When LMF and TMF meet

Towards a Unified Markup Framework (UMF)

Federica Vezzani[1], Giorgio Maria Di Nunzio[1], Ana Salgado[2] & Rute Costa[2]
[1] University of Padova | [2] NOVA CLUNL, FCSH

The interoperability of language resources is crucial for effective communication and data exchange across various computational systems. In this context, the ISO/TC 37 standards, specifically the Lexical Markup Framework (LMF) and the Terminological Markup Framework (TMF), play a vital role by providing a common framework for the modelling, representation, and exchange of lexical and terminological data. The LMF has been deliberately aligned with TMF to facilitate close coordination between the two standards. This paper explores the convergence between LMF and TMF, underscoring the need for a Unified Markup Framework (UMF) that enhances interoperability and effective resource management. We propose a unified meta-model that integrates these frameworks through comparative analysis and real-world examples, facilitating the development of advanced language processing applications and multilingual lexicographic and terminology management. This study not only underscores the opportunities and challenges of such coordination but also sets the groundwork for future research directions in the harmonisation of lexicographic and terminology resources.

Keywords: data modelling, Lexical Markup Framework, lexicographic resources, Terminological Markup Framework, termbases

1. Introduction

Interoperability is a cornerstone of language resources, facilitating seamless communication and data exchange across diverse systems and applications. With the advancement of language technologies, the need for interoperable resources has become increasingly critical (Branco et al. 2023). In this context, ISO/TC 37 'Lan-

https://doi.org/10.1075/term.00084.vez
Terminology 31:1 (2025), pp. 72–109. ISSN 0929-9971 | E-ISSN 1569-9994
Available under the CC BY-NC 4.0 license. © 2025 John Benjamins Publishing Company

guage and Terminology'[1] standards play a pivotal role in ensuring interoperability among language resources by providing a common framework and guidelines for data representation and exchange.

The ISO 24613 series, commonly referred to as the LMF or Lexical Markup Framework, namely ISO 24613-1: 2024 'Language Resource Management — Lexical Markup Framework (LMF) — Part 1: Core Model', and ISO 16642: 2017 'Computer applications in terminology — Terminological markup framework' (TMF), are relevant standards in this regard. LMF focuses on lexical resources, encompassing lexical entries, morphology, syntax, and semantics, while TMF on the organisation and exchange of terminological data, including concepts, terms, definitions, and conceptual relationships. Despite their distinct areas of focus, both standards share a common goal: to facilitate effective resource management and enhance semantic interoperability among language technologies (Caselli and Bos 2023).

The ISO 24613 series[2] is explicitly "designed to coordinate closely with ISO 16642" (ISO 24613-1: 2024, v). This acknowledgement underscores the potential for synergy between LMF and TMF, indicating the possibility of harmonising these frameworks to achieve enhanced interoperability and resource management. Such coordination becomes increasingly significant in the context of computational terminology, where the integration of lexical and terminological data is essential for the development of advanced language processing applications (Bellandi et al. 2023b), such as enhanced information retrieval, question answering, and machine translation.

LMF and TMF provide standardised frameworks for building lexical and terminological resources, respectively, thereby facilitating their effective use in various computational and human-oriented applications. The computational aspect of modelling lexical and terminological data within LMF and TMF is of primary importance. In an era dominated by data-driven approaches and artificial intelligence, the effective representation and processing of linguistic data play a central role in the development of innovative language technologies (Bosque-Gil et al. 2018). LMF provides a standardised framework for constructing lexical resources, offering a rich set of features and functionalities that support the computational processing of linguistic data. From morphological analysis and syntactic parsing to semantic interpretation and text generation, LMF-compliant resources serve as foundational building blocks for a wide range of language processing tasks (Eckle-Kohler et al. 2012). Similarly, TMF plays a crucial role in facilitating the exchange and integration of terminological data, laying the groundwork for consistent ter-

1. https://www.iso.org/committee/48104.html

2. https://www.iso.org/committee/297592/x/catalogue/

minology management and multilingual communication (Vezzani and Di Nunzio 2020). By adopting standardised terminological structures and encoding conventions, TMF-compliant termbases enable interoperability across diverse domains and applications. From terminology extraction and alignment to multilingual terminology management and ontology development, TMF-based resources allow language technology developers and researchers to leverage terminology in innovative ways.

Although LMF and TMF share complementary objectives, their integration presents both significant opportunities and challenges. The structural and conceptual differences between the two frameworks may pose obstacles to consistent integration, requiring careful analysis and alignment. Nevertheless, by identifying commonalities and overlaps between LMF and TMF, it is possible to pave the way towards a unified meta-model that bridges the gap between lexical and terminological research domains. In this paper, we explore the potential coordination between LMF and TMF, with a focus on their structural alignment and computational implications. We aim to propose a unified meta-model that facilitates the integration of these frameworks, thereby enhancing interoperability and synergy among language resources. Through a comparative analysis of LMF and TMF, supplemented by examples and implementations, we seek to identify convergence points and propose a solution for a single cohesive meta-model.

The remainder of this paper is structured as follows: Section 2 provides an overview of the current state-of-the-art, examining seminal studies and implementations leveraging LMF and TMF. In Section 3, we conduct a comparative analysis of the structural, conceptual, and functional aspects of LMF and TMF, elucidating their potential convergence points and challenges. Real-world examples and case studies are presented to provide practical insights into the successes and challenges of implementing these standards. Building upon these foundations, Section 4 introduces the concept of a Unified Markup Framework (UMF) designed to bridge the gaps between LMF and TMF. Finally, Section 5 synthesises our findings, outlines future research directions, and concludes with reflections on the significance of harmonising lexical and terminological resources for advancing language technology applications.

2. Background

The evolving landscape of language technologies has seen significant advancements thanks to standardised frameworks like the LMF and the TMF. These frameworks not only ensure consistency in representing linguistic data but also

foster collaboration among diverse applications, thus advancing language resource management.

2.1 Overview of LMF

The genesis of LMF dates back to the early 21st century, influenced by international initiatives such as Acquilex,[3] Genelex,[4] and Parole,[5] which were supported by the European Commission. The initial expert team aimed to devise a general structure for LMF that would reflect the common features of existing lexicons. Their goal was to establish a consistent meta-terminology to identify and describe the components of these lexicons, thereby crafting a comprehensive model that accurately represents various lexicons and their components. The seminal work by Francopoulo et al. (2006) introduced LMF to the scholarly community. It provided the first comprehensive description of the framework's goals and architecture, emphasising its potential for facilitating interoperability and data exchange among different resources. This framework was formalised as ISO 24613 in 2008. Subsequently, the LMF book (Francopoulo 2013) elaborated on the various components and extensions of LMF, such as the core model, machine-readable dictionary (MRD) model, and etymological extension. This publication has become a pivotal resource for developers and researchers implementing LMF in various projects. More recently, Romary et al. (2019) examine the updates and revisions to the LMF standard, discussing the newer parts like the 'Syntax and Semantics' and 'Inflectional Morphology' extensions. This last presentation demonstrates the continuous evolution of the LMF to address the changing needs of the lexical resource community.

Developed by ISO/TC 37/SC 4/WG 4,[6] LMF is designed to provide a common model for lexical resources, facilitating data exchange and interoperability, primarily in Natural Language Processing (NLP) applications. The original ISO 24613 standard has been updated to include distinct parts:

- 'Core model' (ISO 24613-1: 2024),
- 'Machine-readable dictionary (MRD) model' (ISO 24613-2: 2020),
- 'Etymological extension' (ISO 24613-3: 2021),
- 'TEI serialisation' (ISO 24613-4: 2021),
- 'Lexical base exchange (LBX) serialisation' (ISO 24613-5: 2022),
- 'Syntax and Semantics' (ISO 24613-6: 2024).

3. https://www.cl.cam.ac.uk/research/nl/acquilex/
4. http://www.ilc.cnr.it/EAGLES96/lexarch/node15.html
5. https://cordis.europa.eu/project/id/LE24017
6. https://www.iso.org/committee/297592.html

New parts are currently being developed, namely 'Inflectional Morphology' (ISO 24613-7), and a new one to be initiated, 'Metadata for Lexical Resources' (ISO 24613-8).

The application of LMF is widespread in several real-world projects and applications, such as the Ortolang (Open Resources and TOols for LANGuage)[7] project in France which utilises LMF to organise and provide access to a diverse array of language resources, including corpora, lexicons and terminologies which support research in linguistics and language processing. During the METANET4U Project,[8] a good number of lexicons were converted into LMF in an effort to upgrade existing resources to agreed standards and guidelines.

Platforms like ORTOLANG[9] or resources such as UBY[10] and Apertium[11] have successfully adopted LMF to standardise and manage lexical data, ensuring semantic interoperability and facilitating tasks such as linking lexicons to ontologies, organising language resources, and supporting machine translation.

Additionally, integrating LMF with the OntoLex model[12] has enabled the creation of rich, ontology-based lexical resources for semantic web applications. This integration supports semantic annotation of lexical items, enhancing their readability and interoperability across web platforms. Furthermore, the Apertium Open-Source Machine Translation Platform[13] uses LMF to manage the lexical resources required for translating between multiple language pairs, particularly for under-resourced languages. The standardised structure of LMF facilitates the efficient update and scaling of lexical databases essential for translation accuracy. In this context, one of the masterworks that presents a detailed analysis about the problem of modelling language resources, and in particular how to apply Linked Data principles to Linguistic Data[14] by means of the Ontolex-lemon for lexical resources, is the study by Cimiano et al. (2020).

This paper will primarily focus on ISO 24613-1: 2024, the core model document, which encompasses a core package representing fundamental lexical entry information and interlinked extension packages, allowing for flexibility and reuse of components tailored to specific lexical resources.

---

7. https://www.ortolang.fr/en/home/

8. http://www.meta-net.eu/projects/METANET4U/

9. https://www.ortolang.fr/en/home/about/

10. https://dkpro.github.io/dkpro-uby/

11. https://github.com/apertium

12. https://www.w3.org/community/ontolex/wiki/Main_Page

13. https://www.apertium.org/index.zho.html#?dir=por-cat&q=

14. https://linguistic-lod.org/

2.2 Overview of TMF

On the terminological front, TMF has played a pivotal role in standardising the representation and exchange of terminological data (ISO 16642: 2017). TMF provides a structured framework for representing data within terminological collections, and it includes a meta-model and methods for describing specific Terminological Markup Languages (TMLs), with examples given in XML format.

In this section, we present the state-of-the-art of research papers that focus on the critical role of computational aspects in the management of terminology databases. Starting from the year 2017 (year of the publication of the ISO 16642), we collected the main works that focused on four main aspects:

- The computational methods for implementing and evaluating various approaches to modelling multilingual terminological data, as demonstrated by studies comparing ISO TC 37/SC 3[15] standards and Semantic Web frameworks.
- The computational techniques that enable the assessment of interoperability and reusability of corpora, facilitating the exploration of syntactic and semantic dimensions in language resource interoperability.
- The computational tools that play a crucial role in transitioning between different structural paradigms, such as from concept-oriented to sense-centred data organisations.
- The computational approaches that are guided by a FAIR management (Wilkinson et al. 2016) of terminological data, aiding in standardisation efforts across diverse domains and enhancing the accessibility and usability of terminology resources.

The recent survey by Gromann et al. (2024), despite not being directly linked to TMF, is an excellent starting point for an analysis of linguistic resources. In particular, the authors discuss the problem of multilingualism and Linguistic Linked Open Data (LLOD), emphasising support for various linguistic description levels. One important point is the description of best practices in representing, modelling, and linking linguistic description levels across multilingual LLOD resources.

In Vezzani, Di Nunzio and Costa (2023), the authors discuss the impact of three ISO TC 37/SC 3 standards (mainly ISO 16642: 2017, ISO 12620: 2019 and ISO 30042: 2019) on current research on terminology and, in particular, as the foundation of the FAIR terminology paradigm (Vezzani 2022). In particular, the authors intend to reflect in a critical perspective to highlight some possible limitations of the previously mentioned SC 3 standards in terms of FAIRification

15. https://www.iso.org/committee/48136.html

of terminological data. A similar analysis was performed by Piccini, Vezzani and Bellandi (2023) contrasting approaches to modelling multilingual terminological data (again by means of the ISO TC 37/SC 3 model using TMF as meta-model and the TermBase eXchange (TBX) standard as serialisation) and the Ontolex-Lemon model within the Semantic Web framework.[16] The paper offers a comparative multilevel analysis of these paradigms, aiming to uncover both their disparities and commonalities.

The research proposed by Caselli and Bos (2023) focuses on assessing the reusability of semantically interoperable corpora for events, exploring a dimension where resources share a standard vocabulary but implement separate schemes and guidelines. By starting from the distinction between syntactic and semantic interoperability in corpora, the authors aim to investigate the extent to which the promise of reusability, advocated by language resource interoperability, is fulfilled in this context.

The issues of interoperability among multilingual resources, in particular between terminology resources serialised in TBX and lexicographic resources serialised in Ontolex-lemon, is the focus of the work proposed by Bellandi et al. (2023b, 2023a). The focus is on exploring the theoretical and implementational implications of transitioning from a concept-oriented structure (TBX) to a sense-centred data organisation (Ontolex-lemon), and a methodology, design, and implementation of an interactive converter that will allow terminologists to actively participate in the conversion process.

The same idea of FAIR terminology was discussed and analysed by Vezzani and Di Nunzio (2020); Silecchia, Vezzani and Di Nunzio, (2022); Vezzani, Di Nunzio and Silecchia, (2022). In particular, these research works focus on the standardisation of the structure of the existing or newly created multilingual termbases in the medical domain, for example TriMED (Vezzani and Di Nunzio 2020), in the disarmament domain, for example DITTO (Vezzani, Di Nunzio and Silecchia 2022), or a new terminology database of human rights aiming to provide a contribution in terms of clear representation and simplification of legal language (Silecchia, Vezzani and Di Nunzio 2022).

Finally, Pernes, Romary and Warburton (2017) outline the specification of the TBX by means of the One Document Does it all (ODD) language, a generic specification language. This approach establishes a separation between the specification serialisation and the schema languages.

The research on TMF data modelling is, as can be seen from the volume of work in recent years, as active as the revision activity by the relevant ISO technical committee. Indeed, at present, ISO 16642: 2017 is being updated by TC 37/

16. https://www.w3.org/2019/09/lexicog/

SC 3/WG 3. In this paper, we will refer to the ISO/WD 16642: 2024 version which constitutes a working draft serving as a basis for review and commentary.

3. Comparison between LMF and TMF

Efforts to harmonise the LMF and TMF require a profound understanding of both frameworks. This section aims to delineate their similarities and distinctions, emphasising their potential synergies. By examining both frameworks, we can identify several key points of convergence and divergence, particularly concerning their focus, scope, and application, which impact their practical integration.

LMF is primarily focused on providing a structured approach to lexical data, encompassing aspects such as morphology, syntax, and semantics as previously mentioned. In contrast, TMF is designed for the organisation and exchange of terminological data, focusing on terms, concepts with their definitions, and the relationships among them. While LMF is often applied in the development of lexical resources and language processing tools, TMF is crucial in settings that require precise terminology management, such as in multilingual database systems and termbases.

Both LMF and TMF utilise Unified Modelling Language (UML) diagrams,[17] which are instrumental in conceptualising and visualising their respective structures and conceptual relationships. These diagrams provide a clear depiction of how each framework organises and manages language data, thus serving as a foundational tool for this comparative analysis.

To further enhance our understanding and explicitly illustrate the comparison between these two frameworks, we propose new diagrams using the entity-relationship (ER) model. Originally formulated by Chen (1976), the ER model is a conceptual tool widely used in the design of relational database applications. It allows for a detailed description of the objects of interest within a resource through the creation of an ER schema. The basic elements of this schema include:

1. Entities: which represent the set of objects of interest that have common properties.
2. Properties: which describe the attributes or characteristics of these entities.
3. Relationships: which delineate a logical link between several entities.

By employing this model, we aim to create diagrams that explicitly compare and contrast the structural components and relationships inherent in the LMF and TMF frameworks. This approach will facilitate a deeper insight into their respec-

17. https://www.omg.org/spec/UML/

tive functionalities and interactions, thereby aiding in a more comprehensive analysis of their potential for integration and harmonisation.

The following ER diagrams are designed to reflect the requirements specified within the ISO 24613-1: 2024 standard for the core model of the LMF meta-model for lexical resources, and the ISO/WD 16642: 2024 standard for the TMF meta-model for terminology resources.

From a graphical point of view, entities are represented by rectangles, connected by relationships in the form of diamonds. Additionally, the cardinalities are indicated, that is, the pair of numbers present on each segment that links the entities to the associations and respectively represents the minimum and maximum number of objects of that entity (the one closest to the pair of numbers) that can be linked with the entities involved in that relationship.

The analysis is conducted on two structural levels:

1. Macrostructure of the resource: an overview of the entire framework structure.
2. Microstructure of the entries: a detailed examination of individual entries.

By analysing these structural levels, our goal is to identify similarities, differences, strengths, and weaknesses in the organisation and representation of lexical and terminological data.

3.1 Macrostructures

3.1.1 *LMF: Macrostructure of a Lexical Resource*

In Figure 1, we illustrate the macrostructure of a Lexical Resource according to the LMF meta-model. The three entities represented are defined within ISO 24613-1: 2024. A lexical resource is defined as "a database consisting of one or several lexicons", while a lexicon is "a resource comprising lexical entries for one or several languages". Although lexical entries are fundamental constituents of a lexical resource, the concept of lexical entry is not defined within the standard (particularly in Section 3 of the standard, which is generally dedicated to 'Terms and definitions'). However, in section 5.3.6 of the standard, dedicated to the description of the LMF core package through the UML diagram, the class lexical entry is described as "a container for managing Form and Sense classes"'.

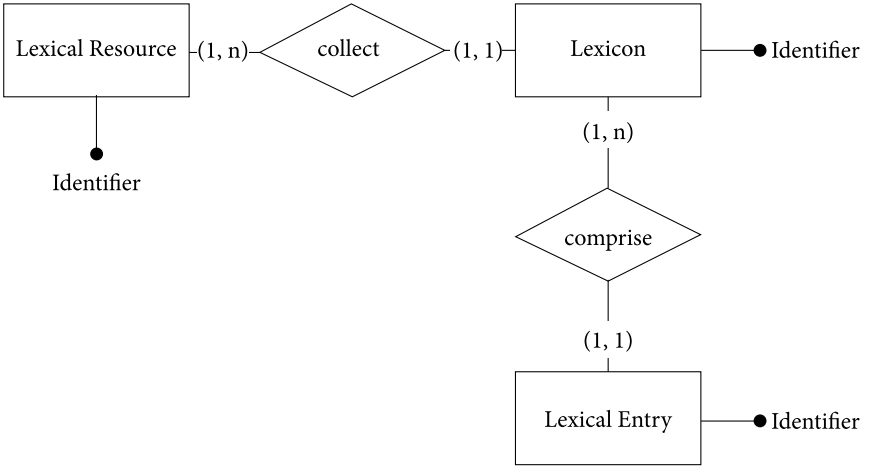

Figure 1. ER schema of a lexical resource macrostructure

Considering the above descriptions, the ER schema in Figure 1 can be read as follows:[18]

- The entity "Lexical Resource" is associated with the entity "Lexicon" through the relationship 'collect'. Each lexical resource, identified through the property *Identifier*, must collect at least one lexicon (cardinality (1, n)).
- The entity "Lexicon", also identified by the property *Identifier*, must be included in a single lexical resource (cardinality (1, 1)). Additionally, "Lexicon" is associated with the entity "Lexical Entry" through the relationship 'comprise'. A "Lexicon" must comprise at least one lexical entry (cardinality (1, n)).
- The entity "Lexical Entry", also having its own *Identifier*, must be included in only one "Lexicon" (cardinality (1, 1)).

3.1.2 *TMF: Macrostructure of a Termbase*

In Figure 2, we illustrate the macrostructure of a termbase according to the TMF meta-model. The three entities represented are defined within ISO/WD 16642: 2024 and ISO 1087: 2019. Specifically, a termbase is a "database comprising a ter-

18. In the description of the ER schemas in the text, we use the double quotes to indicate an entity; the single quotes for relationships; and *italics* for properties. For example, when we write "Lexicon", we are actually indicating the corresponding entity in the schema, while 'comprise' indicates the corresponding relationship and *Identifier* the property of the corresponding entity. The verb that is used in a relationship can be read in the active or in the passive form according to the reading direction: A "Lexical Resource" collects a "Lexicon" and/or a "Lexicon" is collected within a "Lexical Resource".

minology resource" (ISO 1087: 2019). The designation "terminology resource" is considered a synonym of "terminological data collection" used within ISO/WD 16642: 2024, and is defined as a "resource consisting of concept entries with associated metadata and documentary information". Finally, the concept entry is defined as a "part of a terminological data collection which contains the terminological data related to one concept" (ISO/WD 16642: 2024).

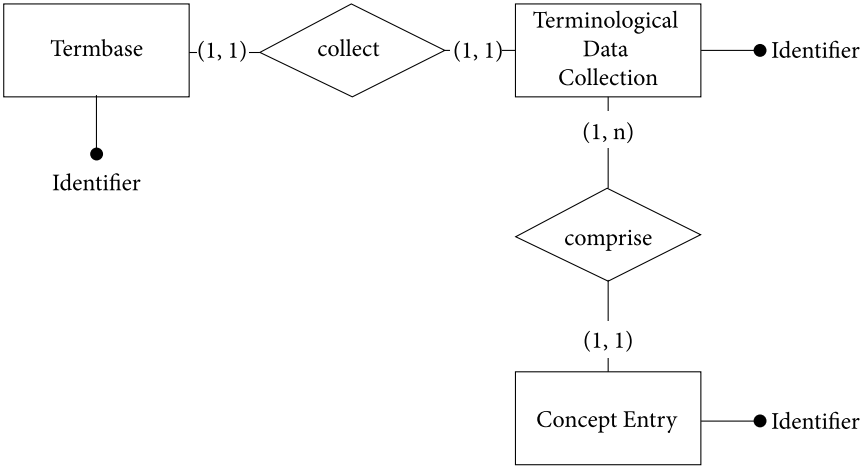

Figure 2. ER schema of a termbase macrostructure

Given these descriptions, the ER schema in Figure 2 can be read as follows:

1. The entity "Termbase" is associated with the entity "Terminological Data Collection" through the relationship 'collect'. Each termbase, identified through the property *Identifier*, must collect at most one terminological data collection (cardinality (1, 1)).
2. The entity "Terminological Data Collection", also identified by the property *Identifier*, must be included in a termbase (cardinality (1, 1)). Additionally, "Terminological Data Collection" is associated with the entity "Concept Entry" through the relationship 'comprise'. A "Terminological Data Collection" must comprise at least one concept entry (cardinality (1, n)).
3. The entity "Concept Entry", also having its own *Identifier*, must be included in only one "Terminological Data Collection" (cardinality (1, 1)).

The two diagrams, Figures 1 and 2, show that the macrostructures of the two resources are almost identical. However, there is one detail worth mentioning: the number of lexicons or terminological data collections that can be collected in a lexical resource or a termbase, respectively. To align the two macrostructures, it would be necessary to modify the wording of the ISO 1087: 2019 (specifically, the

definition of a termbase as a "database comprising a terminology resource") to match the cardinality (1, n); thus, allowing a termbase to collect more than one terminological data collection. In Figure 3, we show an example of a set of elements that correspond to the two ER diagrams in a tree-structure form. In particular, on the left side, a lexical resource collects three lexicons, where Lexicon 1 comprises two lexical entries. On the right side, a termbase containing one terminological data collection that in turn comprises two concept entries. We greyed out the possibility of having two more terminological data collections.

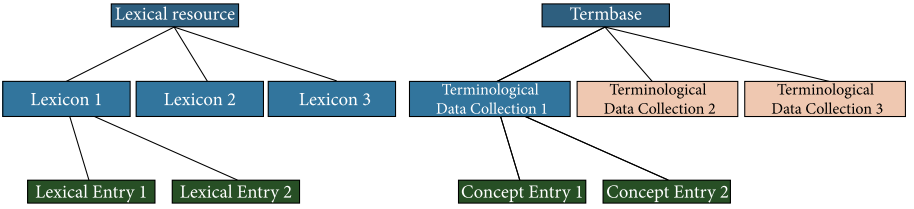

Figure 3. Comparison between the two macrostructures

3.2 Microstructures

3.2.1 *LMF Microstructure of a Lexical Entry*

Based on the descriptions in the LMF standard,[19] we now move to the microstructure level of entries for each of the two types of resources. Specifically, concerning the lexical entries, the ER schema represented in Figure 4 can be interpreted as follows:

1. The entity "Lexical Entry", as described above and identified by its own *Identifier*, is associated with the entities "Form" and "Sense" through the two relationships 'containForm' and 'containSense', respectively. In particular, one lexical entry contains zero or more forms (cardinality (0, n)), the latter being, according to ISO 24613-1: 2024, a class grouping and managing "all the information about the written and spoken forms of a word, multiword expression, root, stem or morpheme". At the same time, a lexical entry contains zero or more senses (cardinality (0, n)), the latter representing "one meaning of a lexical entry" (ISO 24613-1: 2024). Finally, a "Lexical Entry" is also associated with the entity "Lemma" (defined as "word form chosen to represent a lexeme" (ISO 24613-1: 2024)) through the 'hasLemma' relationship, specifically a lexical entry can have at most one lemma (cardinality (0, 1)).

19. For the sake of clarity, we have not included the GrammaticalInformation class and the OrthographicRepresentation class in this ER schema. These classes can be incorporated subsequently without altering the current schema.

2. The entity "Form", identified by its own *Identifier*, must be contained within at least one lexical entry (cardinality (1, n)). The entity "Form" is associated with the entity "Sense" through the relationship 'hasSense'. In the UML diagram of ISO 24613-1: 2024, this relationship is underspecified, as the cardinalities linking the two classes are not expressed. Therefore, in the ER diagram, where it is mandatory to specify the cardinalities, we decided to assign a generic (0, n) cardinality between the two entities. This means that a form can have multiple senses (cardinality (0, n)). Finally, a "Form" is a superset of the entity "Lemma" which means that some forms are considered as lemmas. This is indicated in the diagram with an arrow.
3. The entity "Lemma" is a subset of the entity "Form" and is associated with the entity "Lexical Entry" through the relationship 'hasLemma'. In particular, one element of the entity "Lemma" must be a lemma of only one lexical entry (cardinality (1, 1)).
4. The entity "Sense" identified by its own *Identifier*, must be contained in only one "Lexical Entry" (cardinality (1, 1)). For the aforementioned reason, a sense can be associated with different forms (cardinality (0, n)). Moreover, the entity "Sense" can be organised hierarchically in the following way: a "Sense" can be the subordinate of at most another "Sense" (cardinality (0, 1)) and, at the same time, it may have many subordinate senses (cardinality (0, n)). Finally, the entity "Sense" is associated with the entity "Definition" through the relationship 'hasDefinition'. In particular, a sense can have multiple definitions (cardinality (0, n)). The "Definition" is defined in ISO 24613-1: 2024 as the "narrative description of a sense".
5. The entity "Definition" with its own *Identifier* must be associated with only one "Sense".

At this point, to proceed with the analysis of the comparison between LMF and TMF, it becomes necessary to make some modifications to the schema in Figure 4 regarding three fundamental points:

1. We believe that the participation of the entity "Sense" in the relationship 'containSense' is too restrictive because, for some lexical resources, the same sense can be associated with more than one lexical entry. For example, consider the case of two synonymous forms contained within two separate lexical entries, but share their sense.[20] For this reason, it is more accurate to remove the relationship 'containSense' and retrieve all the lexical entries to which a sense is

20. For example, see the two lexical entries of the *Dicionário da Língua Portuguesa*, where the lemma "parkinson" (https://dicionario.acad-ciencias.pt/pesquisa/?word=parkison) and the lemma "parkinsonismo" (https://dicionario.acad-ciencias.pt/pesquisa/?word=parkinsonismo) share the exact same definition as narrative description of the same sense. Consequently, the same sense is linked with two lexical entries.

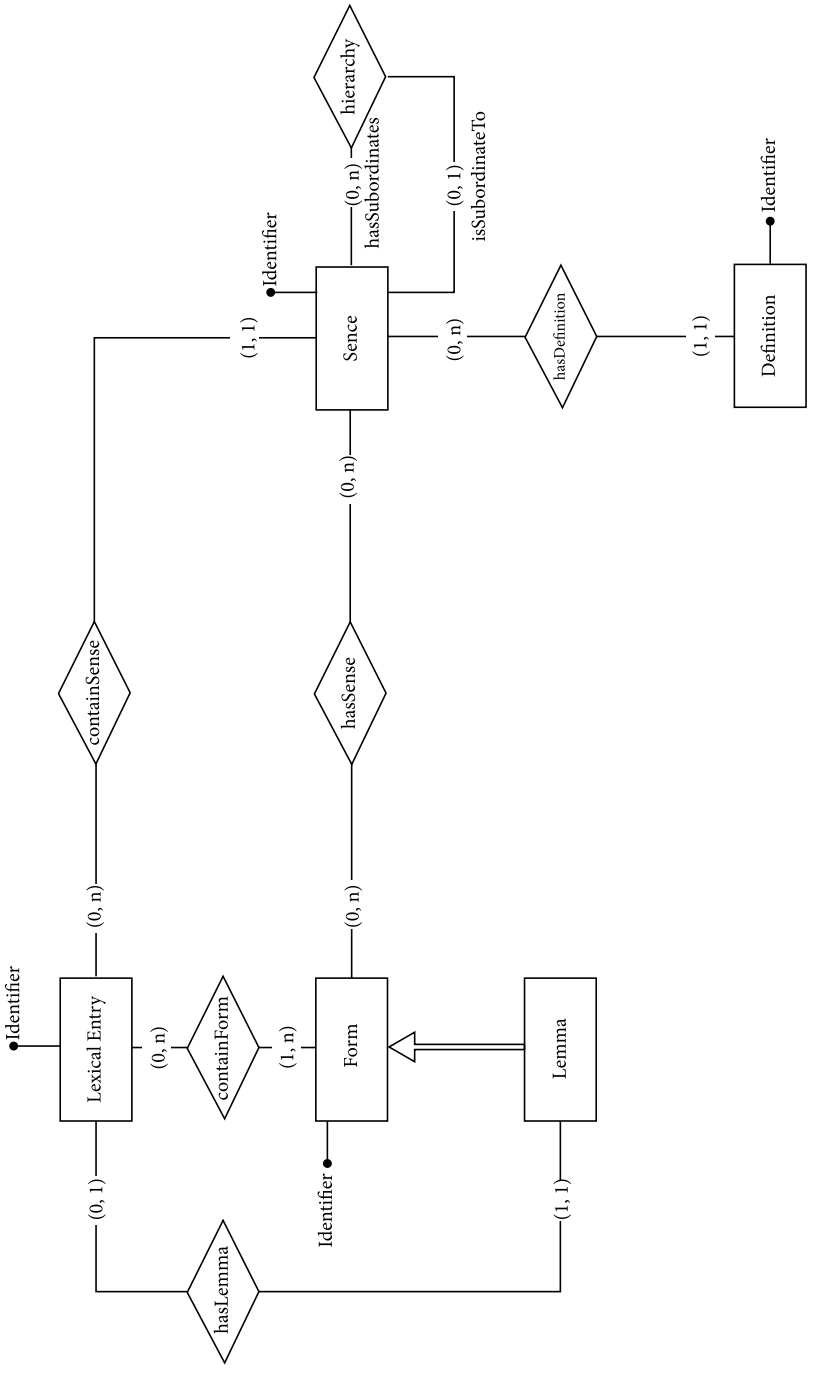

Figure 4. ER schema of a lexical entry microstructure according to LMF

associated through the link with the entity "Form". In Figure 5, we have represented this deletion by reducing the opacity of the corresponding portion of the ER schema (suggesting its removal in the new version of the schema).

2. At the same time, for a sense to be included in a "Lexical Resource", it must be associated with at least one "Form" (senses not associated with any form in a lexical resource cannot exist). We have illustrated this modification with a different colour on the new cardinality (1, n) from the entity "Sense" to the relationship 'hasSense'.
3. Finally, in the documentation of the UML classes of ISO 24613-1: 2024, there is not a class specifically dedicated to the natural language, which is always mentioned in relation to the "Form" and the "Definition" classes. We believe it is more accurate, from a modelling perspective of the entities involved in representing LMF, to add the entity "Language". We have represented this modification with a different colour of the added entities and associations.

As a consequence of the proposed modifications, the description of the new ER schema in Figure 5 has to be changed accordingly:

- The entity "Lexical Entry" does not contain senses but only forms.
- The entity "Sense" is not contained within a lexical entry; however, a sense must be associated with at least one form (cardinality (1, n)).
- The entity "Language", identified by its own *ISO code* (ISO 639: 2023),[21] is associated with the entities "Form" and "Definition" through the relationships 'express' and 'isWritten', respectively. In particular, one language can be used to express any form (cardinality (0, n)) or to write any definition (cardinality (0, n)).
- A "Form" must be expressed in only one language (cardinality (1, 1)).
- A "Definition" with its own identifier must be written in only one language (cardinality (1, 1)).[22]

The decision to remove the 'containSense' relationship reflects our commitment to ensuring that the structure of the model proposed in Section 4 remains consistent with real-world usage scenarios. By requiring that a sense be directly linked to at least one form, we maintain a clear and logical connection between the abstract meaning (sense) and its concrete linguistic realisation (form). This approach sim-

21. https://www.iso.org/standard/74575.html

22. This requirement does not imply the fact that a definition of a sense must be written in only one language. For example, given a sense with the identifier "s1", it can have two definitions identified with "d1" and "d2" respectively. Each definition can be written in a different natural language, for example "d1" in language "l1" and "d2" in language "l2".

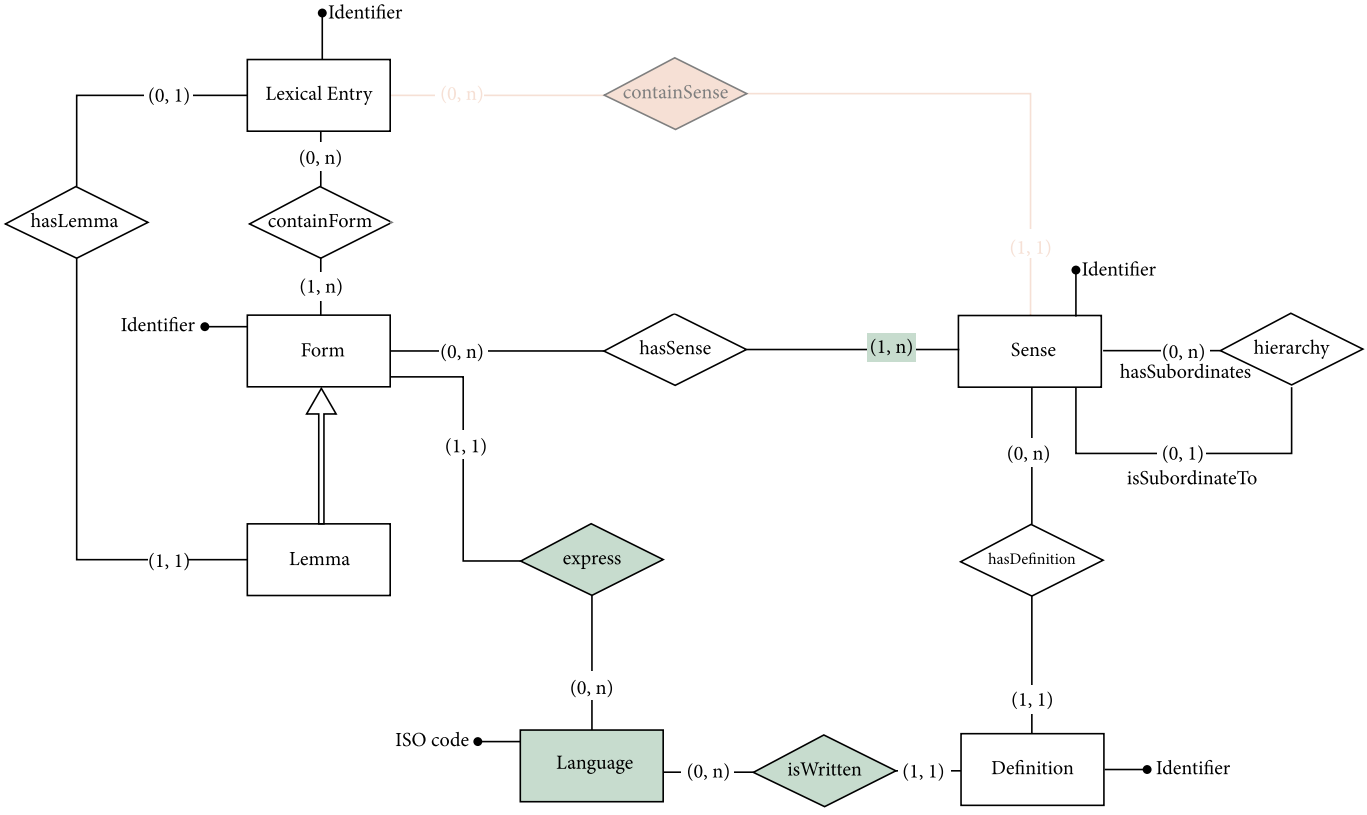

Figure 5. Revisiting the ER schema of the microstructure of a lexical entry

plifies the schema, avoids potential redundancies, and ensures that the model aligns with the way lexical data is typically represented in practice.

3.2.2 *LMF: Microstructure of a Lexicographic Entry*

In the previous sections, we presented an ER schema of the UML diagram of the LMF standard (Figure 4) and its subsequent correction based on requirements that we consider more accurate regarding the landscape of lexical resources (Figure 5). Since our goal is to compare the structure of two specific types of resources, general language dictionaries on one hand and domain-oriented terminology resources on the other, it becomes necessary to further specify the ER schema presented in Figure 5. In particular, in Figure 6, we represent the case of lexicographic resources which are considered as a kind of lexical resource. The specificity of this type of resource is expressed through the variation of three cardinalities illustrated in Figure 6 with the colour green. In particular:

1. A "Lexical Entry" must have at least one form (cardinality (1, n)).
2. A "Lexical Entry" must have a lemma (cardinality (1,1)).
3. A "Form" must have at least one sense (cardinality (1, n)).

3.2.3 *TMF: Microstructure of a Concept Entry*[23]

Following the same logic of the microstructure analysis of the entry and based on the requirements expressed in the TMF standard, Figure 7 illustrates the representation of a specific terminological data collection.[24] The structure of the ER schema for TMF reflects the three-level architecture of the meta-model.

In particular:

1. The entity "Concept Entry" is defined as the "information that pertains to a single concept"' (ISO/WD 16642: 2024). For this reason, it is associated and identified by the related "Concept" through the relationship 'hasConcept', with cardinality (1, 1). A concept entry contains zero, one or more language sections (cardinality (0, n)).
2. The entity "Language Section" is a "container for all the term sections of a concept entry for a given language, as well as information pertaining to the

23. We will use Concept Entry instead of Terminological Entry according to the new edition of the ISO/WD 16642: 2024.

24. For the sake of clarity, we have excluded in this ER schema the UML classes relative to the Terminological Component Section (as the "information about parts of a term such as morphemes, phonemes, syllables, or single words" according to ISO/WD 16642: 2024), and the Structural node associations that allow for a parent-child relationship of each level of the "Entry".

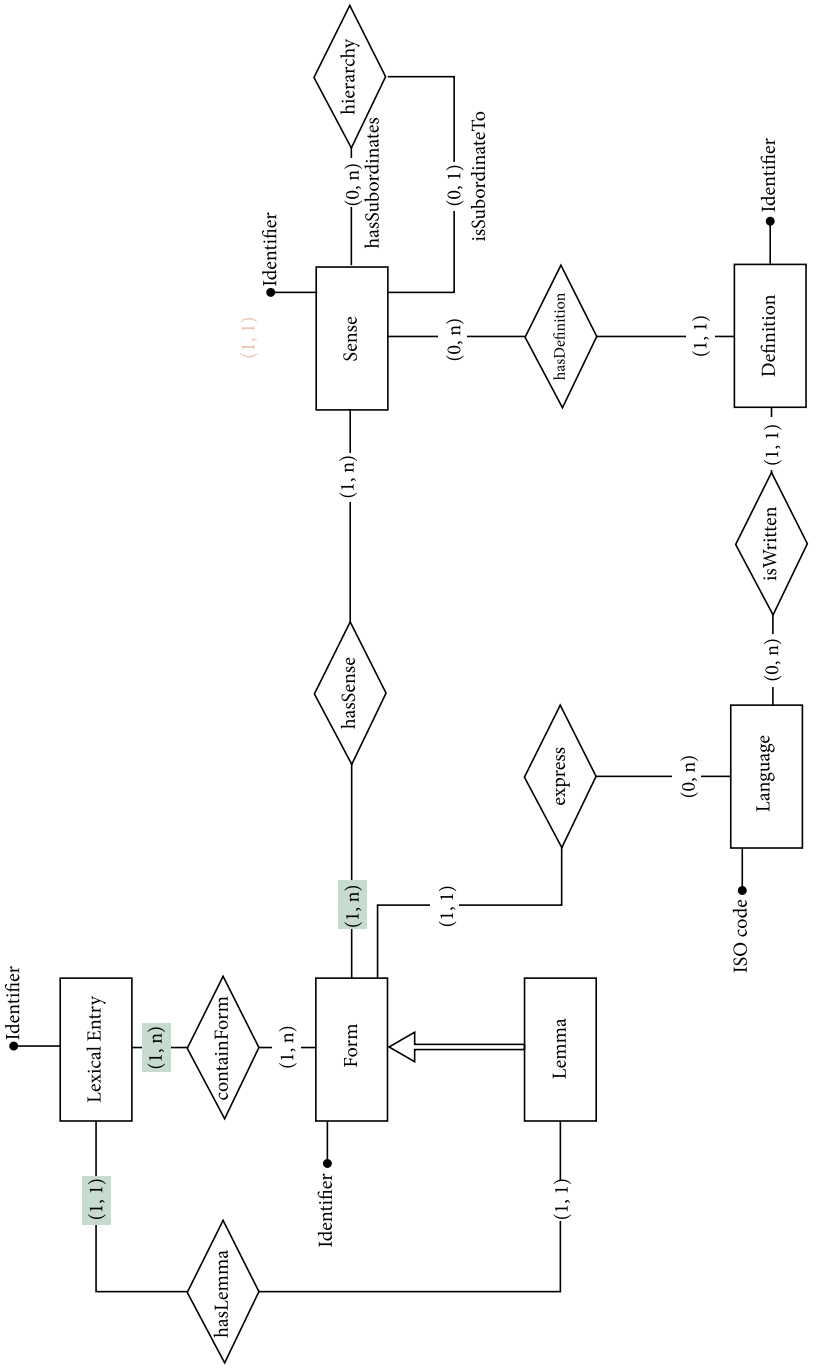

Figure 6. ER schema of a lexicographic entry microstructure

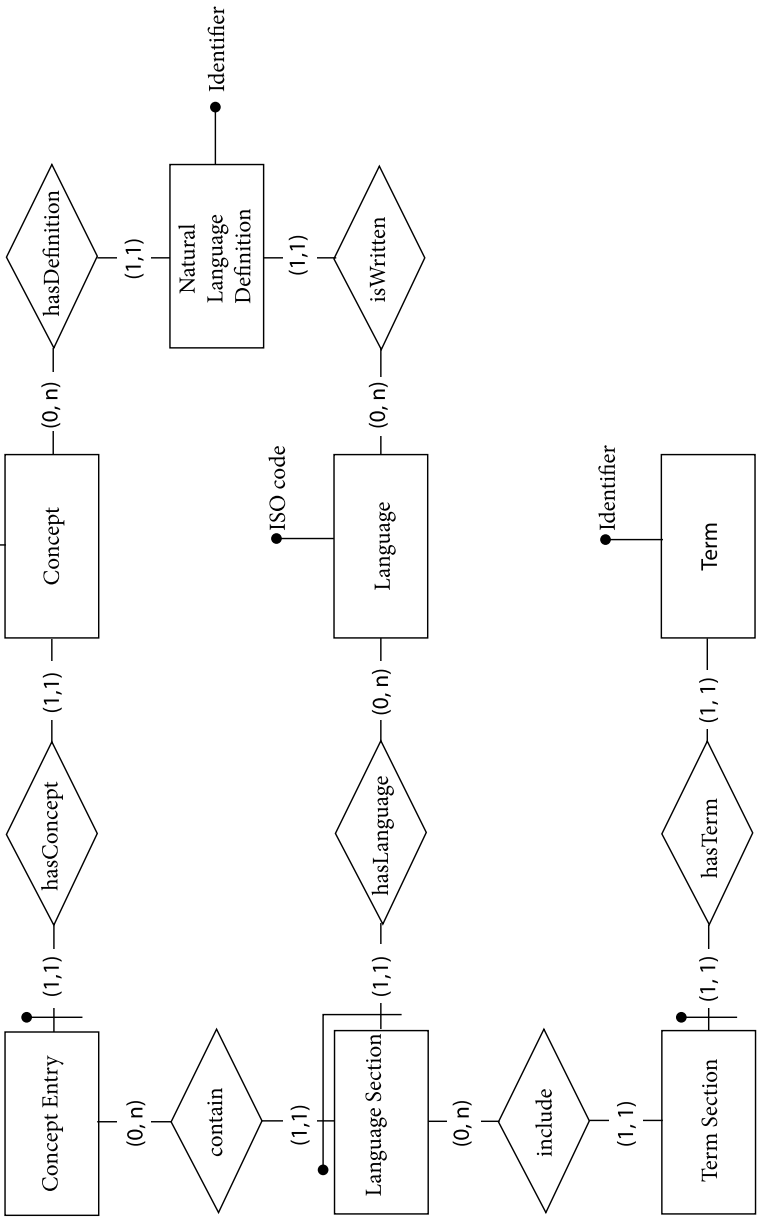

Figure 7. ER schema of a concept entry microstructure

concept in that language" (ISO/WD 16642: 2024). This entity is identified by the pair "Concept Entry" and "Language" through the relationships 'contain' and 'hasLanguage' (both with cardinality (1, 1)). A language section is associated with the "Term Section" entity and can include different term sections (cardinality (0, n)).

3. The entity "Term Section", as described in ISO/WD 16642: 2024, "contains exactly one term, and information about the term". In particular, it is identified by the "Term" itself through the association 'hasTerm' with cardinality (1, 1) and can be included in only one "Language Section" (cardinality (1, 1)).
4. The entity "Concept" (defined in ISO 1087: 2019 as a "unit of knowledge created by a unique combination of characteristics") is identified by its own *Identifier* and must be associated with only one "Concept Entry" through the relationship 'hasConcept' (cardinality (1, 1)). A concept entry is also associated with the entity "Natural Language Definition" through the 'hasDefinition' relationship. Specifically, a concept may have more natural language definitions (cardinality (0, n)).
5. The entity "Language" identified by its own *ISO code* is associated with the entities "Language Section" and "Natural Language Definition" through the relationships 'hasLanguage' and 'isWritten', respectively. In particular, one language can be used in any language section (cardinality (0, n)) or to write any definition (cardinality (0, n)).
6. The entity "Natural Language Definition" with its own *Identifier* must be associated with only one concept and one language through the relationships 'hasDefinition' and 'isWritten' (both with cardinality (1, 1)).
7. The entity "Term" (defined in ISO 1087: 2019 as a "designation that represents a general concept by linguistic means") is identified by its own *Identifier* and must be associated with only one Term Section (cardinality (1, 1)).

3.3 Comparison: Lexicographic Entry vs Concept Entry

In Figure 8 (left side), we rearranged a portion of Figure 6 to match the layout of Figure 7 (shown on the right side of Figure 8). We coloured the entities and relationships to highlight the similarities between the two diagrams.

The form-to-sense approach of the left side (which corresponds to the LMF) can be appreciated due to the fact that the lexical entry points directly to the form. From the form, we can obtain all the other information relative to that form, specifically, the language in which it is verbalised and the sense (or senses) that it expresses.

The concept-to-term approach of the right side (which corresponds to the TMF) is less evident given the more elaborate layout: starting from the top left,

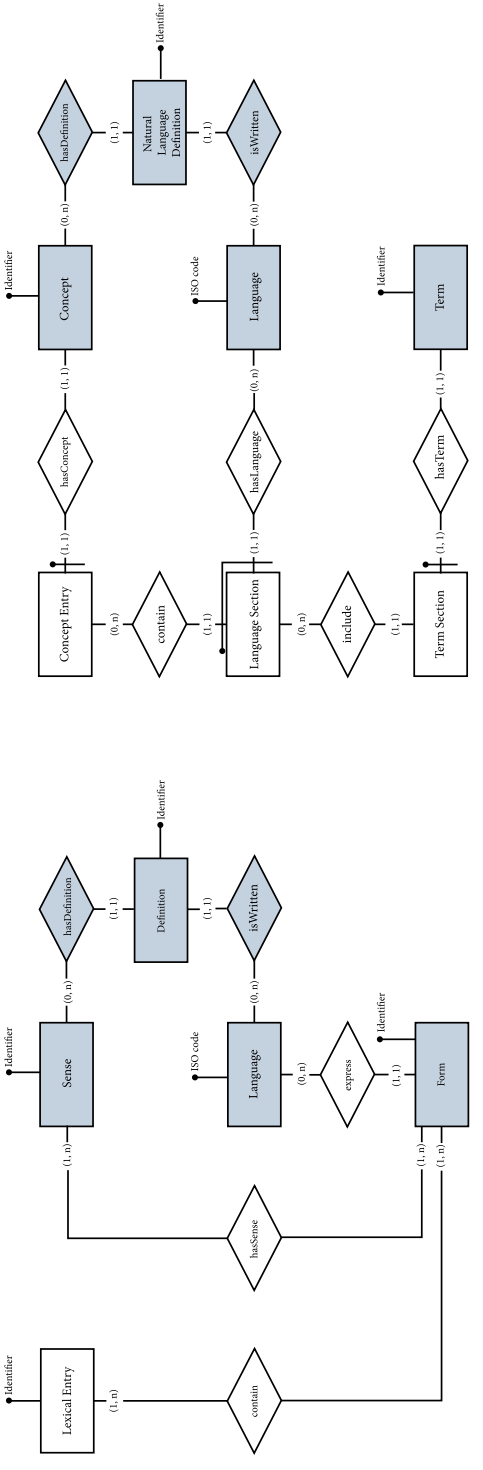

Figure 8. Comparison between the ER schemas of the two microstructures

given a concept entry, we have different language sections, and for each of these sections different terms sections.

From a graphical viewpoint, this distinction is evident from the three-level organisation of the TMF (concept, language, term) compared to a more graph-oriented way of representing the information in LMF.

Nevertheless, it is also possible to notice a good amount of overlap of elements of both diagrams where the entities "Form" and "Term", "Sense" and "Concept", and "Language" are aligned in this schema. Moreover, we have the same exact structure among the "Definition" and "Natural Language Definition" entities. These elements constitute the starting point in conceiving a unified model.

To highlight even more the overlap between these two diagrams, we can add to the TMF diagram two relationships that can be elicited from the three-tier structure itself. In Figure 9, we show in yellow these new relationships 'designate' and 'express' and we grey out the structural elements of the TMF for a better visualisation.

By adding these elements, we express the fact that a "Term" must designate only one "Concept" and must be expressed in at most one "Language". This addition allows us to have an almost perfect overlap with the relationships 'hasSense' and 'express' of the LMF diagram.

In Figures 10, 11, and 12, we illustrate three examples of XML-based serialisation of a lexicographic entry and a concept entry modelled according to LMF (Figures 10 and 11) and TMF (Figure 12). The serialisations correspond to the standards ISO 24613-5: 2022 for Lexical base exchange (LBX), ISO 24613-4: 2021 for Text Encoding Initiative (TEI), and ISO 30042: 2019 for TBX.

In Figure 10, we depict an example of a lexicographic entry in Portuguese with the lemma 'galáxia'.[25] In addition to specifying the grammatical features of the lemma (in this case, noun feminine, which includes part-of-speech and gender), we represent two of the senses attributed to the lemma based on the information contained in the *Dicionário da Língua Portuguesa* (ACL 2024).[26] For each sense, the narrative description is provided in the form of a definition in the Portuguese language.

In Figure 11, we show an example of another lexicographic entry in Portuguese for the lemma 'virus'[27] serialised in TEI. In addition to specifying the grammatical features of the lemma (in this case, masculine noun, singular and

25. https://dicionario.acad-ciencias.pt/pesquisa/?word=galaxia

26. The lexicographic entry available in the dictionary in question contains a total of four senses attributed to the lemma. In Figure 10, we selected only two of them as an example of an entry with more than one sense. The same applies also to the example shown in Figure 11.

27. https://dicionario.acad-ciencias.pt/pesquisa/?word=virus

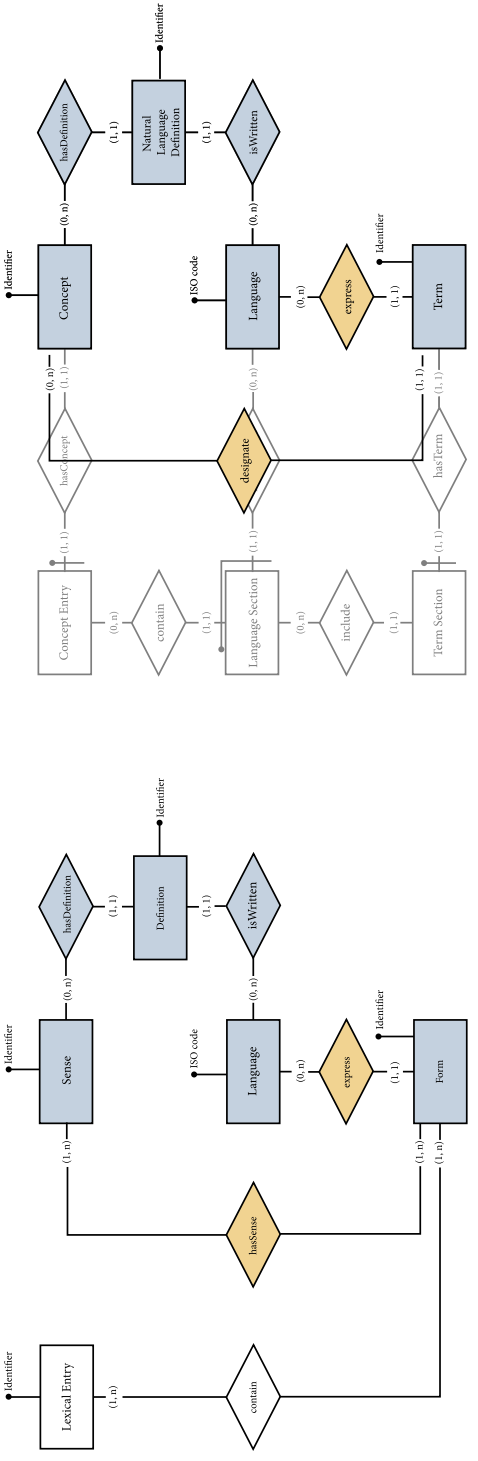

Figure 9. Comparison between the ER schemas of the two microstructures with additional relationships

```
<Entry xml:lang="pt">
    <Lemma>
        <GramFeats>
            <POS>noun</POS>
            <Gender>fem</Gender>
        </GramFeats>
        <FormRep xml:lang="pt" notation="Portuguese">galáxia</FormRep>
    </Lemma>
    <Sense senseNR="1">
        <Def>
            <DefRep xml:lang="pt">sistema astral exterior àquele a que pertence o Sol e
            que pode apresentar forma espiral, elíptica ou irregular</DefRep>
        </Def>
    </Sense>
    <Sense senseNR="2">
        <Def>
            <DefRep xml:lang="pt">mundo distante, por oposição à realidade
            circundante</DefRep>
        </Def>
    </Sense>
</Entry>
```

Figure 10. Example of a LBX-serialised lexical entry

plural), we represent two of the senses attributed to the lemma based on the information contained in the *Dicionário da Língua Portuguesa* (ACL 2024). The narrative description is provided in the <def> tag in Portuguese.

```
<entry xml:id="ACL.DLP.VIRUS" type="mainEntry" xml:lang="pt">
   <form type="lemma">
      <orth>virus</orth>
   </form>
   <gramGrp>
      <gram type="pos" norm="NOUN" expand="nome">n.</gram>
      <gram type="gen" expand="masculino">m.</gram>
      <gram type="num" expand="singular e plural">sing. e pl.</gram>
   </gramGrp>
   <sense xml:id="ACL.DLP.VIRUS.sense.1" n="1">
      <usg type="temporal" expand="antigo">ant.</usg>
      <def>designação atribuída a qualquer germe patogénico, agente de doenças contagiosas</def>
   </sense>
   <sense xml:id="ACL.DLP.VIRUS.sense.6" n="6">
      <usg type="domain" ana="#domain.informatics" resp="#user123"/>
      <def>programa de computador executado independentemente da vontade do utilizador que se
         infiltra em sistemas informáticos, causando danos diversos, como
         corrupção de dados, interrupção do funcionamento normal do computador e replicação
         automática para outros sistemas</def>
   </sense>
   <etym>Do latim <i>vīrus</i>, ‘suco; peçonha, veneno’</etym>
</entry>
```

Figure 11. Example of a TEI-serialised lexical entry

In Figure 12, we depict a concept entry containing the concept identified as c1 related to the domain of astronomy. The concept is subsequently verbalised in two languages (English and Italian), corresponding to two language sections within the concept entry. At the language level, the natural language definition of the concept is provided in both English and Italian with the respective indication of the consulted source. For each language, the terms designating the concept are then visible. In particular, for the English language, we have two term sections con-

taining respectively the designation 'astronomical object' and 'celestial object' as two term variants for the same concept. For the Italian language, the term 'oggetto celeste' is indicated as the verbalization of the concept in question.

```
<conceptEntry id="c1">
    <min:subjectField>Astronomy</min:subjectField>
    <langSec xml:lang="en">
        <descripGrp>
            <basic:definition>physical body of astronomically-significant size, mass, or
            role, naturally occurring in a universe</basic:definition>
            <basic:source>Wikidata</basic:source>
        </descripGrp>
        <termSec>
            <term>astronomical object</term>
            <min:partOfSpeech>noun</min:partOfSpeech>
        </termSec>
        <termSec>
            <term>celestial object</term>
            <min:partOfSpeech>noun</min:partOfSpeech>
        </termSec>
    </langSec>
    <langSec xml:lang="it">
        <descripGrp>
            <basic:definition>qualsiasi corpo o oggetto non appartenente al pianeta
            Terra</basic:definition>
            <basic:source>Wikidata</basic:source>
        </descripGrp>
        <termSec>
            <term>oggetto celeste</term>
            <min:partOfSpeech>noun</min:partOfSpeech>
        </termSec>
    </langSec>
</conceptEntry>
```

Figure 12. Example of a TBX-serialised concept entry

Once the examples of XML-based serialisations have been illustrated, our goal is to demonstrate how the data contained in these serialisations can be easily stored in the corresponding tabular version derived from the ER schema proposed in Figure 9. The transformation of the ER model into a tabular representation follows standard database normalisation principles, based on the Chen Entity-Relationship Model (1976). The algorithm employed organises the entities and relationships into relational tables, preserving referential integrity through the use of foreign keys. A detailed description of this process can be found in Chen (1976).

In Figure 13, we show the tabular representation[28] of the ER schema of the LMF meta-model (left side of Figure 9) with the data presented in Figure 10.

28. In the description of the tabular representation, we use the double quotes to indicate the name of a table, for example "Lexical Entry"; the single quotes for the name of the columns, for example 'Representation'; and underlined text to indicate the primary key of the column (for example, Identifier for the table "Form").

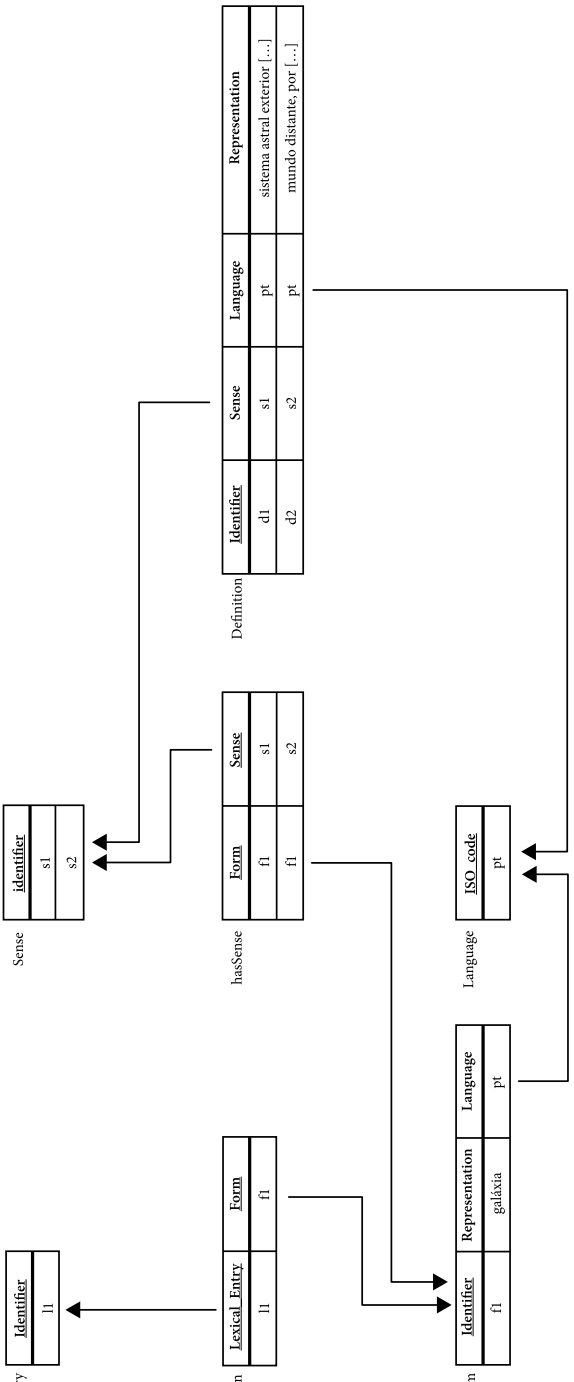

Figure 13. Tabular representation of Figure 9 (left) with data of Figure 10

Arrows indicate the referential integrity constraints between two tables. By referential integrity between two columns of two tables (for example, the 'Lexical_Entry' column of the "contain" table and the Identifier column of the "Lexical Entry" table), we refer to the constraint that the values contained in the referring column (in the example, the 'Lexical_Entry' column of the "contain" table) must be present in the referred column (in the example, the Identifier column of the "Lexical Entry" table).

To represent the data of the lexical entry in Figure 10, we have added a row with the value l1 in the "Lexical_Entry" table to represent this specific entry. The representation of the form *galáxia* has been identified in the "Form" table through the value f1 and associated with the respective entry in the 'contain' table. The language in which that form is expressed refers to the "Language" table through the identifier value of the Portuguese language (ISO code pt). The two senses of the form have been saved in the "Sense" table with identifiers s1 and s2, and their respective definitions, identified with the values d1 and d2, are saved in the 'Representation' field in the "Definition" table.

In Figure 14, we show the tabular representation of the ER schema of the LMF meta-model (left side of Figure 9) with the data presented in Figure 11. In this case, we have used the original identifiers of the entry and the two senses (i.e., ACL.DLP.VIRUS, ACL.DLP.VIRUS.sense.1, and ACL.DLP.VIRUS.sense.6). Each sense has been linked to the corresponding definitions, identified with the values d3 and d4.

In Figure 15, we present the tabular representation of the ER schema of the TMF meta-model (right side of Figure 9) with the data presented in Figure 11. Beginning with the "Concept_Entry", we have added a row with the value c1, referencing the concept already stored in the "Concept" table. The terms designating this concept are stored in the "Term" table with identifiers t1, t2, and t3. The terms t1 and t2 (that is *astronomical object* and *celestial object*) are expressed in English, while the term t3 (*oggetto celeste*) is verbalised in Italian. The two natural language definitions identified with d1, d2 are stored in the "Natural_Language_Definition" table with the corresponding languages.

These three examples give us the starting point to define the basic elements for the design and implementation of a unified model that allows for the homogeneous representation of both meta-models.

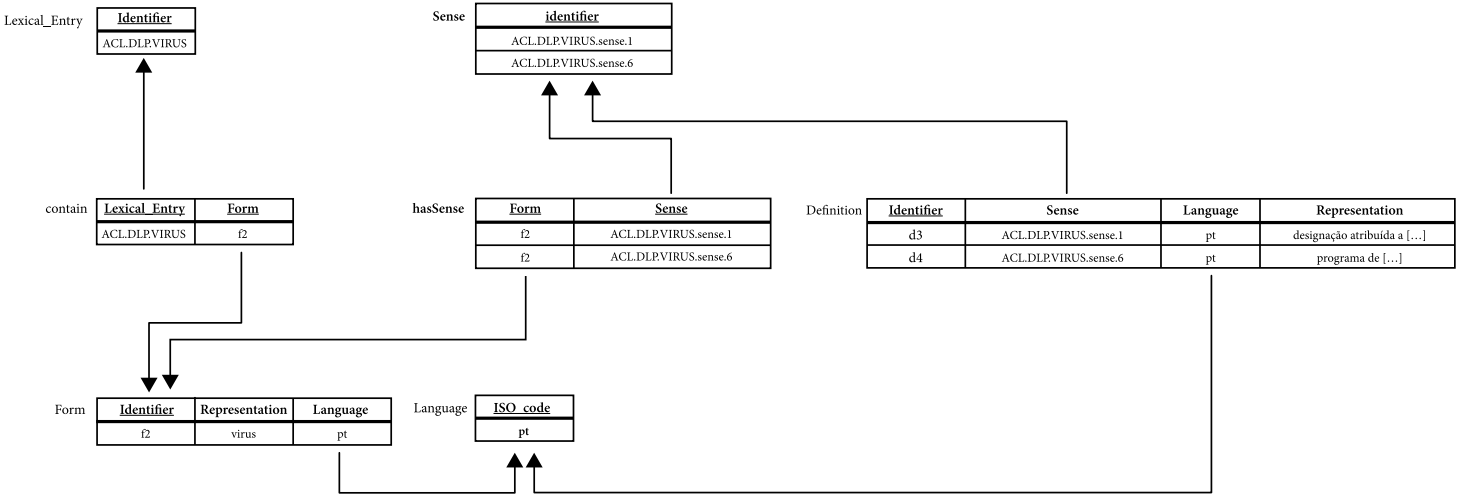

Figure 14. Tabular representation of Figure 9 (left) with data of Figure 11

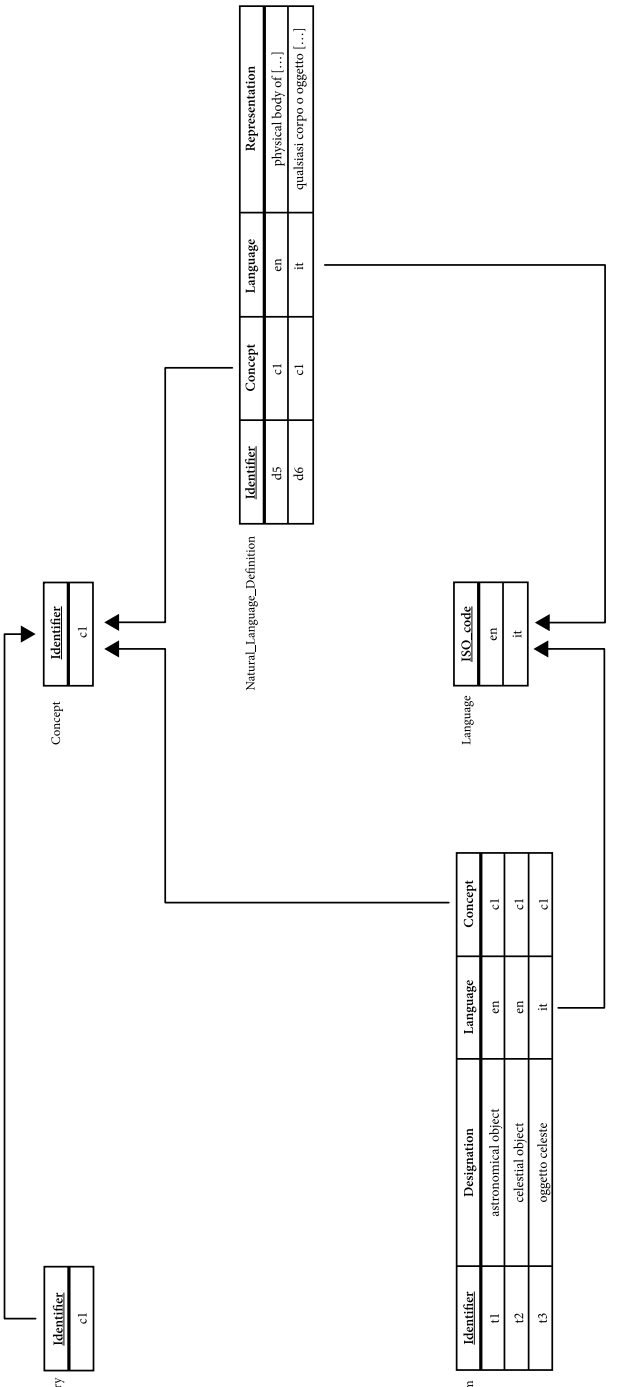

Figure 15. Tabular representation of Figure 9 (right) with data of Figure 12

4. Proposal for a Unified Markup Framework (UMF)

Based on the considerations conducted so far, in this section, we propose a unified structure that contains the basic elements (entities and relationships) necessary to represent both models in a single schema which we name Unified Markup Framework (UMF).

Figure 16 includes the common elements of LMF and TMF both at the macrostructural and microstructural levels. In particular, starting from the highest structural level, a "Resource" is conceptualised as an entity that includes at least one "Collection", which in turn contains at least one "Entry". The entity "Resource" in this case includes both lexical resource and termbase elements. The entity "Collection" includes, in turn, both lexicon elements and terminological data collection elements. Finally, the entity "Entry" groups both lexical entry and concept entry elements. At the macrostructural level of the UMF, we can thus observe that there is a perfect mapping of the shared entities between lexical resource and termbase.

Before proceeding with the description of the mapping at the microstructural level of the "Entry", it is necessary to make a premise about the two parts highlighted in Figure 16 with two different colours. In particular, as shown in Figure 9, the two entities "Sense" for the lexical entry and "Concept" for the concept entry are considered as two entities that are structurally at the same level, thus implying a direct mapping into a single entity.

Now, although this does not present particular problems from a structural point of view, it is necessary to pay attention to the impact that this overlap can have from a theoretical point of view. In this study, we distinguish between the concept designated by a term and the sense conveyed using that term. Adopting an approach that sees terminology as a discipline with a dual dimension of analysis — conceptual and linguistic — (Costa, 2013), the concept adheres to the principles of logic and is an extra-linguistic element intended as a unit of specialised knowledge shared by experts of the domain. The concept is, therefore, what one refers to using a specific term for a domain. The sense, on the other hand, lies rather in a linguistic dimension and concerns the meaning conveyed by the term. In our effort to map elements, we have chosen to merge the "Sense" and "Concept" entities within UMF into a unified entity called "Referent". This entity encompasses both concept-type and sense-type elements. Our goal is not to engage in theoretical debates; rather, we aim to be practical, and "Referent" appears suitable to us as it conveys the notion of something being referred to.

The second issue shown in Figure 16 with a darker colour concerns the two relationships 'contain' and 'hasConcept' involving the entity "Entry". Specifically, from the perspective of LMF, this entity is the entry point to access the "Form"

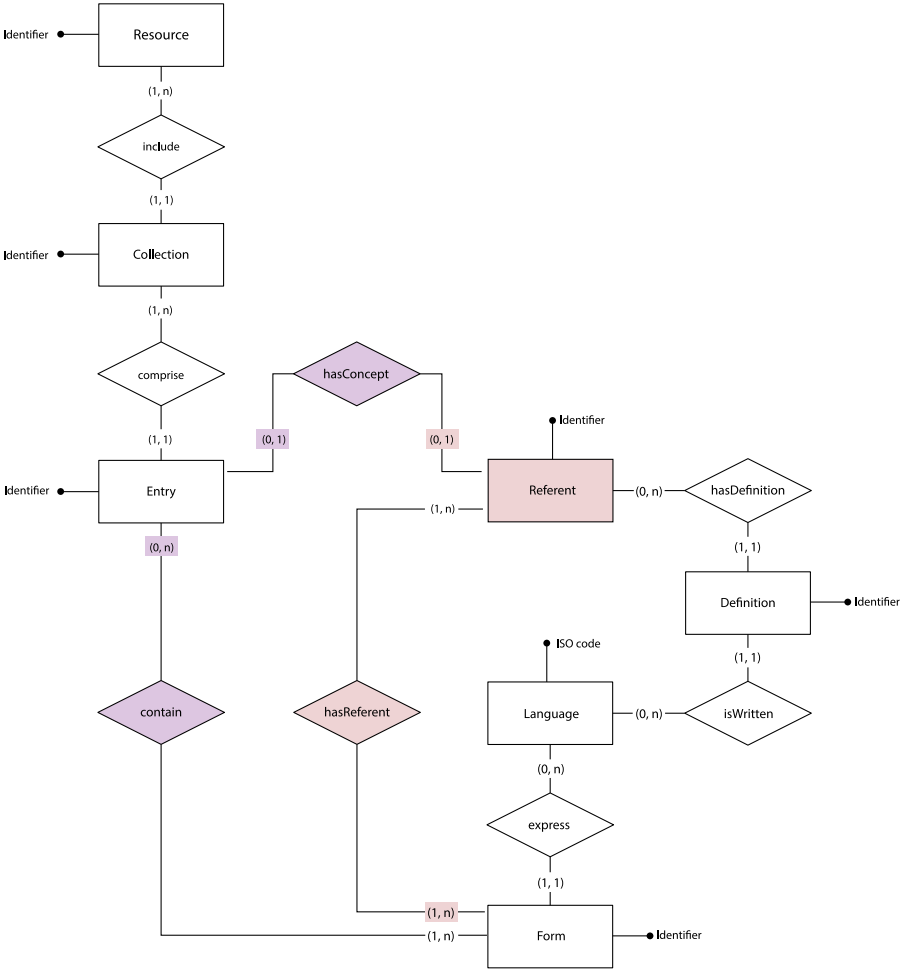

Figure 16. ER Schema of the core UMF model

through the 'contain' relationship, while for TMF, it is the entry point through the 'hasConcept' relationship to access the concept (now indicated with the entity "Referent"). For this reason, these two relationships are the only ones, in this core schema of UMF that we are proposing, to remain distinct and not to be overlapped into a single relationship since they necessarily indicate two different points of access to the information. In particular, an element of an "Entry" may contain forms only if it is a lexical-entry element while it will not have any associated form if it is a concept-entry element (cardinality (0,n)). Similarly, an element of an "Entry" may have a "Concept" associated only if it is a concept-entry element while it will not have any if it is a lexical-entry element (cardinality (0,1)). At the same time, an element of the entity "Referent" will be related to only one

"Entry" if that element is a concept, while it will not be associated with any entry if it is a sense-type element.

The choice to maintain the entity "Form" as an overlap between "Form" and "Term" is due to the fact that a term is effectively a form in the same way that a lemma is a form. Therefore, we did not consider it necessary to rename this entity and reuse the superclass "Form" for both the forms of a lexical entry and the terms that designate the concepts of an entry of type concept.

However, it is necessary to pay attention to the cardinality of the participation of "Form" in the 'hasReferent' relationship. In both cases (whether the sequence is a form in general or a term), a form must necessarily have at least one referent. In cases where it is a form of a lexical entry, there may also be multiple referents (and these referents will be only sense-type elements), while in the case of a term there can be at most one referent (and this referent will be only a concept-type element). This particular constraint cannot be expressed with a single pair of numbers in the ER schema, and it will need to be implemented at a lower level (for example, in the specification of the serialisation) with constraints (cardinality (1, n) or cardinality (1, 1)) that depend on the type of object (respectively a form or a term).

The entity "Definition" contains both the definition of a lexical entry, as the narrative description of a sense, and the natural language definition of a concept. In any case (lexical entry or concept entry), a definition must be related to only one referent and must be written in only one language (to this purpose, see also footnote 24).

In Figure 17, we present the tabular representation of the ER schema of the UMF meta-model shown in Figure 16 with the data presented in Figures 10, 11, and 12. This example shows how the core UMF model can store the data of both LMF and TMF serialisations. In this minimal example, we created three different resources: a lexical resource with identifier r1, a lexicographic resource with identifier r2, and a termbase with identifier r3. Each resource, r1, r2, and r3, includes one collection, respectively coll1, coll2, and coll3. Each collection comprises one entry. The remainder of the tables are basically the collation of all the data already shown in Figures 10, 11, and 12. In particular, the tables "Referent", "Form", "Language", and "Definition" perfectly match the content of the previously presented tables: "Sense" and "Concept", "Form" and "Term", "Language", "Definition" and "Natural Language Definition". The table "hasConcept" has been introduced (differently from the example in Figure 15) to mirror the other table "contain" so to have the information about the concept entries in the first, while the information about the lexical entries in the second. The table "hasReferent" contains both the information about the links between the forms and senses and the terms and con-

cepts. As we have discussed in the previous paragraph, the constraint that a term can only refer to one concept must be implemented at a lower level.

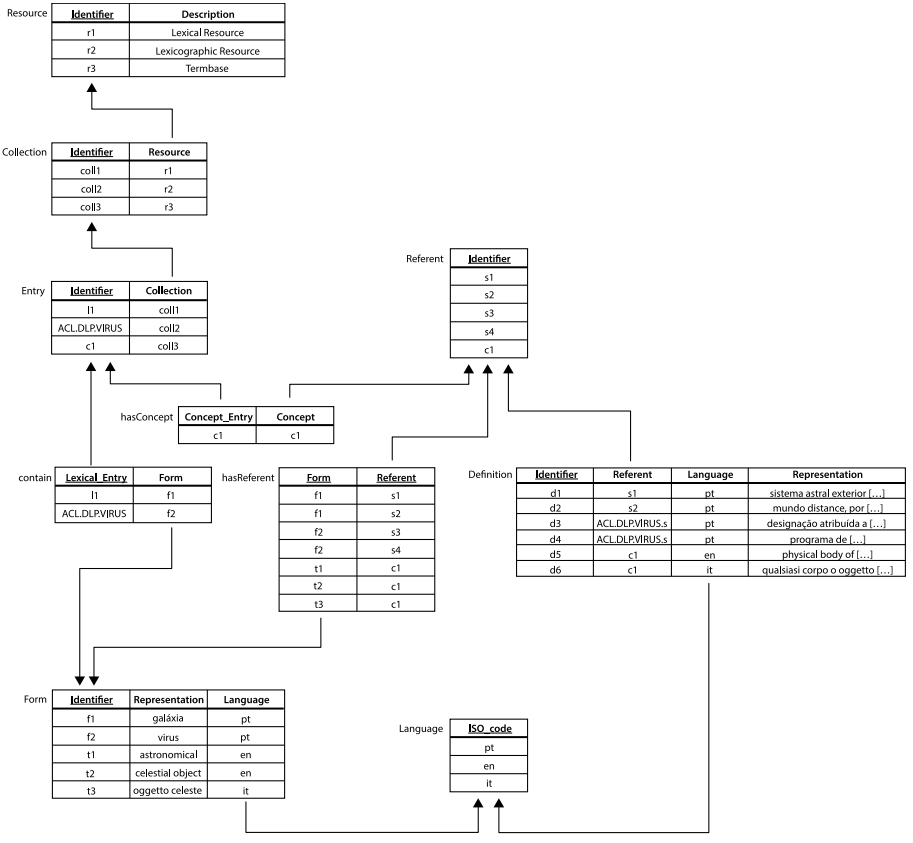

Figure 17. Tabular representation of Figure 16 with data of Figures 10, 11, and 12

This example allows us to verify in detail the possibility of building a unified model that takes into account both LMF and TMF. As already mentioned, the primary aim of this work is to identify the common structural elements between LMF and TMF, while deliberately avoiding deeper theoretical discussions, such as the distinction between sense and concept, which would lead the work to a different level of analysis and outside the proposed objective. Future extensions of the UMF will include mechanisms to handle this distinction more explicitly while preserving the benefits of a unified model.

Finally, another important element to underline is that in this structural core framework, we have deliberately not added fundamental elements for the respective meta-models (for example the orthographic representation for LMF or a data

category specification for TMF) both for a question of a cleaner presentation and for the fact that this specific blocks can be linked to their respective referenced tables as additional elements without modifying the central core.

5. Conclusions and future perspectives

This paper seeks to initiate a dialogue regarding the potential integration of the Lexical Markup Framework — ISO 24613-1: 2024, and the Terminological Markup Framework — ISO/WD 16642: 2024. When comparing the LMF and TMF standards, it is evident that they serve distinct purposes: while the former prioritises lexical resources with a lemma-oriented approach, the latter centres around the organisation and exchange of terminological data with a concept-oriented focus. Despite these significant differences, implying varying methodologies and theoretical frameworks, both standards share the common goal of supporting an efficient management of language resources and providing semantic interoperability within language technologies.

Recognizing the complementary nature of both standards, our intention was to conduct a comparative analysis to identify both their shared areas and their distinctive features with the aim of proposing a Unified Markup Framework (UMF). To conduct comparisons, we set the two standards side by side, focusing on structural analysis of their UML diagrams. These diagrams are crucial for conceptualising and visualising the structures and relationships within each framework. They offer a clear depiction of how language data is organised and managed within each system, forming the premise of our comparative analysis.

To progress our analyses, we chose to employ an ER schema diagram to emphasise the shared characteristics and unique attributes of both the macrostructure and microstructure found in lexicographic (LMF) and concept-based (TMF) entries. To demonstrate both the commonalities and differences, we provide three examples of XML-based serialisation for both lexicographic entries and concept entries. These examples provided us with a foundation to delineate the fundamental components necessary for designing and implementing a unified model capable of seamlessly representing both meta-models.

We finally came up with the ER Schema of the Core UMF model followed by its tabular representation to show how the identified core common elements of LMF and TMF can be represented and serialised with a unified model. Finally, we believe the meta-terminological aspect of the choice of marking names is not secondary. Given the nature of the work resulting from the reanalysis of the two ISO standards, we have chosen to end our work leaving the names of the entities

as proposals to be taken into consideration in a possible formulation of a unified standard at an international level.

Funding

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. This work is also part of the initiatives carried out by the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT. This work is also partially supported by the National Agency for Research–FCT (Foundation for Science and Technology)– within the framework of projects UIDB/LIN/03213/2020 and UIDP/ LIN/ 03213/2020 of the Centro de Linguística da Universidade NOVA de Lisboa (CLUNL).
This article was made Open Access under a CC BY-NC 4.0 license through payment of an APC by or on behalf of the authors.

References

Bellandi, Andrea, Giorgio Maria Di Nunzio, Silvia Piccini, and Federica Vezzani. 2023a. "From TBX to Ontolex Lemon: Issues and Desiderata." In *Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023)*, ed. by Giorgio Maria Di Nunzio, Rute Costa, and Federica Vezzani, Lisboa, Portugal: CEUR (CEUR Workshop Proceedings). Available at: https://ceur-ws.org/Vol-3427/#paper4

Bellandi, Andrea, Giorgio Maria Di Nunzio, Silvia Piccini, and Federica Vezzani. 2023b. "The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon." In *Proceedings of the 4th Conference on Language, Data and Knowledge. LDK 2023*, ed. by Sara Carvalho et al., 646–651, Vienna, Austria: NOVA CLUNL, Portugal. Available at: https://aclanthology.org/2023.ldk-1.70

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2018. "Models to represent linguistic linked data." *Natural Language Engineering* 24(6): 811–859.

Branco, António, Maria Eskevich, Francesca Frontini, Jan Hajič, Erhard Hinrichs, Franciska de Jong, Pawel Kamocki, Alexander König, Krister Lindén, Alexander Constanza Navarretta, Maciej Piasecki, Stelios Piperidis, Olli Pitkänen, Kiril Simov, Inguna Skadiņa, Thorsten Trippel, andreas Witt, and Claus Zinn. 2023. "The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond". *Language Resources and Evaluation*. Available at:

Caselli, Tommaso, and Johan Bos. 2023. "Investigating interoperable event corpora: Limitations of reusability of resources and portability of models". *Language Resources and Evaluation* 57(3): 1107–1137. Available at:

Chen, Peter Pin-Shan. 1976. "The entity-relationship model — toward a unified view of data." *ACM Transaction on Database Systems* 1(1): 9–36.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham, Switzerland: Springer International Publishing. Available at:

Costa, Rute. 2013. "Terminology and Specialised Lexicography: two complementary domains". *Lexicographica* 29: 29–42.

Eckle-Kohler, Judith, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. "UBY-LMF — A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), ed. by N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, 275–282. European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/475_Paper.pdf

Francopoulo, Gil (ed.). 2013. *LMF — Lexical Markup Framework*. London: ISTE/Wiley.

Francopoulo, Gil, Bel Nuria, George Monte, Calzolari Nicoletta, Monachini Monica, Pet Mandy, and Claudia Soria. 2006. "Lexical markup framework (LMF) for NLP multilingual resources." In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, ed. by Witt, A., Sérasset, G., Armstrong, S., Breen, J., Heid, U., Sasaki, F., 1–8. Stroudsburg, PA: Association for Computational Linguistics.

Gromann, Dagmar, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andriusm Utka, and Giedreh Valunaite Oleskeviciene. 2024. "Multilinguality and LLOD: A survey across linguistic description levels", *Semantic Web, Preprint (Preprint)*, 1–44. Available at:

ISO 1087. 2019. *Terminology Work and Terminology Science — Vocabulary — Part 1: Theory and application*. Geneva: International Organization for Standardization.

ISO 16642. 2017. *Computer Applications in Terminology — Terminological markup framework*. Geneva: International Organization for Standardization.

ISO 24613–1. 2024. *Language resource management — Lexical Markup Framework (LMF) — Part 1: Core model*. Geneva: International Organization for Standardization.

ISO 24613–2. 2020. *Language resource management — Lexical Markup Framework (LMF) — Part 2: Machine Readable Dictionary (MRD) model*. Geneva: International Organization for Standardization.

ISO 24613-3. 2021. *Language resource management — Lexical Markup Framework (LMF) — Part 3: Etymological Extension*. Geneva: International Organization for Standardization.

ISO 24613-4. 2021. *Language resource management — Lexical Markup Framework (LMF) — Part 4: TEI serialisation*. Geneva: International Organization for Standardization.

ISO 24613-5. 2022. *Language resource management — Lexical Markup Framework (LMF) — Part 5: Lexical base exchange (LBX) serialization*. Geneva: International Organization for Standardization.

ISO 30042. 2019. *Management of Terminology Resources — Term-Base eXchange (TBX)*. Geneva: International Organization for Standardization.

ISO 639. 2023. *Code for individual languages and language groups*. Geneva: International Organization for Standardization.

Pernes, Stefan, Laurent Romary, and Kara Warburton. 2017. "TBX in ODD: Schema-agnostic specification and documentation for TermBase eXchange", in *LOTKS 2017-Workshop on Language, Ontology, Terminology and Knowledge Structures*. Available at: https://inria.hal.science/hal-01581440/document

Piccini, Silvia, Federica Vezzani, and Andrea Bellandi. 2023. "TBX and "Lemon": What perspectives in terminology?". *Digital Scholarship in the Humanities*, *38* (Supplement_1): i61–i72.

Romary, Luarent, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, George Monte, Mandy Pet, and Piotr Banski. 2019. "LMF reloaded." In *Proceedings of the 13th International Conference of the Asian Association for Lexicography*, edited by Ahmet, M. G., Çiçekler, N. & Taşdemir, Y., 533–539. Istanbul: Istanbul University Department of Linguistics. https://cdn.istanbul.edu.tr/FileHandler2.ashx?f=asialex_proceedings.pdf

Silecchia, Sara, Federica Vezzani, and Giorgio Maria Di Nunzio. 2022. "Knowledge Representation and Language Simplification of Human Rights", in *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places. TERM 2022*, 8–12, Marseille, France: European Language Resources Association. Available at: https://aclanthology.org/2022.term-1.2

Vezzani, Federica. 2022. *Terminologie numérique: conception, représentation et gestion*. Bern: Peter Lang International Academic Publishers.

Vezzani, Federica, and Giorgio Maria Di Nunzio. 2020. "Methodology for the standardization 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.

Vezzani, Federica, Giorgio Maria Di Nunzio, and Rute Costa. 2023. "ISO standards for terminology resources management: Are they FAIR enough?". *Digital Translation* 10(2): 233–252.

Vezzani, Federica, Giorgio Maria Di Nunzio, and Sara Silecchia. 2022. "La fraseologia dei trattati internazionali di disarmo: la risorsa terminologica DITTO". *Umanistica Digitale* (14): 91–117.

Wilkinson, Mark, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, …Barend Mons. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, *3*(1).

Address for correspondence

Federica Vezzani
Department of Linguistics and Literary Studies
University of Padua
Via Elisabetta Vendramini, 13
35137 Padova
Italy
federica.vezzani@unipd.it
https://orcid.org/0000-0003-2240-6127

Co-author information

Giorgio Maria Di Nunzio
Department of Information Engineering
University of Padua
giorgiomaria.dinunzio@unipd.it
https://orcid.org/0000-0001-9709-6392

Ana Salgado
NOVA CLUNL, FCSH
Instituto de Lexicologia e Lexicografia da Academia das Ciências de Lisboa
anasalgado@fcsh.unl.pt
https://orcid.org/0000-0002-6670-3564

Rute Costa
NOVA CLUNL, FCSH
rute.costa@fcsh.unl.pt
https://orcid.org/0000-0002-3452-7228

Publication history

Date received: 25 April 2024
Date accepted: 12 July 2024