Actas del II Congreso Internacional
de la Asociación Ibérica de Estudios de Traducción e Interpretación

XML en la gestión terminológica - Nuevos horizontes

Detlef Reineke
Universidad de Las Palmas de Gran Canaria

Como citar este artículo:

Reineke, Detlef (2005) «XML en la gestión terminológica – Nuevos horizontes», en Romana García, María Luisa [ed.] *II AIETI. Actas del II Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. Madrid, 9-11 de febrero de 2005*. Madrid: AIETI, pp. 343-355. ISBN 84-8468-151-3. Versión electrónica disponible en la web de la AIETI: <http://www.aieti.eu/pubs/actas/II/AIETI_DR_Xml.pdf>.

XML EN LA GESTIÓN TERMINOLÓGICA – NUEVOS HORIZONTES

Detlef Reineke
Universidad de Las Palmas de Gran Canaria

Introducción

En muchos ámbitos, la terminología supone un vehículo indispensable para una eficaz comunicación intralingual (entre especialistas, entre "producto" y usuario, etc.) e interlingual (traducción, localización). La gestión proactiva y el uso correcto de la terminología no sólo facilita la comprensión de contenidos, sino también agiliza los flujos de trabajo en proyectos de traducción o de localización evitando costosos procesos de corrección. Para que la terminología se pueda elaborar, gestionar, procesar y usar adecuadamente, existe una serie de herramientas electrónicas específicas como son las aplicaciones de Trados, STAR, Déjà Vu o SDL.

En la historia de las herramientas electrónicas de gestión terminológica ha habido constantes esfuerzos hacia una mejora en la representación de datos terminológicos y en su intercambio entre los diferentes sistemas. Desde la implementación de las primeras bases de datos terminológicas en los años sesenta del siglo XX en macroordenadores hasta las aplicaciones actuales para el PC, los estamentos interesados en la terminología han ido elaborando distintos estándares como, por ejemplo, el formato MARTIF (Machine-readable terminology interchange format) con el fin de superar las limitaciones impuestas por los formatos particulares que provocan volúmenes prohibitivos de trabajo manual en la conversión de datos.

En los últimos años, y surgido del proyecto SALT (Standards-based Access to Multilingual Lexicons and Terminology), la LISA (Localisation Industry Standards Association) promueve un nuevo estándar prometedor, el formato TBX (TermBase eXchange Format) basado en el lenguaje XML (eXtensible Markup Language). Este formato cuyo núcleo estructural está basado, a su vez, en el ya mencionado estándar MARTIF pretende garantizar la interoperabilidad entre los distintos sistemas y permitir el intercambio de datos sin pérdida de información (*losless roundtrip of information*).

MATER y MicroMATER

A mediados de los años sesenta se instalaron las primeras bases de datos terminológicos en macroordenadores, y a poco tiempo quedó patente que había que crear un estándar para el intercambio de datos terminológicos entre estos ordenadores (Melby et. Al., 2001). Así, en los años setenta se iniciaron los trabajos con el fin de desarrollar el primer formato de intercambio estandarizado que finalmente fue aprobado como ISO 6156 (1986): "MATER (Magnetic tape exchange for terminological/lexicographical records)". MATER fue diseñado para el intercambio de datos tanto terminológicos como lexicológicos por medio de cintas magnéticas de nueve pistas. No obstante, no llegó a implementarse a gran escala, ya que la aparición de la generación de los microordenadores (IBM compatibles y MacIntosh) a principios de los años ochenta precisaba de un nuevo enfoque. Por ello, el formato MicroMATER derivado de MATER se concibió para el intercambio de datos entre macro, mini y microordenadores que permitía, a su vez, la definición de estructuras de datos flexibles.

TEI

En 1987, unos filólogos fundaron el consorcio TEI (Text Encoding Initiative) con el objetivo de desarrollar métodos para la descripción de textos literarios y humanísticos mediante un lenguaje de marcas. El formato para la codificación de textos y para el intercambio de textos llamado TEI estaba basado inicialmente en SGML (Standard Generalized Markup Language = ISO 8879), pero fue adaptado recientemente a XML (eXtensible Markup Language). El enfoque del TEI resultaba igualmente atractivo a los creadores de MATER y MicroMATER, puesto que se presentaban obstáculos similares para el ámbito del intercambio de datos terminológicos como, por ejemplo, el problema de los conjuntos de caracteres o de la codificación de enlaces. Por este motivo, a principios de los años noventa se creó un grupo de trabajo en el seno del TEI con la participación de algunos creadores de MicroMATER, cuyo objetivo radicaba en desligarse de los formatos desarrollados hasta ese momento y en crear un estándar radicalmente nuevo. Con el fin de lograr una aceptación universal de este estándar, se acordó, tras un año de trabajo en el TEI, continuar el diseño de un formato de intercambio de datos terminológicos estandarizado bajo los auspicios de la ISO.

MARTIF

El grupo de trabajo ISO TC37/SC 3/ WG 3[1] prosiguió con los trabajos iniciados en el consorcio TEI para la creación de un formato de intercambio genérico, lo que terminó con la aprobación de la norma ISO 12200 (1999): "Machine-readable terminology interchange format (MARTIF) – Negotiated interchange". MARTIF que está basado en SGML, y que utiliza el conjunto de caracteres de 7 bits definido en la ISO 646, especifica las reglas para la modelación y representación de datos terminológicos.

Un documento MARTIF se compone de un prólogo y de una instancia del documento. El prólogo contiene una referencia a la DTD (*Document Type Definition*) que define la estructura del documento MARTIF, y que es utilizado por un *parser* SGML para su validación. Generalmente, la DTD de MARTIF es modular y consta de tres partes (*framework*, *body* y *sets*):

```
        I. Prolog
[1]     <!DOCTYPE martif PUBLIC "ISO 12200:1999//DTD for MARTIF (framework/ //EN" [
[2]     <!ENTITY % mtf-body "ISO 12200:1999//DTD for MARTIF (body //EN" >
[3]     <!ENTITY % mtf-ents "ISO 12200:1999//ENTITIES for MARTIF (sets //EN" > ] >

        II. Document instance
        …
```

Fig. 1: DTD para MARTIF

La DTD *framework* [1] determina la estructura global de un documento MARTIF y referencia dos elementos, el elemento *body* [2] y el elemento *sets* [3] con sus respectivas DTDs. La manipulación por separado del elemento *body* permite validar documentos, cuya macroestructura (*header*, *text*, *front*, *body*, *back*) coincide con la de un documento MARTIF, pero cuyo contenido del elemento *body* es distinto a la estructura del elemento *body* definido en la ISO 12200. En la DTD *sets* pueden definirse caracteres o signos adicionales no contenidos en la ISO 646 como, por ejemplo, caracteres cirílicos o griegos.

[1] ISO Technical Committee 37 (entonces "Terminlogy – principles and coordination", hoy "Terminology and other language resources"), Sub-Committee 3 (Computer applications), Working Group 3 (Data interchange)

En la DTD se declaran elementos, los atributos asignados a estos elementos, así como las interdependencias entre los elementos. La utilización de comodines (ENTITY) permite representar signos individuales, cadenas o archivos enteros mediante una abreviatura (véase fig. 2).

```
[4]   <!ELEMENT text          - -   (front?, body, back?) >
[5]   <!ELEMENT body          - -   (termEntry+) >
[6]   <!ENTITY %AuxInfo             ‘ descrip ¦ descripGrp ¦ admin ¦
                                    adminGrp ¦ ptr ¦ ref ¦ date ¦ note’ >
[7]   <!ELEMENT termEntry     - -   ((%AuxInfo;) *, (langSet ¦ tig ¦ ntig) +) >
```

Fig. 2: Elementos, atributos y comodines en MARTIF

En la fig. 2, al elemento *text* [4] está asignado (y subordinado) el elemento *body*. Los elementos *front* y *back* son opcionales. El elemento *body* encapsula, a su vez, uno o varios elementos *termEntry* [5]. Los ejemplos [6] y [7] demuestran la función del comodín. En los puntos del documento MARTIF, donde se inserta la abreviatura *%AuxInfo* [6], son válidos todos aquellos atributos definidos en la DTD, en este ejemplo los atributos *descrip, descripGrp, admin, adminGrp, ptr, ref, date* y *note*. En el ejemplo [7], estos atributos pueden ser utilizados entre los elementos *termEntry* y *langSet*.

En la instancia (fig. 3) se documenta la información sobre la base de datos en general y los propios términos. A la declaración [8] acerca del tipo de documento (*martif*) y del idioma utilizado en el documento para la descripción de los objetos (*lang=en*) le sigue el elemento *martifHeader* [9] que comprende informaciones relativas al conjunto de la base de datos terminológicos como, por ejemplo, el nombre del archivo MARTIF, la codificación utilizada o informaciones sobre las revisiones de la base de datos.

```
      I. Prolog
      II. Document instance
[8]     <martif lang=en>
[9]         <martifHeader> (Información sobre la base de datos) </martifHeader>
                <text>
                    <front></front>
[10]                <body> (entradas terminológicas) </body>
[11]                <back> (datos bibliográficos) </back>
                </text></martif>
```

Fig. 3: Instancia de un documento MARTIF

En el elemento *body* [10] se documenta cada una de las entradas terminológicas. El elemento *back* [11] acoge la información bibliográfica completa o referencias a objetos externos, a los que se puede acceder desde las entradas terminológicas por medio de hipervínculos.

Generalmente, una entrada terminológica (*termEntry*) consta de uno o varios bloques de lenguas (*langSet*) (véase fig. 4) que, a su vez, contienen uno o varios bloques de designaciones (*tig* o *ntig*[2]). En el elemento *termGrp* se documentan la propia designación (*term*), así como los datos relativos a la designación.

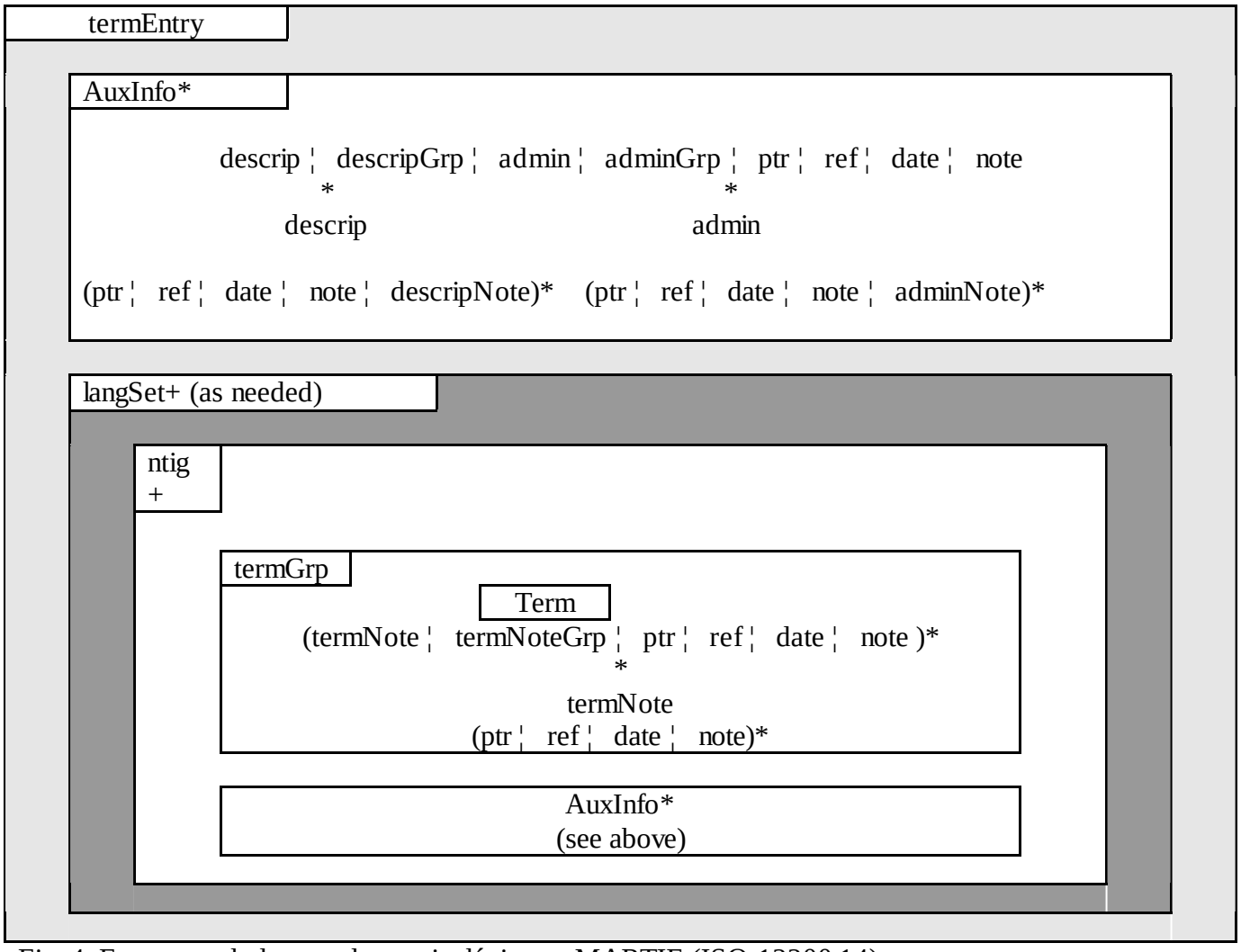

Fig. 4: Estructura de la entrada terminológica en MARTIF (ISO 12200:14)

[2] Si el bloque de designación no integra subniveles, se usa el elemento *tig* (*term information group*). De lo contrario, se utiliza el elemento *ntig* (*nested term information group*).

Representación y modelación de categorías de datos

Las categorías de datos terminológicas recomendadas para su uso en la estructura MARTIF están alistadas en el anexo A de la ISO 12200. La representación y los posibles valores de estas categorías de datos están estrechamente sintonizados con la ISO 12620 (1999): "Computer applications in terminology – Data categories".

En la modelación de la estructura de las entradas terminológicas las categorías de datos no cuelgan directamente de los nudos de la estructura básica de MARTIF, sino son asignadas a metacategorías de datos (*Generic Identifiers*) e instanciadas mediante un atributo *type*.

Position no.	Data category name	MARTIF data category representation (Full normalized form)	Value	Examples
A 2.2.3	grammatical number	<termNote type='grammaticalNumber>	Perm. instance	Common permissible instance include: m, f, n, other …
A 5.1	definition	<descrip type='definition'>	Text	…<descrip type='definition'>substance capable of holding materials together by adhesion</descrip> …

Fig. 5: Especificación de categorías de datos en MARTIF

Este formalismo permite una identificación unívoca de la información terminológica, así como de las interdependencias de las categorías de datos. Además, el uso de pocas categorías de datos garantiza una aplicación flexible de la norma a un número máximo de base de datos terminológicas existentes en la práctica.

El siguiente ejemplo ilustra la manera, en la que se expresan categorías de datos y sus correspondientes contenidos en la instancia de un documento MARTIF:

```
Subject field:          Materialbeschaffenheit
Term:                   Opazität
Part of speech:         Substantiv
Gender:                 feminin
Definition:             Maß für Lichtundurchlässigkeit
Source:                 DIN 6370:1996-05, S. 383

<martif lang=en><martifHeader>… </martifHeader>
<text>
  <body>
    <termEntry id='ID0000073578'
      <descrip type='subjectField'>Materialbeschaffenheit</descrip>
       <langSet lang=de>
         <ntig>
           <termGrp>
            <term>Opazit&auml;t</term>
              <termNote type='partOfSpeech'>Substantiv</termNote>
              <termNote type=' grammaticalGender'>feminin</termNote>
              <descripGrp>
                <descrip type=' definition'>Ma&szlig; f&uuml;r die
                    Lichtundurchl&auml;ssigkeit</descrip>
[11]          <ref type=' sourceIdentifier' target=' DIN-6370.1996-05>S. 383</ref>
              </descripGrp>
         </ntig>
       </langSet>
    </termEntry>
  </body>
</text></martif>
```

Fig. 6: Categorías de datos y sus correspondientes datos en MARTIF

Por motivos de minimización de redundancias, las entradas terminológicas no deberían documentar informaciones bibliográficas o hipervínculos completos. En estos casos son convenientes abreviaturas [11] que referencian los correspondientes datos documentados en el elemento *back* [13]. Caracteres diacríticos u otros signos especiales son expresados mediante representaciones sustituyentes definidos en el anexo D de la norma ISO 8879 (= SGML) [12].

```
        …
        <ntig>
          <termGrp>
[12]        <term>Opazit&auml;t</term>
            …
          <descripGrp>
            <descrip type=‘ definition’ >Ma&szlig; f&uuml;r die
              Lichtundurchl&auml;ssigkeit</descrip>
            <ref type=‘ sourceIdentifier’  target=‘ DIN-6370.1996-05>S. 383</ref>
          </descripGrp></ntig></termEntry>
        </body>
        <back>
          <refObjectList type=‘bibl’ >
            <refObject>
              <item id=‘ DIN-6370.1996-05’ >
[13]          <xref target=‘c:\bibl\normen\DIN-6370\DIN-6370-1996-05.doc</xref>
              </item>
            </refObject>
          </refObjectList>
        </text>
        </martif>
```

Fig. 7: Hipervínculo en MARTIF

Blind interchange

MARTIF es un estándar flexible que ayuda a intercambiar datos procedentes de base de datos terminológicas con distintas estructuras e independientemente del hardware y del software utilizados para su gestión. Esta flexibilidad hace imprescindible un acuerdo previo al intercambio de datos entre los gestores de las correspondientes bases de datos, dado que MARTIF define solamente la estructura de datos, así como las posibles categorías de datos, pero no fija el contenido de ciertas categorías de datos con valores predeterminados. Supongamos que el gestor A asigna a la categoría de datos *grammatical number* los valores "p" y "sg", mientras que el gestor B le asigna los valores "plural" y "singular". Un intercambio de datos sin acuerdo previo (*blind interchange*) sería imposible.

Con el fin de satisfacer la demanda de la industria lingüística por una mayor automatización de los procesos se consideró ampliar la ISO 12200 (MARTIF - Negotiated interchange) por una segunda parte llamado MARTIF - Blind interchange. Por una parte, la

idea consistía en definir también los valores predeterminados de las categorías de datos pertinentes. Por otra parte, era necesario adaptar MARTIF a XML y, de esta forma, superar otros obstáculos como el conjunto de caracteres insuficiente para la representación de lenguas con alfabeto no latino, así como las limitaciones derivados del SGML en cuanto al intercambio y a la representación de datos terminológicos a través de Internet.

TBX

TBX (TermBase eXchange Format) es un formato de intercambio que nació en el marco de un proyecto iniciado por el grupo OSCAR (Open Standards for Container/Content Allowing Re-user group) de la LISA y que se refinó en el marco del proyecto SALT (Standards-based Access Service to Multilingual Lexicons and Terminologies). TBX está basado en XML y se apoya en las normas ISO 12200, ISO 12620 y la recién aprobada ISO 16642 (2003) "Computer applications in terminology - Terminological markup framework (TMF)".

La norma ISO 16642 propone un marco metodológico genérico para la descripción de todo tipo de recursos lingüísticos y las reglas para la especificación de lenguajes de marcas terminológicos (Terminological Markup Language - TML). Los TML se comparan por medio de su especificación y se modelan (para el intercambio) con la ayuda de herramientas genéricas (GMT - Generic Mapping Tool) conservando la totalidad de la información terminológica.

Un TML definido por la norma es MSC (MARTIF with Specified Constraints) que permite describir el formato TBX. MSC está basado en la estructura de MARTIF (véase fig. 8 y 9) y define un subconjunto de categorías de datos de la ISO 12620, tal como estaba previsto para Blind-MARTIF.

Igual que en MARTIF, los elementos y la estructura de un documento TBX pueden definirse en una DTD. El problema de las DTD radica en que solamente permiten tipos de definiciones rudimentarios insuficientes para intercambios de datos complejos vía Internet. Además, las DTD no son expresadas en XML, por lo que son de poca utilidad para aplicaciones basadas en la Web. Por ello, las especificaciones TBX recomiendan uso de XML Schemas para la validación de documentos TBX y de sus especificaciones de categorías de datos (LISA, 2002). Al contrario de las DTD, los XML Schemas permiten definir valores

predeterminados de categorías de datos para un intercambio sin previo acuerdo, así como espacios de nombres (*namespaces*).

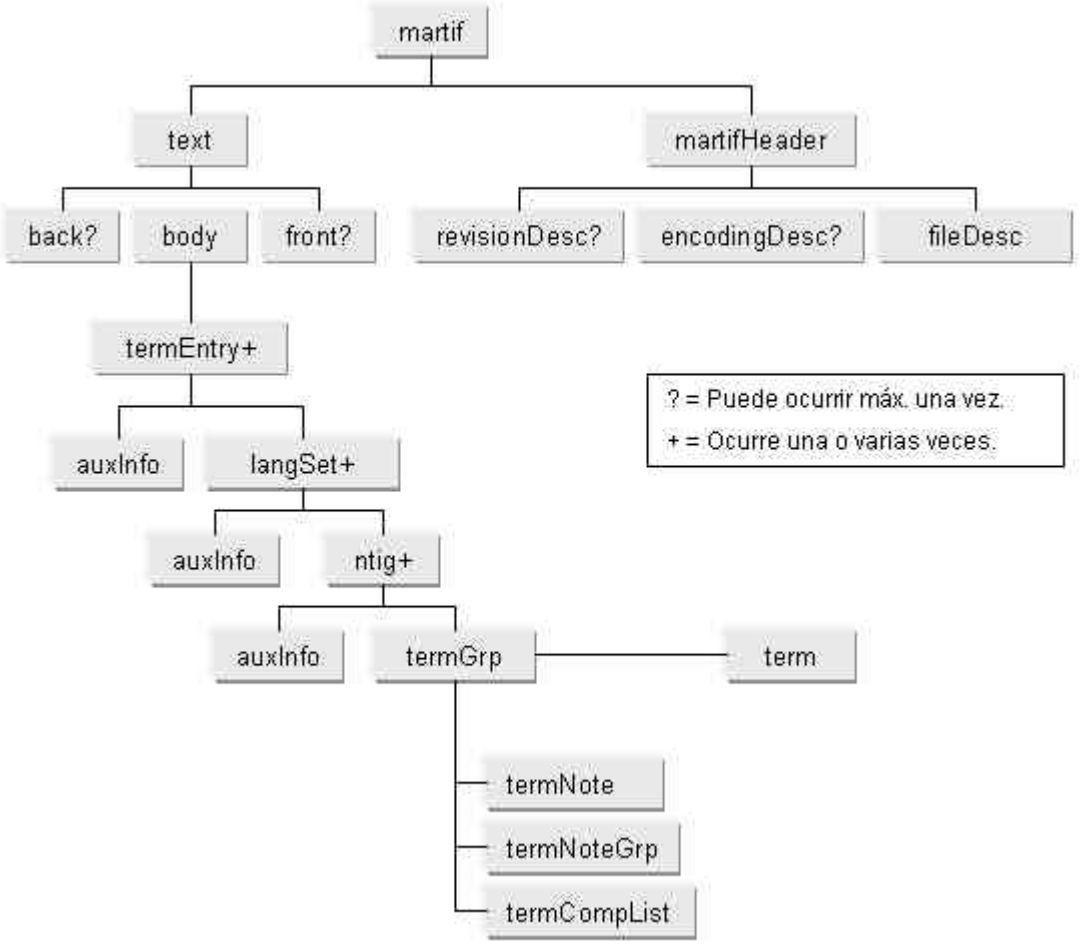

Fig. 8: Estructura TBX

Otra novedad de TBX con respecto a MARTIF consiste en la inclusión del elemento *termCompList* que permite documentar de forma adecuada los componentes de designaciones pluriverbales. La necesidad de la gestión explícita de estos componentes se da sobre todo en la traducción o en la redacción en lenguas analíticas.

La instanciación de las categorías de datos se realiza, como en MARTIF, mediante un atributo *type*.

```
<?xml version='1.0'?>
<!DOCTYPE martif SYSTEM "./TBXcoreStructureDTD-v-1-0.DTD">
<martif lang=en><martifHeader>… </martifHeader>
<text>
  <body>
    <termEntry id='ID0000073578'
        <descrip type='subjectField'>Materialbeschaffenheit</descrip>
        <langSet lang=de>
          <ntig>
            <termGrp>
              <term>Opazität</term>
                <termNote type='partOfSpeech'>n</termNote>
                <termNote type=' grammaticalGender'>f</termNote>
                <descripGrp>
                  <descrip type=' definition'>Maß für die
                    Lichtundurchlässigkeit</descrip>
                <ref type=' sourceIdentifier' target=' DIN-6370.1996-05>S. 383</ref>
                </descripGrp>
          </ntig>
        </langSet>
    </termEntry>
  </body>
</text>
</martif>
```

Fig. 9: Documento TBX

Panorama

Cambiar la estructura de una base de datos terminológicas para adaptarla a un formato estandarizado no sólo requiere conocimientos especializados, sino sobre todo una inversión en tiempo y dinero considerable. Los costes para la conversión de base de datos con centenas de miles de entradas terminológicas pueden alcanzar niveles prohibitivos y, a menudo, resulta muy difícil sensibilizar a los superiores o a los compañeros - muchas veces anclados en la filosofía del beneficio inmediato - de las ventajas a medio o largo plazo que aportan soluciones estandarizadas. La falta de estándares globalmente aceptados y longevos ha dificultado esta tarea aún más. A esto se añade que en la industria lingüística no se hallan

suficientes expertos para implementar los estándares desarrollados. Este hecho es un reto esencialmente para las universidades, en cuyas titulaciones correspondientes, de momento, apenas se contemplan contenidos que forman a los estudiantes para las tareas complejas exigidas. Es de esperar que, en el marco de la conversión hacia los criterios de la declaración de Bolonia, se introduzcan elementos formativos que palien las deficiencias actuales.

¿Cómo se presenta el futuro de TBX? En la industria de la traducción/localización ya se están utilizando con éxito formatos estandarizados basados en XML como TMX (Translation Memory eXchange), XLIFF (XML Localisation Interchange File Format) o OLIF2 (Open Lexicon Interchange Format). Cabe esperar que los miembros de la LISA, entre ellos los fabricantes de software más importantes, agencias de servicios globales y fabricantes de herramientas de traducción/localización, utilicen TBX como su formato de intercambio de datos terminológicos, y que convenzan a sus clientes no miembros de la LISA a optar igualmente por dicho estándar. En cuanto a los programas de gestión terminológica profesionales MultiTerm iX de Trados es, hasta el momento, la única herramienta, desde la cual se puede generar un formato basado en XML y muy similar a las especificaciones TBX. Entre los usuarios hay que mencionar al fabricante de automóviles Ford que ha incorporado TBX como formato de intercambio para sus datos terminológicos.

Actualmente, el grupo OSCAR sigue trabajando en la versión definitiva de TBX. Algunos problemas como, por ejemplo, la codificación estándar o la representación de datos binarios en el elemento *refObject* están aún sin resolver. Por otra parte, se prevé la elaboración de interfaces entre TBX y otros estándares como TMX, XLIFF y OLIF2, puesto que estos formatos también pueden contener informaciones terminológicas reutilizables. Además, se están llevando a cabo investigaciones con el objetivo de modelar de forma más adecuada las entradas terminológicas específicas para la localización de software (Reineke, 2004).

Bibliografía

ISO/IEC 646. 1991. *Information technology – ISO 7-bit coded character set for information interchange*. Ginebra: ISO

ISO 6156. 1986. *Magnetic tape exchange for terminological/lexicographical records (MATER).* Ginebra: International Standards Organization

ISO 8879. 1986. *Information processing – Text and office systems – Standard Generalized Markup Language (SGML)*: Ginebra: International Organization for Standardization

ISO 12200. 1999. *Computer applications in terminology - Machine-readable terminology interchange format (MARTIF). Negotiated interchange*. Ginebra: International Organization for Standardization

ISO 12620. 1999. *Computer applications in terminology - Data categories*. Ginebra: International Organization for Standardization

ISO 16642. 2003. *Computer applications in terminology - Terminological markup framework (TMF).* Ginebra: International Organization for Standardization

LISA (= Localisation Standards Association Industry). *2002. TBX specifications. Working Draft, 5 May 2002*. http://www.lisa.org/tbx, consultado el 23 de octubre de 2004

Melby, Alan K./Schmitz, Klaus-Dirk/Wright, Sue E. 2001. "Terminology Interchange" en Wright, Sue E./Budin, Gerhard (eds.) *Handbook of Terminology Management*. *Volume 2: Application-Oriented Terminology Management*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 613-642

Reineke, Detlef. 2004. "fFpPtTdDcCrRlL??? – Wissenseinheiten in der Softwarelokalisierung (und deren Verwaltung in TVS)" en Mayer, Felix/Schmitz, Klaus-Dirk/Zeumer, Jutta (eds.) *Terminologie und Wissensmanagement*. Akten des Symposions, Colonia 26 – 27 marzo de 2004. Colonia: Deutscher Terminologie-Tag e.V., 193-208

TEI. 2004. *Text Encoding Initiative*. http://www.tei-c.org, consultado 23 de octubre 2004