

Using international standards for terminology exchange

KLAUS-DIRK SCHMITZ

Cologne University of Applied Sciences

KEYWORDS: terminology management, terminology exchange, standards, data categories, TBX

1. MOTIVATION

The creation of high-quality terminology and professional terminology management are both time-consuming and cost-intensive activities. The access to and the reuse of already existing terminological resources can considerably reduce the effort for terminology work. These economic issues are in many cases the main arguments for the interest in terminology exchange.

But there are a lot of other reasons why terminology exchange is so important. Translation business activities between clients, language service providers and freelance translators require a transfer of terminological data in both directions. Companies provide terminology to support translation projects and translators return terminological resources they have updated during the translation job. And the consistent use of terminology in larger projects with several individual translators can only be guaranteed, if all persons involved will have access to the same terminological data resources; but different terminology management tools and individually designed terminology databases complicate the lossless and consistent exchange of terminological data.

A final important scenario for the exchange of terminological data is the migration of terminological data collections from one system to another. This may be necessary if a company or freelancer changes the tools provider and has to convert the terminology from the old terminology management system to the new tool, but also if terminological data are needed in other programs like content management systems or (rule-based) machine translation systems.

But a simple and easy terminology exchange in all these scenarios can be problematic, since the technical requirements for terminology manage-

ment (languages, data categories, data value sets) can be quite different depending on the needs of the respective user groups and the organizational environments.

2. REQUIREMENTS

Exchange routines between two specific terminology databases are typically developed individually, taking into account the existing import and export routines and formats of the two systems involved. It is also necessary to know the data models and the “semantics” of the data categories of both systems. If data exchange is needed between more than two systems, the number of necessary conversion routines will increase rapidly. That is why there is a strong demand for standardized exchange formats, where each of the participating terminology management systems has to provide only a feature for the import of data from the standardized exchange format and only a feature for the export into the standard exchange format.

One of the most famous and often provided formats is CSV (comma-separated values). This format separates information by commas or tabs and data records by returns or paragraph marks. Although individual data are usually separated from each other, the meaning and interpretation of the individual data is unclear and not explicitly laid down in the CSV format.

Fig. 1: Data from a MultiTerm 2011 termbase exported in CSV format

```
[60 → J.·Spiegel→2011-02-24T18:07:26 → J.·Spiegel→2011-02-24T18:07:26
→ Atraumatische·Klemme.jpg → Allgemein Chirurgie|Gefäßchirurgie → vorläufig
→ Atraumatische·Klemme → J.·Spiegel→2011-02-24T18:07:26 → J.·Spiegel
→ 2011-02-24T18:07:26 → sub → f. → sg. → bevorzugt → Der·überschüssige·
Ureter·wird·durch·atraumatische·Klemmen·gefasst·und·abgetrennt. → Die·
atraumatische·Klemme·dient·zur·Gewebepräparation,·aber·auch·zum·Fassen·von·
Gefäßen·beim·Anlegen·einer·Ligatur. → vorläufig → Atraumatic·clamp→J.·
Spiegel → 2011-02-24T18:07:26 → J.·Spiegel→2011-02-24T18:07:26 → sg. → sub
→ bevorzugt → US|UK·The·catheter·should·be·clamped·with·an·atraumatic·clamp·
between·the·catheter·exit·site·and·the·damaged·area·when·damage·occurs·and·must·
remain·clamped·during·repair.·An·atraumatic·clamp·is·a·surgical·instrument·
designed·to·compress·a·structure,·such·as·a·blood·vessel·or·a·cut·end·of·the·
intestine.·Atraumatic·clamps·are·used·to·stop·bleeding·from·the·cut·vessels·and·
are·designed·not·to·damage·the·arterial·wall. → vorläufig → 2.3.2·J.·Spiegel
→ 2011-02-24T18:07:26 → J.·Spiegel→2011-02-24T18:07:26 → ¶
61 → J.·Spiegel→2011-02-24T18:08:02 → J.·Spiegel→2011-02-24T18:21:47
→ Atraumatische·Nadel.jpg|Allgemein Chirurgie → vorläufig → Atraumatische·
Nadel·J.·Spiegel→2011-02-24T18:08:02 → J.·Spiegel→2011-02-24T18:08:02 → sub
→ f. → sg. → bevorzugt → Die·atraumatische·Nadel·besitzt·im·Schaft·eine·
axiale·Bohrung,·in·die·der·Faden·eingebracht·wird. → Die·atraumatische·Nadel·
zählt·zu·den·chirurgischen·Nadeln·und·dient·zur·Gewebevereinigung.·Im·Gegensatz·
zur·Öhrnadel·ist·bei·der·atraumatischen·Nadel·der·Faden·in·das·Nadelende·
```

The CSV format is often used to convert data to and from Excel spreadsheets. Sometimes you will find names of the columns specified in the first line of the exchange file, which allows inferences to be made about

the semantics of the following data. But this can only work if the structure of the terminological data is not too complex and if each record contains the same type and number of data; variable data (e.g. a different number of terms in several languages) will shift the columns and will make it impossible to interpret the data correctly.

Using markup languages such as XML (extensible markup language) circumvents exactly this problem. Each individual piece of information carries the semantic of its meaning in a so-called *tag*, and the grouping and allocation of each data item is uniquely defined by the XML structure.

Fig. 2: Data from a MultiTerm 2011 termbase exported in XML format

```
<concept>60</concept>¶
<transacGrp><transac.type="origination">J. · Spiegel</transac><date>2011-02-
24T18:07:26</date></transacGrp>¶
<transacGrp><transac.type="modification">J. · Spiegel</transac><date>2011-02-
24T18:07:26</date></transacGrp>¶
<descripGrp><descrip.type="Abbildung">Atraumatische ·
Klemme.jpg</descrip><descripGrp><descrip.type="Quelle"><xref ·
ulink="http://www.dimeda.de/pdf/general/kap_13.pdf">http://www.dimeda.de/pdf/gen
eral/kap_13.pdf</xref></descrip><transacGrp><transac.type="origination">J. ·
Spiegel</transac><date>2011-02-
22T17:32:35</date></transacGrp><transacGrp><transac.type="modification">J. ·
Spiegel</transac><date>2011-02-22T17:32:35</date></transacGrp></descripGrp>¶
</descripGrp><descripGrp><descrip ·
type="Fachgebiet">Allgemeinchirurgie|Gefäßchirurgie</descrip></descripGrp>¶
<descripGrp><descrip.type="Qualitätsstatus">vorläufig</descrip></descripGrp>¶
<languageGrp><language.lang="DE" · type="Deutsch"/>¶
<descripGrp><descrip.type="Definition">Die · atraumatische · Klemme · dient · zur ·
Gewebepräparation, · aber · auch · zum · Fassen · von · Gefäßen · beim · Anlegen · einer · Ligatur. ·
</descrip>¶
```

Figure 2 shows the beginning of the CSV exchange file of Figure 1 in XML format, namely in the specific MultiTerm 2011 XML format used for the conversion of MultiTerm databases.

3. SOLUTION

We are now not far away from achieving a well-defined standardized exchange format. It is only necessary to standardize the markup in the angle brackets, the tags and the type attributes, to provide a so-called “blind exchange” between terminology databases. Blind means that you do not need to know the structure, the data categories, and their values in the originating and in the target termbase. This is exactly what the ISO 30042:2008 standard, specifying TBX (Term Base eXchange) attempts to accomplish. TBX is based on ISO 16642:2003, defining a terminological meta-model with several levels and structural elements for the

XML file, and on ISOCAT (www.isocat.org), specifying terminological data categories according to ISO 12620:2009. Figure 3 shows a simplified extract of a terminological entry in TBX representation.

Fig. 3: Simplified example of a TBX file

```
...
<text><body>
  <termEntry id='ID0000073578'>
    <descrip type='subjectField'>Materialbeschaffenheit</descrip>
    <langSet lang=de>
      <descrip type='definition'>Maß für die Lichtundurchlässigkeit</descrip>
      <ntig>
        <termGrp>
          <term>Opazität</term>
          <termNote type='partOfSpeech'>Substantiv</termNote>
          <termNote type='grammaticalGender'>f.</termNote>
        </termGrp>
      </ntig>
    </langSet>
  </termEntry>
</body></text>
```

TBX is specified as a single international standard, but one will recognize when consulting the document and the ISOcat website that there exist different variants of TBX. The standard itself defines something like a family of TBX-compatible formats, where TBX-Default supports the whole set of terminological data categories of the TBX family and TBX-Basic only uses a smaller subset of data categories in order to facilitate simpler import and export routines for specific user groups and less complex databases.

4. CHALLENGES

Although TBX is a well-defined ISO standard, two problems have to be mentioned which may extremely complicate the exchange of terminological data: a lacking strictness of TBX in some areas and the flexibility of modern terminology management systems.

Among others, TBX attempts to standardize the permissible instances of closed data categories, i.e. the values that can occur. This might work properly for certain data categories such as grammatical gender (gender = m / f / n), but is doomed to failure for a data category such as subject field, because there exists no (ideal) subject field classification

that is appropriate for all user needs and all areas of application. Therefore, a completely “blind” data exchange using TBX cannot be guaranteed for all data categories, since TBX cannot specify the possible values for data categories such as subject field.

Terminology management system should provide import and export interfaces supporting the TBX standard. Some system developers state explicitly that their software is able to import and export files according to the TBX format. However, this can only be true if the terminology management system has a fixed predetermined entry structure with preset data categories; only in this case the system developer will know the semantics of the data categories and can adequately represent them in TBX export files or interpret them in TBX import files. But if the terminology management systems requires / allows that the user specifies the data model and the data categories according to his specific needs (e.g. MultiTerm), or at least modifies a preset data modelling (e.g. crossTerm, qTerm or TermStar), the system developer will not be able to provide a clean TBX interface, because he cannot know the semantics of the user-defined data categories. To compensate cases like this, only two solutions (outside of TBX) are possible: either a specific mapping tool helps the user to map his own data categories to the standardized ISOcat categories, or the user creates his own terminology database on the basis of the standardized data categories in ISOcat directly from the beginning.

5. CONCLUSION

ISO/TC37, the international technical committee for the standardization of terminology, has developed and published several standards that directly relates to the exchange of terminological data (e.g. ISO 30042:2008 for TBX). But also other TC37 standards can be used for the selection of terminological data categories and for modelling terminology databases in order to facilitate and ensure the exchange of terminological data.

Since there are fundamental problems with modern terminology management systems with respect to the support of exchange formats such as TBX, the exchange issue has to be taken into account when designing a solution for terminology management for companies and freelancers. Wrong decisions and mistaken selections of data categories will hinder future exchange of terminological data and can only be repaired with great effort. ISO standards such as ISO 12620:2009 and ISO 26162:2012

as well as the Data Category Registry (www.isocat.org) will support users to design professional and state-of-the-art terminology management systems allowing for standard-based terminology interchange.

BIBLIOGRAPHY

- ISO 12620:2009 *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources*. Geneva: ISO.
- ISO 16642:2003 *Computer applications in terminology – Terminological markup framework*. Geneva: ISO.
- ISO 26162:2012 *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of terminology management systems*. Geneva: ISO.
- ISO 30042:2008 *Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)*. Geneva: ISO.
- Melby A. K., Schmitz K.-D., Wright S. E. 2001: Terminology Interchange. – *Handbook of Terminology Management*, Volume 2, Wright S. E., Budin G. (eds.), Amsterdam/Philadelphia: John Benjamins, 613–642.
- Schmitz K.-D. 2004: Terminologearbeit und Terminographie. – *Angewandte Linguistik – Ein Lehrbuch*, Knapp K. et al. (Hrsg.), Tübingen/Basel: Francke, 435–456.
- Schmitz K.-D. 2008: Bedeutung von Normung und Terminologearbeit für die Technische Dokumentation. – *Terminologearbeit für Technische Dokumentation. (= tekom Schriften zur technischen Kommunikation 12)*; Hennig J., Tjarks-Sobhani M. (Hrsg.), Lübeck: Schmidt-Römhild, 11–17.
- Schmitz K.-D., Straub D. 2010: *Successful Terminology Management in Companies – Practical tips and guidelines*, Stuttgart: TC and More.

TARPTAUTINIŲ STANDARTŲ NAUDOJIMAS TERMINOLOGIJOS MAINAMS

Kruopštus terminologijos darbas ir profesionali terminijos tvarkyba įmonėms ir laisvai samdomiems darbuotojams yra daug laiko ir išlaidų reikalaujanti veikla. Ją gerokai palengvintų prieiga prie esamų terminologinių išteklių ir jų panaudojimas. Klientų, vertimo paslaugas teikiančių įmonių ir vertėjų bendravimui ir bendradarbiavimui reikalingas duomenų perdavimas įvairiomis kryptimis. Skirtingos terminologijos tvarkybos sistemos ir nesuderinti, atskiri terminologinių išteklių duomenų modeliai gali trukdyti terminologinių duomenų mainams ar užkirsti jiems kelią.

ISO/TK37 – tarptautinis terminologijos ir kitų kalbos bei turinio išteklių standartizavimo technikos komitetas – parengė ir išleido keletą standartų, skirtų terminologinių duomenų mainams (pvz., TBX mainų formatui skirtas ISO 30042). Pasirenkant tinkamas terminologinių duomenų kategorijas ir modeliuojant terminologinių duomenų bazes, galima remtis ir kitais ISO/TK37 standartais, kurių taikymas palengvintų terminologinių duomenų importą ir eksportą iš kitų išteklių.

Straipsnyje kalbama apie bendruosius terminologinių duomenų mainų tikslus, principus ir problemas, supažindinama su tarptautiniais standartais, skirtais veiksmingus duomenų mainus užtikrinančioms terminologijos tvarkybos sistemoms kurti, aiškinamas TBX mainų formato taikymas.

Gauta 2012-11-18

Klaus-Dirk Schmitz
Cologne University of Applied Sciences
Institute of Translation and Multilingual Communication
Ubierring 48, 50678 Cologne, Germany
E-mail klaus.schmitz@fh-koeln.de