separate character count 'PunctuationCharacterCount' will be maintained for all punctuation characters in the document. Please refer to clause 4.3.4.

A separate character count 'WhiteSpaceCharacterCount' will be maintained for all white space characters in the document.

### 4.2.11 Inline Element Counts

Inline elements give an indication of the complexity of the localization task. Among inline elements, a separate count will be maintained for elements that reference other elements. An additional count of inline elements will be maintained for each text unit's categorization category detailed below. Inline elements with content will be counted as two inline elements.

EXAMPLE:

```
<source>In this <g id="g1">example</g>
the in-line codes do not form part of the word or character counts but constitute a
separate inline element count of 2, because the inline element has content.</source>

<source>In this <g id="g1">exa<x id="x1"/>mple</g>
the in-line codes do not form part of the word or character counts but constitute a
separate inline element count of 3, because we have one element with content and
one without.</source>
```

### 4.2.12 Linking Inline Elements

An additional count of inline elements that link to another text unit will be kept. Linked elements require additional localization effort as the linked text unit needs to be referenced as part of the translation. These elements are also counted as part of the inline element count above.

EXAMPLE:

```
   <source>In this <g id="g1" xid="t2">example</g> the in-line element references another trans-unit
via the xid attribute -
   it forms part of the inline element count as well as the linking inline element count.</source>
```

### 4.2.13 Logographic Scripts

Logographic scripts such as Chinese, Japanese and Korean as well as South Asian languages such as Thai, Lao, Khmer, Myanmar do not use white space characters to delineate words. GMX/V Version 1.0 allowed for the omission of word counts for these languages. Subsequent feedback from GMX/V users resulted in the application of standard factors to the character count that provides an approximate word count based on averages of characters per word. The Unicode TR 29 - Text Boundaries, Section 3 Grapheme Cluster Boundaries rules will still apply to distinguish text from punctuation characters for these scripts. These will be used to provide character counts for these scripts. Please refer to clauses 4.2.8 and 4.2.9 for a detailed explanation.

The following factors are to be applied to the base character count for each of the following languages in order to provide a word count:

1) Chinese (all forms): 2.8.

2) Japanese: 3.0.

3) Korean: 3.3.

4) Thai: 6.0.

These factors are based on acknowledged best practice within the Localization Industry.

There are no currently accepted word count factors for Lao, Khmer and Myanmar. For these languages only character counts will be required.

All other counts are also relevant for these scripts.

### 4.2.14 Localization specific counts

The following count categories are specific to localization industry tasks.

#### 4.2.14.1 Qualitative Text Unit Categorization

A typical translatable document will contain a variety of types of text units. Some of these will require translation and some will not, while other text units will require only proofing since they have been matched against a leveraged translation memory database. Regardless of the agreement between a translation supplier and a customer, a count for overall text units, as well as translatable and non-translatable, will be provided for both word and character counts. In the following XLIFF based examples the canonical form is that of the <source> element.

EXAMPLE 1:

```
<trans-unit id="t1" translate="yes">
    <source>This is an example of translatable text</source>
    <target>This is an example of translatable text</target>
</trans-unit>
<trans-unit id="t2" translate="no" state-qualifier="x-alphanumeric">
    <source>10AB1024</source>
    <target>10AB1024</target>
</trans-unit>
<trans-unit id="t3" translate="no" state-qualifier="x-punctuation">
    <source>-</source>
    <target>-</target>
</trans-unit>
<trans-unit id="t4" translate="yes">
    <source>matched sentence</source>
    <target state-qualifier="leveraged-tm">zdanie dopasowane</source>
</trans-unit>
```

Prior to sending the translation project to the translation supplier, the customer may have analyzed the source document against a translation memory in order to retrieve previously-translated segments. In this instance, an additional word count of the segments that are found in the translation memory may be provided.

EXAMPLE 2:

```
<trans-unit id="t1" translate="yes">
    <source>This text unit has been matched against a leveraged matched database.</source>
    <target state-qualifier="leveraged-tm">To zdanie zostalo dopasowane z bazy danych.</source>
</trans-unit>
```

Certain text units, such as numeric or measurement-only text units, may be converted automatically by software into the target language.

EXAMPLE 3:

```
<trans-unit id="t1" translate="no" state-qualifier="x-numeric">
    <source>10,000.00</source>
    <target>10.000,00</target>
</trans-unit>
<trans-unit id="t2" translate="no" state-qualifier="x-measure">
    <source>10.50 mm</source>
    <target>10,50 mm</target>
</trans-unit>
```

It is up to the translation supplier and customer to agree on the exact nature and type of non-translatable text units. Text unit categorization will not be mandated, merely offered in the standard as an option.

### 4.2.14.2 Unqualified Text Units

Unqualified text units are text units that require translation. Any text unit that is not qualified as per clause 4.2.14.1 and has no exact or leveraged match is deemed to be unqualified.

This classification is important as any auto text (see clause 4.2.17) or inline linking (see clause 4.2.11) and non-linking (see clause 4.2.12) counts are only applied to unqualified text units.

### 4.2.14.3 Translatable Text Counts

One of the main aspects of the present document is to produce an unambiguous industry accepted figure for the total quantity of words and characters within an electronic document. For translation tasks the minimum required by the specification, is that a volume metric is produced for words and characters that includes the following:

- Total Word Count:
  - This includes all words as defined in clause 4.2.8.
- Total Character Count:
  - This the character count for all words as defined in clause 4.2.9.

Over and above this minimum conformance for translation tasks (see 4.3.9. Conformance) required by the present document, it is up to the customer and supplier to decide how the other count categories will be applied to the translation task at hand. The present document provides a flexible and comprehensive vocabulary for customizing this calculation, but does not attempt to mandate any one given solution.

### 4.2.15 XML Entity References

The built-in XML entity reference characters <, >, &, ", and ' will be counted as a single character for character counting purposes. It is a requirement of the GMX-V Canonical Form that all XML entity references are resolved prior to counting.

### 4.2.16 User Defined Entity References

When counting XML documents user defined entity references may be encountered. The GMX-V Canonical Form canonical form requires that all user defined entities must be fully resolved prior to counting.

### 4.2.17 Auto Text

Within a document it is possible to identify text segments that can be handled automatically. Items such as numeric values, e.g. 10 or 10,000.00, measurement units e.g. 10,5 mm, standard phrases or acronyms e.g. WYSIWYG or trade names e.g. "Weapons of Mass Destruction™".

It is possible to maintain word and character counts for such treatable text. This can be used to identify and charge for these categories at a different price. The categories will follow closely those defined in clause 4.2.14.1, although they will apply ONLY to 'unqualified' text units which do not have fuzzy matching.

Unqualified in this sense relates to text units that are not already covered by a qualitative category. This count category may be referred to as auto text for short.

### 4.2.18 Repetition Counts

Within a document the same unmatched text units may occur multiple times. This fact can be exploited by translation software to automatically populate subsequent repeating text units once the first one has been translated. Subsequent occurrences can therefore be automatically qualified as 'repeat' text in the same manner as leveraged matched text. Auto text metrics can therefore only be applied to the first occurrence of the text unit and not those qualified as 'repeat' text units.

Repetition counts should be based on the non-normalized form of the text unit in order to insure that the repetition is based on the exact form of the text. For example text units with identical text but different inline elements are not repetitions.

### 4.2.19 Current Commercial Practice

Current commercial practice varies from product to product. There is no unified method of providing an industry wide set of metrics. The GMX-V specification provides a level of detail which provides an adequate way of reconciling GMX-V with metrics provided by commercial practice which is based on accepted standards such as Unicode TR29-9, SRX and XLIFF.

The actual makeup of the count is up to the supplier and customer.

## 4.3 Counts

The count element type attribute contains the count category. The following count categories are provided for:

- Word Count Categories.
- Auto Text Word Count Categories.
- Character Count Categories.

- Auto Text Character Count Categories.
- Inline Element Count Categories.
- Linking Inline Element Count Categories.
- Text Unit Counts.
- Other Count Categories.

Auto Text comprises automatically identifiable text such as numeric values, date and trade marks or model names that can be automatically processed or ignored during translation. Inline elements relate to elements that occur within the text units that are being counted. Linking Inline Element Count Categories relate to linking elements that occur within the text units that are being counted such as HTML 'a' elements.

### 4.3.1 Word Count Extension Mechanism

The count categories specified in the present document are designed to accommodate the most frequent types of count encountered. While these cover the most commercial requirements it is envisaged that they will not be sufficient to cover all instances required. To this end an extension mechanism is provided. All extension count types will start with 'x-'. The following extension conventions apply:

- All custom word count type attributes must end with the text: 'WordCount'.
- All custom character count type attributes must end with the text: 'CharacterCount'.
- All non-word and non-character custom count type attributes must end with the text: 'OtherCount'.
- All non-translatable custom count categories must contain the text: 'NonTranslatable' before the count type text that ends the count type.
- All translatable custom count categories must contain the text: 'Translatable' before the count type text that ends the count type.

An example of the use of the extension conversions would be:

'x-SomeDescriptionOtherNonTranslatableTextUnitWordCount'.

These conventions are designed to facilitate the accumulation of appropriate count categories automatically.

The following count concepts are fundamental to the GMX-V standard.

### 4.3.2 Word Count Categories

The following word counts will be provided by symbolic name:

TotalWordCount:

- Total word count - an accumulation of the word counts, both translatable and non-translatable, from the individual text units that make up the document.

Count categories for non-translatable words:

This is an accumulation of the word counts from the non-translatable categories of text units within the document. The following possible categories are proposed:

- ProtectedWordCount:
  - An accumulation of the word count for text that has been marked as 'protected', or otherwise not translatable (XLIFF text enclosed in <mrk mtype="protected"> elements).
- ExactMatchedWordCount:
  - An accumulation of the word count for text units that have been matched unambiguously with a prior translation and thus require no translator input.

- LeveragedMatchedWordCount:
  - An accumulation of the word count for text units that have been matched against a leveraged translation memory database.
- RepetitionMatchedWordCount:
  - An accumulation of the word count for repeating text units that have not been matched in any other form. Repetition matching is deemed to take precedence over fuzzy matching.
- FuzzyMatchedWordCount:
  - An accumulation of the word count for text units that have been fuzzy matched against a leveraged translation memory database.
- AlphanumericOnlyTextUnitWordCount:
  - An accumulation of the word count for text units that have been identified as containing only alphanumeric words.
- NumericOnlyTextUnitWordCount:
  - An accumulation of the word count for text units that have been identified as containing only numeric words.
- MeasurementOnlyTextUnitWordCount:
  - An accumulation of the word count from measurement-only text units.

The following is an example of the use of the extension mechanism for user defined word counts defined in the clause 4.3.1:

x-OtherNonTranslatableTextUnitWordCount:

- An accumulation of the word count for text units that have been identified as containing only other user-defined non-translatable words. The actual definition of the content and naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and ends with the text NonTranslatableTextUnitWordCount.

x-TranslatableTextUnitWordCount:

- An accumulation of the word count for text units that have been identified as containing only other user-defined translatable words. The actual definition of the content and naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and end with the text TranslatableTextUnitWordCount.

The actual translatable text count can be obtained by subtracting all of the above categories from the TotalWordCount, with the exception of LeveragedMatchedWordCount, FuzzyMatchedWordCount and RepetitionMatchedWordCount. These last three categories can be used to qualify the translation count itself.

### 4.3.3 Auto Text Word Count Categories

The following auto text categories are applicable to text from unqualified text units with the exception of fuzzy matched text units.

The following word counts will be provided by symbolic name:

SimpleNumericAutoTextWordCount:

- An accumulation of the word count for simple numeric values, e.g. 10.

ComplexNumericAutoTextWordCount:

- An accumulation of the word count for complex numeric values which include decimal and/or thousands separators, e.g. 10,000.00.

MeasurementAutoTextWordCount:

- An accumulation of the word count for identifiable measurement values, e.g. 10,50 mm. Measurement values take precedent over the above numeric categories. No double counting of these categories is allowed.

AlphaNumericAutoTextWordCount:

- An accumulation of the word count for identifiable alphanumeric words, e.g. AEG321.

DateAutoTextWordCount:

- An accumulation of the word count for identifiable dates, e.g. 25 June 1992.

TMAutoTextWordCount:

- An accumulation of the word count for identifiable trade marks, e.g. "Weapons of Mass DestructionTM".

The following are examples of how to use the user defined extension mechanism for auto text word count categories defined in the clause 4.3.1:

x-OtherAutoTextWordCount:

- Other auto text word counts. The actual naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and ends with the text OtherAutoTextWordCount. This is an extension mechanism.

### 4.3.4 Character Count Categories

The following character counts will be provided by symbolic name:

TotalCharacterCount:

- An accumulation of the character counts, both translatable and non-translatable, from the individual text units that make up the document. This count includes all non white space characters in the document (please refer to clause 4.2.7 for details of what constitutes white space characters), excluding inline markup and punctuation characters (please refer to clause 4.2.10 for details of what constitutes punctuation characters).

PunctuationCharacterCount:

- The total of all punctuation characters in the canonical form of text in the document that DO NOT form part of the character count as per clause 4.2.10.

WhiteSpaceCharacterCount:

- The total of all white space characters in the canonical form of the text units in the document. Please refer to clause 4.2.7 for a detailed explanation of how white space characters are identified and counted.

OverallCharacterCount:

- The total of all of the three main character counts (TotalCharacterCount + PunctuationCharacterCount + WhiteSpaceCharacterCount) in the canonical form of the text units in the document.

Count categories for non-translatable characters:

This is an accumulation of the character counts from the non-translatable categories of text units within the document. The following possible categories are proposed:

- ProtectedCharacterCount:
  - An accumulation of the character count for text that has been marked as 'protected', or otherwise not translatable (XLIFF text enclosed in <mrk mtype="protected"> elements).
- ExactMatchedCharacterCount:
  - An accumulation of the character count for text units that have been matched unambiguously with a prior translation and require no translator input.

- LeveragedMatchedCharacterCount:
  - An accumulation of the character count for text units that have been matched against a leveraged translation memory database.
- RepetitionMatchedCharacterCount:
  - An accumulation of the character count for repeating text units that have not been matched in any other form. Repetition matching is deemed to take precedence over fuzzy matching.
- FuzzyMatchedCharacterCount:
  - An accumulation of the character count for text units that have a fuzzy match against a leveraged translation memory database.
- AlphanumericOnlyTextUnitCharacterCount:
  - An accumulation of the character count for text units that have been identified as containing only alphanumeric words.
- NumericOnlyTextUnitCharacterCount:
  - An accumulation of the character count for text units that have been identified as containing only numeric words.
- MeasurementOnlyTextUnitCharacterCount:
  - An accumulation of the character count from measurement-only text units.

The following is an example of the use of the extension mechanism for user defined character counts defined in the clause 4.3.1:

x-OtherNonTranslatableTextUnitCharacterCount:

- An accumulation of the character count for text units that have been identified as containing only other user-defined non-translatable words. The actual definition of the content and naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and ends with the text NonTranslatableTextUnitCharacterCount.

x-TranslatableTextUnitCharacterCount:

- An accumulation of the character count for text units that have been identified as containing only other user-defined translatable words. The actual definition of the content and naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and ends with the text TranslatableTextUnitCharacterCount.

The actual translatable text count can be obtained by subtracting all of the above categories from the TotalCharacterCount, with the exception of LeveragedMatchedCharacterCount, FuzzyMatchedCharacterCount and RepetitionMatchedCharacterCount. These last three categories can be used to qualify the translation count itself.

### 4.3.5 Auto Text Character Count Categories

The following auto text categories are applicable to text from unqualified (see clause 4.2.14.2) text units.

The following character counts will be provided by symbolic name:

SimpleNumericAutoTextCharacterCount:

- An accumulation of the character count for simple numeric values, e.g. 10.

ComplexNumericAutoTextCharacterCount:

- An accumulation of the character count for complex numeric values which include decimal and/or thousands separators, e.g. 10,000.00.

MeasurementAutoTextCharacterCount:

- An accumulation of the character count for identifiable measurement values, e.g. 10,50 mm. Measurement values take precedent over the above numeric categories. No double counting of these categories is allowed.

AlphaNumericAutoTextCharacterCount:

- An accumulation of the character count for identifiable alphanumeric words, e.g. AEG321.

DateAutoTextCharacterCount:

- An accumulation of the character count for identifiable dates, e.g. 25 June 1992.

TMAutoTextCharacterCount:

- An accumulation of the character count for identifiable trade marks, e.g. "Weapons of Mass DestructionTM".

The following is an example of the use of the extension mechanism for user defined auto text character counts defined in the clause 4.3.1:

x-OtherAutoTextCharacterCount:

- Other auto text character counts. The actual naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-' and ends with the text AutoTextCharacterCount.

### 4.3.6 Inline Element Count Categories

The following counts will be maintained for non-linking inline elements by symbolic name:

Translatable inline element count:

- TranslatableInlineCount:
  - The actual non-linking inline element count for unqualified (see clause 4.2.14.2) text units.

Please refer to clause 4.2.11 for a detailed explanation and examples for this category.

### 4.3.7 Linking Inline Element Count Categories

The following count will be maintained for inline elements by symbolic name:

Translatable linking inline element count:

- TranslatableLinkingInlineCount:
  - The actual linking inline element count for unqualified (see clause 4.2.14.2) text units.

Please refer to clause 4.2.12 for a detailed explanation and examples for this category.

### 4.3.8 Text Unit Counts

The following count will be maintained for inline elements by symbolic name:

Translatable linking inline element count:

- TranslatableLinkingInlineCount:
  - The actual linking inline element count for unqualified (see clause 4.2.14.2) text units.

Please refer to clause 4.2.12 for a detailed explanation and examples for this category.

### 4.3.9 Other Count Categories

The following other counts can be provided by symbolic name:

FileCount:

- The total number of files.

PageCount:

- The total number of pages.

ScreenCount:

- A count of the total number of screens.

x-OtherCountCategories:

- Other count categories. The actual naming of the attribute is up to the supplier and customer with the one requirement that they begin with the sequence 'x-'. This is an extension mechanism.

### 4.3.10 Project Specific Count Categories

The following project specific counts can be provided by symbolic name:

ProjectRepetionMatchedWordCount:

- The word count for text units that are identical within all files within a given project. The word count for the primary occurrence is not included in this count, only that of subsequent matches.

ProjectFuzzyMatchedWordCount:

- The word count for fuzzy matched text units within all files within a given project. The word count for the primary occurrence is not included in this count, only that of subsequent matches.

ProjectRepetionMatchedCharacterCount:

- The character count for text that is identical within all files within a given project. The character count for the primary occurrence is not included in this count, only that of subsequent matches.

ProjectFuzzyMatchedCharacterCount:

- The character count for fuzzy matched text within all files within a given project. The character count for the primary occurrence is not included in this count, only that of subsequent matches.

The ProjectRepetionMatched counts relate to identical text units that are repeated within a project. The First occurrence of the text unit is not counted. Only the second and subsequent occurrences are counted within the ProjectRepetionMatched counts. In a similar fashion the ProjectFuzzyMatched counts also relate to fuzzy matched counts that are repeated. The first base occurrence of the fuzzy text unit is not counted, only subsequent matches.

### 4.3.11 Conformance

A minimum conformance level will encompass the provision of the following categories of GMX-V for translation related tasks:

1) TotalWordCount.

2) TotalCharacterCount.

Over and above the minimum level of conformance for translation related tasks it is up to the tool supplier to provide the level of detail that is required from their product.

It is recommended that the full levels of detail are provided for both word and character counts, although it is acknowledged that this may depend on the capabilities of individual tools. For instance a given tool may not support auto text (see clause 4.2.17) and so would not be able to support auto text count categories (see clause 4.3.5).

For non translation related Global Information Management tasks there is no minimum level of conformance apart from the need to provide at least one count metric.

### 4.3.12 Validation

Any measurement standard must have a reference implementation as well as an authoritative body that tests and validates the measuring instruments. In the USA, this is provided by the National Institute of Standards and Technology. In order to be successful, GMX-V must provide for a certification authority that will (1) maintain reference documents with known metrics and (2) provide an online facility to test given XLIFF documents. In this way, both customers and suppliers can be safe in the knowledge that GMX-V provides an unambiguous and reliable way of quantifying a Global Information Management task.

The validity of both the embedded namespace version and the stand alone version of GMX-V metrics can be validated by means of any standard XML parser. The stand alone version cannot be validated regarding word and character counts as the text to be counted is not present.

## 4.4 General Structure

The GMX-V document structure is designed to exist as a namespace so that it can be embedded into any document.

GMX-V can also be used in a stand alone XML document that expresses the metrics for an individual file or resource, or for a whole project. When used in a stand alone document it is recommended that the '.gmx' extension is used.

GMX-V comprises the following elements:

Metrics:

- This is the top level element for GMX-V.

Stage:

- GMX-V counts can be maintained for each stage in the workflow.

Count-group:

- This is the main count group identifier. There are separate count-group elements for verifiable and non-verifiable counts.

Count:

- The individual categorization, units of measure and values are declared in count elements.

Project:

- The project element allows the grouping of counts for individual resource components into one project count.

Resource:

- The resource element is used to hold the counts for individual project resources when accumulating metrics for an overall project count.

GMX-V can be used to define the metrics for an individual file or resource, or for an accumulation of resources at the project level.

The following is an example of a GMX-V instance for a single file/resource:

```
<metrics:metrics version="1.0" source-language="en-GB" tool-name="XYZ Tool" tool-version="1.23">
    <metrics:stage phase="initial" date="20041218T13:06:52Z">
        <metrics:notes from="auser@company.com">
            Initial count based on source document.
        </metrics:notes>
        <metrics:count-group name="non-verifiable">
            <metrics:count type="x-TestingFilesOtherCount" value="99"/>
            <metrics:count type="x-DTPFilesOtherCount" value="99"/>
            <metrics:count type="ScreenCount" value="99"/>
        </metrics:count-group>
        <metrics:count-group name="verifiable">
            <metrics:count type="TotalWordCount" value="99"/>
            <metrics:count type="TotalCharacterCount" value="99"/>
            <metrics:count type="TranslatableLinkingInlineCount" value="99"/>
        </metrics:count-group>
    </metrics:stage>
</metrics:metrics>
```

The following is an example of a GMX-V instance for a project within a stand alone document, where the use of the metrics namespace in not required:

```
<metrics version="1.0" source-language="en-GB" tool-name="XYZ Tool" tool-version="1.23">
  <project identifier="Project ABC">
    <resource identifier="file-1.abc">
      <stage phase="initial" date="20041218T13:06:52Z">
        <notes from="auser@company.com">
            Initial count based on source document for the first resource.
        </notes>
        <count-group name="non-verifiable">
            <count type="x-TestingFilesOtherCount" value="99"/>
            <count type="x-DTPFilesOtherCount" value="99"/>
            <count type="ScreenCount" value="99"/>
        </count-group>
        <count-group name="verifiable">
            <count type="TotalWordCount" value="99"/>
            <count type="TotalWordCount" value="99"/>
            <count type="ProjectRepetionMatchedWordCount" value="99"/>
            <count type="ProjectFuzzyMatchedWordCount" value="99"/>
            <count type="ProjectRepetionMatchedCharacterCount" value="99"/>
            <count type="ProjectFuzzyMatchedCharacterCount" value="99"/>
        </count-group>
      </stage>
     </resource>
    <resource identifier="file-2.abc">
      <stage phase="initial" date="20041218T13:06:52Z">
        <notes from="auser@company.com">
            Initial count based on source document for the first second resource.
        </notes>
        <count-group name="non-verifiable">
            <count type="x-TestingFilesOtherCount" value="99"/>
            <count type="x-DTPFilesOtherCount" value="99"/>
            <count type="ScreenCount" value="99"/>
        </count-group>
        <count-group name="verifiable">
            <count type="TotalWordCount" value="99"/>
            <count type="TotalWordCount" value="99"/>
            <count type="ProjectRepetionMatchedWordCount" value="99"/>
            <count type="ProjectFuzzyMatchedWordCount" value="99"/>
            <count type="ProjectRepetionMatchedCharacterCount" value="99"/>
            <count type="ProjectFuzzyMatchedCharacterCount" value="99"/>
        </count-group>
      </stage>
     </resource>
   </project>
</metrics>
```

Please note that there are a variety of count types of which only some basic examples are present in the examples above. A full list of count types is provided in the specification of the <count> element type attribute.

There are four type attributes that are specific to project specific counts:

1) ProjectRepetionMatchedWordCount.

2) ProjectFuzzyMatchedWordCount.

3) ProjectRepetionMatchedCharacterCount.

4) ProjectFuzzyMatchedCharacterCount.

The ProjectRepetionMatched counts relate to identical text units that are repeated within a project. The First occurrence of the text unit is not counted. Only the second and subsequent occurrences are counted within the ProjectRepetionMatched counts. In a similar fashion the ProjectFuzzyMatched counts also relate to fuzzy matched counts that are repeated. The first base occurrence of the fuzzy text unit is not counted, only subsequent matches.

### 4.4.1 Metrics Element

The <metrics> element is the top level of the hierarchy. It signals the start of the metrics namespace DOM tree. Its direct children are one or more <stage> elements, or one or more <project> elements when exchanging project based metrics. It is possible to maintain metrics for one or more stages, or one or more projects.

### 4.4.2 Project Element

The <project> element is used to accumulate the individual <resource> elements for a specific count when exchanging metrics for a whole project. This element is only used for project based counts.

### 4.4.3 Resource Element

The <resource> element holds the individual <stage> elements with the specific counts when exchanging metrics for a single file in a project. This element is only used for project based counts.

### 4.4.4 Stage Element

The <stage> element is used hold the <count-group> elements for a specific count stage, as well as one or more optional <notes> elements.

### 4.4.5 Notes Element

The <notes> element is used hold optional comments about the metrics stage.

### 4.4.6 Count Group Element

The <count-group> element is used to contain verifiable or non-verifiable <count> elements.

### 4.4.7 Count Element

The individual <count> elements hold the values of the count and identify the type of count.

## 4.5 Detailed Specification

### 4.5.1 GMX-V Namespace Declaration

The GMX-V document structure can exist either as a namespace within another document such as an XLIFF document, or within a standalone file. If GMX-V data is embedded within another document then a namsepace declaration will be required. The mandated namespace to be used is 'metrics'.

The GMX-V namespace declaration has the following form:

```
xmlns:metrics="urn:lisa-metrics-tagshttp://www.etsi.org/lis/gmx-v/2.0"
```

If the GMX-V document exist within its own file, then by convention the file should have a 'gmx' extension. No namespace declaration is required in this instance.

## 4.5.2 Elements

Elements <metrics>, <project>, <resource>, <stage>, <notes>, <count-group>, <count>

### 4.5.2.1 Metrics Element

The topmost GMX-V element has the following format:

**<metrics>**

The <metrics> element encloses all the other GMX-V elements of the document.

Required attributes:

- version - the fixed GMX-V current version identifier, currently "1.0".
- source-language - the language in which the document is authored.
- tool-name - the name of the tool that generated the metrics.
- tool-version - the version identifier of the tool that generated the metrics.

Optional attributes:

- target-language - the target language for the document. Only relevant if any translation memory matching has been done for a particular target language.
- reference - a reference to an external file or identifier if the metrics are not embedded in the file that is being counted.

Contents:

- One or more <stage> elements, or one or more <project> elements.

### 4.5.2.2 Project Element

The optional project element has the following format:

**<project>**

The <project> element can be used to group individual <resource> counts to provide a metric count for a complete project.

Required attributes:

- NONE

Optional attributes:

- identifier - the project identifier.
- One or more <resource> elements.

### 4.5.2.3 Resource Element

The resource element has the following format:

**<resource>**

The <resource> element contains the <stage> count elements for each resource.

Required attributes:

- identifier - the resource identifier, e.g. the file name of the resource.

Optional attributes:

- NONE
- One or more <stage> elements.

#### 4.5.2.4 Stage Element

The Stage element has the following format:

**<stage>**

Required attributes:

- phase - the identifier for the stage.
- date - the date that the stage count was created.
- source-language - the source language for the stage.

Optional attributes:

- target-language - the target language for the stage.

Contents:

- One or more <count-group> elements.

#### 4.5.2.5 Notes Element

The Notes element has the following format:

**<notes>**

Required attributes:

- NONE

Optional attributes:

- from - the email address or other identifier of the creator.
- date - the date that the notes element was created.

Contents:

- Comments text, no elements.

#### 4.5.2.6 Count Group Element

The Count Group element has the following format:

**<count-group>**

Required attributes:

- name - the count group name. This will have two possible values: verifiable or non-verifiable.

Optional attributes:

- state - the count group state. This can be used to create count groups for different states during translation.

Contents:

- one or more <count> elements.

#### 4.5.2.7 Count Elements

The Count element has the following format:

**<count>**

Required attributes:

- type - the count type.
- value - the quantity value.

Optional attributes:

- category - the fuzzy match count category, e.g. "93-95".

Contents:

- EMPTY.

### 4.5.3 Attributes

This clause lists the attributes used in the metrics elements. An attribute is never specified more than once for each element. Along with some of the attributes are the "Recommended Attribute Values". Values for these attributes are case sensitive. These lists are purely informative; the goal is to specify a preferred syntax so tools can have some level of compatibility.

| | |
|---|---|
| attributes | category, date, from, identifier, name, phase, reference, source-language, state, target-language, tool-name, tool-version, type, value, version |

#### 4.5.3.1 GMX-V Attribute

**category**

The category of fuzzy match. This is the percentage category within which the match falls, e.g. "99-95".

Value description:

- The fuzzy match category value.

Default value:

- Undefined

Used in:

- <count>.

**date**

The date attribute indicates when a given element was created or modified.

Value description:

- Date in ISO 8601 [6] Format. The recommended pattern to use is: YYYYMMDDThhmmssZ
- Where: YYYY is the year (4 digits), MM is the month (2 digits), DD is the day (2 digits), hh is the hours (2 digits), mm is the minutes (2 digits), ss is the second (2 digits), and Z indicates the time is UTC time. For example:

```
date="20020125T210600Z"
is January 25, 2002 at 9:06pm GMT
is January 25, 2002 at 2:06pm US Mountain Time
is January 26, 2002 at 6:06am Japan time
```

Default value:

- Undefined.

Used in:

- <stage>

**identifier**

The project or resource identifier.

Value description:

- The identifier for the project or resource.

Default value:

- Undefined.

Used in:

- <project>, <resource>.

**from**

The email address or other identifier of the creator of a given notes element.

Value description:

- The identifier of the creator of this notes element.

Default value:

- Undefined.

Used in:

- <notes>.

**name**

The name of the count-group.

Value description:

- Must have the value verifiable or non-verifiable.

Default value:

- Undefined.

Used in:

- <count-group>.

**phase**

The phase name of the stage.

Value description:

- Can have the value initial, final or user defined.

Default value:

- Undefined.

Used in:

- <stage>.

**reference**

An identifier to the external file or identifier if the metrics relate to an external file.

Value description:

- The file name or other identifier relating to an external file to which the metrics relate.

Default value:

- Undefined.

Used in:

- <metrics>.

**source-language**

The language for the main <metrics> element.

Value description:

- A language code as described in the RFC 4646 [7]. For more information see the section on xml:lang in the XML specification, and the erratum E11 (which replaces RFC 1766 [10] by RFC 4646 [7]).

Default value:

- Undefined.

Used in:

- <metrics>.

**state**

State - The optional count-group state qualifier. Separate count-group elements can be maintained for the different states of the target elements that correspond to the counted source element content in an XLIFF file.

Value description:

- The pre-defined values are based on the state attribute values from the XLIFF specification document.

| Value | Description |
|---|---|
| final | The count-group for XLIFF trans-units with target elements with a status attribute of 'final'. |
| needs-adaptation | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-adaptation'. |
| needs-l10n | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-l10n'. |
| needs-review-adaptation | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-review-adaptation'. |
| needs-review-l10n | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-review-l10n'. |
| needs-review-translation | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-review-translation'. |
| needs-translation | The count-group for XLIFF trans-units with target elements with a status attribute of 'needs-translation'. |
| new | The count-group for XLIFF trans-units with target elements with a status attribute of 'new'. |
| signed-off | The count-group for XLIFF trans-units with target elements with a status attribute of 'signed-off'. |
| translated | The count-group for XLIFF trans-units with target elements with a status attribute of 'translated'. |

- In addition, XLIFF user-defined values can be used with this attribute. A user-defined value must start with an "x-" prefix.

Default value:

- Undefined.

Used in:

- <count-group>.

**target-language**

The target language for the main <metrics> element.

Value description:

- A language code as described in the RFC 4646 [7]. For more information see the section on xml:lang in the XML specification, and the erratum E11 (which replaces RFC 1766 [10] by RFC 4646 [7]).

Default value:

- Undefined.

Used in:

- <metrics>.

**tool-name**

The identifier of the tool used to create the metrics.

Value description:

- The name of the tool used to perform the metrics count.

Default value:

- Undefined

Used in:

- <metrics>.

**tool-version**

The version identifier of the tool used to perform the metrics count.

Value description:

- The version identifier of the GMX-V count tool.

Default value:

- Undefined

Used in:

- <metrics>.

**type**

The count type.

Value description:

- Can have any of the following values, or a user defined type beginning with x- as defined in clause 4.3.1:
  - ScreenCount.
  - FileCount.
  - PageCount.
  - TextUnitCount.
  - TotalWordCount.
  - WhiteSpaceCharacterCount.
  - OverallCharacterCount.
  - PunctuationCharacterCount.
  - AlphanumericOnlyTextUnitWordCount.
  - MeasurementOnlyTextUnitWordCount.
  - NumericOnlyTextUnitWordCount.
  - ExactMatchedWordCount.
  - LeveragedMatchedWordCount.
  - RepetitionMatchedWordCount.
  - ProtectedWordCount.
  - ProjectRepetionMatchedWordCount.
  - FuzzyMatchedWordCount.
  - ProjectFuzzyMatchedWordCount.
  - TotalCharacterCount.
  - AlphanumericOnlyTextUnitCharacterCount.
  - MeasurementOnlyTextUnitCharacterCount.
  - NumericOnlyTextUnitCharacterCount.
  - ExactMatchedCharacterCount.

- LeveragedMatchedCharacterCount.
- RepetitionMatchedCharacterCount.
- ProjectRepetionMatchedCharacterCount.
- FuzzyMatchedCharacterCount.
- ProjectFuzzyMatchedCharacterCount.
- ProtectedCharacterCount.
- SimpleNumericAutoTextWordCount.
- ComplexNumericAutoTextWordCount.
- MeasurementAutoTextWordCount.
- AlphaNumericAutoTextWordCount.
- DateAutoTextWordCount.
- TMAutoTextWordCount.
- SimpleNumericAutoTextCharacterCount.
- ComplexNumericAutoTextCharacterCount.
- MeasurementAutoTextCharacterCount.
- AlphaNumericAutoTextCharacterCount.
- DateAutoTextCharacterCount.
- TMAutoTextCharacterCount.
- TranslatableInlineCount.
- TranslatableLinkingInlineCount.

Default value:

- Undefined.

Used in:

- <count>.

**value**

The numeric value of the count record.

Value description:

- The value of this count.

Default value:

- 0

Used in:

- <count>.

**version**

The current GMX-V version number.

Value description:

- The version number of this metrics document:

Fixed value:

- 1.0

Used in:

- <metrics>.

# Annex A (normative): GMX-V Document Structure

The following figure shows the possible structure as a tree. Each element is followed by notation indicating its possible occurrence according to the corresponding legend.

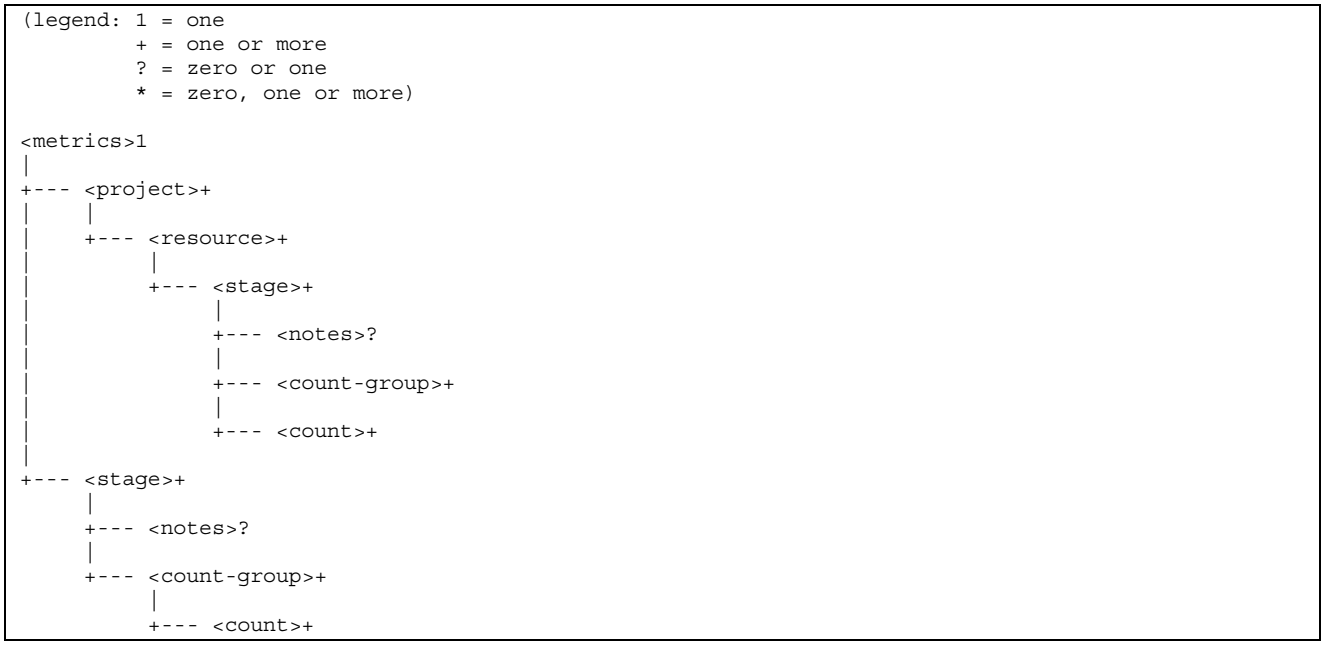


**Figure A.1: GMX-V Document Structure**

# Annex B (informative): GMX-V Schema

The XML schema for GMX-V is available at: http://www.xtm-intl.com/manuals/gmx-v/gmx-v.xsd.

# Annex C (informative): Authors & contributors

The following people have contributed to the present document:

**Rapporteur:**
Andrzej Zydroń <azydron@xml-intl.com>

**Other contributors:**
Arle Lommel alommel@gala-global.org

# Annex D (informative): Bibliography


- IANA Names for Character Sets. IANA (Internet Assigned Numbers Authority), October 2011.

NOTE: Available at http://www.iana.org/assignments/character-sets

- ISO 639-2: "Codes for the representation of names of languages".

NOTE: Available at http://www.loc.gov/standards/iso639-2/langcodes.html

- ISO 3166: "Codes for the representation of names of countries and their subdivisions".

NOTE: Available at http://www.iso.org/iso/country_codes

- SRX 2 Specification. LISA (Localization Industry Standards Association), 26 Feburary 2007.
- OASIS Translation Web Services TC.

NOTE: Available at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=trans-ws

- W3C Recommendation 26 November 2008 EeXtensible Markup Language (XML) 1.0 (Fifh Edition)".

NOTE: Available at http://www.w3.org/TR/REC-xml/

- W3C Recommendation 8 December 2009: "Namespaces in XML 1.0 (Third Edition)".

NOTE: Available at http://www.w3.org/TR/REC-xml-names/.

- A .Net (C#) implementation of the LISA GMX/V standard.

# History

| Document history | | |
|---|---|---|
| V2.0.0 | July 2012 | Publication |
| | | |
| | | |
| | | |
| | | |