
LISA
The Localization Industry
Standards Association
Erasing borders. Respecting boundaries.

TBX-Basic

Table of Contents

1. Document information

- Owner: LISA Terminology Special Interest Group (SIG)
- http://www.lisa.org/term
- Last updated: January 23, 2009
- Nature of last update - added source of context as a data category
- Document status: Final
- Send enquiries to: tbxbasic@lisa.org

Conventions used in this document

- **bold** - used for the human readable names of data categories
- *italics* - used for sample terms
- `courier font` - used for the xml elements and attributes

Terms used in this document

For the purposes of this document, the following terms are used with the meaning indicated below.

data category
result of the specification of a given data field (ISO 1087-2:2000), i.e. a type of data field, such as **definition** (ISO 16642)
Note: A data category can be considered a type of data element. However, the concept of a data category is independent of its implementation in an XML environment (e.g. using an element or attribute)[1].

terminology resource
file, database, or other collection, containing terms and information about terms

terminological markup language
TML
XML application for describing a terminological data collection conforming to the constraints expressed in ISO 16642 (Terminological Markup Framework) (ISO 16642)
Note: TBX-Basic is a TML

Acknowledgements

LISA would like to thank the members of the LISA Terminology Special Interest Group (SIG) for developing TBX-Basic, a simple terminology markup language based on TBX. For more information about the Terminology SIG, including a list of its members, visit the following Web site:

The LISA Terminology SIG would like to thank Sue Ellen Wright, of Kent State University Institute of Applied Linguistics, for the content on copyright considerations, and Klaus-Dirk Schmitz, of Cologne University of Applied Sciences, for the picklist values for the **Term location** data category applicable to Windows environments.

2. Introduction

This document identifies the types of terminology data that are permissible in a terminology data collection that complies with TBX-Basic, and how this data shall be structured. TBX-Basic is recognized by the LISA Terminology Special Interest Group (SIG) as a terminological markup language for the localization industry. TBX-Basic is intended to be a lighter version of TBX (TermBase eXchange), which is now an ISO standard for terminology data (ISO 30042) . It is particularly suited to small or medium sized language industries. While the primary audience is localization service

1 Internationalization Tag Set (ITS) 1.0 - http://www.w3.org/TR/its/#def-datacat

providers (LSPs), the format is also suited for any language application that requires a lightweight approach to terminology management, such as some implementations of controlled authoring.

An inventory of types of data (called "data categories" by terminologists) for recording terminology has been defined in *ISO 12620:1999 - Computer applications in Terminology - Data categories*. However, language industries typically only require a small subset of the data categories that are defined in that standard. Studies have been conducted by LISA in order to determine what types of data the localization industry actually needs for managing terminology, and those studies have formed the basis for TBX-Basic[2].

TBX-Basic is a terminological markup language (TML) that allows a limited set of data categories and adheres to a basic entry structure. It is intended for terminology resources that are commonly developed to support translation and localization processes. The purpose of TBX-Basic is to formalize the translation and localization industry's needs for terminology markup in an XML standard, in order to increase the ability to exchange terminology resources between users and to use these resources in various computerized environments.

In addition to defining the data categories and entry structure for TBX-Basic, this document also includes some guidelines for using the data categories, and some general best practices for managing terminology.

TBX-Basic is not a standard. TBX-Basic is recognized as a TBX-compliant TML in ISO 30042.

For more information about TBX (ISO 30042), refer to the following Web site:
http://www.lisa.org/Term-Base-eXchange.32.0.html

For more information about TBX-Basic, refer to the following Web site:
http://www.lisa.org/TBX-Basic.926.0.html

Send any comments about TBX-Basic to: tbxbasic@lisa.org

3. Data categories

This section describes the data categories that are permissible in a terminology data collection that is compliant with TBX-Basic.

Mandatory data categories

To allow for very simple terminology files to be encoded in TBX-Basic, mandatory data categories were kept to a minimum. The only data categories that are mandatory in TBX-Basic are the data categories for the term itself, and for the language in which the term and its associated information is represented. Thus, only the `<term>` element, and the `xml:lang='xx-XX'` attribute on the `<langSet>` element, are mandatory.

Several of the remaining optional data categories are very important and should be included in a terminology data collection as a general best practice. The important data categories are described in the best practices parts of this document, and they include, for instance, definitions, contexts, part of speech, and subject fields. However, the most important of these which the LISA Terminology SIG wishes to emphasize is the data category **part of speech**. The **part of speech** data category is required for the following purposes:

- To differentiate homonyms, for instance, *port* is actually two terms in English: a noun, and a verb. In a terminology database, each would be recorded in its own entry. Without a part of speech value in the entry, it can be difficult to determine which term the entry represents.
- To permit machine processing. Without a part of speech value, it is impossible to carry out certain software-assisted tasks in an automated fashion, such as importing a set of entries into an existing database. Importing would require human intervention to properly insert homonyms into their correct place in the target collection. Without a part of speech value, it is impossible to apply grammatical filters to facilitate search and export of

2 (a) Terminology Survey Results (2001) - http://www.lisa.org/LISA-Terminology-Sur.460.0.html; (b) LISA Terminology Survey for the Localization Industry - http://www.lisa.org/LISA-Terminology-Sur.463.0.html; (c) LISA Terminology Management Survey 2005 - http://www.lisa.org/Terminology-Manageme.461.0.html

data for specific purposes. And it would be impossible to use the terminology in extended applications such as spell checking programs.

- To enable blind interchange. Blind interchange refers to the ability to receive a terminology file and integrate it into a target system, such as a Computer-Assisted Translation (CAT) tool, without having to contact the originator of the file in order to understand its contents. When there is no part of speech value, it becomes necessary to discuss many of the entries with the originator in order to disambiguate their content.

Further guidelines about the requirements for the part of speech are included in the *Compliance* section.

Data type

The data type column indicates what kind of data the element can contain. This value is usually taken from TBX, however, in several cases the data type in TBX-Basic is more precise than it is in TBX. This has been done to facilitate data exchange. For example, in TBX, the part of speech data category (`<termNote type="partOfSpeech">`) takes as its data type `plainText`. In TBX-Basic, the data type has been refined to picklist. By using a more constrained form of the data type, TBX-Basic complies fully with TBX.

The value `plainText` refers to data that can only contain text (PCDATA). The value `noteText` is richer than `plainText` in that `noteText` can have embedded elements in it such as for hilighting and formatting. So, for example, a discursive data category such as **Definition** uses `noteText` to allow for inline elements such as links to other terms. But a data category that takes simple content, such as the **Subject field** data category, supports only `plainText` in order to prevent any internal markup.

Levels and order of data categories

The order of the data categories in the *Table of Data Categories* is not significant. However, the order of data categories in a terminological entry *is* significant and is defined in the Core Structure DTD. For instance, a `<descrip...>` element cannot come before a `<termNote...>` element. Data categories are arranged in the terminological entry in three different levels:

C = termEntry (the entry or *Concept* level)
L = langSet (the language level)
T = tig (the term level)
* = any (any level)

The level or levels at which a data category can appear in an instance of a TBX-Basic file are indicated in the Level column of the table below. Only the levels of `<descrip...>` elements can be formally constrained by the XCS file. Note that the levels are not formally constrained by the RelaxNG schema that is referred to in the section *XML files and validation scenarios*.

Table of TBX-Basic data categories

The Name column is a general name of the data category that can be used as labels on database fields, and when generally referring to the data categories. The designations in the Name column can be modified according to your preference in individual databases. For instance, you could choose the field label Graphic instead of Image for a terminology database. The Representation column provides the XML representation of the data category, as taken from TBX (ISO 30042). The XML representation must not be modified.

Note that for simplicity purposes the closing tag is not shown in the Representation column, for instance, the complete representation of the first item would be `<descrip type="subjectField">...</descrip>` where the content of the element would be a Subject Field value expressed in plainText.

Name	ISO 12620	Representation	Data type	Level	Comment
Subject field	ISO12620A-04	<descrip type= "subjectField">	plainText	C	The use of a picklist instead of plainText is recommended.
Image	ISO12620A-050501	<xref type="xGraphic" target="file_location">	plainText	C	Reference (URI, URL, or local file path) to a graphic file that is external to the TBX document instance. The reference to the graphic file is recorded as the value of the target attribute of the <xref...> element. The type value "xGraphic" identifies this external reference as pointing to a graphic file. The content of the <xref...> element is the name or description of the file for display purposes. For example: <xref type="xGraphic" target= "bat.jpg"> cricket bat</xref>
Language	ISO12620A-100701	<langSet xml:lang='xx-XX'>	Language code	L	Mandatory. Attribute for the language section. The language code must be taken from ISO 639-1, ISO 639-2, or ISO 639-3 as recommended in BCP-47: http://www.rfc-editor.org/rfc/bcp/bcp47.txt
Note	ISO12620A-08	<note>	noteText	*	Any kind of note, including usage notes, explanations, instructions, and so forth.
Term	ISO12620A-01	<term>	basicText	T	Mandatory.
Source of term	ISO12620A-1019	<admin type= "source">	noteText	T	Source of the term.
External cross-reference	ISO12620A-101807	<xref type= "externalCrossReference" target="externalId">	plainText	C, T	Pointer to external reference information or explanatory text about the term or concept, such as a Wikipedia article.
Creation date					See next section, *Transactions*
Created by					See next section, *Transactions*
Last modified date					See next section, *Transactions*
Last modified by					See next section, *Transactions*
Term type	ISO12620A-0201	<termNote type= "termType">	picklist	T	Only one value is permitted for each term. Permissible values: fullForm acronym abbreviation shortForm variant phrase
Part of	ISO12620A	<termNote type=	picklist	T	Only one value is permitted for each term.

Name	ISO 12620	Representation	Data type	Level	Comment
speech	-020201	"partOfSpeech">			Permissible values: noun verb adjective adverb properNoun other **Note:** This data category is mandatory under certain conditions. Refer to the section *Compliance* for details. TBX uses *plainText* as the data type for the part of speech data category. The use of a picklist in TBX-Basic is in compliance with TBX, since a picklist is more constrained than *plainText*. The “other” value can be used for terms of the “phrase” type, which include multiple words having different part of speech values.
Gender	ISO12620A -020202	<termNote type= "grammaticalGender">	picklist	T	Permissible values: masculine feminine neuter other
Definition	ISO12620A -0501	<descrip type= "definition">	noteText	C, L	See the section *Definitions and contexts* for guidelines on what constitutes a definition.
Source of definition	ISO12620A -1019	<admin type="source">	noteText	C, L	The source of the definition. It is a recognized best practice to document the source of the definition. **Note:** This element shall occur in a <descripGrp> element so that it can be associated with a definition, as in the following example: <descripGrp> <descrip type="definition">This is a sample definition. </descrip> <admin type="source">Webster's </admin> </descripGrp>
Context	ISO12620A -0503	<descrip type= "context">	noteText	T	A sample sentence that contains the term. See the section *Definitions and contexts* for guidelines on what constitutes a context.
Source of context	ISO12620A -1019	<admin type="source">	noteText	T	The source of the context. It is a recognized best practice to document the source of the context. **Note:** This element shall occur in a <descripGrp> element so that it can be associated with a context, as in the following example: <descripGrp>

Name	ISO 12620	Representation	Data type	Level	Comment
					<descrip type="context">This is a sample context. </descrip> <admin type="source">New York Times</admin> </descripGrp>
Usage status	ISO12620A-020903	<termNote type= "administrativeStatus">	picklist	T	This data category is useful for controlled authoring and controlled translation purposes, to mark whether a term is approved or not recommended for use. Only one value is permitted for each term. Permissible values: preferred admitted notRecommended obsolete The above values are recommended as display values for ease of use. However, the values in the XCS file must be as follows, taken from TBX: preferredTerm-admn-sts admittedTerm-admn-sts deprecatedTerm-admn-sts supersededTerm-admn-sts
Geographi-cal usage	ISO12620A-020302	<termNote type= "geographicalUsage">	plainText	T	This data category should be implemented as a picklist whenever possible. If the values required correspond to countries, use the ISO 3166 country codes. If the values correspond to locales, use the codes from IETF RFC 4646 or its successor, as identified in IETF BCP 47.
Term location	n/a	<termNote type= "termLocation">	plainText	T	This data category refers to a location in the corpus where the term frequently occurs, such as a user interface object (in software), a packaging element, a component in an industrial process, and so forth. A picklist is recommended. The appendix includes a set of picklist values which are recommended for software user interface locations in a Windows environment.
Cross reference	ISO12620A-1018	<ref type= "crossReference" target="elementId">	plainText	C, T	A generic linking mechanism to point to another entry or to a term in another entry in the same TBX document instance.
Customer	ISO12620A-100301	<admin type= "customerSubset">	plainText	T	This data category can be used to identify terms that are required for specific customers.
Project	ISO12620A-100303	<admin type= "projectSubset">	plainText	T	This data category can be used to identify terms that are required for specific jobs or projects.

Transaction-related data categories

Transaction-related data categories are administrative in nature, and thus are often automatically generated by the terminology management application. They are also handled by a common set of elements (`transacGrp`, `transac`, `transacNote`, and `date`), therefore, for simplicity purposes they are grouped in this section rather than being listed separately in the previous table.

Dates

Two types of dates are permissible in TBX-Basic.

- Creation date - Date that the parent element was created.
- Last modified date - Date that the parent element was last modified.

Both are handled by the same ISO 12620 data category, **date (**ISO12620A-1001).Dates can occur at any level of the entry (C, L, T). The required format from ISO 8601 is: YYYY-MM-DD, where YYYY is the year, MM is the month, DD is the day.

Responsible persons

The **responsibility** data category (ISO12620A-100202) is used to record the names of individuals who have worked on the entry. Two types are permissible in TBX-Basic

- Created by - Person who created the parent node (entry, language, term, sub-term data fields).
- Last modified by - Person who last modified the parent node (entry, language, term, sub-term data fields).

The responsibility data category can occur at any level of the entry (C, L, T).

For these data categories, the way to distinguish whether the date or person pertains to a creation or modification activity is determined by the value of the `<transac>` element, which must be either *origination* or *modification*, as shown in the following two samples:

```
<transacGrp>
   <transac type= "transactionType">origination</transac>
   <transacNote type="responsibility" target="CA5365">John Harris</transacNote>
   <date>2008-05-12</date>
</transacGrp>

<transacGrp>
   <transac type="transactionType">modification</transac>
   <transacNote type="responsibility" target="CA5365">John Harris</transacNote>
   <date>2008-05-12</date>
</transacGrp>
```

Note that a `transacGrp` element can contain either one `transacNote` element, or one `date` element, or both.

4. Clarification of several data categories

This section describes how to use several of the data categories that require some further clarification. Using the data categories as specified below will help to ensure that your terminology data is consistent with other databases and with international standards. For data interoperability, databases that claim to be compliant with TBX-Basic must include, in fields that use the XML representations described in this document, the types and nature of data that are described for those XML representations in this document.

Term type picklist values

The **Term type** data category is optional. You do not have to specify a term type for a term. When a term has no term type specified, it is assumed to be an ordinary entry term which is not an abbreviation or a variant of another term nor does it have an abbreviation.

full form
The complete representation of a term for which there is an abbreviated form.

abbreviation
An abbreviated form formed by omitting letters from a longer form.
Example:
full form: adjective
abbreviation: adj

short form
An abbreviated form that includes fewer words than the full form.
Example:
full form: Intergovernmental Group of Twenty-four on Monetary Affairs
short form: Group of Twenty-four
full form: United States of America
short form: United States

acronym
An abbreviated form made up of the initial letters of the components of the full form or from the syllables of the full form.
Example:
full form: United Nations
acronym: UN
full form: Extensible Markup Language
acronym: XML

variant
An alternative form of a term other than an abbreviated form. Variants can include words that have an alternative spelling, punctuation, capitalization, word formation, or even a numeric representation.
Examples:
term: Yahoo
variant: Yahoo!
term: soft switch
variant: softswitch

phrase
Any group of two or more words that are frequently expressed together and that consist of more than one concept. The individual words in a phrase usually function in more than one grammatical category (part of speech) within the syntax of a sentence. Although a phrase comprises more than one concept, they are often stored in single concept entries in terminological databases to address a need to record their translations for end-users.
Examples:
send feedback
work offline

Note that there is no term type value of "synonym" or "translation." This is because all terms in a concept entry within the same language section are synonyms, and all terms in a concept entry in different language sections are assumed to be possible translations of each other.

Use the following examples as guidelines:

- The term *blue box*, which is a recycling box in Canada, does not have any abbreviated forms, and it is not an alternate form of another term, therefore you do not need to specify a **Term type** value.

- The term *cell phone* is a short form of the more correct form *cellular phone*. Therefore, *cellular phone* should have a **Term type** value of **full form** and *cell phone* should have a **Term type** value of **abbreviation**.
- The term *application program interface* and *application programming interface* are interchangeably used. However, the more common form of the term is *application programming interface*. The term *application program interface* could therefore be considered a type of spelling variant. You do not need to specify a**Term type** value for *application programming interface*, but for *application program interface* you should specify a **Term type** value of **variant**.

Usage status picklist values

The **Usage status** data category is used to differentiate between terms in a multi-term entry, for controlled authoring or translation urposes. That is, you use it when you have a set of synonyms, and you want the database to guide people to use certain terms and not others.

preferred
Indicates the term, among a set of synonymous terms, that is the most recommended for use.

admitted
Indicates that the term is acceptable for use.

not recommended
Indicates that you should not use the term.

obsolete
Indicates terms that are no longer used, usually because a more modern term has replaced it.

5. The structure of a TBX-Basic entry

Regarding entry structure, like TBX, TBX-Basic complies with ISO 16642. However, for TBX-Basic, certain restrictions have been placed on the entry structure to simplify it and to make it reflect the selection of data categories. For instance, it is not necessary to allow, in the TBX-Basic entry structure, nodes that are provided in TBX for data categories that are not actually supported in TBX-Basic. Simplifying the entry structure, by adding restrictions rather than actually modifying the core structure, makes interchange easier, while continuing to maintain full compliance of TBX-Basic to TBX. The following sections describe the hierarchical entry structure of TBX-Basic.

Concept level

The concept level of an entry refers to the elements whose immediate parent is or can be <termEntry>. They are, in this order:

1. descrip - used for subject fields and definitions. Subject fields should only occur at the concept level. Definitions can occur at other levels, as described in the following sections.
2. descripGrp - used instead of descrip to document a definition and its source. Contains: one descrip and one admin element. If the source of the definition is not required or available, use just a descrip.
3. transacGrp - used for administrative information, such as the date that the entry was created and the name of the person who created it. Contains: one transac, and either one or both of transacNote, and date.
4. note - any concept-level note information.
5. ref - used for an internal reference, with the target attribute pointing to the concept ID of another entry.
6. xref - used for an external cross-reference, such as a URL, or to point to an external graphic file.

Language level

The language level of an entry refers to the elements whose immediate parent is or can be <langSet>. They are, in this order:

1. descrip - used for a definition that one wishes to document at the language level. This position therefore allows for definitions in different languages.
2. descripGrp - used instead of descrip to document a definition and its source. Contains: one descrip and one admin element. If the source of the definition is not required or available, use just a descrip.
3. transacGrp - used for administrative information about the language, such as the date that the language section was created and the name of the person who created it. Contains: one transac, and either one or both of transacNote, and date.
4. tig - a nesting element for the term level elements. Each tig contains information about one term.

Term level

The term level of an entry refers to the elements whose immediate parent is or can be <tig>. They are, in this order:

1. term - contains the term.
2. termNote - contains information about the term, such as the part of speech, or term type value.
3. descrip - used at this level only for the context sentence. Do not use this element to record a definition at the term level.
4. descripGrp - used instead of descrip to document a context and its source. Contains: one descrip and one admin element. If the source of the context is not required or available, use just a descrip.
5. admin - used to document the source of the term, or a customer or project that the term is associated with.
6. transacGrp - used for administrative information about the term, such as the date that the term was added to the entry and the name of the person who added it. Contains: one transac, and either one or both of transacNote, and date.
7. note - any note about the term or any of the term-related data categories.
8. ref - used for an internal reference, with the target attribute pointing to the ID of a term in another entry.
9. xref - used for an external cross-reference providing term-related information, such as a URL.

Backmatter

The backmatter of a TBX-Basic file is significantly simplified compared to TBX. It is only used to record the names and contact information for people who are responsible for creating or updating the terminology entries. The following is a sample of the markup allowed in the backmatter. As with TBX, the values of the type attribute for the <item> elements are taken from the vCard standard; more values are available, such as to record telephone numbers.

```
<back>
      <refObjectList type="respPerson">
            <refObject id="US5001">
                  <item type="fn">Jane Doe</item>
                  <item type="email">jane_doe@mymail.com</item>
                  <item type="role">approver</item>
            </refObject>
            <refObject id="US5002">
                  <item type="fn">John Smith</item>
                  <item type="email">john_smith@mymail.com</item>
                  <item type="role">inputter</item>
            </refObject>
      </refObjectList>
</back>
```

6. Differences between the core structure of TBX and of TBX-Basic

The core structure of TBX-Basic complies fully with the core structure of TBX. Both are expressed as a DTD file. However, the core structure of TBX-Basic is slightly simpler than that of TBX. Therefore, a customized version of the TBX DTD has been provided for TBX-Basic. It has the non-supported elements commented out, and brief comments added to explain the differences. The following lists the key differences:

- TBX supports both <tig> and <ntig> for term information groups. TBX-Basic only supports <tig>.
- Documenting term components (the individual parts of terms) is not supported in TBX-Basic. Therefore, the following elements are not supported: <termComp>, <termCompList>, <termCompGrp>, and <termGrp>
- TBX-Basic does not support the following grouping elements and their child elements: <adminGrp>, <termNoteGrp>, <itemSet> and <itemGrp>. Of the grouping elements, only <descripGrp> and <transacGrp> are allowed.
- In TBX-Basic, the <descripGrp> element is only used to associate a source to a definition or to a context. The following child elements are not supported: <descripNote>, <admin>, <adminGrp>, <note>, <ref>, <xref>.
- In TBX-Basic, the attribute values "DCSName" and "XCSContent" are not supported on the paragraph tag in the <encodingDesc> element.

7. Compliance

A terminology resource (database, file, or repository) is compliant with TBX-Basic if it meets all of the following conditions:

- It validates, without errors, against the TBX-Basic Core Structure DTD, which is available at the following Web site: http://www.lisa.org/fileadmin/standards/tbx_basic/TBXBasiccoreStructV02.dtd. It also must validate against the TBX Core Structure DTD, which is available at the following Web site: http://www.lisa.org/fileadmin/standards/tbx/TBXcoreStructV02.dtd
- It uses only the data categories that are defined in Section 3 and in the TBX-Basic XCS file (See Appendix 2), and uses them according to their descriptions in this document in terms of the nature and type of data.
- The data categories are inserted at the correct level of the entry structure as specified in Section 3.
- It respects the usage guidelines and best practices outlined in this document.
- Each entry contains at least one language section (<langSet xml:lang='xx-XX'>) and at least one **Term** (<term>)
- One of the following conditions has been met:
  - If the resource is intended to be submitted to any form of machine processing (see definition below), each term level (<tig>) has a **Part of speech** explicitly indicated through a <termNote type="partOfSpeech"> element.
  - If the resource is only intended for human consultation, the Part of speech may be omitted if either a Definition or a Context is provided.

8. Validating TBX-Basic document instances

Note: The files discussed in this section are available from the LISA Web site. The exact locations are given in Appendix 2.

The core structure of a TBX file is defined in the DTD `TBXcoreStructv02.dtd`. It is published in Annex A of the TBX specification (ISO 30042). The core structure of TBX-Basic is slightly simpler than the core structure of TBX, and therefore, it has its own DTD: `TBXBasiccoreStructV02.dtd`. This TBX-Basic DTD is a subset of the TBX DTD and therefore fully complies with it.

The set of data categories used in a terminological markup language is called a Data Category Selection (DCS). The DCS for TBX and TBX-Basic is defined in an XML file called an Extensible Constraints Specification (XCS) file. The default XCS file for TBX is called `TBXXCSV02.xcs`. TBX-Basic, comprising a subset of the default DCS, requires its own XCS file, called `TBXBasicXCSV02.xcs`. The XCS files are validated by using the DTD `tbxxcsdtd.dtd`.

ISO 30042 includes the Core Structure DTD, the XCS DTD, and the default TBX XCS files as appendices. LISA distributes the TBX standard free of charge.

It should be noted that, since the XCS formalism as a second layer of constraints on top of the Core Structure DTD is not a standard XML practice, off-the-shelf XML validators cannot read an XCS file. Therefore, you cannot use an off-the-shelf XML validator to check that your TBX-Basic document instance complies with an XCS file. An off-the-shelf XML validator can, however, check that the TBX-Basic document instance complies with the core-structure DTD. The XCS file is actually intended as a machine and human readable record of the data categories used in a TML, primarily for interchange purposes.

If you want to validate a TBX-Basic document instance against the core-structure DTD and the XCS file, you must use the TBX-Checker. The TBX-Checker is a validation tool specifically designed for checking a TBX-document instance against both the core-structure DTD *and* an XCS file. This tool is available free of charge from SourceForge (see Appendix 2).

Alternatively, you can express the constraints that are in the core-structure DTD and in the XCS file in an integrated schema, such as the RelaxNG schema supplied by LISA. Then you can use any off-the-shelf validator that supports your chosen schema language to validate your TBX-Basic document instance against all the constraints.

If you use the RelaxNG schema supplied by LISA, you need to use an XML validator that supports the RelaxNG and Schematron languages, such as oXygen.

Validation limitations of the TBX-Checker

Currently, the TBX-Checker cannot validate the level constraints for the elements `xref`, `ref`, and `admin`, because these level constraints are not expressable in the XCS file format. Therefore, these level constraints will not be checked by the TBX Checker. The TBX-Checker may be updated in the future to address this limitation.

Using the validation files to check your own TBX-Basic document instance

Before you can use the files supplied by LISA to validate your own TBX-Basic document instance, you need to declare the languages that are used in your TBX-Basic document in those files.

If you are using the TBX-Checker

You need to add the languages that are used in your TBX document to the header of the XCS file as shown below. The following sample shows how to document English and German, using a simple two-character code. You can use any language code format that is supported by IETF RFC 4646 or its successor.

```
    <languages>
        <langInfo>
            <langCode>en</langCode>
            <langName>English</langName>
        </langInfo>
        <langInfo>
            <langCode>fr</langCode>
            <langName>French</langName>
        </langInfo>
    </languages>
```

If you are using an off-the-shelf XML validator

If you are using the RelaxNG schema provided by LISA, you can use any off-the-shelf validator that supports the RelaxNG and Schematron languages, such as oXygen. In this case, in order to permit validation of the languages in your TBX file, you should modify the definition of the xml:lang attribute.

Currently, the definition of the xml:lang attribute indicates that it can contain any kind of text. This section starts on line 415.

```
   <define xmlns="http://relaxng.org/ns/structure/1.0" name="langSet.localattributes">
      <rng:attribute name="xml:lang">
         <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/annotations/
1.0">Indicates the language of the language section. This attribute is required for the
langSet element. See also the description on the martif element.</a:documentation>
         <text/>
      </rng:attribute>
   </define>
```

To declare the specific values of the xml:lang attribute that are used in your TBX document, follow the example below. This example shows how to declare English and German using a simple two-character language code, but you can use any format that complies with IETF RFC 4646 or its successor.

```
   <define xmlns="http://relaxng.org/ns/structure/1.0" name="langSet.localattributes">
      <rng:attribute name="xml:lang">
         <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/annotations/
1.0">Indicates the language of the language section. This attribute is required for the
langSet element. See also the description on the martif element.</a:documentation>
            <rng:choice>
                  <rng:value>en</rng:value>
                  <rng:value>de</rng:value>
            </rng:choice>
      </rng:attribute>
   </define>
```

9. Usage guidelines

The following are some general good practices to follow when creating terminology entries.

Recommended standards

Since terminology in a database will be reused for many purposes, you should follow international standards when creating the database and when entering data into the database.

In addition to ISO 30042 (TBX), ten of the most relevant terminology standards published by the International Standards Organization (ISO) are listed below:

- ISO 704:2000 - Terminology work – Principles and methods

- ISO 860:1996 - Terminology work – Harmonization of concepts and terms
- ISO 1087-1:2000 - Terminology work – Vocabulary – Part 1: Theory and application
- ISO 1951:1997 - Lexicographical symbols and typographical conventions for use in terminography
- ISO 1087-2:2000 - Terminology work – Vocabulary – Part 2: Computer applications
- ISO 10241:1992 - International terminology standards – Preparation and layout
- ISO 12200:1999 - Computer applications in terminology – Machine-readable terminology interchange
- ISO 12616:2002 - Translation-oriented terminography
- ISO 12620:1999 - Computer applications in terminology – Data categories
- ISO 16642:2003 - Computer applications in terminology – Terminological markup framework

In addition, ISO and many national standards organizations have published virtually hundreds of standardized monolingual and multilingual terminologies for a wide range of domains. Visit the ISO website (http://www.iso.org/) for more information.

Definitions and contexts

Although many localization service providers often collect and record minimal data in their terminological entries, such as just a source language term and a target language term, it is very useful to translators to have some type of conceptual description in the entry, so that they can be sure of the meaning of the term and whether or not the suggested translation is suitable for the text they are translating. In TBX-Basic, two data categories are provided for this purpose: **Definition** and **Context**.

A definition describes the meaning of a term, and a context is a sample sentence containing the term. Definitions provide meaningful explanatory information, whereas sometimes context sentences do not. It is recommended to include definitions in terminological entries, however, they are time-consuming to prepare. If you do not have the time to prepare definitions, include a context sentence in the entry. Context sentences can be automatically extracted by some term extraction software and some translation alignment software. It is widely recognized in the localization industry that a definition *or* a context sentence should be provided whenever possible. Thus for TBX-Basic, either a definition or a context sentence shall be present in all entries.

Context sentences

Not all context sentences are equally valuable. Some are very useful while others are not worth recording at all. Context sentences must always be "authentic," that is, they must be retrieved from an actual document or other communication medium; the terminologist should never create a context sentence. Context sentences should also be retrieved from original (not translated) documents, and they should be a complete sentence. Context sentences serve the following purposes:

- They prove that the term actually exists in real language.
- They can shed light on the meaning of the term.
- They can provide additional information of an encyclopedic nature about the term that is not in the definition (the who, why, when, where, how).
- They can illustrate how the term is used in discourse (collocations, register, etc.). For instance, a context sentence could alert the translator that the term is colloquial.
- They can provide grammatical information (such as gender), stylistic clues (such as hyphenation or capitalization) as well as alternate forms (abbreviations and so forth).
- The requirement to include a context sentence for the target language term helps to prevent the terminologist from simply translating the source language term, by requiring him or her to find an equivalent designation of the concept actually in use in the target language. This helps to ensure authenticity of the target language term and helps to reduce influence of the source language on the target language.

The following is a description of the various types of context sentences, arranged in descending order of preference for terminological entries.

1. Definitional context

A definitional context explains the meaning of a term. It has the information contained in a definition, but not necessarily the rigid form of a definition. This is sometimes called a terminological context because it is considered the best type for

terminological entries, since it can be used to help the terminologist write a definition. The following is an example of a definitional context for the term *Connection Manage*r :

> The Connection Manager is a utility for managing all of the registered connections to workspaces and repositories.

2. Encyclopedic context

An encyclopedic context provides some information about the meaning of the term, but not enough to fully understand the concept. For example, rather than telling you what something is, it may tell you where or how it is used. This is sometimes called an explanatory context. The following is an example of an encyclopedic context for the term *navigation controller* :

> The navigation controller allows navigation from panel to panel, data mapping across the process, and execution of operations in response to certain events.

3. Associative context

An associative context has little or no information about the meaning of the term, but reveals the subject area of the term by virtue of the other associated words in the context. The following is an example of an associative context for the term *transaction posting engine*, which enables us to determine that the term comes from the field of banking:

> For a bank teller application, access to the services of these entities (for example, to conduct a withdrawal transaction) requires delivery channels and a transaction posting engine that can handle the many tasks involved with transaction processing.

4. Metalinguistic context

A metalinguistic context describes some linguistic feature of the term. It is a sentence about the term itself, not about the concept. The following is an example of a metalinguistic context of the term *Blue box* :

> When the term Blue Box designates the recycling receptacle, it should be written with a capital B.

5. Discursive context

A discursive context only shows that the term has actually been used. It provides no explanatory or linguistic information. The following is a discursive context of the term *virtual tester* :

> Use the technique described here to enable multiple virtual testers.

Context sentences that are automatically extracted from a corpus should be reviewed manually before they are imported into a terminology database. Automated extractors usually do not have any selection criteria for determining which contexts are better than others. Furthermore, some automatically extracted contexts may not be full sentences, and may even just constitute the term itself, such as a simple user interface label. These poor context sentences should be replaced by better ones.

Definitions

The definition is arguably the data category that offers the most value to users apart from the terms themselves. The ISO 704:2000 standard, *Terminology work– Principles and methods*, provides comprehensive guidelines for writing terminological definitions. Following the ISO model, a definition is a sentence that explains the meaning of a term by a) identifying the class to which the term belongs and b) describing the characteristics that distinguish this term from other terms in this class.

Example:
pacemaker
implantable medical device that treats abnormal heart rhythms

As this example illustrates, a terminological definition is not the same as a lexicographical definition, which aims at providing a complete description of all senses of a term.

Subject fields

The majority of localization companies that manage terminology also collect categorical information, such as subject fields, product identifiers, and so forth. Therefore, for translation and localization business processes, this kind of information is important.

Well-organized and consistently applied categorical information can play an important role in disambiguation. Disambiguation refers to the process of clarifying the meaning of a term, where the term is used, or how the term is differentiated from other terms. Categorical information, combined with grammatical and contextual information, can help to narrow the scope of a term. In this sense, it has great practical value in the translation process.

Categorical information can also be used for filtering and sorting. Many localization companies need to filter their terminology data for use in specific translation environments or for reviewing purposes, and they use categorical fields to do this.

At least one field with categorical information per concept (entry) is recommended for terminology databases. A widely used categorical data field is "subject field" or "domain." ISO 12620 defines a subject field as "an area of human knowledge to which a terminological record is assigned."

Typically, categorical information like subject fields are chosen from a picklist. The use of a picklist ensures that the values are entered correctly and that they are standardized.

Multiple subject fields can be assigned to a concept, both on the same level (biology + chemistry) or in a hierarchical structure (technical -> heavy machinery -> wheel loaders).

10. Copyright Considerations

The relationship of terminologists to copyright and issues of intellectual property involves their rights and responsibilities both as authors and as users. They produce potentially copyrightable data collections and they use potentially copyrighted resources in documenting their collections. The copyrightability of types of information varies depending on the type of data in question:

- Terms and broader lexical units (collocations), with the exception of trademarks, service marks, and the like, are not subject to copyright and are viewed as in the public domain.
- Information per se (as opposed to specific texts used to express information) is not subject to copyright.
- Short excerpts from documents quoted as definitions, contexts, explanations, notes, etc. are usually acceptable under fair use rules, but should definitely be attributed in order to avoid the potential for charges of plagiarism, which is not the same thing as copyright infringement.
- Use of database material in European or international environments may be subject to the 1996 EU Database Directive, which could in some cases limit reusability for information derived from database resources.
- Issues involving fair use can arise in the event that terminologies or glossaries are published or sold. Legal advice is appropriate when contemplating the publication of any resource that utilizes any significant amount of quoted, copyrighted information.
- Photos and other graphics are subject to far more stringent copyright rules than is text.

Copyright covers a wide variety of genres and media. It applies to the form in which ideas are expressed and not to the ideas themselves, but the copyrighted material must be preserved on some sort of tangible medium to be valid. Copyright can be held individually or by corporate entities. Copyright is valid upon its preservation on any type of medium and does not require formal registration with official copyright authorities, although official registration is critical in the event of legal disputes.

Unless otherwise expressly cited when specifying scope of work, terminology resources created by translators and localizers remain the property of the creators and are not subject to contractual agreements involving the translated or

localized product. Terminologies and glossaries are compilations, in contrast to translations, which are derivative works. Parties to such contracts (vendors and clients alike) are wise to clarify ownership of such resources (along with translation memories) at the beginning of a contractual arrangement and to specifically cover these issues in their agreements. Clients who request or demand the production of terminology resources should be prepared to pay a separate fee for this service.

Vendors who wish to disseminate, reuse, or otherwise exploit terminology resources containing proprietary information obtained from clients should carefully weigh issues involving client confidentiality and good will before divulging such information to third parties.

Terminologists who work together in teams or who collaborate with clients and other third parties to compile terminologies should establish agreements outlining how the work should be done and where ownership lies in order to avoid future misunderstandings or disagreements.

Copyright laws vary from country to country, therefore for specific activities it may be advisable to consult national copyright legislation.

Appendix 1 - Term location picklist values for Windows user interface objects

The following picklist values are recommended for software user interface locations in a Windows environment. [3]

Note: In TBX-Basic, these should be written in camel case, such as "menuItem".

- Menu item
- Dialog box
- Group box
- Text box
- Combo box
- Combo box element
- Check box
- Tab
- Push button
- Radio button
- Spin box
- Progress bar
- Slider
- Informative message
- Interactive message
- ToolTip
- Table text
- User defined type

3. From the Dandelion project led by Klaus-Dirk Schmitz, of Cologne University of Applied Sciences.

Appendix 2 - Files and Resources

A **TBX-Basic package** is avaliable from LISA to help implementers and users learn about and use TBX-Basic. It can be downloaded from the following Web site.

http://www.lisa.org/TBX-Basic.926.0.html

At the time this report was written, the package contained the following resources, however, additional resources may be added at any time.

- TBX-Basic specification (the current document)
- TBX-Basic XCS file and the DTD required to validate it
- TBX-Basic Core-Structure DTD
- TBX-Basic integrated RelaxNG Schema
- Sample TBX-Basic document instance, showing various types of terminological entries
- Sample TBX-Basic document showing errors

The TBX-Basic files required for validation purposes are also available at the following URLs. These URLs are provided by LISA as persistent identifiers that can be used to point to the files directly on the LISA Web site from within software applications:

http://www.lisa.org/fileadmin/standards/tbx_basic/TBXBasiccoreStructV02.dtd

http://www.lisa.org/fileadmin/standards/tbx_basic/TBXBasicXCSV02.xcs

http://www.lisa.org/fileadmin/standards/tbx_basic/TBXBasicRNGV02.rng

http://www.lisa.org/fileadmin/standards/tbx/tbxxcsdtd.dtd

For more information about TBX, refer to the following Web site:

http://www.lisa.org/Term-Base-eXchange.32.0.html

LISA also provides persistent identifiers for TBX files:

http://www.lisa.org/fileadmin/standards/tbx/TBXcoreStructV02.dtd

http://www.lisa.org/fileadmin/standards/tbx/tbxxcsdtd.dtd

http://www.lisa.org/fileadmin/standards/tbx/TBXXCSV02.xcs

http://www.lisa.org/fileadmin/standards/tbx/TBX_RNGV02.rng

The TBX Checker is available from SourceForge:

http://sourceforge.net/projects/tbxutil/