Uta Seewald-Heeg

Terminology Exchange without Loss?

Feasibilities and Limitations of Terminology Management Systems (TMS)

Abstract

The present article gives an overview over exchange formats supported by Terminology Management Systems (TMS) available on the market.

As translation is one of the eldest application domains for terminology work, most terminology tools analyzed here are components of computer-aided translation (CAT) tools.

In big corporates as well as in the localization industry, linguistic data, first of all terminology, have to be shared by different departments using different systems, a situation that can be best solved by standardized formats.

The evaluation of seven widely used TMS shows, however, that formats other than the standards proposed by organizations like LISA currently dominate the picture. In many cases, the only way to share data is to pass through flat structured data stored as tab-delimited text files.

1 Workflow and Interchange Scenarios

In the brief history of terminology management since the 1960s, when the first databases for terminology work were developed, terminology management has become a key resource, not only for the language industry, but also for globally acting industrial firms.

Usually, different departments within a company have access to the terminology resources, and if freelancers or translation service providers come into play, terminology interchange with external partners has to be organized as well.

At least in an architecture where corporate terminology has to be accessed from different applications under different circumstances – this is, for example, the case in corporates like SAP or DaimlerChrysler – questions of terminology interchange and supported formats arise. The need of interchange formats that guarantee the identification of data categories in different environments becomes obvious (Alder 1998). Here, standards come into play that map local system data categories to data categories specified in an open standard (Fig. 1), provided that developers of NLP tools make use of such standardized formats.

Fig. 1: Mapping local system categories to categories specified in a standard (following Alder 1998:12)

2 Interchanging Terminological Data – Standards

The need for terminology interchange has long been recognized by industrial users of TMS. Consequently, the past 15 years have seen several standardization initiatives aimed at developing standardized formats. One of these initiatives led to the CLS Framework (Melby/Wright 2000) which deals with the structure and content of terminological databases (Fig. 2). The CLS Framework (CLS stands for Concept-oriented with Links and Shared references, cf. Melby/Wright 1998) is based on the ISO 12620 standard "Computer applications in terminology – Data categories" which was published in 1999. CLS provides explicit data models for all types of terminological databases by structuring the items in a term

entry according to theory and practice in concept-oriented terminology. The framework specifies the structure of a term entry and the relationships among data items in an entry using as one of the formats describing the structure of a terminological entry the Machine-Readable Terminology Interchange Format (MARTIF).

The development of the MARTIF standard, which formed the starting point for the CLS framework, was actually preceded by the development of OLIF (Open Lexicon Interchange Format), a more machine oriented standard, originally focussing on Machine Translation. The XML-compliant OLIF2 standard published in 2002 defines a large number of lexical features, but does not make statements about their structural embedding (Wittenburg/Gibbon/Peters 2001). Although OLIF2 aims at integrating data of Machine Translation and of Terminology Management Systems, OLIF has been of little importance in the field of Terminology Management Systems so far.

Another standard released to the public in 2002 by the Localization Industry Standards Association (LISA) is the TermBase eXchange Format (TBX) worked out by the LISA working group for the development and maintenance of open standards for the language industry, OSCAR (Open Standards for Container/Content Allowing Re-use). TBX, which is also based on XML, is only slowly being integrated into commercial terminology systems.

3 Terminology Management Systems (TMS)

3.1 Conceptual Features of TMS

Despite the existence of standards, commercial TMS still seem to be far away from the expressed goal of CLS, which is preservation of data when interchanging terminology (Alder 1998:6).

TMS not only differ in the formats they store lexical or terminological data, but also in their conceptual features. They can be classified by their

– **language concept** specifying whether a system is monolingual, bilingual, or allows multilingual data;
– **entry structure** which either can be predefined, definable or free, that is entirely specifiable by the user;
– **entry model** distinguishing systems only allowing a lemma-oriented structuring of the terminological database from systems allowing concept-oriented keeping of data;

Regarding the conceptual features of TMS the difference in the entry structure turns out to be one of the key problems.

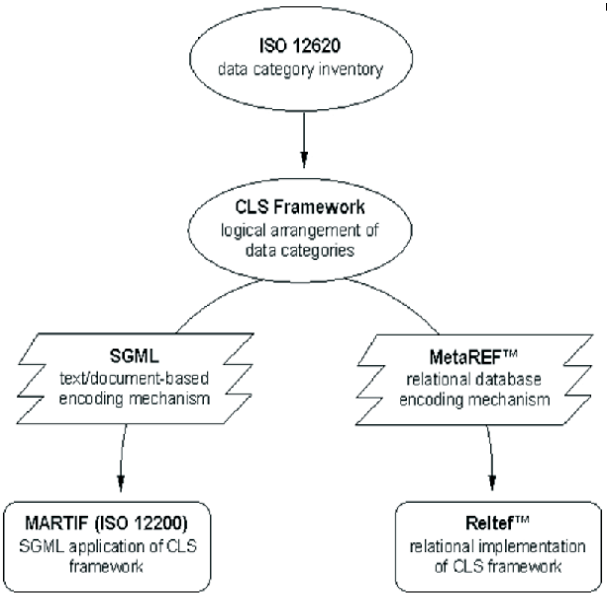

Fig. 2: Structure of the CLS Framework (Melby/Wright 2000)

3.2 Systems

In order to give an idea of the variety of differences concerning the conceptual features as well as the supported formats of existing commercial products, 7 systems have been selected. The following sections contain a discussion of their interchange functionalities according to the list below:

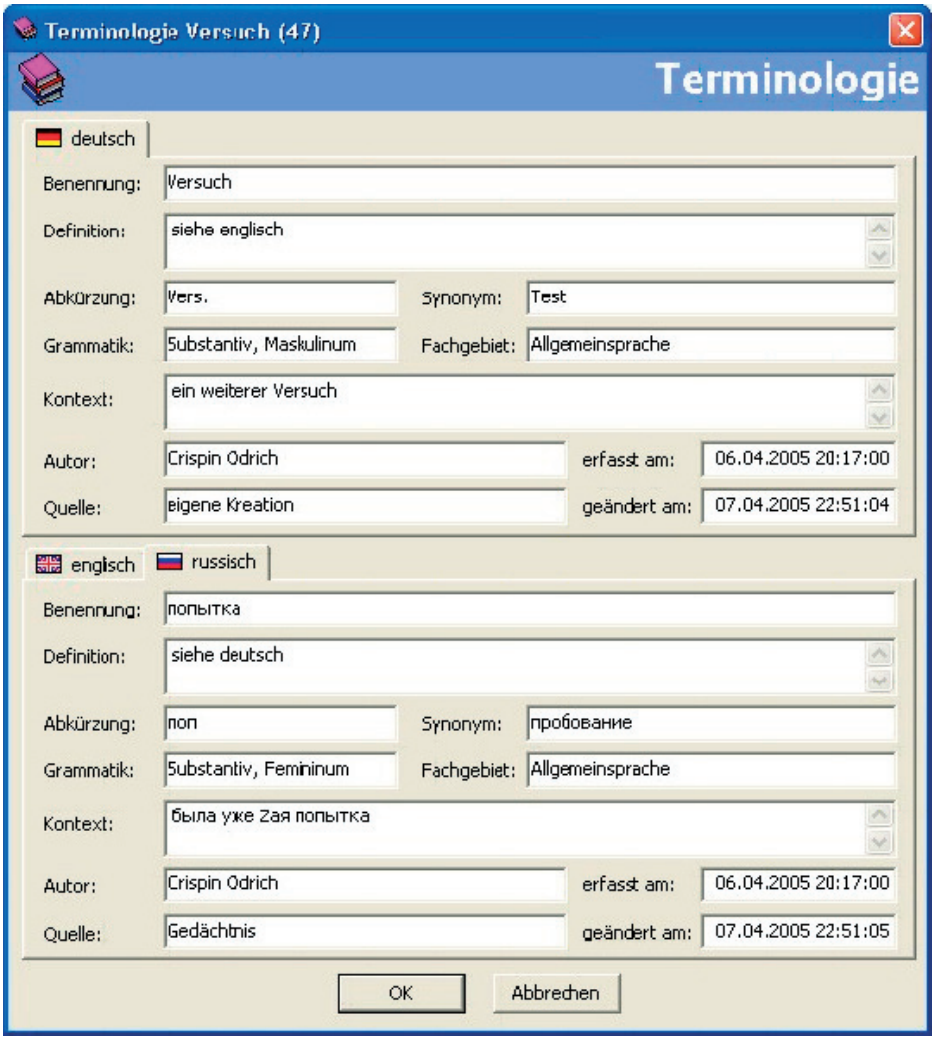

Fig. 3: GFT DataTerm interface

GFT DataTerm by GFT (*www.gft-online.de*).
UniTerm by Acolada (*www.acolada.de*).
Déjà Vu Terminology by Atril (*www.atril.com*).
SDL TermBase by SDL (*www.sdl.com*).
MultiTerm iX by SDL Trados (*www.trados.com*).
TermStar XV by Star (*www.star-group.net*).
crossTerm by across (*www.across.net*).

3.2.1 Standalone Systems

The first system mentioned here, **GFT DataTerm** (Fig. 3), is a standalone system in the sense that it does not provide interfaces to tools like Translation Memories (TM) or other applications. It is a lemma-oriented system, even if multiple language pairs can be stored in a single entry. Descriptive categories can only be assigned to individual terms; other levels of specification, e.g. a concept level linking different terms to a given concept do not exist. For import, GFT DataTerm provides tab-delimited text file format as well as the Excel XML spreadsheet format. Formats provided for the export of terminology are Excel and XML-based MARTIF.

Another standalone system is the **UniTerm** tool (Fig. 4) from which terminological data can also be exported as text file or as XML together with a DTD[1]. It has a definable entry structure and allows multilingual conceptual information. Term describing fields as well as fields containing conceptual information can be selected among a predefined set of categories which can be labelled individually. Furthermore, for different purposes of terminological work different editing patterns are available.

3.2.2 Integrated Systems

In contrast to the standalone systems mentioned so far, most terminology systems are actually integrated into TM environments. Thus, across, Déjà Vu, SDLX, Star, and Trados all have more or less powerful terminology components. In

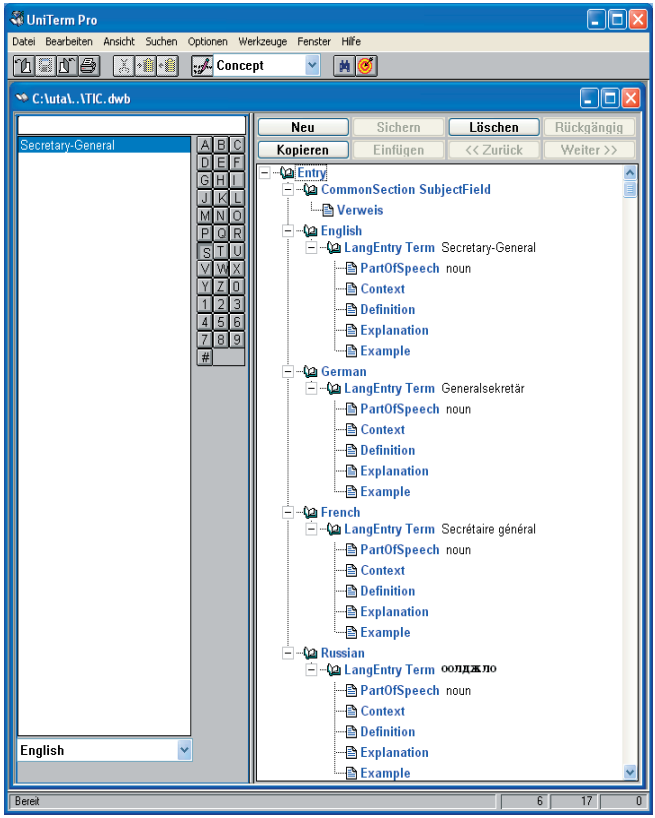

Fig. 4: UniTerm interface

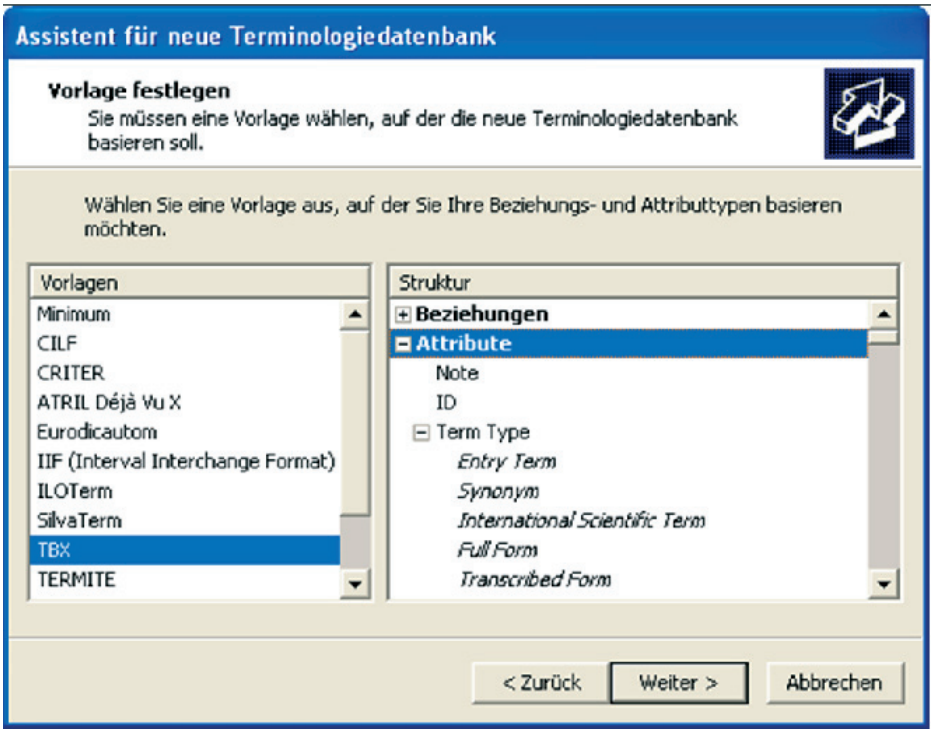

Fig. 5: Pattern selection for the structure of entries in Déjà Vu

the case of the Star and the Trados products, i.e. TermStar and MultiTerm, the terminology components can even be purchased separately.

Part of the **Déjà Vu** TM-System is a so-called terminology database which is mainly lemma-oriented. To create a termbase, Déjà Vu provides templates to determine the entry structure for a new database. One of them reflects the structure and categories of TBX (Fig. 5) although TBX is not supported for import or export. Déjà Vu allows the import of text files, Excel and Access files as well as TermStar files. The same file types can also be exported.

When terminology has to be imported from an Excel file, the Excel column headers have to be assigned to Déjà Vu fields, a common way to map the content of the spreadsheet file to the terminology system where the user has to determine the fields to be imported and to specify whether filters shall be applied.

The **SDL TermBase** (Fig. 6), a component of the SDLX TM system, is structured very similarly to the Déjà Vu terminology component. As far as the multilinguality and the treatment of synonyms are concerned, the structuring of the data is concept-oriented. But one misses a conceptual level allowing the specification of non-redundant information valid for the concept, that is, for all terms of a given entry. For the import and export of terminology, apart from the proprietary format, tab-delimited text files as well as files in Trados MultiTerm 5 format can be imported.

The Trados terminology component **MultiTerm iX** is one of the two terminology systems

which provide interfaces to other components of a translation memory environment, but which can also be used without launching the TM system.

MultiTerm provides a concept-oriented storage of data (Fig. 7) and has a hierarchical structure with three different levels, one level to specify concept-related information, another one for language-specific terminological information, and a third one to describe an individual term. It has a definable entry structure, but provides also predefined termbase templates in which the fields are already specified, and the entry structure is already defined. The structure of the termbank and the terminological data are stored in separate files. For import, MultiTerm supports Excel and tab-delimited text files which first have to be converted by MultiTerm Convert (Fig. 8). For export, MultiTerm provides as format its own XML format which follows the main structuring principles of TBX although it proved to be incompatible with TBX in the evaluated version (Trados 7). Apart from its own XML format, MultiTerm IX provides two other formats for terminology export, MultiTerm 5 and tab-delimited text file format.

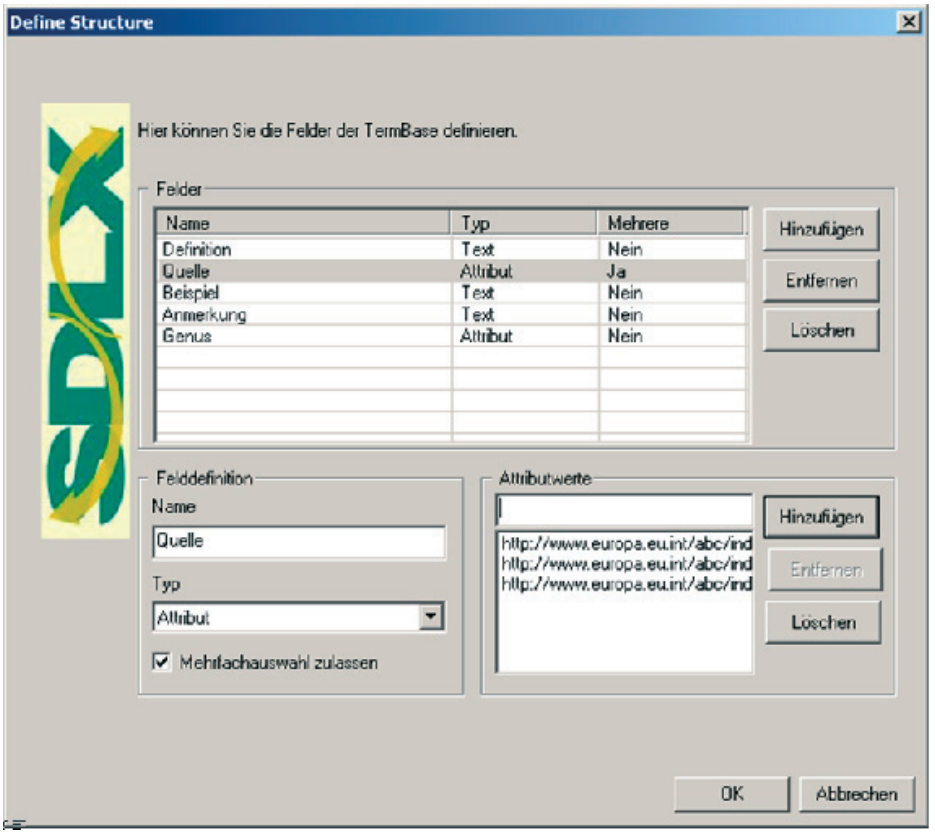

Fig. 6: Definition of termbank structure in SDLX

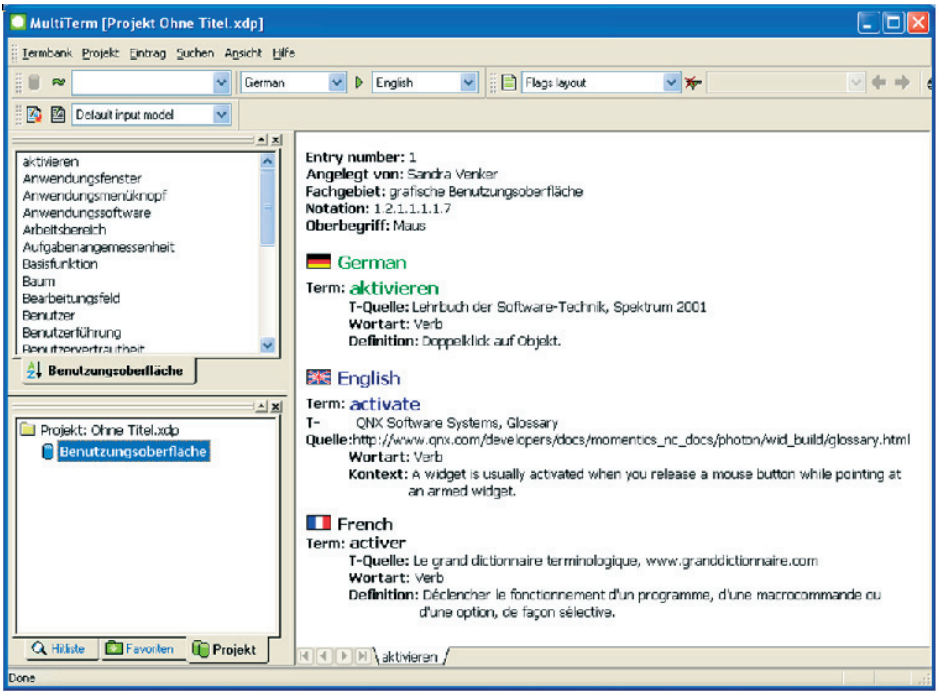

Fig. 7: MultiTerm iX interface

The Star terminology system, **TermStar XV**, is the other TMS which provides interfaces to other components of a translation environment, and which can also be used as a standalone system, i.e. independent of a translation memory environment. TermStar has a definable entry structure, however with a predefined set of possible data categories which can be named according to the need of the users.

Similar to MultiTerm, TermStar (see Fig. 9) distinguishes different description levels: The header of an entry is meant to store conceptual information. Terms can be described depending on the individual language, and an intermediate information level can be used to store information for all terms of a given language.

For the import of terms TermStar provides, apart from its proprietary formats of different TermStar versions, an XML-based MARTIF and for everything else an import dialogue for so

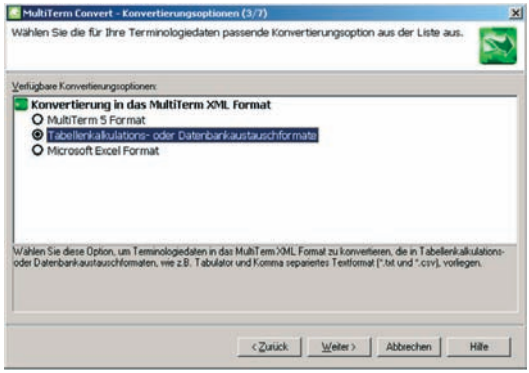

Fig. 8: Format conversion using MultiTerm Convert

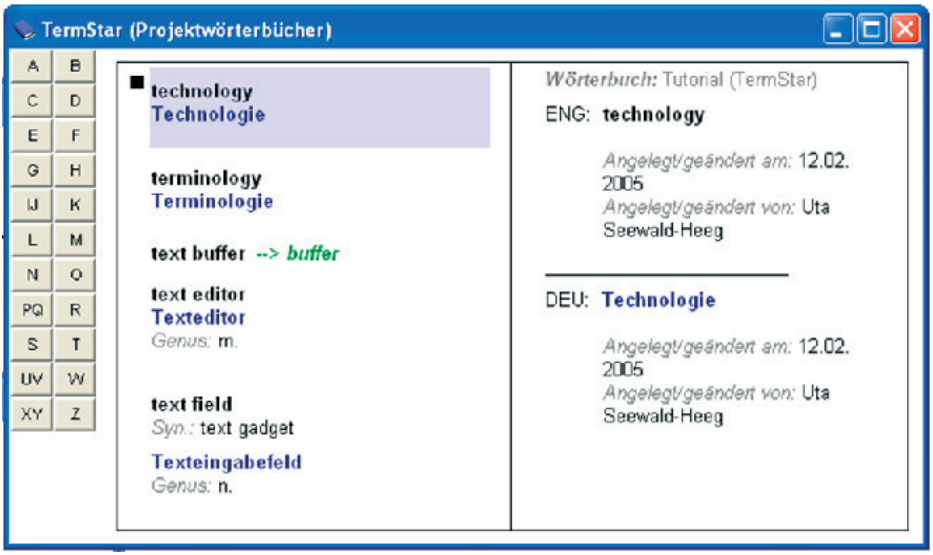

Fig. 9: TermStar XV interface

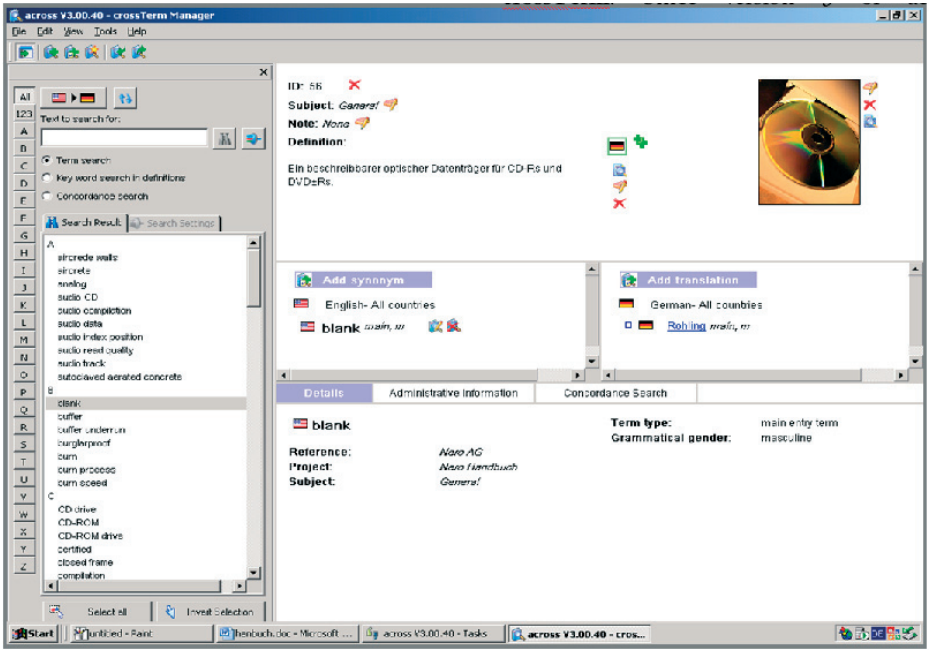

Fig. 10: crossTerm user interface

called "user defined formats", which allows, for example, to configure the import of Excel and MultiTerm 5 files. If proprietary formats are not considered, the export from TermStar is restricted to XML MARTIF.

Among the systems mentioned here, the most recent system on the market is across, a translation management environment which also provides a terminology component called **crossTerm**. Since version 3 of across, crossTerm allows concept-oriented data storage. Concept-relevant information can be stored in the head of an entry which is visually separated from the bilingual view of an entry (Fig. 10). The across developers have avoided using a proprietary terminology format. In crossTerm, terminology is stored in TBX format, which is also the only format provided for export. To import data crossTerm provides in addition to CSV-format, the Langenscheidt electronic dictionary format, Trados MultiTerm 5, and the Star MARTIF format.

4 Supported formats

The evaluation has shown that all the systems analyzed so far allow import from Excel files or file formats such as CSV or TXT that can be generated by Excel. As Trados – at least until its acquisition by SDL – has dominated the TM and TMS market, several products also support MultiTerm format. However, instead of supporting MultiTerm iX, they usually support the text based format formerly used by Trados 5. The support of formats can be visualized as illustrated below (see Fig. 11).

5 Exchange of data

As shown in Figure 11, Excel or Excel-derived formats like CSV and tab-delimited text are in many cases the only formats allowing the interchange of data between two or more systems. Thus, the question arises whether all of the data intended to be transferred are actually transferred or interchanged completely and correctly using Excel

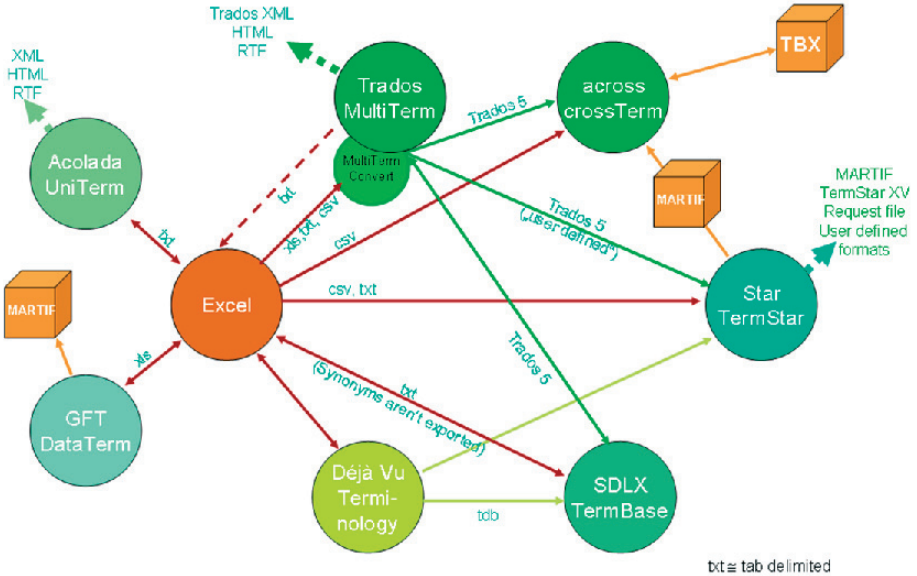

Fig. 11: Exchange formats supported by TMS

	A	B	C	D	E	F
1	English	Deutsch	Français	AS-Datei	Plattform	OS
2	Unknown	Unbekannt	Inconnu	TXT	Windows	Microsoft Plus! XP
3	Unknown Album	Unbekanntes Album	Album inconr	TXT	Windows	Microsoft Plus! XP
4	Unknown Artist	Unbekannter Künstler	Artiste inconr	TXT	Windows	Microsoft Plus! XP
5	Unknown Genre	Unbekannte Stilrichtung	Genre inconn	TXT	Windows	Microsoft Plus! XP
6	Unknown Title	Unbekannter Titel	Titre inconnu	TXT	Windows	Microsoft Plus! XP

Fig. 12: Multilingual glossary in Excel format

files. To answer this question the structural layer comes into play, because each system presupposes a defined structuring of the stored data. And as data interchange also has to guarantee the correct interpretation of the content, we also have to consider the semantic, or representational layer.

To gain insight in this question, we now will have a closer look at the import and export of terminology stored in Excel files as well as the interchange of these data between different TMS.

The starting point will be an Excel file containing a simple multilingual glossary (Fig. 12) in the form glossaries are provided by Microsoft with some additional information.

In order to get these data into **MultiTerm iX**, they first have to be converted by MultiTerm Convert into MultiTerm-compatible format. During this process, the Excel column headers have to be assigned to MultiTerm fields, and the entry structure has to be defined. The result of

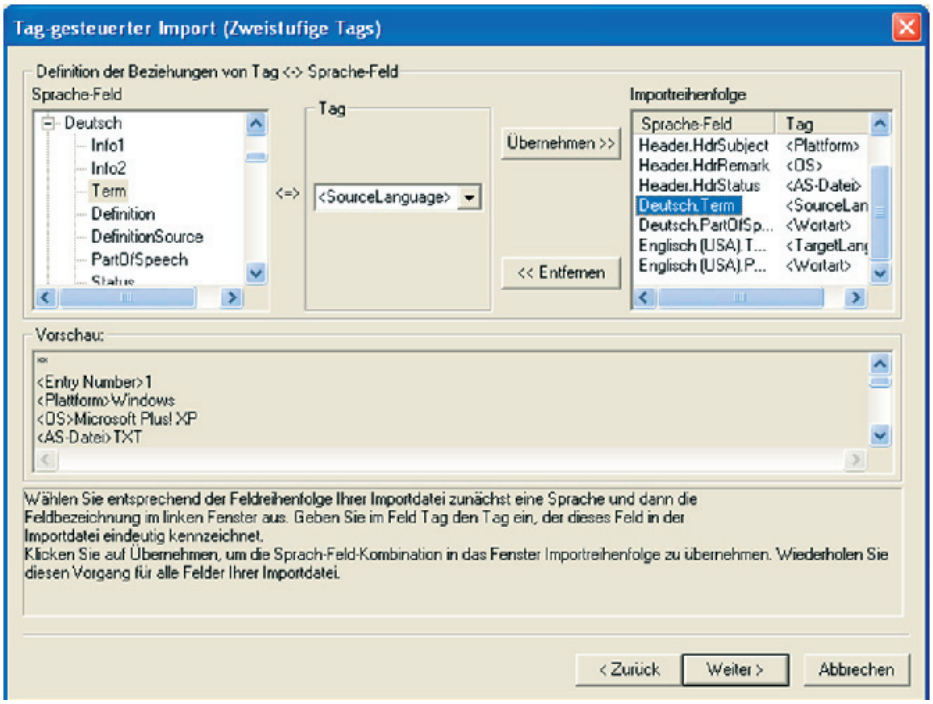

Fig. 13: Import dialogue in TermStar

the conversion is a termbank definition file, and an XML file containing the terminological data. This XML file finally can be used to create a new termbase and to launch the default process for the import. Importing data from an Excel file produces a satisfactory result, since all information can be transferred completely and correctly.

Furthermore, some of the term-related information may not be present in all of the entries. In this case, the use of MultiTerm is problematic, because the MultiTerm export functionality creates files where descriptive fields, which are used only in part of the entries, are ignored when writing the tab-delimited text file. As a result, the system generates columns with different type of content in their respective cells. In this case, the resulting files turn out to be unusable for further handling. Another kind of problem is caused by line breaks in definition texts. As line breaks split up an entry on different lines, an import where one line corresponds to one entry is not possible any more.

The export of data in tab-delimited format does not necessarily suffer from these limita-tions. If unused fields of the entry structure are exported as empty fields (this is, for exam-ple, the case when exporting data in tab-delimited text format from the **SDL TermBase**) the structure of the content can be preserved, so that the export file allows further handling of the exported data and import in other systems supporting tab-delimited format.

Another export scenario is the exchange of terminology between **MultiTerm iX** users and users of other systems supporting MultiTerm 5 format including terminologists still working with Trados 5. A closer look at the MultiTerm 5 export functionality provided by MultiTerm iX revealed that this functionality supports only bilingual export. As a result, a multilingual termbase can only be exported selecting different language pairs with one language as reference. Therefor, n languages require n-1 export procedures, and certainly also n-1 import procedures on the side of the receiving system.

Saved as an ANSI-encoded tab-delimited text file, an Excel glossary can also be imported in

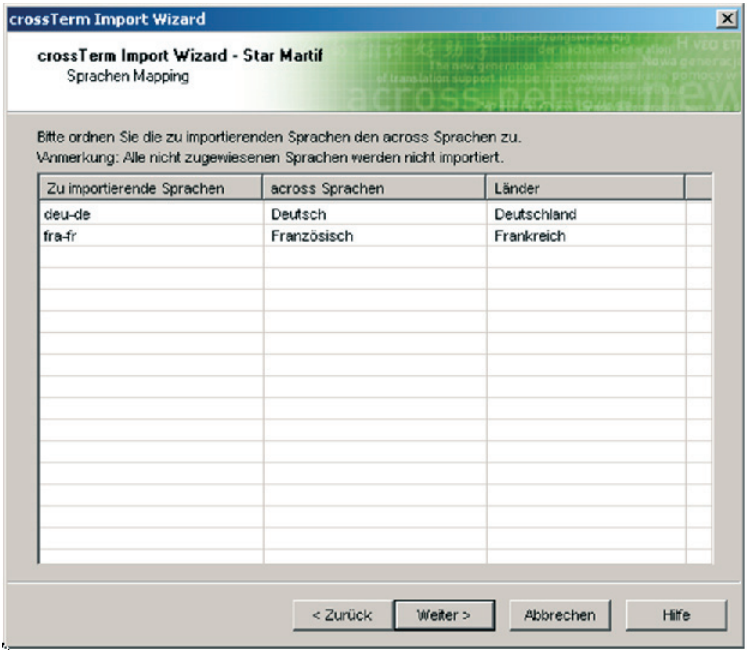

Fig. 14: crossTerm Import Wizard for Star MARTIF

TermStar XV, where the column headers have to be assigned to TermStar fields (Fig. 13). No information is lost during this process; the only inconvenience is that the column headers of the text file are imported in TermStar as first entry.

The import of MultiTerm 5 files to TermStar XV has to pass through the conversion of the MultiTerm 5 text file in ANSI format because – at least in the build analyzed here – Unicode-encoded MultiTerm files are not supported which already restricts the type of languages which can be interchanged with this format. The MultiTerm 5 import in TermStar transfers the entire information to TermStar.

The import of the Excel file in **crossTerm** leads to a satisfactory result as it did for the previously mentioned systems.

The import of a Star MARTIF file into crossTerm does not differ substantially from the Excel import, i.e. the field names of both representations have to be mapped to each other (Fig. 14). Here again, the result is quite satisfactory.

From a purely technical point of view, terminological data can be imported, exported, and interchanged using tab-delimited text files. However, as systems like MultiTerm allow a certain descriptive field to be used at different levels and related to distinct fields, the information of the embedding of categories disappears when mapping entry structures to flat rows and columns so that this kind of information cannot be maintained transferring data between different systems using tab-delimited text format.

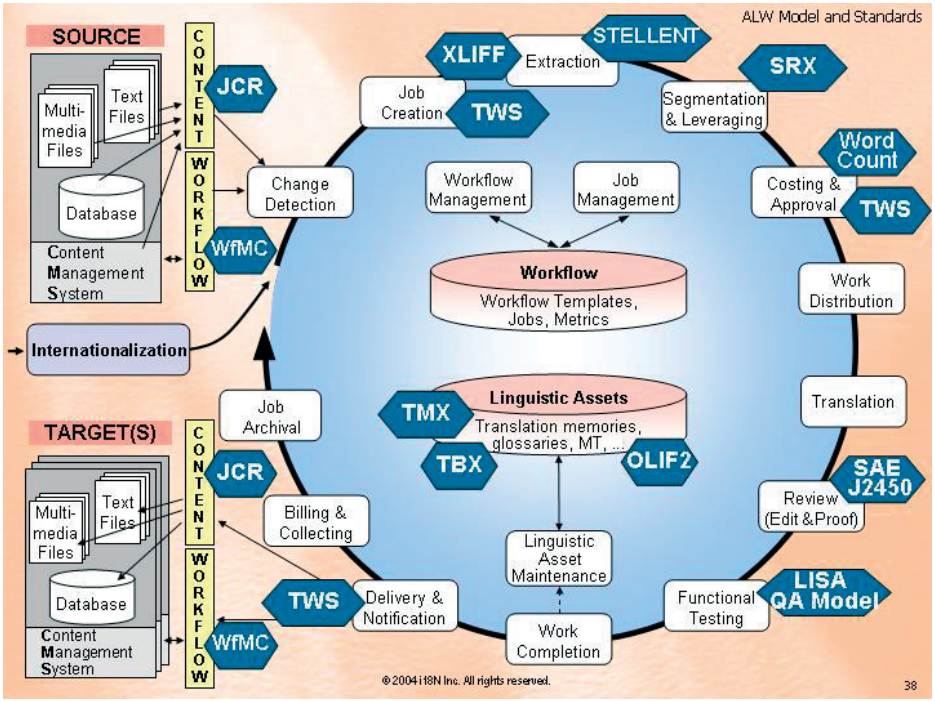

Fig. 15: The role of standards in an automated workflow

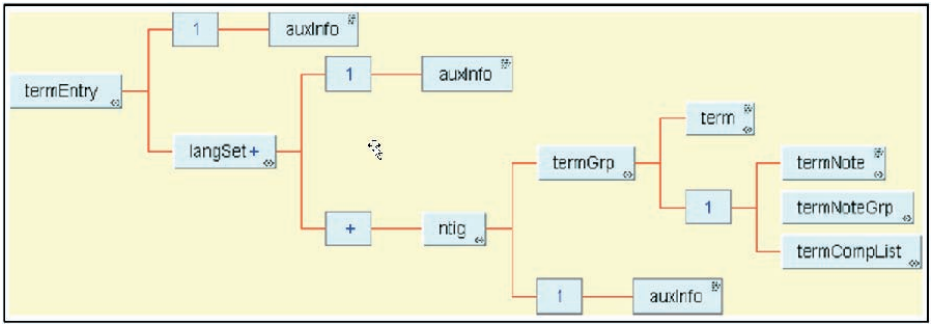

Fig. 16: Structure of a terminological entry in TBX

6 The Role of Standards in an Automated Workflow

The interchange scenario described in the previous sections calls for standardized interchange between NLP systems. There are already workflow scenarios where the only way of cost-effective and efficient transfer of data from one tool to another and from one phase to another consists of using standardized formats. This is, for example, the case in software localization where standards play a predominant role in the localization process (see Fig. 15)[2].

Concerning terminology interchange the Localization Industry Standards Association (LISA) propagates TBX. TBX is an XML-based terminology markup format that is consistent with ISO 12200 (MARTIF).

A TBX file consists of a header that describes the file, a set of entries, one per concept in the termbase, and a set of terms for each concept, which designate the concept, and which are grouped by language. Thus, the structure of a terminological entry in the body of a TBX document distinguishes three levels (see Fig. 16): the entry level (<termEntry>), the language level (<LangSet>), and the term level (<ntig>). TBX therefore provides all prerequisites for supporting concept-oriented terminology work and guarantees a number of benefits for terminology exchange provided that it is supported by more than one commercial system.

7 Conclusion

We have to conclude that standardized interchange formats for platform-independent terminology interchange are still rarely supported by commercial systems. Regarding the supported import formats of terminology systems, CSV instead of TBX turns out to be a quasi-standard at least if we use the number of systems supporting this format as an indicator. The export to CSV or tab-delimited files may, however, be problematic when line breaks occur in descriptive text fields, or when the number of descriptive fields used differs between several entries, as could be seen in the case of the MultiTerm iX export. Here, reusable data are only generated if the type and number of information describing an entry is homogeneous over all entries. Another problem may occur if the structuring and the number of fields used in the entry structure of one system is not compatible with the number of fields allowed in the receiving system.

There is no doubt that standards are indispensable, not only from the point of view of the user, but also with respect to complex workflow scenarios. Perhaps, new industrial alliances as they were formed in 2005 will enforce the support of open source formats. From the point of view of the terminologist as well as from the point of view of the company which has to handle terminology in complex workflow situations the li-

mited use of standards in terminology exchange by commercial systems is rather disillusioning.

References

Alder, A. C. (1998). "An Experiment in Blind Terminology Interchange: Developing and Testing Conversion Algorithms for Externally Supplied Data". Master Thesis, Brigham Young University.

LISA (Localization Industry Standards Association). *http://www.lisa.org* [13.01.2006].

Melby, A./ Wright, S. E. (1998). "The CLS Framework Overview". *http://www.ttt.org/clsframe/overview.html* [19.01.2006].

Melby, A. / Wright, S. E. (2000). "The CLS Framework". *http://www.ttt.org/clsframe/index.html* [9.11.2005].

OLIF (Open Lexicon Interchange Format).*http://www.olif.net/* [19.01.2006]

OSCAR (Open Standards for Container/Content Allowing Re-use). *http://www.lisa.org/sigs/oscar/* [13.01.06].

TBX (TermBase eXchange). http://www.lisa.org/standards/tbx/ [13.01.2006].

Wittenburg, P. / Gibbon, D. / Peters, W. (2001): "Metadata Elements for Lexicon Descriptions". IMDI Technical Report. *http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf* [16.01.2006]

Zenk, W. (2006): "UniTerm – Formats and Terminology Exchange". In: Geldbach, St., Seewald-Heeg U. (eds.): "Exchange of Lexical and Terminological Resources", LDV-Forum 21(1), pp 19-26.

TMS Vendors

Acolada (*www.acolada.de*) [13.01.2006].

across (*www.across.net*) [13.01.2006].

Atril (*www.atril.com*) [13.01.2006].

GFT (*www.gft-online.de*) [13.01.2006].

SDL (*www.sdl.com*) [13.01.2006].

Star (*www.star-group.net*) [13.01.2006].

Trados (*www.trados.com*) [13.01.2006].

Endnotes

1 For a detailed discussion of UniTerm, see also the contribution by Zenk in this volume.

2 This model of the localization process was created by Pierre Cadieux, president of i18N Inc. (*www.i18n.ca*).