Multi-Modal Terminology Management

Corpora, Data Models, and Implementations in TermStar

Enrico Giai[1], Nicola Poeta[2], and David Turnbull[3]

[1] STAR7, S.p.A., Terminology and Language Technologies Expert
[2] STAR7, S.p.A., Global Content, Service Line Leader
[3] STAR7, S.p.A., Language Lead

Abstract. Terminology is a key part of the translation process. Nonetheless, the benefits of implementing a terminology management workflow using specialist tools and processes is sometimes disregarded, as the benefits in terms of ROI are not always easy to evaluate. As a result, the use of spreadsheets and other inappropriate tools leads to fragmented and inefficient terminology management processes.

In this paper we set out to describe an efficient terminology management workflow which has been developed for real terminology projects. We will also assess the benefits of implementing a proper terminology management workflow where all stakeholders (terminologists, linguists, authors, and end users) are involved. We will highlight the benefits of using a Terminology Management System (TMS) such as TermStar, which can make use of parallel corpora and collaboration functions to streamline the entire process, from terminological extraction to glossary approval and maintenance.

Keywords: Terminology Management, TMS

1 Introduction

Computer-Assisted Translation (CAT) has been at the core of the localisation industry for over three decades. Using CAT tools, linguists can translate more efficiently thanks to Translation Memory (TM) suggestions: CAT tools can leverage TMs to pre-translate content that has been translated in the past or offer 'fuzzy match' suggestions for similar source texts. Consequently, texts translated using a CAT tool are usually more consistent and can be delivered in less time.

While the importance of TMs in terms of quality assurance and economic profit is self-explanatory and can easily be calculated, the added value of setting up a TermBase (TB) is not always evident.

A TB can be defined a "a database comprising information about special language concepts and terms designated to represent these concepts, along with associated conceptual, term-related, and administrative information." [3]. This definition is based on the strict definition of 'term' as being "an expression that designates a particular concept within a given subject field" [9]. As such, it comes as no surprise that assessing

the benefits of investing in terminological work is a hard task: not all organisations make use of highly specialised terminology in their texts, especially in the case of marketing and e-commerce, where the need for technical terminology is scarce.

In this context, the concept of 'termhood' (i.e., the degree to which a term is justified being included in a TB [10]) can be broadened to include a range of words that are vital to corporate communication, despite not being part of a specialised language. These may include product names, organisation and entity names, slogans, frequently used words, or words that appear in sensitive contexts, to name just a few.

Another pain point is the format in which the TB is presented. Commonly, terminological entries are not stored in specialised Terminology Management Systems (TMSs); rather, they are collected in text document lists or in spreadsheets, at best. This is a great obstacle when it comes to organising and sharing terminological assets.

At STAR7 we are aware of the value of a well-structured, centralised TB. Ideally, this can be accessed by all stakeholders in different modes. STAR Group's TermStar has been acknowledged as a TMS that can meet the needs of everyone in the information lifecycle: terminologists, who can take advantage of the highly customisable data model; linguists, who can use TermStar in STAR's CAT tool Transit to have morphology-based term suggestions and use the right terms for each context; authors, who can use TermStar in their authoring tool; and clients, who can access the terminology online via WebTerm – STAR7's solution for online terminology management.

In this paper, we will present STAR7's terminology management workflows and tools aimed at extracting terminology from bilingual corpora, adapting our data models to best fit each term entry, and facilitating the validation and distribution processes for all stakeholders.

2 Related Work

The importance of Terminology Management has been clear since the early days of modern terminology studies as pioneered by Wüster [14]. The onomasiological approach is still a founding pillar of terminology work, and data models in terminography have been shaped to accommodate this concept [10], [11].

While these assumptions are still valid, in recent years the focus has shifted towards a more pragmatic approach. The role of the Corporate Terminologist [11] has surged, and a question has been raised with it: what is a term in a corporate context?

Warburton [10] broadens Pavel's definition [6] of term to "any lexical unit that might help a potential consumer of the termbase". The Terminology for Large Organizations Consortium (TerminOrgs) builds on that by stating:

"To support the communicative aims of large organisations, the notion of a 'term' extends beyond the conventional view to include any expression that, if it is managed according to the methods outlined in this document, brings some benefit to the organization such as improved communication and reduced translation costs. This includes,

sometimes, words from general language, marketing slogans, short sentence fragments, and so forth." [9]

Lexicology and its lexicographic applications have developed significantly over recent decades [4], thanks to corpus research [8] and increasingly powerful technology. However, terminology management in CAT tools often plays second fiddle to translation memory management. Terminology can be confined to easy to use but poorly organised TBs. Specialised terminology is often considered monosemic, but even the most specific term needs contextual details. Technology can offer suitable solutions, such as structured entries, examples, definitions and images.

Despite the high number of TMSs available, not all of them are flexible enough to allow the end user to harness the benefits of the system, especially when used with a CAT tool [5]. TermStar has been praised for its highly customisable data model, which can be adapted to the glossary's needs and even used for lexicography work [7].

3 Methodology

While previous literature on the topic has been the basis for our enquiry, the findings shown in this paper are the result of processes developed empirically over the years while working on actual terminology projects. These involved several different domains, including automotive and agriculture, luxury and fashion, finance and banking, sport and fitness, and pharma. Overall, STAR7 manages over 400 termbases in TermStar and 150 termbases in other TMSs, counting more than 200,000 data records ranging from bilingual to 36-language entries.

Text types also vary accordingly: owner's manuals, service manuals, marketing leaflets, product catalogues, websites, financial reports, and many other text types were used as source texts in the terminology extraction process.

Despite the different nature of these contents, the workflows described in this paper can still be considered valid. The process has been validated internally and well received by all stakeholders. Improvements have been made based on clients' and linguists' feedback.

We have identified five steps which contribute to successfully completing a terminology project. These are described in further detail in the next chapter.

4 Results

4.1 Preliminary analysis

The first step consists of analysing the scope of the project. This can be done by considering the elements listed below with their reasoning:

- Domain: Each domain has its own lexicon and specialised terminology. Determining the domain helps in limiting the scope of the project.

- Text type: Identifying the text type helps in setting the termhood level for the project. The termhood bar for technical documentation might be higher than that for marketing material.
- Languages: Helps in identifying the number of language resources to be involved in the project.
- Budget & Timeframe: Budget is key in determining the resources that can be spent on glossary creation in terms of number of records and data granularity.
- Reference material: Parallel corpora facilitate the terminology extraction process, enabling linguists to extract terms that are actually in use. When not available, open-source corpora can be used. Existing glossaries can also be used as a basis for the terminology work.
- Final audience: Considering the end users is key to understanding how the glossary is to be published. If the glossary is for linguists, it can be implemented in a CAT tool; if the end user is the client, it can be published online.

4.2 Data model setup

Once the project scope is clear, the next step is to understand which data model to adopt. TermStar offers a high level of customisation – the result of lexicography and terminography studies.

A TermStar terminological card follows the traditional onomasiological approach, in that each card represents a single concept. However, TermStar's data record structure allows for a deeper level of content organisation: each term can have sub-entries defined as abbreviations, synonyms, irregular forms, alternatives and disallowed terms.

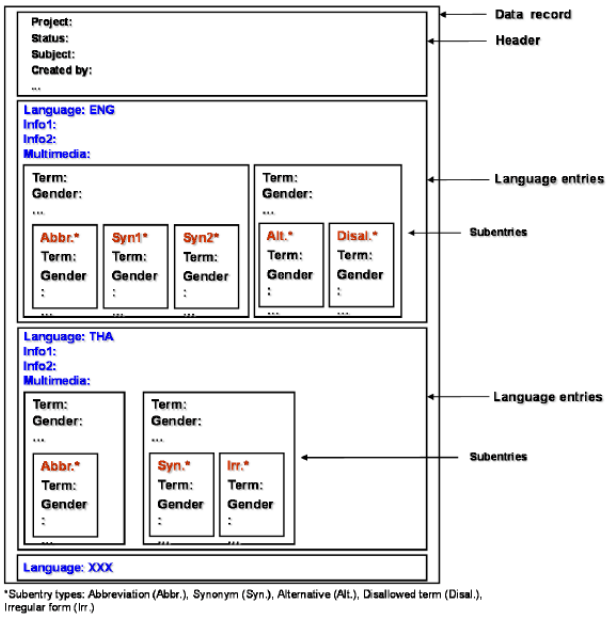

Fig. 1. TermStar data model

This approach is deeply embedded into Transit, whose morphological search capabilities makes it possible for terms to be recognised in texts even if they appear declined or conjugated, while being classified in their base form in the glossary.

In addition, each language entry can be classified using a number of different attributes, including status, data source, definition, definition source, gender, remark, subject, part of speech, and many others. This level of detail is particularly useful to clarify the use of homographs or to distinguish term use based on context (e.g., one term should be used in technical documents and another in marketing texts). Pictures can also be inserted to better clarify complex terms.

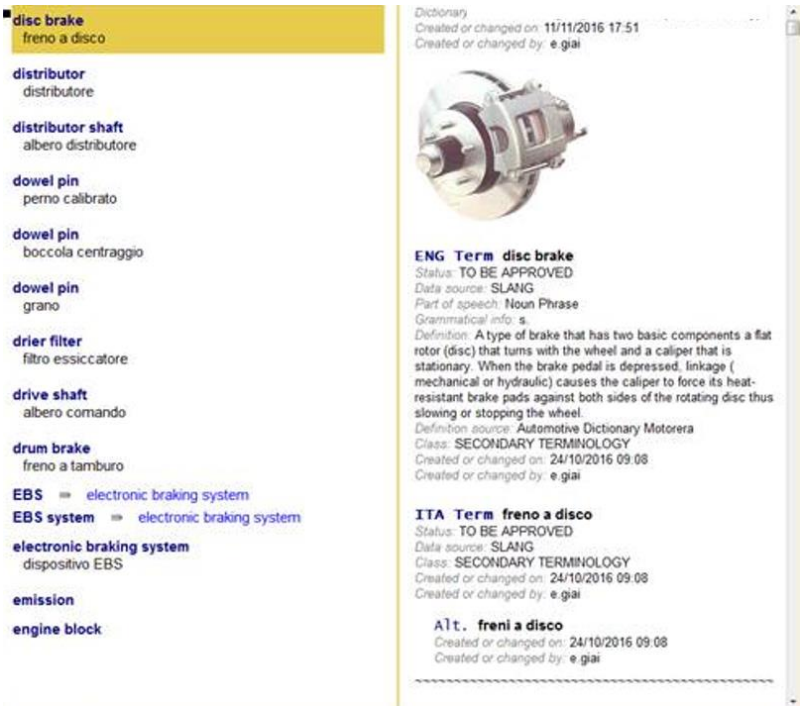

Fig. 2. TermStar data record sample

4.3 Terminology extraction

The terminology extraction step is the most time-consuming part of the process. It can be divided into three steps: (1) source term and (2) target term extraction; and (3) term tagging and consolidation.

Based on budget and time constraints, the terminologist can agree with the customer the number of terms to be extracted and the level of additional information that can be collected.

Despite the number of (semi-)automatic terminology extraction tools on the market, their effectiveness is still far from satisfactory. Most tools are based on frequency and stop-words rules, and even if contexts are offered for each candidate term, the risk of not grasping the correct context or not considering a term in its entirety is high.

For these reasons, source term extraction is usually performed manually, by reading the source texts in their entirety and extracting terms in the process. For us, this is the most effective approach, since terms are not extracted in isolation, but directly from the texts. This also makes it easier to collect context and usage notes.

The whole process takes place in the CAT tool: the terminologist can import source files and create an empty termbase, which will be used for the entire workflow. Terms will be added to the TB, which can be configured to ease the work of the terminologist

(e.g., by setting input verification rules to maintain consistency in the attributes used for each label).

While reading the texts and extracting source terms, the terminologist is able to fine-tune the termhood level and get the most out of the source material. The following table lists possible terms that can be included in a selection of domains.

Table 1. Possible terms in a glossary based on selected domains

Domain/text type	Candidate terms
Luxury & Fashion	Product names, colour names, taglines.
Law / Finance & Banking	Law names, entity and body names.
Corporate communication	Division names, corporate role names.
IT & Software	Button names, menu items.
Technical documentation	Acronyms, abbreviated forms, technology names

Once the source terms have been extracted, the CAT tool can be used to leverage existing parallel corpora (TMs) to facilitate the work of translators. Linguists will be able to run 'concordance searches' to look up source terms in the TM and get a list of already translated sentences. From there, translators can extract any matching term in the target language and insert it in the data record in just a few clicks.

4.4 Terminology validation

Once the glossary is completed, the validation step can take place. This is an essential part of the workflow: without subject-matter expert validation, the glossary cannot be considered as complete.

Usually, validation is performed by clients, or by different client branches around the world. Performing such a task in a spreadsheet would not be efficient. For this reason, STAR7 offers clients a terminology validation process in WebTerm – TermStar's web interface. With it, clients are able to see the glossary without the need for a TMS and can easily add comments and suggestions that can be read by the terminologist and implemented in real time.

WebTerm can also be offered in 'read and write mode', meaning that clients can make changes directly in each data record and changes are immediately available for all stakeholders.

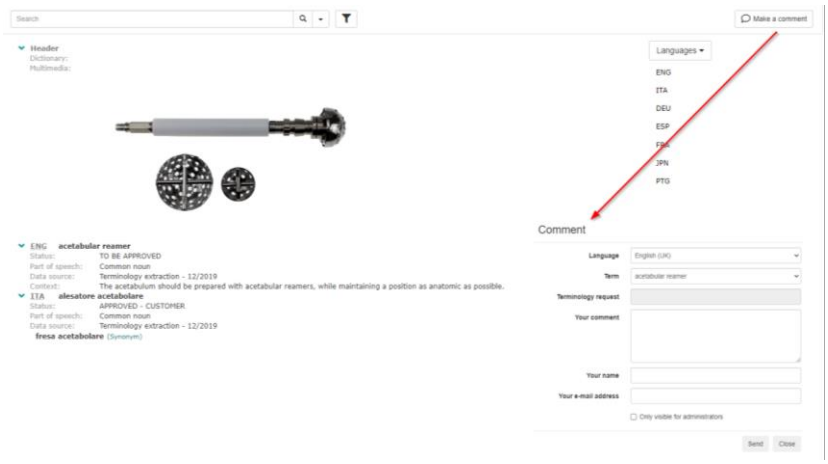

Fig. 3. WebTerm7 comment function

At the end of the validation step, any 'status' metadata associated with approved terms should be updated consequently.

4.5 Termbase deployment and update

Finally, the termbase can be deployed to all stakeholders. When using STAR7's technologies, the TermStar TB can be accessed during the entire information lifecycle:

- Technical authors using selected authoring tools can connect to the TermStar database, or look up terms in WebTerm;
- Linguists using Transit as their CAT tool have direct access to TermStar;
- Clients and reviewers using WebTerm can look up terms, insert comments, make terminology requests, or even edit data records.

Terminology is never static, but it constantly evolves. Technological changes in technical texts, new products launched in marketing material, and changes in term use and preferences should be all recorded as updates in the termbase. For this reason, it is vital to plan a TB update schedule that, based on the available budget and expected workloads, can either be triggered for each new project, for any project which may be particularly important or belonging to a new domain, or on a monthly/half-yearly basis.

5 Conclusion

In this paper we have described in detail a standardised process for implementing a terminology workflow for all use cases. A glossary shared among all stakeholders (clients, authors, linguists, reviewers, etc.) is beneficial in terms of:

- consistency, as a centralised termbase helps to reduce the use of variants;
- prescription, as non-allowed words can be noted;

- time, as linguists can look up terms in a single source instead of multiple, often unreliable sources;
- overall quality, as the corporate terminology will be used instead of general words.

That said, quantifying the benefits in terms of time and money is difficult, as not all texts may contain the terms mapped in the glossary. General productivity can also depend on external factors such as TM quality and linguists' experience and know-how in the subject.

Nonetheless, implementing a terminology management process is still widely recognised as important. We would point out that the fundamental research performed during the TermStar project has laid the basis for further projects within the group. An example of this is the **StarPrinting** project that took advantage of the new terminology management techniques for performing further research on user profiling, with the crucial goal of providing a new and better printing and delivery experience to users.

References

1. Cruse, D.A.: Lexical Semantics, Cambridge University Press, Cambridge (1986).
2. ISO 12616-1:2021: Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography.
3. ISO 30042:2019: Management of terminology resources — TermBase eXchange (TBX).
4. Landau, S.: Dictionaries. The Art and Craft of Lexicography, 2nd Edition, Cambridge University Press, Cambridge (2001).
5. Magris, M., Musacchio, M. T., Rega, L., Scarpa, F.: Manuale di Terminologia. Aspetti teorici, metodologici e applicative. Ulrico Hoepli Editore, Milano (2017).
6. Pavel, S., Nolet, D.: Handbook of Terminology. Minister of Public Works and Government Services Canada, Ottawa (2001).
7. Poeta, N.: Terminologia, corpora e contesto negli strumenti di traduzione assistita. In: Collesi, P., Serpente, A., Zanola, M. T.: Terminologie e ontologie. Definizioni e comunicazione fra norma e uso, pp. 87–94. EDUCatt, Milano (2013).
8. Sinclair, J. (ed.): Corpus, Concordance, Collocation, Oxford University Press, Oxford (1991).
9. TerminOrgs.: Terminology Starter Guide, http://www.terminorgs.net/, last accessed 2022/07/26
10. Warburton, K.: A Practical Approach to Terminology: Developing lexical resources for companies, http://www.ccaps.net/blog/lets-talk-terminology/, last accessed 2022/07/27
11. Warburton, K.: The Corporate Terminologist. John Benjamins Publishing Company, Philadelphia (2021).
12. Wright, S. E.: Data Categories for Terminology Management. In: Wright, S. E., Budin, G.: Handbook of Terminology Management, Vol. 2, pp. 552–571, John Benjamins Publishing Company, Philadelphia (2001).
13. Wright, S. E.: Terminology Management Entry Structures. In: Wright, S. E., Budin, G.: Handbook of Terminology Management, Vol. 2, pp. 572–599, John Benjamins Publishing Company, Philadelphia (2001).
14. Wüster E.: The Machine Tool – An Interlingual Dictionary of Basic Concepts, Comprising an Alphabetical Dictionary and a Classified Vocabulary with Definitions and Illustrations. Technical Press, London (1968).