Journal of Digital Terminology and Lexicography
Vol. 1. No. 1 (2025) DOI: 10.25430/pupj.jdtl.1751631260

How to create and manage terminology resources: a practical guide from two termbases

Natascia Ralli[1*] and Dóra Mária Tamás[2*]

[1]Eurac Research, Institute for Applied Linguistics, Viale Druso 1, Bolzano/Bozen, 39100, Italy.
[2]HUN-REN Hungarian Research Centre for Linguistics, Benczúr u. 33, Budapest, 1068, Hungary.

*Corresponding author(s). E-mail(s): natascia.ralli@eurac.edu; tamas.dora.maria@nytud.hun-ren.hu

Abstract

The objective of this paper is to provide theoretically grounded, practical advice for the creation, management, and reuse of terminology resources, whether starting from scratch or working with an existing dataset. This approach facilitates the swift and efficient design of new terminology resources tailored to specific parameters (e.g. scope, domain, working languages, user group(s) and user situations). The paper highlights the importance of data interchangeability and reuse in today's fast-moving world, emphasising the need for careful planning and integration into quality evaluation processes. Moreover, it highlights the evolution of modern termbases to meet diverse needs, supporting interoperability and collaborative work for terminologists. The theoretical framework is mainly grounded in ISO standards, which outline principles for thoughtful design and quality management. Practical examples of a Hungarian national termbase and the Information System for Legal Terminology *bistro* illustrate the application of these principles, offering insights into the challenges and considerations involved in developing and managing terminology resources.

Keywords: terminological principles, modelling of termbases, handling of terminological datasets, quality issues, national termbase, bistro

1 Introduction

Designing and managing terminology resources is a multifaceted task that requires both theoretical knowledge and practical expertise in terminology work. In this regard, it is worth exploring this concept in more detail before moving on to other aspects.

ISO 1087 (2019) defines terminology work as "the systematic collection, description, processing and presentation of concepts and their designations". This also includes the management of terminology resources, terminological planning, harmonisation of concepts and terms, and term creation. Drewer and Schmitz (2017) expand this definition to include term extraction from texts and the incorporation of terms into texts. In recent years, the English term 'terminology management' has been established as

synonymous with ‘terminology work’. This synonymic relationship is also confirmed by ISO 1087 (2019). It is interesting to note that the term ‘terminology management’ stems from corporate practice and can be described even more broadly (Warburton, 2021). In the organisational-corporate context, it really encompasses the use of tools and the integration of terminology workflows into corporate processes. Even project management measures are inherent parts of building terminology databases, hereafter referred to as ‘termbase’. A termbase collects and organises terminological data (ISO 26162-1, 2019) and is typically part of a terminology management system (TMS). According to ISO 1087 (2019), a TMS is a “software tool with a metadata structure specifically designed for collecting, maintaining, and accessing terminological data”. A TMS may vary according to functionality and platform (Schmitz, 2025). However, it would be appropriate that they support the ISO standard XML format (ISO 30042, 2019) and other formats like Microsoft Excel for importing and exporting data (Warburton, 2021). This technical aspect is relevant for data exchangeability (Früh & Tamás, 2021).

In today’s fast-moving world, the interchangeability, interoperability and reuse of data have become increasingly important. Terminological data can be reused in various fields of information science (Warburton, 2021), including integration into computer-assisted translation (CAT) tools, term injection into machine translation (MT) tools, and the creation of ontologies and knowledge graphs. Compatibility with other data structures facilitates easy fusion with other datasets (Früh & Tamás, 2021).

As Schmitz (2025) highlights, the quality of a termbase is determined not just by the design of the data model, its suitability for various applications, and the user-friendliness of the software, but also by the quality of its terminological data. Thus, the handling of data should be carefully planned and also integrated into quality evaluation. Accurate and well-thought-out conceptual modelling of the data structure, metadata, and terminology collections is crucial before populating a terminology resource (Vezzani, 2022). Such a foundation supports subsequent workflow stages, as outlined by Chiocchetti, Lušicky and Wissik (2023), which include needs analysis, design and implementation, documentation, term extraction, terminology entry creation, verification and quality assurance, maintenance, and dissemination.

Against this background, this paper aims to provide theoretically grounded, practical advice for designing and managing terminology resources as well as evaluating terminological data. Special attention will be given to the presence or absence of a terminological dataset, as it significantly influences the design of a termbase and the management of terminological data. Throughout this paper, we will use the term ‘terminology resource’ to broadly refer to terminological data collections, TMS and termbases.

2 Theoretical framework

The design and management of termbases are deeply rooted in the principles and methods of terminology work. This section delves into the main aspects that underpin these processes: terminological principles (2.1), the terminological metamodel (2.2), and data categories (2.3). These elements are essential for creating well-structured, consistent, and reliable terminology resources, as highlighted by relevant international standards, which will be referenced throughout the paper.

2.1 The terminological principles

The organisation and management of terminological data should adhere to four key principles (Arntz, Picht & Schmitz, 2021; Drewer & Schmitz, 2017; ISO 16642, 2017; ISO 26162-1, 2019):

- Concept orientation: each concept entry should contain all relevant information about a given concept. This includes, for instance, domain, designations, definitions, contexts and equivalents in other languages (in the context of multilingual terminological work). For example, the concept ‘bat’ should be recorded in two different entries: one for the nocturnal flying mammal and one for the sports equipment, as this designation refers to different concepts.
- Term autonomy: all terms, including synonyms and orthographic variants, are treated as an independent sub-unit. Consequently, they should be documented with the same set of data

categories. For example, treating synonyms as an attribute of the main term would violate this principle.

- Data granularity: each data category should be precisely described to identify individual information and facilitate efficient and accurate use. A typical example is splitting the data category /Grammar/ into three distinct data categories: /part of speech/, /grammatical gender/, and /grammatical number/ (Drewer & Schmitz, 2017).
- Data elementarity: each data category should contain only one piece of information. For example, recording a term and its grammatical gender in the same field (e.g. *avvocato, m.*, lawyer, masculine) would contravene this principle.

Compliance with these principles ensures that terminological data collections are well-organised and consistent across languages and domains. Additionally, they facilitate automatic data processing, machine readability and smooth updating of specific information.

2.2 The terminological metamodel

Termbases comprise terminological data collections and "have a logical structure that is reflected in a fundamental hierarchical data model, containing various levels at which data categories can be anchored" (ISO 26162-1, 2019). This structure should align with the terminological metamodel as outlined in ISO 16642 (2017), serving as a prerequisite for transitioning to the TBX framework,[1] which facilitates data exchange and reuse, such as training MT tools or large language models (LLMs). The terminological metamodel comprises two levels of abstraction (ISO 16642, 2017; Vezzani & Di Nunzio, 2020): the metamodel level and the data model level. The first facilitates analysis, design, and exchange, independent of any specific implementation or software. The second includes the necessary data categories to represent a specific collection of terminological data.

According to this metamodel, a concept entry consists of the concept level (concept entry), the language level (language section) and the term level (term section) (ISO 26162-1, 2019). All these levels are interconnected and create a nested structure (Figure 1).

The concept level provides administrative data and language-independent terminological information that pertains to the entire concept entry, such as /modification date/, /domain/ or /project/. The definition can be recorded at this level unless it has already been allocated at the language level (Section 3.1.2). Each concept entry pertains to a single concept and can be expressed in *n* languages. This level contains the language sections.

The language level includes all term sections and concept-related information for each of the languages involved. It also includes language-specific information, such as culture-dependent illustrations (Drewer & Schmitz, 2017). The definition can also be allocated at this level unless it has already been placed at the concept level (Section 3.1.2).

The term level contains all term-related information, such as /part of speech/, /context/, /term status/, etc. It corresponds to the term section and ensures the implementation of term autonomy. This section can include term component sections for providing "linguistic information about the components of a term" (ISO 26162-1, 2019), such as morphemes or single words from a multiword term (Drewer & Schmitz, 2017; Vezzani & Di Nunzio, 2020).

Terminological information is distributed across these levels and organised into data categories, which we will discuss in the following section.

2.3 Data categories

Data categories are classes of information like /definition/ or /part of speech/ and are typically implemented as fields in a termbase. They are identified according to specific parameters (Section 3.1).

As mentioned at the beginning of Section 2.2, data categories are anchored to a specific level of the terminological metamodel (concept, language or term level). Some data categories may appear at different levels, like /definition/. Conversely, others may occur only at a specific level. For example, /term

[1] A more in-depth discussion of TBX goes beyond the scope of this paper. For more information, see, for example, ISO 16642 (2017), ISO 30042 (2019) and Vezzani (2022).

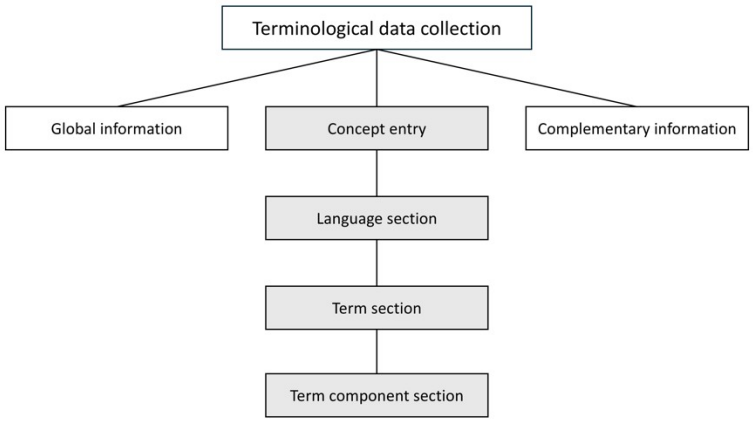

Fig. 1 Terminological metamodel (Simplified schematic view based on ISO 16642, 2017 and Drewer & Schmitz, 2017)

status/ may occur only at the term level, as it "indicates the acceptability rating of a term" (DatCatInfo, n.d.).

Depending on the type of content allowed, data categories can be open or closed (Drewer & Schmitz, 2017; ISO 26162-1, 2019; Warburton, 2021). Open data categories encompass any text that fits their definitions. For example, /definition/ is considered an open data category because the text recorded to describe a concept is unpredictable (ISO 26162-1, 2019). In contrast, closed data categories are restricted to a finite set of permissible values, presented as a picklist. For instance, /geographical usage/ may consist of a picklist with values corresponding to specific countries or regions. By using picklists, we can select the appropriate value without manually typing it, which helps prevent the introduction of misspellings or new variations, thereby ensuring consistency throughout the termbase (Drewer & Schmitz, 2017; ISO 26162-1, 2019). This consistency is essential for optimising the performance of search, filter, and other data management functions (ISO 26162-1, 2019). Warburton (2021) also lists a third type of data category, namely "constrained", because it is "restricted to a certain pattern or format" like date fields (e.g. /modification date/).

To facilitate exchange and interoperability, standardised data categories are available in recognised repositories, such as DatCatInfo. When these repositories do not contain suitable data category names that meet the scope of the terminological data collections or the user's needs, it is possible to create custom ones. For data interoperability and reuse, such custom data categories should be equipped with traceable information, including, among others, a unique persistent identifier (PID), a unique and stable mnemonic identifier, a unique canonical data category name and the data category type (e.g. open, closed) (ISO 12620-1, 2022).

Therefore, careful planning is essential to effectively identify the type of content needed in the termbase. If data categories are not suitably calibrated for the type of information, terminological data may be recorded incorrectly, or disparate kinds of information may be combined into a single data field (ISO 26162-1, 2019). This would violate the principles of data elementarity and granularity. Conversely, treating all data categories as open could decrease productivity and compromise the consistency of terminological information, ultimately undermining the termbase's usability (ISO 26162-1, 2019).

3 Initial considerations

This section consists of three parts: the first (3.1) outlines the planning stage, the second addresses the presence or absence of a terminological dataset (3.2) and the third (3.3) presents a short criteria catalogue of the essential functionalities and additional features a terminology resource should include.

3.1 The planning stage

In the following, we describe four interrelated parameters that shape the design of termbases and the

selection of data categories, affecting how information is structured, prioritised and accessed: scope (3.1.1), domain (3.1.2), working languages (3.1.3), user group(s) and user situation(s) (3.1.4).

3.1.1 Scope

The primary function of a terminological data collection is to enhance communication across one or more languages (Sager, 1990). This can be achieved on multiple levels, such as facilitating cross-border or national communication, improving organisational communication, or supporting language planning and translation. The scope is pivotal for the type of terminology work, whether descriptive or prescriptive. A descriptive terminology work "aims at documenting designations as they are used in contexts without favouring preferred usage" (ISO 12616-1, 2021). In contrast, prescriptive terminology work "aims at deciding on preferred usage of designations" (ISO 12616-1, 2021). The latter can occur after a descriptive phase, for example, when the descriptive terminology work reveals a need for standardisation (Chiocchetti et al., 2013; Drewer & Schmitz, 2017). In this case, terminological data will require specific data categories to indicate the standardisation status of a term by an authoritative body. An example of such a category is /authoritative status/, which may include values like "legal", "regulated", or "standardised" (DatCatInfo). Treating this category as closed would be advisable to ensure consistency in the input of the corresponding values.

If the terminological data collection is intended to support translation, the terminology work can be either prescriptive or descriptive, depending on the goal of the translation project. A data category like /note/ would be helpful to highlight any translation or terminological gaps or to point out any discrepancies between the languages involved. This category may be placed at the concept level and treated as open, as its content is unpredictable and may vary according to the information required for that concept entry. However, in practice, for specific terms — such as those within the legal domain in a translation-oriented termbase — a prescriptive approach may be necessary to ensure uniform usage and clear communication, notably where legal and material consequences are involved (e.g. modalities of contract dissolution). On the other hand, for other terms that operate as context-dependent terms (e.g. 'officer', 'official in charge', 'person in charge', 'responsible person', 'responsible officer', 'case administrator' or 'please contact'), a descriptive approach may be more appropriate, allowing for the flexibility needed to capture the nuances of different contexts accurately.

In the modern digital age, terminological data collections can be used, for example, to customise machine translation. Some MT producers offer the option to upload glossaries to ensure consistency and accuracy throughout translation projects (Nesbigall, 2025). Alternatively, such terminological data can serve to adapt and evaluate machine translation, for example in the context of minority languages (e.g. South Tyrolean German) (Contarino & De Camillis, 2023). To this end, accurate data categories at the term level, such as /term status/ with values like "admitted", "deprecated", or "preferred", can facilitate the preparation and processing of terminological data.

3.1.2 Domain

According to ISO 1087 (2019), a domain[2] is a "field of special knowledge". This data category is typically placed at the concept level since it applies to the entire concept entry. Its role in terminology work is crucial as it helps distinguish designations from one another (Drewer & Schmitz, 2017). This is particularly relevant for homonyms and polysemes (see example 'bat' in Section 2.1), allowing the principle of univocity to be applied: one term, one concept.

To ensure consistency, this data category should be treated as closed, containing a comprehensive list of domains relevant to the terminological data collection's scope. For instance, in a legal terminological data collection, the data category /domain/ would include a picklist of the domains covered by the termbase, such as civil law, criminal law, procedural law, and others. Although domains are highly specific to organisations and applications (Drewer & Schmitz, 2017), it is advisable to use

[2]The term 'domain' is also referred to as 'subject field'. Both terms are considered synonymous by ISO 1087 (2019) and ISO 704 (2022), with a preference for 'domain'. Conversely, ISO 12616-1 (2021) views them as synonyms but favours 'subject field'. For consistency throughout this paper, we will align with ISO 1087 (2019) and ISO 704 (2022) by using the term 'domain'.

existing public domain classification systems, such as EuroVoc or Lenoch, whenever possible. This would enhance reuse and interoperability (Warburton, 2021).

From a data modelling perspective, the domain influences the structure of the concept entries in multilingual terminology work. Domains with shared cognitive background or internationalisation (Sandrini, 1996) allow definitions to be placed at the concept level, provided that an anchor language is chosen. For example, the domain of physics is characterised by universally recognised concepts like "gravity": its definition applies in all countries, regardless of language and culture. Conversely, domains lacking these aspects, such as religion, education, or law (cf. Sandrini, 1996), require definitions at the language level to account for nuanced conceptual differences across languages and cultures (Drewer & Schmitz, 2017).

3.1.3 Working languages

Will the termbase be monolingual or multilingual? This question may sound rhetorical, but it is not. Multilingual terminology work requires special attention to pluricentric languages like English, French or German. These languages are officially recognised in at least two countries as state languages, co-state languages, or regional languages (Muhr, 2016) and have multiple standard varieties linked to specific national or regional contexts (Ammon et al., 2016). For example, English includes British, American, Australian, Canadian and other varieties.

Based on this, it is crucial to assess whether the terminological data collection will focus on a specific language variety (e.g. British English) or multiple varieties (e.g. American English, Australian English, British English). In this regard, Ralli (2025) discusses how their treatment impacts database structure and provides different strategies for representing them: 1) as an attributive data category, 2) through language-level encoding and 3) as an alternative workaround.

If the language variety is considered an attribute of a term (Strategy 1), a closed data category (e.g. /geographical usage/) can be added at the term level. The picklist values should be based on language or country codes from ISO 639 (2023), ISO 3166-1 (2020) or ISO 3166-2 (2020). Additionally, fields such as /definition/ or /context/ should be differentiated by inserting a language or country code within the data category (e.g. /definition GB/, /definition US/). This method facilitates data filtering and exporting while clearly indicating which field corresponds to a specific language variety. However, the principle of term autonomy (Section 2.1) might not be entirely fulfilled, as it can be challenging to label a standardised or preferred term for each language variety (Ralli, 2025).

If the language variety is stored at the language level, it has its own language section, to which one or more term sections are anchored (Strategy 2). This representation allows the principle of term autonomy to be fully satisfied and facilitates the anchoring of definitions at the language level, which is essential for domains lacking the same cognitive background or internationalisation (Section 3.1). Storing a language variety at the language level also allows standardised data categories to be used. In the presence of a language identifier (e.g. en-GB, en-US), this representation ensures smooth reuse, interchange and interoperability.

However, how can language varieties be managed if they still lack a language identifier or are not supported by the TMS? Consider Hungarian as an example of a pluricentric language. This language has seven different varieties since it is spoken in the neighbouring countries to Hungary as a minority language (Austrian, Slovakian, Ukrainian, Rumanian, Serbian, Croatian, and Slovenian varieties). In these cases, two strategies can be applied: treating the language variety as an attributive data category (Strategy 1) or using an alternative workaround through language-level encoding (Strategy 3). For the latter, the language variety is assigned an existing but unused language identifier within the termbase (Ralli, 2025). While the principle of term autonomy is fully satisfied and terms of language varieties are not confused with synonyms of the reference pluricentric language, the *xml:lang* attribute would contain a language identifier that does not correspond to the stored language variety. This approach requires additional adjustments in the case of data interchange and interoperability.

3.1.4 User group(s) and user situation(s)

Thirty-five years ago, Sager (1990) stated that "every speaker or writer of a special subject language is a user of terminology and every learner of a special subject, be it in school, college, university or an industrial training course, is a learner of terminology" (p. 197). He identified seven types of users based on the type and combination of information they regularly seek in a terminology resource. These users included domain experts, professional communication mediators (e.g. technical writers, translators, interpreters), specialist lexicographers and terminologists, information and documentation specialists (e.g. librarians, indexers), language planners, language users (e.g. publishers, language teachers, researchers in applied linguistics), and the general user (Sager, 1990). Despite the passage of time, this classification remains relevant.

Each user group has distinct needs because they consult termbases for different purposes. Translators and interpreters typically require multilingual resources that can be integrated with CAT tools (Warburton, 2021). They often seek ready-made translation, definitions and term validation (Chiocchetti, 2023; Warburton, 2021). Content producers, such as technical or marketing writers, generally look for terms in the source language to verify spelling, meaning, usage, etc. (Warburton, 2021).

Some users can take on multiple roles: domain experts, like legal experts, may need terminology resources to find sources related to the target legal system(s) and comparative notes (Chiocchetti, 2023). At the same time, they might also be involved in compiling concept entries. Consider a national termbase as an example. In such a case, the focus could be on language planning based on a specific strategy, where domain experts ensure a professional vocabulary in the mother tongue or a minority language despite the overwhelming presence of English. Also terminologists can have multiple roles (Kranebitter & Ralli, 2022): they are both 'developers' and 'curators' as well as effective 'users' of a termbase. On the one side, they design the termbase, select the data categories, populate it with content, and perform quality controls. On the other side, they may consult the termbase to gain knowledge about a specific concept, for example, for responding to a terminological request coming from outside the own organisation.

Educational qualification, work experience, and the potential user's knowledge of the domain affect the type of information the user might need (Ralli & Andreatta, 2018). Therefore, it is necessary to identify the situations in which a terminological data collection will be used.

Studies conducted on user situations within the functional theory framework in the field of lexicography (Bergenholtz & Tarp, 2010; Tarp, 2008b) can also be extended to terminology resources since user situations are similar.

User situations can be categorised into cognitive, communicative and operational situations (Tarp, 2008a, 2008b). Cognitive situations arise when users need to acquire new knowledge, such as seeking additional information on a specific topic for translation or to better understand a text (Tarp, 2008b). In this context, data categories such as /definition/ or /note/ can be beneficial for expanding or verifying knowledge. Communicative situations involve scenarios where users need assistance during text production, reception, translation, marking, revision, or proofreading (Tarp, 2008b). To this end, data categories such as /degree of equivalence/ or /linguistic context/ can be particularly useful for translation purposes. Meanwhile, data categories like /language register/ or /documentation type/, along with the indication of the framework of the communication situation, can be relevant for text production. Operational situations relate to the user's knowledge and skills concerning a specific subject or task. For example, a legal expert may have extensive knowledge in their native language but might struggle to explain legal concepts in a foreign language. Conversely, translators may have limited specialised legal knowledge but possess "operational" skills that enable them to approach legal texts for translation effectively (Ralli & Andreatta, 2018; Tarp, 2008b).

These situations are not mutually exclusive and can co-occur. Taking the aforementioned example, terminologists can use the termbase to understand the distinction between two concepts (cognitive situation), retrieve terminological data for preparing a glossary (communicative situation), extract terms for injection into an MT tool, or evaluate terminological consistency (operational use). Hence, they may need specific data categories like /personal notes/ to report any doubts or issues related to the concept entries they are working on, which will be visible internally to the working group for discussion. However, such data categories should be restricted from being visible to an external audience.

Identifying the user group(s) and user situation(s) is essential, as they significantly affect the information addressed in the termbase and how it is provided. To this end, Kranebitter and Ralli (2021) suggest considering whether the potential user falls into one or more user groups. Hence, it is necessary to determine if the termbase will target a homogeneous audience or, rather, a heterogeneous one. In the case of a diversified audience, two possibilities arise (Kranebitter & Ralli, 2021). One approach is to structure the concept entries from the beginning based on a representative user archetype and select which data categories and picklist values should be present or not to meet their needs. Alternatively, multiple data categories can be provided, acknowledging that some will be of particular interest to specific user groups while others may not be as relevant. In this case, it is worth considering whether providing different users with various search criteria could better satisfy their needs.

3.2 Presence or absence of a terminological dataset

In designing a termbase, two scenarios may arise:

- the termbase is empty, containing no terminological data;
- the termbase contains terminological data;
- a preexisting dataset is available, but not contained in a termbase.

If a terminological data collection is not yet present, the termbase will be structured and populated from scratch. While this offers the flexibility of defining everything without being bound by prior decisions, it also introduces complexity and uncertainty as new decisions must be made without the benefit of experience from existing terminological data (Kranebitter & Ralli, 2021). An existing dataset would provide specific insights or potential issues that must be addressed.

To understand how to navigate the design of an empty termbase, some guiding questions might be (Kranebitter & Ralli, 2021):

- What information should be included and presented to the user group(s)?
- Will the termbase be managed centrally or locally?
- Should the dataset include special data types that require a specific approach (e.g. descriptive, prescriptive, both descriptive and prescriptive, translation-oriented)?
- What kinds of languages (e.g. pluricentric languages, minority languages) will be covered in the termbase? Do these languages have a language identifier?
- Which type of data categories should be included (e.g. also phraseologism for law, images for medicine)?
- Where should definitions be placed within the entry structure?
- Should the concepts be defined in an anchor language or all working languages?
- Which categories should be open data categories and closed data categories?
- Which types of information are best suited for closed data categories?
- Is data exchange expected, for example, between two or more institutions or offices? If so, how will terminological data be exchanged?
- Will terminological data be used for MT tools or LLMs?

If a terminological data collection is already present, similar reflections from an absent terminological dataset apply. However, existing terminological data greatly affect the design of a new termbase and require further consideration. In this regard, some guiding questions might be (Kranebitter & Ralli, 2021):

- How were the terminological data managed in the past (locally or centrally)?
- Was a single approach adopted for data processing, or was a mix of approaches used (e.g. descriptive or prescriptive or both)? Should the same method be maintained for the future?
- What is the structure of the existing terminological datasets, and what elements, if any, need to be changed?
- What formats are the terminological data currently in, and are they uniform or in various formats (e.g. XML, TBX, .docx, .xlsx)?
- Was an anchor language used for all data categories, or are they distinct according to the working languages?

- Will the new termbase cover language varieties or minority languages without a language identifier?
- Were the principles of concept orientation, term autonomy, data elementarity and data granularity observed?
- Do the existing terminological data appear uniformly, or are there duplicates?
- Is data exchange expected, for example, between two or more institutions or offices? If so, how will terminological data be exchanged?
- Will the new terminological data collection be used for MT tools or LLMs?

A more detailed set of guiding questions will be provided by Part 4 of ISO 26162 on *Management of terminology resources — Termbases — Part 4: Quality,* which is currently (May 2025) under development (Schmitz, 2025).

3.3 A short criteria catalogue for designing a terminology resource

In the practical field, considerable discussion exists about what qualifies as nowadays an appropriate instrument to record, edit and publish easy searchable terminological data. It is worth briefly reviewing the main possible technical aspects of a modern tool, whether a commercial one, one's own development, or a hybrid one. A well-designed structure and functionalities allow a quick reuse and easy exchange of data. While not exhaustive, the following table (Table 1) outlines the essential functionalities and additional features that a terminology resource should include (cf. also Drewer & Schmitz, 2017; Fóris, Somogyi & Papp, 2024; Fóris & Somogyi, 2024; Kranebitter & Ralli, 2021).

4 Conscious handling of data in termbases: issues on quality

This section focuses on evaluating existing and newly created termbases. To this end, we interpret 'quality' as the level of consistent compliance with predefined needs and expectations (Section 3.1.4), which is in harmony with the ISO definition of quality: "degree to which a set of inherent characteristics [...] of an object [...] fulfils requirements" (ISO 9000, 2015). We agree with Chiocchetti, Lušicky and Wissik (2023), who, by analysing multilingual legal termbases, noted that: "Quality is neither an absolute nor entirely objective variable but is ultimately determined by the stakeholders, users, and applications". From this perspective, the early consideration of the main parameters described in Section 3.1 can prevent costly changes later (Kranebitter & Ralli, 2021). In this regard, for instance, an audit to evaluate the quality of terminological data is based on establishing whether these requirements align with the output and which modification proposals are needed for improvement. In fact, an audit usually contains "a report, comments of non-conformities and recommendations for improvement" (Früh & Tamás, 2021).

Proper quality management can lead to achieving the desired outcome, which can include establishing policies, aims and processes through quality planning, quality assurance, quality control and quality improvement (ISO 9000, 2015). Quality planning is concentrated on the setting of quality objectives, specifying operational processes and resources to achieve the aims of the quality set (ISO 9000, 2015). Quality assurance (QA) is a "proactive process to prevent quality non-conformity of a terminological product", which reveals and fixes the sources of feasible quality problems, and it is relevant during the preparation phase and while operating the termbase (Früh & Tamás, 2021). Quality control is, by comparison, a reactive process focused on the results, the fulfilling of quality requirements (ISO 9000, 2015), and "making sure that the termbase complies with the requirements for the intended use" (Chiocchetti, Lušicky & Wissik, 2023). Quality improvement is concentrated "on increasing the ability to fulfil quality requirements" (ISO 9000, 2015) and achieving a higher quality level through gradually increasing compliance. These elements of quality management need to be observed when building a termbase.

On the one hand, termbases are "products (databases) that implement requirements formulated along the lines of process quality, database data quality, and data model quality. On the other hand, they are also services, allowing, for example, querying, filtering, collaborative work, etc. Quality objectives ideally address both functions" (Chiocchetti, Lušicky & Wissik, 2023). There are many factors to consider when achieving the proper quality of product and service. First, the TMS should be appropriate as a tool,

Table 1 Short criteria catalogue for designing a terminology resource

Internal editor's interface	
Basic functionalities	**Plus elements**
• Stable handling of large and heterogenous data (up to 40 data categories including text, numbers, data and multimedia) and languages/language varieties (with anchor language indication) • Format handling for data import, export and exchange (xlsx, CSV, TBX) • Free choice of XML-based data categories • Free choice between mandatory and optional data categories • Metadata in multiple languages • User-friendly for terminologists (quick import/export, merge/split/clone, filter duplicates, easy Batch Edit) • Changeable database definition file on the entry, language and term level • Easy handling of external and internal links • Automated saving and backups • Information about the number of terms and entries available	• Built-in workflow with different rights (data life history, validation process) • AI term extraction • AI formatting of definitions • Classical and/or AI-generated and visualised concept map system, etc. • AI-readable data
External user interface	
Basic functionalities	**Plus elements**
• Modern interface with public access • Customisable for user-friendly version • Simple and advanced search options (i.e. ignore upper/lower case in term search, exact/partial match, abbreviations) with filtering of language domains, geographical use; hitlist with basic information • Configurable entry (show basic data/show full entry option) • Feedback option • Basic administrative information (legal disclaimer, copyright, cookies, contact)	• Reset search settings • User guide (text/multimedia) and tutorial (video) • Externally downloadable data for reuse (e.g. pdf, xlsx, sdltb, TBX, app) • Concept map search with reciprocal links to entries • AI results for terms and concept maps indicated • Additional information (FAQ, chat service forum for proposals and questions, technical information about the number of entries and queries, partners, news, publications, term of the week, declaration of policies).

which, with technical development and time, has become more complex to meet newly emerging needs. It requires the right design to fulfil aims to satisfy the needs of a user group and domains, and, nowadays, it must ensure interoperability for the exchange and reuse of data, not to mention the application of AI. Additionally, it has to offer workflow management, which should be well-defined for collaborative work. The workflow to fulfil its functionalities requires proper coordination, including guidelines, clear roles, work phases, regular maintenance and feedback for improvement.

The field of lexicography, boasting a long history, has various classification and evaluation systems to assess the lexicographic tools. In comparison, modern terminology science has recently been trying to formulate appropriate classification criteria, already adapting them to modern technology (cf. Drewer & Schmitz, 2017; ISO 26162-3, 2023). Tamás and Sermann (2019) limited their evaluation criteria to online surfaces of larger organisations, while Früh and Tamás (2021), extended them to internal ones, both intended to create a tentative system for examining and evaluating termbases according to relatively

objective and comprehensive criteria, which can serve the comprehension, description, classification, evaluation and review of termbases. The last experimental classification revolves around the four main categories, which are closely interrelated:

1) environment
2) technical parameters
3) structure and content
4) usability and features of the termbase.

For instance, the main category of structure and content is influenced by different aspects of the main category of environment, which itself contains the subcategories of tendencies (terminology policy aims, translation orientation, and standardisation proposals) or the type of terminology work (descriptive or prescriptive, monolingual or multilingual, systematic or ad hoc). Nonetheless, a categorisation facilitates systematisation and offers a higher awareness of the handling of tools.

The main category of structure and content is subdivided into general features and distinguishes between simple, traditional and complex termbases (Tamás & Sermann, 2019). The detailed structure subcategory is concentrated on (Früh & Tamás, 2021):

a) the megastructure (e.g. the availability of a user guide);
b) the macrostructure (e.g. the search options with different filtering possibilities or the configuration of the hitlist);
c) the microstructure (e.g. data categories);
d) the mesostructure (e.g. cross-references).

Observing the editing principles of concept orientation, term autonomy, data elementarity, and data granularity (Section 2.1) leads to systematically structured content on the interface, which, if associated with the proper metadata, will create a clear structured termbase. This is important because the "display of data categories and the clarity of entries expressed by the order and labelling of data categories can also be a quality indicator" (Früh & Tamás, 2021).

In practice, a few mandatory and a high number of systematically selected optional data fields can help provide flexibility[3] but ensure that the structure can be maintained over the long term. For instance, for legal terminology, the relevant optional fields can be country codes, and for language varieties, regional codes. As for medical terminology, even being optional fields, information on documentation types such as final reports or referral care presenting different term uses as well as the framework for professional communication are of high importance. This is especially relevant since the nature of communication, whether it is a) scientific, b) inter-professional, c) scientific inter-professional and inter-professional-lay, differs significantly in each case. The medical language for special purposes is substantially characterised by stratification according to its scope of use and the language use context (Kuna & Ludányi, 2023).

5 Two applied examples

In the following sections, we will present two concrete examples to bridge the gap between the theoretical framework and the practical implementation: the Hungarian Terminology Strategy project, which was launched on December 1, 2023, under the Science for the Hungarian Language National Program by the Hungarian Academy of Sciences implemented by the HUN-REN Hungarian Research Centre for Linguistics (5.1) and the Information System for Legal terminology *bistro* (n.d.), developed by the Institute for Applied Linguistics of Eurac Research in South Tyrol, Italy (5.2). These examples illustrate how theoretical considerations are operationalised in a real-world digital terminology resource, shedding light on challenges encountered during the design of the termbase and the evaluation of terminological data, thereby contributing to a more nuanced understanding of theory-in-practice.

The reflections carried out in Section 3 will be considered for the Hungarian Terminology Strategy project (5.1), while the considerations from Section 4 and especially the Früh and Tamás (2021)

[3] Heinisch (2023) also emphasises flexibility concerning usability and multi-purpose applications according to user groups. Similarly, Fóris et al. (2024) highlight the importance of tailoring tools to meet specific user needs.

classification will be taken into account for the Information System for Legal Terminology *bistro*[4] (5.2).

5.1 The Hungarian Terminology Strategy project

The Hungarian Terminology Strategy project is an ongoing project, which was launched on December 1, 2023, under the Science for the Hungarian Language National Program by the Hungarian Academy of Sciences. The four-year initiative aims to realise three main aims for a national terminology infrastructure, namely (Lipp & Prószéky, in press):

a) the creation of a national terminology portal as a term research engine and a Hungarian national termbase, initially working as a research-supporting termbase, uploaded with data in collaboration with the scientific sections of the Academy and the institutes of the HUN-REN Hungarian Research Network;
b) the unification of educational terminology in the Carpathian Basin in close collaboration with the Termini Research Network through the realisation of a nine-language term collection to be integrated into the Hungarian national termbase;
c) the creation of a bibliographic database of specialised dictionaries in collaboration with the Terminology Documentation Centre of Pécs.

The project is implemented by the Institute for Lexicology of the HUN-REN Hungarian Research Centre of Linguistics, whose researchers started from the premise of two main scenarios. On the one side, the preservation of the Hungarian language in scientific discourse and higher education should be promoted by the multilingual termbase aimed to support research and gradually expand its scientific vocabulary. More specifically, in the use of academic languages with the increasing dominance of English, even in Hungarian, the priority is given to managing English as a *lingua academica* in a spirit of added bilingualism (Fóris, 2024b). On the other side, there is a focus on promoting the Hungarian minority speakers living in the seven neighbouring states of Hungary. The exercising of linguistic rights of these minorities after the historical changes in 1920 led to a fragmented linguistic landscape and is determined by the different policies of the majority nations, resulting in country-specific concepts and terms of a pluricentric Hungarian language (Lanstyák, 2023; Prószéky et al., 2023). The aim is to allow minority language users to exercise their right by using their mother tongue and realise mobility in education between Hungary and regional areas. The recording of Hungarian and of the foreign languages spoken in minority areas as the state language (Austrian German, Slovakian, Ukrainian, Rumanian, Serbian, Croatian, Slovenian) and of seven Hungary language varieties adds to the complexity of a TMS and requires conscious handling of the languages and language varieties.

For the above reasons and considering the parameters described in Section 3.1, the terminological data collection to be realised has the task of including a wide range of areas in domains, languages and language varieties, and this requires keeping the main functions and the core structure of the termbase consistent while at the same time enabling to cover the different needs of the user groups. This necessitates the careful selection of mandatory and optional data categories as well as the emphasis on different aspects in the workflow of terminology management. In fact, a well-designed, but flexible structure enhances the quality of terminological data (Agrario & Castagnoli, 2010).

The project includes the elaboration of domains without a preexisting terminological dataset, such as educational terminology of primary, second level and higher education (Section 5.1.1), and of scientific domains with an already present terminological dataset, namely within the scientific discourse with the first demands for data collection deriving from natural sciences like forestry, meteorology, microscopy and geographical denominations (Section 5.1.2).

Warburton (2021) mentions examples of corporate contexts, but we agree with her that scenarios have a direct impact on the design of termbases, and it is essential to have a clear understanding of users and needs as, for instance, a specific data category may be common for certain types of domains but not for others. User needs, based on a needs assessment, can even deviate, to a certain extent, from principles of terminology management (Heinisch, 2023). In the case of a national termbase, the scope is to ensure the

[4] https://bistro.eurac.edu/

exercise of fundamental linguistic rights, such as the right to use the mother tongue, notably the cultivation of disciplines at a high level, which is achievable through the development of an elaborated TMS (Papp, 2023) and has different aims as a translation-oriented termbase (Fóris, 2024a; Fóris & Somogyi, 2024). Free online access to such tools contributes to improving professional communication for a large heterogeneous user group (Fóris et al., 2024) and requires countless decisions, including the careful handling of copyright issues, the use of plain language[5] and the extent of the termbases' visibility. Continuous maintenance is also necessary to ensure the terminological data remains up-to-date (Wissik, 2024). We agree with Nilsson (2009) that a national termbase is part of a larger national terminology infrastructure and requires a cooperative network based on a well-defined and appropriate terminology policy.[6]

5.1.1 Absence of a terminological dataset

The first example of educational terminology presented in this section lacked a terminological dataset. However, in some of the seven regional minority areas, different terminological data collections or vocabularies have been published (Benő & Péntek, 2023). Still, a comprehensive global initiative for Hungarian educational terminology, including the languages of the neighbouring countries and English as a means for fostering international communication and mobility, plus all minority language varieties was still missing. The inclusion of harmonised English equivalents of Hungarian educational terms has the advantage to become a reference point for the consistent use of terminology by authorities and universities in the task of issuing and interpreting diplomas.

The working out of the educational terminology started with the extraction of educational terms forming part of the main educational laws of Hungary[7] completed by the selection of country-specific terms in the seven external minority regions by linguists and researchers of the Termini Hungarian Research Network. In a second step, the Educational Authority, with its experts, has been invited to review the data. The experts checked the definitions of the Hungarian concepts of primary and second-level education and the equivalents in English in national public education. In the case of higher education, the cooperation focused on the selection of Hungarian terms, the creation of definitions in Hungarian and the working out of English equivalents. The participation of the Educational Authority and its promise to perform maintenance work in the future on account of the rapid changes in educational terminology contributes enormously to the reliability of the results (for the characteristics of higher education terminology cf. also Papp & Fóris, 2018).

The data on educational terminology are centrally managed, with editors residing mainly in regional areas. The terminological data collection is elaborated from scratch and treated with an onomasiological approach. The teamwork required different training sessions for the more than 23 experts involved, the creation of an editing guideline, and regular online consultations to clarify various questions of experts emerging during the elaboration process.

Regarding the selection of data categories, not only terms, but also much useful information has been included, not always as mandatory data categories but often as optional ones: concept ID, domains, anchor language, term sources, definition and source, country and regional codes, equivalence, related terms and type of relations (i.e. synonyms, superordinate terms, archaic label), editing person, validating person, last modification date and notes.

The user group of educational terminology is wide and heterogeneous: from students, teachers of all educational levels, organisational background of education, media, other types of domain experts like researchers or language experts (e.g. linguists and translators), and even authorities. Therefore, definitions in all main languages, and optionally for the minority varieties, must be professionally

[5] "Communication in which wording, structure and design are so clear that intended readers can easily find what they need, understand what they find, and use that information" (ISO 24495-1, 2023).

[6] Cf. Fóris (2025) and Fóris (2024c) for a general description about the history and recent trends of terminology in Hungary focusing on the 20th and 21st century.

[7] In the first year, the Act CXC of 2011 on National Public Education (2011) and, in the second year, the Act CCIV of 2011 on National Higher Education (2011) and the Act LXXXI of 2023 on the promulgation of the Global Convention on the Recognition of Qualifications concerning Higher Education (2023).

accurate yet written in plain language.

A special, optional data category included is ISCED Code (n.d.), which is the International Standard Classification of Education framework used for comparing education systems internationally. This helps in the unified handling of educational concepts. Educational terms are often country-specific terms, making equivalence a key issue. For instance, the Hungarian term *főiskola* (a higher education institution with typically 3-4 years of education) can be equivalent to the English term 'college', but 'college' itself covers a broader range of concepts (cf. Boronkay-Roe, 2020). The complexity of the terminological data collection and structure is increased by the nine languages and seven language varieties. To avoid confusion with synonyms, the regional languages are placed at the language level. Due to relatively fast-changing designations and concepts and the inconsistency in educational terminology, the concept entries contain a note stating that the terminological data are for "information purposes only," placed at the concept level.

The domain system consists of a tested and duly modified 5-level classification, allowing detailed data filtering. Bibliographical sources are clickable, but to avoid copyright issues, definitions from laws, official sites, and those created by the editors (which may be "based on an x source") are preferred. As also mentioned at the beginning of Section 3.2, in cases where a terminological dataset is absent, terminologists are not bound by previous decisions; however, the lack of specific information makes recording more challenging because of unforeseen needs and characteristics.

5.1.2 Presence of a terminological dataset

In the case of natural sciences, the HUN-REN Hungarian Research Centre for Linguistics collaborates with the different Scientific Committees of the Hungarian Academy of Sciences and Institutes of the HUN-REN Hungarian Research Network and their partners.

The domains of natural sciences focus on a different public than education. Among the domains of forestry, microscopy, meteorology and geographical proper names, the last two are perhaps interesting for a wider audience, while microscopy is a narrower and more specific one. The main scope is to support professional language use in higher education and research, more specifically the professional and scientific language use, text production, translation and terminology planning of Hungarian terms in specific areas. An additional aim is to achieve bilingualism (Fóris, 2024b), rather than the elimination of English. English serves internationally as a *lingua academica* and this approach ensures that an appropriate professional vocabulary is available even in Hungarian.

As Nilsson (2009) states, for the building of a national termbase, a national inventory of existing resources is very useful as there are many high-quality collections of various organisations that do not deal with terminology in a consistent way. Also in this case, as in Sweden (Nilsson, 2009), different types of terminological information, for instance, lexicons, vocabularies or glossaries, are available in different formats (e.g. printed, docx, pdf), which have to be re-edited according to the terminological approach, enlarged and updated in cooperation by the technical experts, language experts (i.e. linguists specialised in orthography, translators) and terminologists. The preexisting vocabularies were usually edited with a semasiological approach or sometimes, but not consciously, with an onomasiological approach (e.g. meteorology[8]). The experts, as editors, have the highest level of professional knowledge, but they usually lack a terminological approach and need appropriate support and training.

The management of data is more locally, but with central coordination to achieve uniform editing since data will be part of a centralised termbase, besides some paper-based publications and electronically stored glossaries. The selection of different data categories must be weighted in the light of remaining adherent to a central structure. Natural sciences should include a definition at least in Hungarian because, in most cases, a full equivalence is already present as a result of an internationally unified interpretation of concepts (e.g. meteorology) or the identical material reality (e.g. microscopy and forestry). However, the existing differences of equivalence must also be noted (i.e. different types of flora or fauna or technology or different classifications adopted in forestry). This is not requested by

[8] Czelnai & Szepesi (1986)

geographical designations, primarily investigated by the study of onomastics, which disposes of common areas with terminology (Bölcskei & Fóris, 2022). It does not form part of the classical concept-based elaboration, but recording is recommended in termbases to have standardised variants. The choice of English equivalents has to consider the frequency and prevalence of use. In this case, the emphasis is put on the data categories, the main term and source, the related terms and types of relation, the anchor language and the note, serving as labels of archaic expression or other interesting linguistic features, and indicating the orthography of professional language.

The terminology of natural sciences is not so rapidly changing as in education, although periodical maintenance is necessary the same, but for other reasons: for instance, changes in the field of meteorology are attributable to new climatic phenomena ('climate stripes' or in Hungarian *klímacsík*) or due developments of technology in every two-three years in the field of microscopy. A lack of unified Hungarian equivalents of English is typical for microscopy. This professional vocabulary has to be developed by domain-experts with the support of linguists. The microscopy domain is also characterised by several abbreviations, which need to be treated equally as /term/ in the termbase structure to ensure term autonomy (Section 2.1) and, therefore, to be searchable. In natural sciences, numerous terms can be categorised in different subdomains, which should be indicated in the termbase as well; sometimes, they cover a different concept (e.g. the term species in flora and fauna) or indicate the same component as for a microscope. The use of pictures is also typical for natural sciences. For instance, meteorology includes pictures, formulas, forestry and microscopy concept maps, and illustrations (e.g. in forestry of flora and fauna, in microscopy components).

The languages of these domains include mainly Hungarian and English and, in the case of forestry, also German, Rumanian and the language variety of Szekler forming part of the region of Transylvania, where the most significant number of Hungarian minorities live.

The HUN-REN Hungarian Research Centre for Linguistics in the above-outlined project focuses on coordination. The first data collections were launched, which required training, guidelines, and the selection of data categories typical for given domains, following clear strategic aims. Nonetheless, the current project can be seen as a pilot providing valuable experience for the creation of a national termbase, which can only function properly if regular maintenance is carried out as requested by the Centre and agreed with partners. In the future, the workflow for experts can be an object of further automatisation, and there are possibilities for setting up audit committees. Currently, the aim is, therefore, to lay the foundations for a national termbase while collecting data of different domains, publishing the scientific results and promoting the project to the audience. The termbase will later be incorporated into the website of the research centre and form part of a national terminology structure with a range of additional information.

5.2 The Information System for Legal Terminology *bistro*

The Information System for Legal Terminology *bistro* is an online application developed by the Institute for Applied Linguistics of Eurac Research. It contains legal terminology in Italian, German (South Tyrolean, Austrian, German, Swiss, European Union and international law varieties) and Ladin (Val Gardena and Val Badia varieties).

Initially launched in 2001 as a support tool for communication, writing and translations within the legal context, *bistro* underwent reprogramming efforts between 2013 and 2016. These efforts were carried out in collaboration with the Department for Information Technologies of Eurac Research and the Office for Language Issues of the Autonomous Province of Bolzano, and were financed by the Office for Information Technologies of the Autonomous Province of Bolzano. The goal was to create a flexible and reliable tool capable of meeting the diverse needs of various user groups, including lawyers, translators, students, or anyone seeking reliable support for understanding and translating legal texts and documents.

The extensive work on redesigning the termbase and restructuring and cleaning the terminological data is thoroughly detailed in Ralli and Andreatta (2018) and Kranebitter and Ralli (2022). Based on the considerations outlined in Section 4, in the following sections, we will focus on the *bistro*'s structure and content to provide an evaluation (5.2.1). Additionally, we will discuss the quality control

and validation of its terminological data (5.2.2) to give a concrete overview of the necessary checks to ensure high-quality content.

5.2.1 *bistro*'s structure and content

After analysing the broader environment and the basic technical parameters, it is worth having a closer look at the structure and content of *bistro*. At the megastructure and macrostructure levels, we must distinguish between the TMS Trados MultiTerm[9] (internal editor's interface), where the terminological data is compiled and exported in XML format, and the online system *bistro* (external user interface), into which the exported terminological data are uploaded, making them accessible to the public.

Regarding the megastructure level, *bistro* features a feedback function available for each concept entry. Through this function, users can send comments to the *bistro* team regarding existing concept entries, suggest changes, or propose new terms for inclusion. Legal information is available on a dedicated page in Italian, German, and Ladin, while instructions are provided through video and PDF files. The website also features information about the collection of terminological data (in Italian, German, Ladin, and English), partners, news, and publications about *bistro* and its terminological data.

At the macrostructure level, *bistro* enables user groups to utilise different search options, such as simple search, advanced search (exact search, search by source language and geographical usage, target language and geographical usage, legal domain, and search by combining all these parameters), and searching in lists of standardised terms for South Tyrol. Furthermore, it allows results to be filtered.

At the microstructure level, the terminological data collection is compiled and managed in Trados MultiTerm according to the terminological principles (Section 2.1) and the method of legal comparison (Mayer, 2000; Sandrini, 1996). It consists of more than 22,000 concept entries, and more than 13,000 are published online. The concept entries present a three-level structure according to the terminological metamodel (Section 2.2). The terminological data collection contains 79 terminological data categories: 55 are open, and 24 are closed. Moreover, 11 data categories are for the internal use of terminologists and are not visible to the external audience (Section 3.1.4). Their name and their related values are specific to each legal system. This means that all fields related to the Italian legal system are listed in Italian (e.g. /Grammatica/, /Definizione/). Similarly, all fields related to German-speaking legal systems (including the Italian legal system in German for South Tyrol) are listed in German (e.g. /Grammatik/, /Sprachgebrauch/). Accordingly, the Ladin language section (i.e. the Italian legal system in Ladin for South Tyrol) contains all fields listed in Ladin (e.g. /Gramatica/, /Adoranza linguistica/).

German and Ladin language varieties are recorded as an attributive data category at the term level since South Tyrolean German and the two Ladin language varieties, *Gherdëina* and *Badiot*, lack a language identifier (Section 3.1.3). Treating the language variety as an attributive data category has a domino effect on those data categories recorded at the term level: in the German language section, data categories such as /definition/ or /context/ are distinguished by adding a country code (e.g. /AT/, /DE/) to the data category name, wherever possible (e.g. /Definition AT/, /Definition DE/). Accordingly, the Ladin language section indicates the language variety the data category refers to, e.g. /Definiziun Val Badia/ or /Definizion Gherdëina/. Such ad-hoc data categories facilitate filtering and exporting data and make it immediately apparent to the user group to which legal system/language variety the information pertains. However, they hinder data exchange and interoperability since they do not entirely adhere to the terminological metamodel (Ralli, 2025). For this reason, *bistro*'s structure will undergo a new redesign to be fair and fully compliant with the respective ISO standards for interoperability and data exchange (i.e. ISO 12620-1, 2022; ISO 12620-2, 2022; ISO 30042, 2019; ISO 16642, 2017).

At the mesostructure level, cross-references among entries are recorded only in the Italian part of the concept entry for technical reasons. Some sources contain external links and take users directly to the cited webpage.

Regarding the usability and features of *bistro*, definitions, contexts, and notes at the concept level are documented by reliable and authoritative sources from legislation, handbooks, case law and websites

[9] https://www.trados.com/it/product/multiterm/

of public institutions, and they are assigned to the appropriate legal system (Ralli & Andreatta, 2018; Ralli & Kranebitter, 2017). When no definition or context is found, term sources are recorded. All sources are presented in a short form and are clickable. Users can access the complete bibliographic information of the short form, in the language of the respective source, including details like legal system, type of source, date, etc. *bistro* is regularly updated, at least once a month. It is freely accessible and serves as a valuable tool to ensure legal certainty. On the one hand, it promotes the use of correct and uniform legal and administrative terminology, not only within the South Tyrolean administration but throughout the entire province of South Tyrol (Ralli & Andreatta, 2018; Ralli & Kranebitter, 2017). On the other hand, it facilitates communication and understanding between citizens and institutions, both at a national and international level.

5.2.2 Quality control and validation in *bistro*

Concept entries are compiled according to internal guidelines, which provide clear instructions on defining concepts, selecting appropriate contexts, recording sources, and managing information in notes at the concept or term levels, etc. Guidelines also exist on how to shorten and record sources from handbooks, normative texts, courts and websites according to the legal system in a dedicated database. Both sets of guidelines ensure consistency in terminology work and correct data entry in the termbase. Concept entries follow a pre-defined structure using a customised input model template that outlines the order of the data categories to be filled with content. These guidelines serve as a benchmark for reviewing concept entries to check for linguistic quality, completeness, accuracy, and relevance.

Quality control and validation of the concept entries are conducted regularly to ensure high-quality content and consistency in terminological data and information. The primary reliance is on the internal guidelines, followed by ISO 26162-3 (2024) and the classification of Chiocchetti, Lušicky and Wissik (2023). According to the latter, validation is performed at three levels (Chiocchetti et al., 2013; Chiocchetti et al., 2023; Heinisch, 2023): formal, linguistic and content.

At the formal level, verification is carried out to ensure that concept entries are complete and that information has been entered into the appropriate data categories. This includes checking for correctness within the data categories (e.g. ensuring that the correct values are selected in closed data categories), identifying missing sources, and detecting concept duplicates or embedded hard line breaks. Additionally, the language/language variety or legal system entered is checked for accuracy. Recurring oversights are addressed, such as missing geographical usage indications or grammatical information. Non-active cross-references between concept entries are corrected, inactive URLs replaced, and citations of bibliographic sources verified. Some errors can be resolved automatically using the Batch Edit function in Trados MultiTerm, which allows for changing a multitude of data simultaneously within the data categories and at each level. For larger find-and-replace tasks, the data are exported in Excel or XML format, edited, and then re-imported into the termbase.

At the linguistic level, spelling is checked, typos corrected, and the appropriateness and naturalness of the language used are verified. This control is crucial for the findability of terms in *bistro*: if a term contains a typo, it cannot be found in the online system.

At the content level, definitions are evaluated for correctness, up-to-dateness, appropriateness, and relevance to the legal domain. It must be ensured that additional information is recorded in the correct data category. Contexts should be appropriate and illustrative and include the term. Terms entered in the /term/ data category must accurately represent the defined concept. Furthermore, terms are assessed for currency (e.g. updates following legal reforms) and the effective synonymy or equivalence of terms within the same concept entry.

These tasks are merely examples and not exhaustive, but they provide a clear idea of the comprehensive work involved in maintaining and ensuring the quality of the termbase so that terminological data are systematically organised and effectively support communication within the legal domain.

6 Concluding Remarks

In our paper, we started from the theoretical framework and focused on the initial considerations at the planning stage, followed by aspects for the conscious handling of data in termbases. The examples provided illustrate the necessary considerations for designing, maintaining, and evaluating a terminology resource, whether starting from scratch or working with existing terminological datasets. They also emphasise the importance of a structured and flexible framework to address the parameters outlined previously. A strong theoretical foundation is essential for designing resources that comply with established terminological principles and methods. Relevant literature and ISO standards play a pivotal role in this endeavour. In this context, the standards developed by ISO TC 37 "Language and Terminology"[10] are particularly crucial, as they provide a common framework and guidelines for representing, evaluating and exchanging data (Schmitz, 2025; Vezzani et al., 2025).

In a fast-moving world, the design of new terminology resources and data modelling requires a well-designed and flexible structure to facilitate data exchange and interoperability while ensuring that terminological data meets quality criteria. By meticulously planning data categories and adhering to international standards, it is possible to achieve a high level of data quality, interoperability, and usability.

Declaration on Generative AI

In preparing this work, the authors used Grammarly and Microsoft Copilot for initial grammar and spelling checks and proofreading. The content was then reviewed and edited with assistance from a native English speaker. The authors take full responsibility for the content of this publication.

References

Act CCIV of 2011 on National Higher Education. (2011). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2011-204-00-00.42 (English translation: https://www.mab.hu/wp-content/uploads/Nftv_angol_2Sept2016_EMMI-forditas.pdf)

Act CXC of 2011 on National Public Education. (2011). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2011-190-00-00 (English translation: https://natlex.ilo.org/dyn/natlex2/natlex2/files/download/106832/act_national_education.pdf)

Act LXXXI of 2023 on the promulgation of the Global Convention on the Recognition of Qualifications concerning Higher Education. (2023). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2023-81-00-00

Agrario, C., & Castagnoli, S. (2010). EOHS Term: Una knowledge base multilingue in materia di sicurezza sul lavoro. In F. Bertaccini, S. Castagnoli, & F. La Forgia (Eds.), *Terminologia a colori* (pp. 121–161). Bononia University Press.

Ammon, U., Bickel, H., & Lenz, A. N. (Eds.). (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen* (2nd ed.). De Gruyter.

Arntz, R., Picht, H., & Schmitz, K.-D. (2021). *Einführung in die Terminologiearbeit* (8th ed.). Olms.

Benő, A., & Péntek, J. (2023). A terminológiastratégia szintjei és feltételei Erdélyben. In G. Prószéky, Á. Fóris, A. Bölcskei, E. B. Papp, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 237–252). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.12

Bergenholtz, H., & Tarp, S. (2010). LSP lexicography or terminography? The lexicographer's point of view. In P. A. Fuertes-Olivera (Ed.), *Specialised dictionaries for learners* (pp. 27–37). De Gruyter.

[10] https://www.iso.org/committee/48104.html

bistro (n.d.). *Information System for Legal Terminology*. Institute for Applied Linguistics, Eurac Research. Retrieved May 5, 2025, from https://bistro.eurac.edu/de/

Bölcskei, A., & Fóris, Á. (2022). A névtudomány, a nevezéktan és a terminológia viszonya, érintkezési pontjai. *Névtani Értesítő, 44*, 59–78. https://doi.org/10.29178/NevtErt.2022.5

Boronkay, Zs. (2020). „Ki volt tanítód? Hol jártál iskolába?" Az oktatással kapcsolatos angol nyelvű okiratok fordításáról. In Sz. Szoták (Ed.), *A hiteles fordítás mint közfeladat*. OFFI. Retrieved May 8, 2025, from https://www.offi.hu/offi-akademia/kiadvanyok/a-hiteles-forditas-mint-kozfeladat

Chiocchetti, E. (2023). Legal terminology work for local-only minorities: The example of German in South Tyrol. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *49*(2), 345–363.

Chiocchetti, E., Heinisch-Obermoser, B., Löckinger, G., Lušicky, V., Ralli, N., Stanizzi, I., & Wissik, T. (2013). In E. Chiocchetti & N. Ralli (Eds.), *Guidelines for collaborative legal/administrative terminology work*. EURAC Research. Retrieved May 8, 2025, from https://cordis.europa.eu/docs/projects/cnect/7/270917/080/deliverables/001-D33Guidelinesforcollaborativelegaladministrativeterminologywork.pdf

Chiocchetti, E., Lušicky, V., & Wissik, T. (2023). Multilingual legal terminology databases: Workflows and roles. In Ł. Biel & H. J. Kockaert (Eds.), *Handbook of Terminology: Volume 3. Legal Terminology* (pp. 458-484). John Benjamins Publishing Company. https://doi.org/10.1075/hot.3.mul1

Contarino, A. G., & De Camillis, F. (2023). Domain-adapting and evaluating machine translation for institutional German in South Tyrol. In M. Izquierdo & Z. Sanz-Villa (Eds.), *Corpus use in cross-linguistic research: Paving the way for teaching, translation and professional communication* (pp. 173–194). John Benjamins Publishing Company. https://doi.org/10.1075/scl.113.10con

Czelnai, R., & Szepesi, D. (Eds.). (1986). *Meteorológia* (Műszaki értelmező szótár sorozat, No. 56). Akadémiai Kiadó. Retrieved May 8, 2025, from https://real-eod.mtak.hu/11421/

DatCatInfo – Data Category Repository. (n.d.). Retrieved May 10, 2025, from https://datcatinfo.net/

Drewer, P., & Schmitz, K.-D. (2017). *Terminologiemanagement: Grundlagen – Methoden – Werkzeuge*. Springer Vieweg.

Fóris, Á. (2025). Terminology in Hungary: From standard Hungarian to terms and scientific names. In K. Warburton & J. Humbley (Eds.), Terminology throughout history: A discipline in the making (pp. 508–529). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.24.24for

Fóris, Á. (2024a). Választások és döntések a terminológiakezelésben. *Alkalmazott Nyelvészeti Közlemények, 17*(1), 23-38.

Fóris, Á. (2024b). A felsőoktatás és a tudomány nemzetköziesedésének hatása a magyar anyanyelvű oktatásra és terminológiára. *Anyanyelv-pedagógia, 18*(4). https://doi.org/10.21030/anyp.2024.4.1

Fóris, Á. (2024c). The history and recent trends of terminology in Hungary in the 21st century. *Terminologija / Terminology, 31*, Article 3. https://doi.org/10.35321/term31-03

Fóris, Á., & Somogyi, Z. (2024). A magyar terminológiastratégia megvalósíthatósága: A szoftveres keretrendszer kiválasztásának szempontjai a Magyar Nemzeti Terminológiai Adatbázis tervezése során. *Glossa Iuridica*, *11*(1–2), 277–297.

Fóris, Á., Somogyi, Z., & Papp, E. B. (2024). Magyar Nemzeti Terminológiai Adatbázis tervezése: Általános kérdések. *Alkalmazott Nyelvtudomány, 24*(2), 21–41.

Früh, B., & Tamás, D. M. (2021). Quality evaluation of termbases. *Edition. Fachzeitschrift für Terminologie, 2/21*, 11–23. Retrieved May 2, 2025, from http://dttev.org/images/edition/ausgaben/edition-2021-2-e-version.pdf

Heinisch, B. (2023). Terminological usability – Adapting terminological databases to different user groups according to usability principles: The case of UniVieTerm. In Ú. Bhreathnach, N. Nissilä, & A. Velicu (Eds.), *Terminology Science & Research / Terminologie : Science et Recherche*, *26*, 24–44. Retrieved June 7, 2025, from https://journal-eaft-aet.net/index.php/tsr/issue/archive

ISO 639 (2023). *Code for individual languages and language groups*. International Organization for Standardization.

ISO 1087 (2019). *Terminology work and terminology science — Vocabulary*. International Organization for Standardization.

ISO 3166-1 (2020). *Codes for the representation of names of countries and their subdivisions — Part 1: Country code*. International Organization for Standardization.

ISO 3166-2 (2020). *Codes for the representation of names of countries and their subdivisions — Part 2: Country subdivision code*. International Organization for Standardization.

ISO 9000 (2015). *Quality management systems — Fundamentals and vocabulary*. International Organization for Standardization.

ISO 12616-1 (2021). *Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography*. International Organization for Standardization.

ISO 12620-1 (2022). *Management of terminology resources — Data categories — Part 1: Specifications*. International Organization for Standardization.

ISO 12620-2 (2022). *Management of terminology resources — Part 2: Repositories*. International Organization for Standardization.

ISO 16642 (2017). *Computer applications in terminology — Terminological markup framework*. International Organization for Standardization.

ISO 24495-1 (2023). *Plain language — Part 1: Governing principles and guidelines*. International Organization for Standardization.

ISO 26162-1 (2019). *Management of terminology resources — Terminology databases — Part 1: Design*. International Organization for Standardization.

ISO 26162-3 (2024). *Management of terminology resources — Part 3: Content*. International Organization for Standardization.

ISO 30042 (2019). *Management of terminology resources — TermBase eXchange (TBX)*. International Organization for Standardization.

ISCED (n.d.). *International Standard Classification of Education (ISCED)*. Retrieved May 2, 2025, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCE)

Kranebitter, K., & Ralli, N. (2021). Piccola guida per sviluppare strumenti terminologici. In C. Grimaldi & M. T. Zanola (Eds.), *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari* (pp. 113–123). Firenze University Press. Retrieved April, 18, from https://library.oapen.org/bitstream/id/d97dac1d-18b9-4029-aecb-be18fbb8f042/20500.pdf

Kranebitter, K., & Ralli, N. (2022). Quanto può influire l'utente nello sviluppo di uno strumento terminologico? L'esperienza di bistro. In E. Chiocchetti & N. Ralli (Eds.), *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia* (pp. 102–116). Eurac Research. https://doi.org/10.57749/wtfr-y339

Kuna, Á., & Ludányi, Zs. (2023). Terminológiai elvek az orvosi szaknyelvben és a gyógyító kommunikációban. Problémák, tendenciák, ideológiák. In G. Prószéky, Á. Fóris, A. Bölcskei, E. B. Papp, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 109–134). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.7

Lanstayák, I. (2023). Nyelvmenedzselés-elmélet és terminológia. In G. Prószéky, Á. Fóris, E. Papp, A. Bölcskei, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása: Zöld könyv* (pp. 253–276). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.13

Lipp, V., & Prószéky, G. (in press). Terminology planning in Hungary. In F. Steurs & R. Rosella (Eds.), *Handbook of terminology: Terminology planning in Europe*. John Benjamins Publishing Company.

Mayer, F. (2000). Terminographie im Recht: Probleme und Grenzen der Bozner Methode. In D. Veronesi (Ed.), *Linguistica giuridica italiana e tedesca / Rechtslinguistik des Deutschen und des Italienischen* (pp. 295–306). UNIPRESS.

Muhr, R. (2016). The state of the art of research on pluricentric languages: Where we were and where we are now. In R. Muhr, K. E. Fonyuy, I. Zeinab, & M. Coreyr (Eds.), *Pluricentric languages and non-dominant varieties worldwide* (Vol. 1, pp. 9–32). Peter Lang Verlag.

Nesbigall, J. (2025). Glossarerstellung für MT und KI. Warum viel nicht immer viel hilft. In P. Drewer et al. (Eds.), *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025,* (pp. 51–60). Deutscher Terminologie-Tag e.V.

Nilsson, H. (2009). The realisation of a national term bank – how and why? In *Proceedings of the ELETO – 7th Conference "Hellenic Language and Terminology"*, Athens, Greece, 22–24 October 2009. Retrieved April 28, 2025, from https://www.eleto.gr/download/Conferences/7th%20Conference/7th_26-26-NilssonHenrik_Paper_V03.pdf

Papp, E. (2023). A terminológiastratégia kérdései Európában. In G. Prószéky, Á. Fóris, E. Papp, A. Bölcskei, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 39–56). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.4

Papp, E. B. & Fóris, Á., & (2018). Planning a multilingual database of higher education terminology. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *44*(2), 595–610. Retrieved April 26, 2025, from https://hrcak.srce.hr/en/file/318304

Prószéky, G., Fóris, Á., Bölcskei, A., B. Papp, E., & Lipp, V. (Eds.). (2023). *A magyar terminológiastratégia kialakítása. Zöld könyv*. Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023

Ralli, N. (2025). Managing language varieties: Examples from legal terminology work. In *Proceedings of the 4th International Conference on "Multilingual Digital Terminology Today. Design, Representation Formats and Management Systems" (MDTT 2025)*. Thessaloniki.

Ralli, N., & Andreatta, N. (2018). bistro – ein Tool für mehrsprachige Rechtsterminologie. *trans-kom*, *11*(1), 7–44.A. (2023).

Ralli, N., & Kranebitter, K. (2017). Rechtsterminologie und Rechtssicherheit: das neue bistro. *eDITion: Fachzeitschrift für Terminologie*, *1*(17), 5–10.

Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins B.V. https://doi.org/10.1075/z.44

Sandrini, P. (1996). *Terminologiearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers*. TermNet.

Schmitz, K.-D. (2025). Wie gut ist meine Termbank? Die Normenreihe ISO 26162 hilft bei der Qualitätbewertung. In P. Drewer et al. (Eds.), *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025* (pp. 179–190). München et al.: Deutscher Terminologie-Tag e.V.

Tamás, D. M., & Sermann, E. (2019). Evaluation system for online terminological databases. *Terminologija/Terminology*, *26*, 24–46. https://doi.org/10.35321/term26-02

Tarp, S. (2008a). *Lexicography in the borderland between knowledge and non-knowledge_ General lexicographical theory with particular focus on learner's lexicography*. Max Niemeyer Verlag.

Tarp, S. (2008b). The third leg of two-legged lexicography. *HERMES – Journal of Language and Communication in Business*, *21*(40), 117–131. https://doi.org/10.7146/hjlcb.v21i40.96785

Vezzani, F. (2022). *Terminologie numérique : Conception, représentation et gestion*. Peter Lang International Academic Publishers. https://doi.org/10.3726/b19407

Vezzani, F., & Di Nunzio, G. M. (2020). Methodology for the standardisation of terminological resources: Design of TriMED database to support multi-register medical communication. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *26*(2), 265–297. https://doi.org/10.1075/term.00053.vez

Vezzani, F., Di Nunzio, G. M., Salgado, A., & Costa, R. (2025). When LMF and TMF meet: Towards a unified markup framework (UMF). *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *31*(1), 72-109. https://doi.org/10.1075/term.00084.vez

Warburton, K. (2021). *The corporate terminologist*. Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.21

Wissik, T. (2024). Dimensions of sustainability in terminology practice in institutional settings. In Ú. Bhreathnach, N. Nissilä, & A. Velicu (Eds.), *Terminology Science & Research / Terminologie : Science et Recherche*, *27*, 93–116. Retrieved June 16, 2025, from https://journal-eaft-aet.net/index.php/tsr/issue/archive