Tiptoeing towards TBX

Strategies for terminology management at a language services provider

David Calvert,
TransForm Gesellschaft für Sprachen- und Mediendienste mbH
d.calvert@transformcologne.de

Introduction

This paper describes some of the strategies which a language services provider is compelled to adopt to deal with the issue of terminology. It is based on my experience at TransForm GmbH, a small LSP based in Cologne, Germany.

Translators soon learn about terminology—how useful it is, how infuriating it can be, and its high potential for causing addictive, time-wasting behaviour (sometimes referred to as "terminological research on the Internet"). Terminology work is a part of what translators do—and not necessarily because we want to.

Terminology can be defined as "the language discipline dedicated to the scientific study of the concepts and terms used in specialized languages."[1] This quote is taken from the Pavel Terminology Tutorial—which is an excellent place to start learning about terminology and its application. There is, of course, a wide range of other literature, courses, books etc. available on the subject of terminology. But it's really about organizing and structuring data. Terminology is defined in terms of the subject area—what might be called the general context of the term. Terms represent concepts in use within particular subject areas.

Here's another quote from Pavel: "The importance of correct terminology to accurate and effective communication in special languages has become increasingly obvious … as has the need to standardize the terminology used by groups of individuals and organizations with common interests."[2]

The structure of terminological data

The basic unit for terminology work is the concept. Multiple terms—synonyms—can represent the same concept within one subject area. Such synonyms can be distinguished by their various criteria such as their degree of acceptance or conformity to particular standards, or their use by a particular company or department. They may also be used differently in different geographical regions. For example, the terms "context", "terminological context" and "example of use" are synonyms in the subject field terminology work. The first is a short form, the second the full form, and the third, a non-technical synonym. Their use in discussing terminology might be allowed, preferred and discouraged, respectively.

This conceptual orientation is distinct from the alphabetic organization of a dictionary or a glossary. An English dictionary entry normally contains definitions of one or more concepts that can be represented by a single headword.[3] A thesaurus is also alphabetically organized, listing synonyms for each individual concept. A monolingual glossary is a simple, alphabetically organized list of one-to-one correspondences between terms and the concepts which they represent.

A terminology database, in contrast, is organized by concept. However, all of the information required to generate dictionaries, thesauri and glossaries is contained within a terminology database, even though the primary form of indexing in the database is different.

Because terms and their use vary according to factors such as region, company or product line, usage labels are necessary to distinguish between them. Administrative information is needed, for example to document who created or last modified a term, and whether the term is admitted, preferred or deprecated, for example. And grammatical information may also be useful, for example to determine if the term is a noun or a verb. The number and nature of textual supports and usage labels is determined by the purpose for which the terminological data is being stored.

So we need a concept-oriented structure with various types of usage labels to distinguish the various terms associated with a concept. The lexical arrangement of a dictionary is not suitable for the storage of terminological data, as it does not completely reflect the necessary conceptual organization.

Translators and terminology

So far, I've described terminology work as a monolingual discipline. As such, it is best applied on the authoring and documentation side. As far as translators are concerned, this kind of monolingual terminology work is of limited application. It is, however, very relevant to our work, especially when it has not been adequately carried out, as it is one of the main factors governing the quality of the original document.

The terminological data in use at an LSP is multilingual, and thus also has to take account of differences between the conceptual structures of the source and target languages. Different types of equivalence, such as inclusion as opposed to complete equivalence, play a greater role, and the terminologist's job, which is already hard enough, becomes much more difficult. For this reason, multilingual terminology records require more sophisticated usage labels to differentiate between domains, regions and customers, for example, and to incorporate additional information.

Added to the complexity of multilingual terminology work is the fact that individual translators and small translation companies are frequently faced with an extremely wide range of subject material. It is quite normal to work on a jeweller's customer magazine, a popular science magazine covering the work of an international company's corporate research department and an automotive company's sustainability report, all on the same day. The demands on any collections of terminology are equally varied.

Translators have a love-hate relationship with terminology. Terminology is concerned with meaning—the translator's essential interest—yet the resources available for terminological research are strictly limited. So translators and LSPs need access to short cuts such as specialist dictionaries and customers' in-house terminology.

In practice, LSPs need a wide range of terminological data, the exact scope of which is very dependent on the type of work being carried out. This data can be supplied by the customer or generated internally. Whereas an individual translator may only need a simple glossary containing source and target terms for a specific customer, any work involving machine translation is going to require the production of domain-specific dictionaries containing a great deal of linguistic information.

However, while some customers will have their own terminological resources and will be willing and able to make them available to the translators, other customers will either regard their terminology as top secret intellectual property or have no awareness of what terminology is.

Terminology for LSPs in practice

Terminology work is best carried out by terminologists working with the developers, documenters and product managers. However, although it is reasonable to expect that customers have developed their own terminology for their products, this isn’t always the case. This is despite the fact that terminology is a good investment for companies producing and exporting technology. One case study showed a return on investment of 172 per cent over the first three years of a project to establish terminology for a client[4], and a clear approach to quantifying costs and ROI on such projects[5] has also been presented. Nonetheless, many companies still haven’t developed monolingual terminology for internal use, let alone made it available for translation. Even when terminology has been established, there may be administrative, technical or political reasons why external service providers do not have access to it.

When such a situation arises, the LSP has no choice but to start thinking about how it is going to ensure at least a minimal degree of terminological consistency. From the LSP's viewpoint, this terminology work is both a significant cost factor and an investment, albeit in many respects a very speculative one.

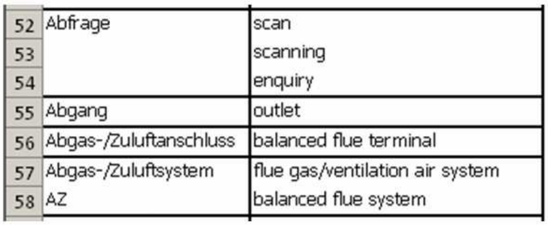
The individual terms in terminology records should be stemmed or lemmatised. The capitalization should be correct for the term as it would appear in context—i.e. in the middle of a sentence in the target language. The singular form should be used, unless the term is always encountered in the plural form. Terminology records for use by a LSP are subject to a number of special considerations. For example, it may also be necessary to store an inflected form if this form is going to produce more and/or more accurate hits for the automatic terminology recognition module of a CAT suite.

Sources of terminological data

Common sources of terminological data for translators are the customers themselves, either knowingly when they provide material or otherwise, when Google is let loose on their websites; monolingual and multilingual dictionaries on paper or online; encyclopaedias and general interest publications on paper or online; specialist publications on paper or online; self-citation from previous jobs; and whatever the translator’s magpie brain has picked up over the years.

Customers’ existing terminology

Customers’ existing terminology can be of extremely variable quality. It can be lexically or conceptually organized, and delivered in any imaginable form. Such externally supplied terminology will usually be processed in some form of text-based file such as Word or Excel before being imported into the appropriate terminology database for the LSP’s preferred CAT tools suite. OCR or file format conversion may be necessary before this processing can be carried out. The processing itself may be also be complex and difficult, involving a substantial amount of manipulation to convert complex and disordered data into usable structures, and to correct, unify and lemmatise the resulting terminology records.



52	Abfrage	scan
53		scanning
54		enquiry
55	Abgang	outlet
56	Abgas-/Zuluftanschluss	balanced flue terminal
57	Abgas-/Zuluftsystem	flue gas/ventilation air system
58	AZ	balanced flue system

Figure 1: Excerpt from an Excel list of terminology supplied by a customer

The example shown is an excerpt from an Excel table with more than 3,500 rows. The first three rows (52–54) show one German term corresponding to three English terms. Row 55 is a simple one-to-one correspondence, as is row 56, although in this case it doesn’t look like it. Rows 57 and 58 contain two German and two English terms. In this excerpt, the heavy borders do indicate the concepts, although this was not true throughout the file. Fortunately, Excel provides reasonably sophisticated string functions which can be used to tackle such problems. These functions can, for example, be used to mark characters such as commas or parentheses which frequently indicate multiple terms or term forms in a single column.

Collection of terminology

In-house, LSP employees can research terminology in advance using existing material such as supplied TMs, literature or any supplied bilingual corpora, and by actively searching for material to build a bilingual corpus for the specific job. They can also collect “terminology on the fly” during the translation stage and during the revision and review steps

External suppliers such as freelance translators also carry out terminology work when translating texts. The challenges for an LSP here are, on the one hand, to ensure that the translator conforms to the use of established terminology and, on the other, to ensure that the benefits of any terminology work carried out by the external supplier are also available to the LSP and are communicated to any other suppliers and personnel working on the same project.

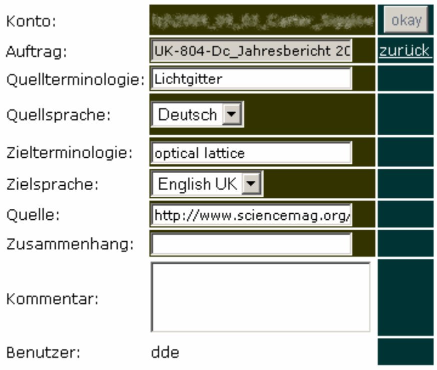
People don’t always appreciate the lengths to which translators will go to turn in a good job. One of our regular freelancers sends us his terminology research notes cut-and-pasted from the results of various searches. These can run to 35 pages in a 1.5-MB file—for a 2,200-word translation.

Our strategy when collecting terminology can be summed up as “record terminology which you feel the need to research.” The theory is that if one translator needs to look it up, we can save the next translator the effort. Conversely, translators sometimes fail to record terms which might take other translators hours to track down. In such a case, the reviser or reviewer applies the same test of “Do I need to research this?”. All of this is, of course, subject to the pressure of deadlines and dependent on the likelihood of repeat work from the customer in question.

Training and assistance with issues such as lemmatisation, what information to put in which columns, and how to effectively research the information in the first place are essential, both for in-house personnel and for external suppliers if they are to cooperate in this area. Over the years, we have developed simple forms for glossaries containing the minimum necessary information to establish a terminology record. We supply freelancers with a simple MS Word template file in A4 landscape page format containing a style called Wordlist, which has tabs set for source term, target term, note and term source.

Wordfast’s ability to support multiple glossaries in the form of tab-delimited text files is also very useful when working with and as an external supplier. We have defined our use of the columns in a Wordfast glossary as source term, target term, note, term source, context sentence and context source. The translator can, for example, define glossary 1 as supplied, job-specific terminology, glossary 2 as new terminology and glossary 3 as existing general terminology. New terms can be added to one or more glossaries as the terms are researched, and glossary 2 returned with the job.

In-house, we use a simple form accessed through our intranet-based job-planning database to record terminology for specific projects. This form deliberately has space for only a limited amount of information. All terminology collected in this way is instantly available to any in-house personnel assigned to the project and can be validated and exported in MultiTerm format.



Konto:	[redacted]	okay
Auftrag:	UK-804-Dc_Jahresbericht 20	zurück
Quellterminologie:	Lichtgitter	
Quellsprache:	Deutsch ▾	
Zielterminologie:	optical lattice	
Zielsprache:	English UK ▾	
Quelle:	http://www.sciencemag.org/	
Zusammenhang:		
Kommentar:		
Benutzer:	dde	

Figure 2: Intranet term entry screen

Affordable strategies for terminological research in an LSP

Affordable strategies for terminological research in an LSP are extremely constrained. Internet discipline, the use of known research sources such as bilingual corpora, and a structured approach to using Google are essential. Sites based on user-generated content are useful when approached with caution. Where sourced terminology is not available, a combination of informed creativity and consistent practice has to fill the gap.

Particularly useful free sources can be divided into three categories, institutional databases and corpora, user-supported sites and search engines.

Institutional terminology databases and corpora include the EU’s IATE and the Canadian government’s Termium Plus, both of which offer public access to large databases of multilingual terminology, and EUR-Lex, the EU’s website of legal and other public documents, which can be regarded as a multilingual corpus. Searching the corpus for a source language term and then calling up a bilingual display of the document is relatively easy.

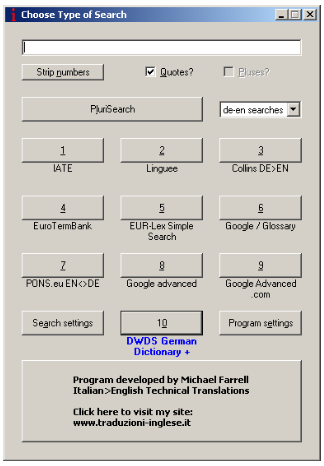
User-supported multilingual dictionaries and forums such as LEO and dict.cc can be helpful, but must be approached with caution, as the quality of much user-supported material online can vary between excellent and appalling.

Linguee is a new development combining search engine technology with multilingual corpora and user feedback. Founded by an ex-Google postdoc and a colleague, it presents itself as “Linguee—The Web as a dictionary.” It currently offers German and English only, but the company states that other languages are planned and funded. Linguee crawls the Web looking for multilingual documents and also uses donated corpora. The material is rated and verified both automatically and by humans. User-supported features are also provided for editing and rating results.

Linguee has rapidly become one of our regular search tools. It is, however, currently noticeable that a large proportion of the bilingual corpora used originate from .de domains. This results in a significant proportion of the results found appearing to be the product of non-native speakers or even machine translation. The in-house human and user-supported editing and verification provide mechanisms to alleviate this problem, but the use of a larger, and thus more balanced German-English corpus will undoubtedly further improve the quality of the results.

What Linguee appears to do is similar in many ways to a structured approach to using Google. Simply including the name of the target language along with the source term will return any indexed multilingual Websites with the target language, and will also bring up many user-supported sites with appropriate entries. Such hits include forums such as ProZ’s Kudos system, which is mostly used by professional translators, and is of corresponding quality. Unfortunately, “search engine optimization” means that some hits are for dictionary and forum sites which do not contain any useful information about the term, but invite you to add such information. It is easy to query Google with logical expressions, or expressions and phrases, while filtering by file type wanted. Searches can also be limited by domain, which can prove very useful when trying to establish whether a target term simply looks strange or is a direct translation by a non-native, non-specialist translator. This feature can also be used to get an impression of the regional prevalence of a term, although the relative size of the domains search has to be taken into account when doing this. Other options on the Google Advanced Search screen are searches by Language, which is not particularly useful, as a large proportion of the Web seems to be declared as the default language, regardless of what the site actually contains.

With so many possibilities for searches, the issue of cost vs. benefit becomes significant. One tool which makes it possible to define a number of searches with different parameters on different websites, then to execute them simultaneously on a selected term by means of a hot key combination is IntelliWebSearch, a compiled AutoHotKey script by Mike Farrell. This utility copies and cleans up a selected search term and sends it to a number of preset websites.



Program developed by Michael Farrell
Italian>English Technical Translations

Click here to visit my site:
www.traduzioni-inglese.it

Figure 3: IntelliWebSearch search dialog

It is provided as freeware, but requires significant customization. The author offers paid training via the Web.

Validation of terminology

Terminology records generated during the course of a translation job should be validated before storage. Terminology established at the pre-translation stage should be validated at this stage, before it is distributed. When working with a translation process according to EN 15038, it makes sense to validate any terminology captured during the translation and revision stages at the revision stage. Final adjustments may be necessary at the review stage. The appropriate strategy is very dependent on the nature of the material to be translated. Where a substantial body of validated terminology already exists, pre-translation terminology work is usually unnecessary.

Validation of a terminology record for translation use involves making sure that the data is useful, that it is correctly recorded and that its content is correct.

The first step in validating a terminology record for an LSP has to be the question: “Is it useful?”

Translators, especially if they lack experience of terminology work, sometimes record terms that are not especially useful, such as relatively common words used in conventional ways. Such records can be discarded before wasting more time on them.

Correct recording requires that the term itself be recorded in a standardized form—lemmatized or stemmed. The correct form for recording a term will be specific to each language and part of speech, but will normally represent the simplest form. For an English noun, this is the singular form, written in lower case—unless it is a proper noun or a trade name written in capitals. For English verbs it is conventional to drop the infinitive particle. German nouns must be recorded singular, capitalized and in the nominative case.

Correct recording also requires that all of the information necessary for a terminology record be collected.

Our strategy here is to collect the minimum amount of data to make the record useful, while ensuring that we have sufficient data to deduce or research additional information should we require it.

In addition to verifying the formal correctness and completeness of the terminology record, the correctness of the data must be verified. If the translator has understood the term and its context, and correctly recorded the source, then this element of the validation is basically a "sanity check". If, however, the translator has failed to find a convincing source, or if the reviser is unhappy with the use of the target term, then the reviser should take corrective measures. The same applies to a reviewer, should the service specifications require a review stage.

It is important to bear in mind that, for an LSP, the source and target terms are treated differently. The source term may need to be lemmatized or stemmed and, if obvious problems exist, a query raised with the customer, but it must basically be taken as correct and recorded. The target term, on the other hand, must be additionally validated as an equivalent term, and this equivalence documented.

Constraints

The obvious constraints on terminology work in an LSP are time, money and expertise, and are interrelated.

The justifications for carrying out terminology work at an LSP are that it: generates cost savings by eliminating duplication of effort; improves the quality of the finished product; and can be sold to the customer as a value added service. All of these justifications are subject to the constraints previously mentioned.

Further constraints on terminology work in an LSP include software and file format compatibility, and access to specific packages. Issues here can be as basic as the character encoding. Non-Unicode software still exists and is still in widespread use. We still sometimes occasionally encounter problems with extended characters in RTF files becoming corrupted between Windows and Mac OSX applications, and with issues of unsupported or differently encoded glyphs in typefaces for DTP, the euro symbol and typographers' quotes being good examples of the latter issues. The software in use also limits the nature of the terminology records that can be stored. Wordfast, for example, permits the storage of source and target terms in a simple glossary along with some additional information. MultiTerm, in contrast, provides an open-ended, user-definable, concept-oriented database.

The information necessary for a terminology record will vary according to the purpose for which the terminology is being collected. For purposes of automatic terminology recognition and human translation, it is not always necessary to record gender and part of speech, although this information is essential if the data is being collected for machine translation.

There are also arguments for storing terminological data in a form appropriate to the nature of the processes utilizing it—normally the form most likely to generate the highest proportion of useful hits. In some cases, it may be sensible to store a plural or otherwise inflected form of a term instead of or as a synonym in addition to the lemma. These additional forms will depend on the language, the part of speech and the program used. At present, most CAT software uses fuzzy matching for automatic terminology recognition. Such systems have problems dealing with significantly different plural forms. One approach to this issue is exemplified by Wordfast glossaries, which can be prepared for "manual fuzzy terminology recognition" by using asterisks as wildcard characters. Such an approach conflicts with the terminological policy of storing the term as its simplest form and raises the question of whether to maintain formally correct terminology records, relying on an element of fuzzy and/or linguistic processing in the software, or to store the terminology in a form that is immediately suitable for delivering the highest number of useful hits in the software environment used.

There is, however, a strong trend toward incorporating a greater element of linguistic processing in CAT systems, trading language-independence for increased capabilities such as subsegment matching, and a degree of machine translation capability. These developments suggest that following the terminologists' approach of storing the term as its simplest form and with additional linguistic information will pay off in the medium term. At present, however, it can be appropriate to maintain glossaries and terminology databases for use with specific software.

Formats, incompatibilities, solutions

The result of the situation described above is a flood of data in a wide range of formats. Different solutions tend to be implemented, often on an ad hoc basis, and terminological data with varying degrees of reliability is collected in a number of different forms, not all of which permit easy interchange. The terminology applications of the CAT tools in use at an LSP may not all be available to everyone involved in the translation and terminology research processes, the applications may not be compatible, and they may not be capable of exchanging data.

Taking TransForm GmbH as an example, our main in-house TM system at present is SDL Trados 2006. This is not a completely satisfactory combination, but porting more than ten year's worth of Workbench and MultiTerm data to new systems is not a trivial task, and we have put it off as long as we could. MultiTerm is a very flexible, concept-oriented database, and we have historically based our terminology records on the default structure. The data is organized by customer, with differing attributes for different customers, so although the basic structures of data for customer A and customer B are broadly similar, the detailed structures are not. There has, however, been a tendency to simplify the amount of usage data collected over the years. We send out presegmented Word files with inserted terminology for translation in Trados or Wordfast. For many years now, we have recommended that our regular freelancers use Wordfast—in the form of the Classic, Word-based version.

External translators supply terminology in MS Word glossaries. In-house translators use the intranet-based system to capture terminology. Customers deliver whatever they think best.

Our reasons for sticking with such an old terminology system are a combination of its compatibility with Translator's Workbench, its utility as a flexible concept-oriented terminology database system, and an element of lock-in due to MultiTerm's proprietary database structure and the fact that migrating terminological data between systems has been considered notoriously difficult. Although Wordfast is an excellent TM tool, it only supports simple glossaries. While this approach has much to commend it, especially for individual freelancers, it is inadequate for our terminological needs as an LSP.

Practical consequences

The consequences of this situation include restricted interoperability, a need to run concurrent, incompatible systems, and pressure to upgrade to extremely expensive server-based solutions.

The market for TM systems is also becoming increasingly fragmented, with a wide range of systems in use. One result of these developments is an increasing need for a means of consolidating and maintaining terminology independently of proprietary formats. The emergence of TBX as an open standard more or less enthusiastically adopted by many of the main TM system vendors has opened the door to fulfilling this need.

TBX and TBX-Basic

TBX has been developed by OSCAR, the open standards body of LISA, the Localisation Industry Standards Association. TBX is an open, XML-based standard for the exchange of terminological data. It is soon to be published by ISO as an international standard. It grew out of the Machine-Readable Terminology Interchange Format (*MARTIF*), which itself built on earlier initiatives to promote the interchange of terminological data[6]. The TBX format offers substantial advantages to the user:

As an open standard, it is effectively future-proof

As an open standard, there is a degree of pressure on tool vendors to support the format

It is clearly defined

It is relatively easy to work with because it is a form of XML, itself a well-defined, simple, well-understood text-based format

It is available for use without licensing fees

TBX offers a further advantage in the form of TBX-Basic. Because TBX is capable of much more than handling the relatively simple terminological markup required by small and medium-sized language industry applications, the LISA Terminology Special Interest Group came up with a lightweight version of TBX known as TBX-Basic. This is a terminological markup language (TML) aimed at users of the sort of terminology resources that are commonly developed to support translation and localization processes.

TBX-Basic has a simple three-level structure. As in MultiTerm, all terms grouped together in a single concept are considered synonyms. The terms themselves are grouped by language. The highest or concept level can hold information such as subject description, a definition with or without a source and cross-reference to, e.g. an image file. The language level can also hold a definition with or without a source, and the term level holds all the information specific to an individual term. Administrative and transactional information can be present on any level, as can notes.

It is fully compliant with TBX, but offers a restricted subset of those TBX features considered most useful for smaller applications

Two types of compliance with TBX and TBX-Basic can be distinguished. These are compliance on the structural and syntactic level and compliance on the content level. Compliance on the structural or syntactic level is relatively easy to check by using a suitable validation program such as tbxcheck, which is available from http://sourceforge.net/projects/tbxutil/

Compliance on the content level is another matter, as it can depend on the purpose for which the terminological data is maintained. For example, in TBX-Basic, each term in any data intended to be submitted to any form of machine processing must have a part of speech explicitly indicated. if the data is only intended for human consultation, the part of speech may be omitted, provided that either a definition or a context is present. Furthermore, both definitions and contexts are defined in greater detail.

Our answer: a TBX-compliant terminology repository

I decided that the best way forward was to use TBX or TBX-Basic to store all of our terminological data. By converting all of our data to a defined, open format with significant and increasing support in the industry, we would ensure future compatibility. The discipline enforced by conforming to a standard would also tend to improve the quality of the data over the medium and long terms. By making the system capable of handling TBX, we automatically made it possible to use TBX-Basic for our existing data without precluding changes in the terminological markup language—the dialect or subset of TBX—at a future date.

The use of a widely supported open standard to store terminological data also opens up the possibility of generating different subsets of that data in different forms, for example as glossaries in the format for use with a specific program, or in the shape of an automatically formatted online HTML or PDF dictionary.

This capability enables decoupling the repository data from the current state of data subsets specifically generated for use with a specific program. Although the data subset required for a specific use is stable, the individual data records can be expanded within the standard defined by TBX-Basic and TBX without affecting the mechanisms used to generate the application-specific file. Or, to rephrase that, we can use as much or as little of the stored XML data as we require for a specific purpose without compromising the stored data.

The restricted subset of TBX features in TBX-Basic actually fits with the reduced amount of usage data that we have moved toward collecting, which encouraged me to plan on using TBX-Basic as the format for storing all of our terminology records. As our terminological data has been collected over at least 16 years, immediate full compliance with TBX-Basic on the content level will only be possible for data collected once the new system has been implemented and integrated into our operational processes. Legacy data will need additional manual input to be brought to compliance. Careful use of the implicit information mentioned previously will help to mitigate these issues.

This approach enables future extension of the legacy data to include, for example, grammatical gender and part of speech.

It also offers a solution to the dilemma of whether to store inflected forms or forms with wildcard characters as used by some Wordfast glossaries, for example, although the latter step in particular would require the use of extensions not covered by TBX-Basic and would raise compliance issues. One possible way forward here would be to store the additional information necessary to generate such non-standard datasets as metadata and generate the datasets by an export process as required.

Data conversion

The main issue in converting data between terminological database formats is mapping. TBX-Basic uses a concept-oriented, three-level structure with concept, language and term levels. The information stored on each level is constrained, both in terms of what may be stored and what must be stored. Both explicit and implicit information must be handled by the mapping

Implicit terminological information

Whether supplied by customers or external partners, or generated in-house, terminology records contain explicit information and often have implicit information such as their source or the domain associated with them. Due to the structure of the collected and supplied terminology data, some of this information may be implicit for one half of a source-target pair and explicit for the other. At TransForm GmbH, our legacy terminological data is organized into termbases for specific customer organizations, which themselves may contain usage labels for distinctions by organizational units such as subsidiary or department on the one hand, and by geographical region on the other. We have also commonly recorded usage labels for terminological data for the target language terms, but not for the source language terms.

This asymmetric nature is a feature of terminological data captured during the translation process. Translators tend to find glossaries the most useful way to deal with terminology. The existence of a concept is implicit in the existence of a pair of source and target-language terms, but the concept is not explicitly stored in a Wordfast glossary, for example.

The process of importing existing or newly recorded terminological data into a data repository must aim to capture all such implicit information. The consolidation of different databases into a repository also requires special attention to usage labels to preserve the information implicit in the original arrangement of separate databases and to ensure customer confidentiality.

Handling implicit and explicit information in terminology records

Implicit information can be derived from existing data, or can be manually entered during the import process. Establishing which implicit information falls into which of these categories is an important part of defining the mapping to be carried out when converting the data.

Some source language term usage information can be inferred from explicit information available for the target term. For example, the source of a source language term can reasonably be taken to be the customer who has sent the job for translation. This can be obtained from the database from which the term record originates. Some of the usage information stored for the target language term may also be applicable to the source language term; for example, the project information stored may provide information on a specific organizational unit.

In the case of our existing MultiTerm databases, this information must be derived or manually entered when the terminology records are transferred to the new system. WordFast glossaries and other similar terminology lists are usually related to a specific job, so the customer, customer account and job also represent important implicit information which has to be captured at the time of data import. Supplementary information on terminology captured in our intranet database system can be derived from the job planning database.

Terminology recorded in the intranet database is linked to customer and account information, so a substantial amount of usage information can be derived when the data is ported.

We are currently developing the use of the input and output templates to take account of such information in the import and export processes.

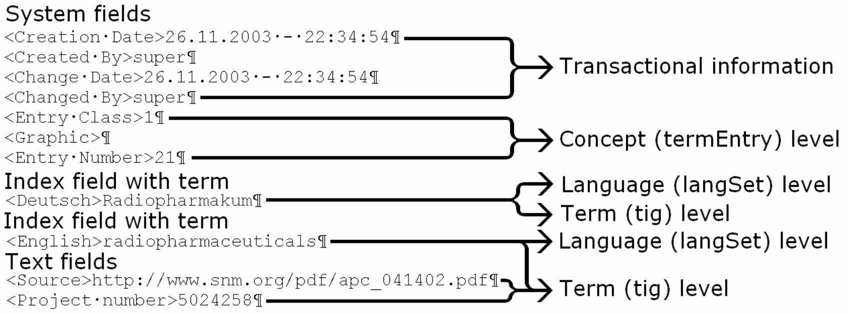
Examples of mappings to TBX-Basic

MultiTerm 5.5

MultiTerm 5.5 was the last file-based version of Trados' original product. It is a flexible concept-oriented system which makes use of four different types of field—index fields, which are defined as languages and contain the terms themselves; attribute fields, which can hold values selected from user-defined picklists; text fields; and system fields holding administrative information.

The order, number and relationships of attribute and text fields are not constrained. This flexibility can make it difficult to predict what's coming next when parsing a MultiTerm export file.

The basic mapping for a simple MultiTerm 5.5 export file to TBX-Basic is shown in Figure 4.



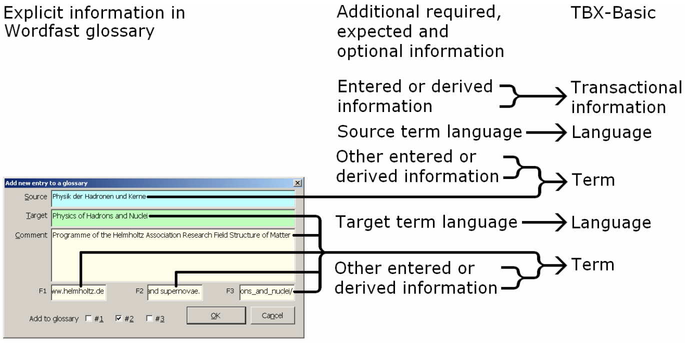
```
<Creation·Date>26.11.2003·-·22:34:54¶
<Created·By>super¶
<Change·Date>26.11.2003·-·22:34:54¶
<Changed·By>super¶
<Entry·Class>1¶
<Graphic>¶
<Entry·Number>21¶
<Deutsch>Radiopharmakum¶
<English>radiopharmaceuticals¶
<Source>http://www.snm.org/pdf/apc_041402.pdf¶
<Project·number>5024258¶
```

Figure 4: Simple MultiTerm to TBX-Basic mapping

The system fields are mapped to transactional information and to the concept level. The date fields require a format conversion. The index fields contain both the language, which must be translated to the language subtag of the XML language tag and mapped to the language level, and the term itself, which must be mapped to the term level.

The text fields are mapped to the term level.

Wordfast



Add new entry to a glossary

Source: Physik der Hadronen und Kerne

Target: Physics of Hadrons and Nuclei

Comment: Programme of the Helmholtz Association Research Field Structure of Matter

F1: ww.helmholtz.de F2: and supernovae. F3: ons_and_nuclei/

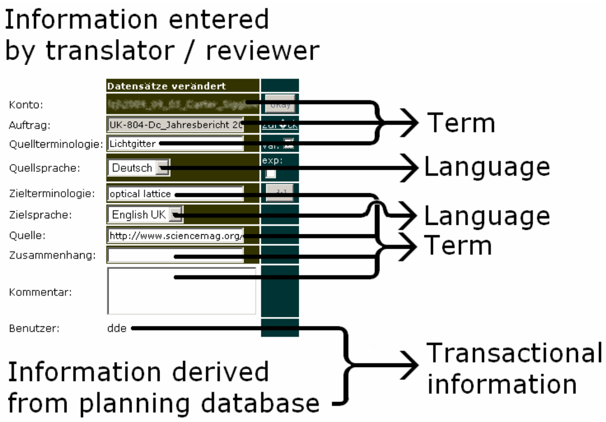
Add to glossary ☐ #1 ☑ #2 ☐ #3

Explicit information in Wordfast glossary
Additional required, expected and optional information
TBX-Basic
Entered or derived information
Transactional information
Source term language
Language
Other entered or derived information
Term
Target term language
Language
Other entered or derived information
Term
OK
Cancel

Figure 5: Simple Wordfast to TBX-Basic mapping

The simple glossary format means that much of the information required to create a TBX entry must be either derived or input when the data record is converted. Source and target languages, project information and some transactional information must be entered.

Intranet terminology



Datensätze verändert

Konto: [redacted]

Auftrag: UK-804-Dc_Jahresbericht 20

Quellterminologie: Lichtgitter

Quellsprache: Deutsch

Zielterminologie: optical lattice

Zielsprache: English UK

Quelle: http://www.sciencemag.org/

Zusammenhang:

Kommentar:

Benutzer: dde

Information entered by translator / reviewer
Okay
zurück
var.
exp:
Term
Language
Language
Term
Information derived from planning database
Transactional information

Figure 6: Simple intranet to TBX-Basic mapping

The mapping to TBX is relatively simple here, as most of the data is either present in its final form or can be derived from other tables in the planning database.

Customers' data

Customers' terminological data arrives in a wide variety of forms, many of them involving Excel tables. Excel data is converted to tab-delimited text and imported using a modified version of the procedure for dealing with Wordfast glossaries.

The software

Working together with Wolf Dietrich von Loeffelholz, who has been maintaining and developing software for TransForm GmbH on a freelance basis since 2001, we have used an open standards-based approach to develop a TBX-compliant terminology repository with the ability to import, manage, store and export terminological data. The system was designed for compatibility with our existing software and systems. It is entirely built on open-source software and XML, and is TBX-compliant. It also supports LDAP authentification. The existing input/output formats are MultiTerm 5.5, Wordfast glossaries, SQL, XML and TBX. The data storage format is XML/TBX, with metadata in a database. XML templates are used to support the import and export of the widest possible range of data formats. This makes it relatively easy to add new import/export formats. The hosting system can be Linux, Windows, etc. and the major databases are supported. The XML-based multilingual front end utilizes Joomla CMS as the view layer and is designed for easy localization.

Concept-based management of stored terminology

The system is designed to enable the concept-based management of the stored terminology, including the combination and separation of terms and the maintenance of concept/term histories and ratings.

Export capabilities

One of the features of the approach chosen is that defining an import template effectively defines an export template too. And of course XML is extremely versatile and relatively easy to transform, so that exporting subsets of the data in any format that the system can import, and as TBX data, is straightforward, as is the automatic production of glossaries and dictionaries in a range of forms.

Current status

The system is Web-based, and is accessed via a SSL connection. The site is self-certified. Initial tests of import and export functions have been carried out. Beta testing by TransForm GmbH and further development of the filters and user interface are now in progress. Once the results of these tests have reached an appropriate level, we intend to start further testing by third parties and with data using more complex character sets, and will proceed with the development of templates for the import and export of other formats.

Future strategies for terminology management at TransForm

Our plans call for the import of all of our existing terminological data and its conversion and storage as TBX. Step-by-step validation of existing terminology records where necessary, and the consolidation of non-confidential terminology into domain-specific terminology databases, should prove particularly helpful in eliminating duplication of terminological research. We will also then have all of our existing terminological data in a form in which it can be maintained and preserved without fear of incompatibility with future developments in TM software.

[1] Pavel's Terminology Tutorial

[2] Pavel's Terminology Tutorial

[3] Wikipedia

[4] Wittner, Janaina, Unexpected ROI from terminology — Corporate Importance of Terminology Management and ROI Returns, presentation at the Tekom conference, Wiesbaden, 2008, summary (in English) in Jahrestagung 2008 in Wiesbaden, Zusammenfassung der Referate, p. 305

[5] Maier, Dr. Elisabeth, Improving the ROI of Corporate Terminology Development, presentation at tekom conference, Wiesbaden, 2008, summary (in English) in Jahrestagung 2008 in Wiesbaden, Zusammenfassung der Referate, p. 307

[6] Robin Bonthrone, R: MARTIF Lite: User-driven Terminology Interchange, The Globalization insider, 01 1998