Bridging the Gap Between Ontologies and Dictionaries: Requirements and Implementation of a New Core for OntoLex-Lemon

John P. McCrae[1], Jorge Gracia[2], Fahad Khan[3], Philipp Cimiano[4]
[1] Research Ireland Insight and ADAPT Research Centres,
Data Science Institute, University of Galway, Ireland
[2] University of Zaragoza, Spain
[3] CNR-ILC, Italy
[4] Cognitive Interaction Technology Center (CITEC), Bielefeld University, Germany
john@mccr.ae, jogracia@unizar.es,
fahad.khan@ilc.cnr.it, cimiano@cit-ec.uni-bielefeld.de

Abstract

This paper presents the requirements and implementation details for a new core module of the OntoLex-Lemon model, representing the first major evolution of the de-facto standard since its 2016 release. While the original model successfully bridged ontologies and dictionaries through "semantics by reference," community adoption has identified critical gaps in handling lexicographic structures and retrodigitized resources. We detail a community-driven methodology that identified fifteen key requirements and we present a proposed architecture for the new OntoLex core, which integrates elements from the Lexicography module and addresses both semantic web and lexicography use cases. Further, we improve interoperability with standards like DMLex and TEI-Lex0 while maintaining strict backwards compatibility for existing users of the model.

Keywords: lexicography, linked data, ontologies, standardisation

1. Introduction

The OntoLex-Lemon Model (McCrae et al., 2017) has emerged as the de-facto standard for modelling dictionaries, lexical resources, terminologies and other resources as Linked Data. The official specification of the model was released in May 2016, almost 10 years ago at the time of writing, as a result of intense discussions of the W3C Ontology-Lexica Community[1] on "Lexicon Models for Ontologies". The model itself drew inspiration from previous models such as LMF (Francopoulo et al., 2006), LexOnto (Cimiano et al., 2007), LIR (Montiel-Ponsoda et al., 2011) and lemon (McCrae et al., 2011).

Since the emergence of the specification in 2016, the OntoLex-Lemon model has been widely used and adopted. As a result of this extensive usage, several gaps have been identified and new requirements have emerged. This paper summarises the key requirements that have been identified via community discussions. Current key limitations that have been identified include the inability of the current model to handle complex dictionary structures including nested entries, senses that can not be directly linked to one ontology element, as well as problems regarding the representation of retrodigitised dictionaries.

Building on top of this legacy, the W3C Ontology-Lexicon group is developing a new core module for the OntoLex-Lemon model that maintains backwards compatibility with existing resources published using OntoLex. This transition is driven by a community-led effort to bridge existing gaps, such as the need for more flexible dictionary structures and the representation of senses that do not align perfectly with ontological entities. It is important to note that the updates described in this paper, including the structural refinements and the integration of specialised modules into the core, represent a work-in-progress. These proposed changes are currently being evaluated through iterative discussions and remain subject to final approval by the OntoLex Community Group to ensure they meet the diverse needs of the Linguistic Linked Open Data community. The goal of this paper is twofold: to offer a comprehensive overview of the background and motivation of the new updated model and to serve as a reference to stimulate discussion in the Ontolex community.

The rest of this paper is structured as follows: Section 2 provides background on the existing OntoLex-Lemon model and its current ecosystem of modules. Section 3 details the community-driven methodology used to gather requirements and identify gaps in the current specification. Section 4 presents a comprehensive analysis of the fifteen key requirements identified during the consultation process, ranging from linking senses to forms to the representation of data quality. Section 5 describes the implementation of the proposed new OntoLex core, including the integration of lexicographic el-

[1] https://www.w3.org/community/ontolex/

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 69–79
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ements and structural refinements. Section 6 discusses the evolution of the model in comparison to other standards. Finally, Section 7 concludes the paper with a summary of benefits for dictionary creators and an outlook into next steps for adoption.

2. Background: The OntoLex-Lemon Model

As mentioned in the introduction, the OntoLex-Lemon model (McCrae et al., 2017) has successfully established itself as the de facto standard for publishing lexical data as linked data. Its primary objective is to bridge the gap between ontologies and dictionaries by providing a formal framework for representing lexical entries, their forms, and their corresponding meanings. The model is architected around a core module that defines the fundamental relationships between a *Lexical Entry*, its morphological *Form*s, and its *Lexical Sense*s, which are linked to ontology entities in a process called "semantic by reference." This core is further supported by a robust ecosystem of modules. In addition to the core module, the proposed update of which is the focus of this paper, the OntoLex-Lemon was released including the following modules:

Syntax and Semantics (SynSem): This module manages the mapping between syntactic frames and semantic arguments.

Decomposition (Decomp): This module provides the structure for modelling multi-word expressions and the decomposition of lexical entries into their constituent components.

Variation and Alignment (VarTrans): This module is used to describe lexical and semantic relations, such as synonymy or translations between different languages.

Metadata (Lime): This module provides a vocabulary for expressing the linguistic metadata of a dataset, such as its language coverage and the density of links (Fiorelli et al., 2015).

After the initial 2016 release of the Ontolex-Lemon model, the following modules have been developed and are released or nearing completion:

Lexicography (lexicog): This module was introduced to address the complexities of dictionary structures, specifically handling the nesting of entries and the specific ordering requirements of senses and lexical entries (Bosque-Gil et al., 2017).

Frequency, Attestation and Corpus (FrAC): Developed to provide mechanisms for linking lexical data to corpus observations, this module supports the inclusion of frequency data and citation mechanisms for corpus-based lexicography (Chiarcos et al., 2022a).

Morphology (morph): This module provides a more granular framework for representing the internal structure of words and complex morphological patterns (Klimek et al., 2019; Chiarcos et al., 2022c,b).

In addition, the current OntoLex-Lemon ecosystem relies on LexInfo (Cimiano et al., 2011) to provide the necessary data categories, such as part-of-speech values (e.g., `lexinfo:noun`), which are typically mapped to standardised URIs. It is not a requirement for datasets to use LexInfo alongside OntoLex, and some implementations, such as the Global WordNet Association formats (McCrae et al., 2021; Bond et al., 2016) use different models. The LexInfo ontology is created as an open-source repository, where the main elements, such as catalogues of part-of-speech values, are editable as CSV files on the GitHub repository. This allows the community to manage and propose changes to extend the set of categories in this model easily.

The adoption of OntoLex-Lemon has moved beyond academic frameworks to become the backbone of several large-scale, widely-used lexical resources. Notably, **Wikidata** has integrated the core classes of the model into its Wikibase ontology (Lindemann et al., 2023) to represent its Lexeme entity type, making it perhaps the largest existing deployment of OntoLex-Lemon with millions of entries across hundreds of languages. Similarly, **BabelNet** (Navigli and Ponzetto, 2012) utilised the model to publish its multilingual semantic network as linked data (Ehrmann et al., 2014), effectively bridging the gap between encyclopedic and lexicographic knowledge. In the domain of computational lexicons, the **Open English WordNet** (McCrae et al., 2019, 2020b,a) and various initiatives within the **Open Multilingual Wordnet** (Bond and Foster, 2013) ecosystem have adopted the model to enrich traditional synset-based structures with granular morphological and syntactic descriptions.

Beyond general-purpose lexicons, OntoLex-Lemon has seen significant application in the field of terminology. There is an ongoing shift from traditional XML-based standards, such as **TBX** (Term Base eXchange), towards RDF representations to improve interoperability. This is exemplified by the **IATE** (Interactive Terminology for Europe) database, where research has focused on converting its complex terminological entries into OntoLex-Lemon to allow for better integration with other Linguistic Linked Open Data (LLOD) resources (Martín-Chozas et al., 2025; Ibarbia et al., 2025). These applications demonstrate the model's versatility in supporting both the "concept-centric" view of terminology and the "lemma-centric" view

of traditional lexicography.

3. Methodology

The development of the new OntoLex core followed a community-driven approach aimed at identifying the gaps in the existing model and gathering evolving requirements from various stakeholders. The process began with an open call for requirements distributed via the W3C Ontology-Lexicon Community Group mailing list. This initial consultation allowed the community to highlight specific limitations encountered during the implementation of the original model in diverse projects.

To ensure a comprehensive analysis, three specialised subgroups were formed to focus on key areas of lexical representation:

1. **Lexicography:** Focusing on the requirements of professional and historical dictionaries, specifically regarding entry nesting and the representation of complex senses.

2. **Terminology:** Addressing the needs of terminological resources and the alignment with standards such as TBX and IATE.

3. **Relation to other models:** Investigating the interoperability and alignment between OntoLex-Lemon and emerging standards, most notably the ISO/TC 37 LMF standard, the *DMLex* (Data Model for Lexicography) and TEI Lex-0.

The findings and requirements from these subgroups were systematically reported and aggregated. These requirements were then subjected to rigorous discussion during a series of regular teleconferences, where members of the community evaluated the proposed changes. This iterative process ensured that the resulting updates to the core model were both technically sound and representative of the needs of the broader Linguistic Linked Open Data (LLOD) community.

4. Requirement Analysis

The main requirements that were obtained from the community consultation procedure are summarised along with the actions proposed in Table 1 and are described in more detail as follows:

Sec.	Requirement	Status
1	Linking Senses to Forms	*Import*
2	Multiple POS Values	*New Modelling*
3	Usage Examples	*Import*
4	Diachronic/Diatopic Links	*No Change*
5	Ordering	*Import*
6	Senses w/o Ontologies	*Axiomatic Change*
7	Definitions	*New Modelling*
8	Cross-references	*No Change*
9	Literal POS Values	*New Modelling*
10	Usage Notes	*Axiomatic Change*
11	Sources	*No Change*
12	Reliability and Status	*New Modelling*
13	POS Property	*New Modelling*
14	Inflected Form	*New Modelling*
15	Module Integration	*Partial Move*

Table 1: Summary of the requirements and the proposed solution. Solutions involved either introducing new modelling, importing modelling from a module, changing the axioms of existing concepts, partial move or no changes.

4.1. Linking Senses to Forms

The first requirement addresses the necessity of associating a lexical sense with a specific grammatical form or colligation of a lexical entry rather than the entry as a whole. While this need is present in general computational lexicons, it is particularly prevalent in lexicographic resources where certain meanings are restricted to specific inflections, such as the plural form “airs” signifying a condescending manner or “games” in specific athletic contexts. Historically, this has been addressed in the OntoLex Lexicography (`lexicog`) module through the `FormRestriction` class, which allows for the explicit narrowing of a sense to a particular form. Current discussions for the new OntoLex core centre on whether to migrate this functionality into the core module to provide a more direct link, similar to the “subject lexeme form” property used in Wikidata or to continue relying on property-based restrictions. This requirement is fundamental for accurately representing “pluralia tantum” or senses tied to suppletive forms (e.g., *cow* vs *cattle*) where the semantic value is inseparable from the morphological realization.

4.2. Entries with Multiple Part-of-Speech Values

This requirement was identified by multiple initiatives and addresses a structural limitation in current lexical models where a single entry is often restricted to a single part-of-speech (POS). In traditional and retrodigitised dictionaries, it is common for one headword to encompass multiple grammat-

ical roles[2]; forcing the creation of separate lexical entries for each POS creates a disparity with the original source and complicates the modelling of languages like Basque, where nominals may function as both nouns and adjectives with identical morphological behaviour. However, interaction with other modules around syntax and morphology requires that a lexical entry has a single part-of-speech value, as the morphology or frame semantics of a word are usually different if it is a member of a different part of speech. To resolve this, the proposed OntoLex core architecture will introduce an `Entry` superclass, previously proposed as part of the OntoLex lexicography module, that can act as a container for multiple `LexicalEntry` components or support more generalised grammatical categories. This solution effectively bridges the gap between the computational need for strict POS tagging and the lexicographic reality of multi-functional headwords, while also clarifying the distinction between a high-level dictionary `Entry` and a specific `LexicalEntry`.

4.3. Usage Examples

Lexicographic models have a fundamental need to include usage examples, which are ubiquitous in both modern and legacy, retrodigitised lexicographic resources. While current implementations often attach examples exclusively to a specific sense, the discussion for the new OntoLex core has brought up the need for a more flexible approach where examples can be associated at both the entry and sense levels. Furthermore, this model supports many-to-many relationships, enabling a single example to illustrate multiple senses or even different entries simultaneously through linking properties similar to Wikidata's "subject sense"[3].

4.4. Diachronic and Diatopic Links

A requirement on the necessity of representing regional (diatopic) and historical (diachronic) variations, such as the differing definitions of *fanny* in UK versus US English, was raised. While this is a common feature in lexicographic resources, the consensus, which has arisen for discussion of the issues through teleconferences and on GitHub, is that the existing OntoLex-Lemon model already provides the necessary mechanisms for this[4].

4.5. Sense Ordering and Lexical Entry Ordering

Many dictionaries and lexical resources order senses by frequency, historical precedence, or other criteria. An important requirement for the OntoLex-Lemon model is thus to allow for ordering senses and lexical entries[5], a feature which was in large part implemented in the `lexicog` module. The community has proposed to move this feature from the `lexicog` module into the core module.

4.6. Senses without Ontologies

"Semantics by reference" was raised as a limitation in traditional and retrodigitised lexicography, where a single dictionary sense may not correspond to a single, clearly defined ontological concept[6]. To resolve this, the proposed evolution of the new OntoLex core involves relaxing the strict modelling constraints, specifically by removing the axiom that requires every `LexicalSense` to have exactly one reference to an `rdfs:Resource`. In particular, it has been suggested that the following axiom could be removed from the core model:

$$\text{LexicalSense} \sqsubseteq 1\ \text{reference.Resource}$$

4.7. Definitions

A need for more robust representation of definitions in lexicographic and terminological resources has been identified. The current OntoLex-Lemon specifications only give explicit guidance in cases where a dictionary sense corresponds to a lexical concept (the latter being a kind of `skos:Concept`): in which case, the use of `skos:definition` to link a lexical concept to a gloss is proposed. However, as we saw in the discussion of sense by reference in Section 4.6, this may not always be the case. Furthermore, a single string literal is often insufficient to capture complex metadata such as definition references, provenance, or internal notes. To resolve this, the new OntoLex core module proposes the use of reified definitions, treating a definition as a resource rather than a simple literal . This allows the textual content to be stored in an `rdf:value` property while supporting additional properties for source citations or semantic links, ensuring that the model can handle the high granularity required for

[2] One obvious example here is the word *youth* which in many languages is frequently listed both as noun and adjective under one common dictionary entry, e.g., this is often the case for Romance languages: *giovane*/*jovem*/*joven*.

[3] https://www.wikidata.org/wiki/Property:P6072

[4] However, there is a necessity for better documentation here to help users understand how they can do this via e.g., language tags.

[5] This might be as simple as a metadata statement which gives the kind of ordering which has been adopted for the dictionary.

[6] For instance, this is very clearly the case with conjunctions such as *that*, but it also causes problems for retrodigitised or philological dictionaries where what is listed as a single sense might not correspond to a single neatly defined concept.

professional lexicography and large-scale terminological databases such as IATE[7].

4.8. Cross-references

The community has raised the need to cross-reference other entries in a dictionary entry. This is particularly the case for entries that primarily serve to point to other headwords. The current consensus suggests that these relationships can be effectively modelled using existing properties like `rdfs:seeAlso` or by extending LexInfo with specific lexicographic properties together with better documentation, rather than requiring structural changes to the OntoLex core.

4.9. Literal Part-of-Speech Values

This requirement highlights a critical need in the digitization of historical and printed dictionaries: the ability to preserve the exact wording or visual representation of grammatical information as it appears in the original source. While computational models typically map parts of speech to standardised categories (such as `lexinfo:noun`), retrodigitised resources often require the retention of the specific string used by the original lexicographer, such as "Substantiv" or "n. f.". To address this requirement, the new OntoLex core will introduce a property like `ontolex:partOfSpeechString`, which allows for a literal representation of grammatical data alongside the standardised URI. However, it is unclear if this need is general enough to justify an extension to the core model or whether this should be handled by introducing a specific property into the LexInfo model.

4.10. Usage Notes

This requirement, which was obtained from multiple sources, addresses the necessity of including usage notes, recommendations, and domain-specific data, which are essential components of authoritative terminology records and traditional dictionaries. Such notes often exceed simple literal descriptions, requiring a combination of textual recommendations and links to external resources or provenance information via standard vocabularies such as PROV-O (the Provenance Ontology)[8]. Current modelling in OntoLex-Lemon is limited by the fact that the `ontolex:usage` property is restricted to the domain of `LexicalSense`. To provide the flexibility required for professional lexicography and terminology, such as applying "archaic" or "dialectal" labels to specific forms, morphemes, or entire entries, the proposed update for OntoLex involves removing these domain restrictions. This allows `ontolex:usage` to serve as a versatile property for reified usage nodes that can capture complex metadata, ensuring that language professionals can model nuanced usage recommendations across all levels of a lexical resource.

4.11. Sources

One requirement underscores the necessity of maintaining traceability across lexical and terminological resources, particularly when automated processes or multiple contributors are involved. OntoLex will emphasise distinguishing between "Original Sources" (such as corpora or specific individuals) and "Intermediate Sources" (such as information providers like IATE). While existing modules, such as FraC, provide citation mechanisms for corpus observations, they do not fully cover the metadata needs of notes and definitions. The current consensus suggests that PROV-O is sufficient for this purpose, provided that the elements, such as a `ConceptDefinition`, are also typed as `prov:Entity`.

4.12. Reliability and Record Status

Two requirements address the qualitative metadata of an entry, specifically its *Reliability* and *Record Status*. *Reliability* refers to the confidence rating terminologists assign to a term, often following a standardised system like the IATE four-star scale. While `lexinfo:confidence` currently exists, there is a proposed move toward a more standardised `ontolex:confidence` property, potentially supported by a "proxy" module for catalogues of values to avoid overloading the core ontology. The closely related requirement on *record status* indicates whether an entry is "finalized", "provisional", or "superseded". Although some aspects of status, such as being "superseded", can be handled via `dc:isReplacedBy` or PROV-O, the community is exploring a unified categorization that aligns reliability and status to distinguish between validated data and data that is work-in-progress.

4.13. Part-of-Speech Property

This requirement addresses the absence of a core part-of-speech property in the current OntoLex model, a feature that is already present in related standards such as DMLex. To enhance the model's utility for defining lexical entries, there is a proposal to migrate the widely-used `lexinfo:partOfSpeech` property directly into the OntoLex core as `ontolex:partOfSpeech`, potentially maintaining compatibility through an `owl:equivalentProperty` axiom. This move would support the introduction of formal axioms,

[7] https://iate.europa.eu/home
[8] https://www.w3.org/TR/prov-o/

e.g., requiring a `LexicalEntry` to have exactly one part-of-speech, and opens the door to incorporating standardised values like Universal POS Tags to improve cross-linguistic interoperability.

4.14. Inflected Form

This requirement regards the necessity of introducing a specific `inflectedForm` property to distinguish morphological inflections from other non-lemma forms. While the current OntoLex model provides `ontolex:otherForm`, this property is often viewed as having a broader interpretation that encompasses various types of non-canonical forms.

Take, for example, the Old English verb *bregdan* meaning, among other things, 'to move quickly' and 'to pull'. This lemma form (the infinitive) has a variant form *brægdan*. It also has inflected forms such as *bregdeþ* (third person singular indicative present) and *brægd* (third person singular indicative preterite). Interestingly, both of these two forms also have variants that occupy the same cell in the inflectional table for the verb (i.e., *brēt*, *brīt* for the former and *brǣd* for the latter). The definition of form in the current guidelines is as follows[9]:

> A form represents one grammatical realization of a lexical entry.

However in our example these forms encode both separate grammatical *and* phonetic variations (the result of different processes) as well as different corpus distributions. It isn't clear how we could model this properly, taking into consideration the correspondences between grammatical, phonetic and other properties in the case of different forms, using the current guidelines.

Among other things, the discussion for the new version weighs the benefit of adding a dedicated `inflectedForm` subproperty (something which could assist in clarifying situations such as those described in the preceding paragraph) against the potential increase in modelling complexity. A key consideration for this requirement is identifying specific use cases where a formal distinction between inflections and other lexical variations is essential for computational or lexicographic accuracy.

4.15. Integration of Modules and the Evolution of Lexicography Module

The original release of the OntoLex-Lemon specification introduced a modular architecture where a central core was accompanied by four initial modules (SynSem, Decomp, VarTrans, and Lime). Subsequent developments to the ecosystem, such as the *lexicog* and *FrAC* modules, were released as independent modules with a separate documentation to address specialised needs. However, the community consultation for the new OntoLex core module revealed that the current fragmentation makes it hard for newcomers to understand the model, leading to a requirement for a more integrated core.

The main focus here is the **lexicog** module, which was originally designed to handle the structural complexities of lexicography. To determine how key features of the **lexicog** module should be integrated into the core module, three distinct strategies were considered:

Option 1: Full Absorption: This approach would involve migrating all classes and properties from the `lexicog` namespace to the `ontolex` namespace, effectively merging the two specifications into a single document.

Option 2: Lexicog as a Core Module: Following this approach, *lexicog* would be maintained as a distinct section within the main specification, but all elements would retain their original URL in the `lexicog` namespace to maintain backward compatibility.

Option 3: Partial Move: This strategy involves a selective promotion of elements. High-utility components currently in the module would move to the `ontolex` core, while highly specialised lexicographic structures would remain in a separate module.

The community has expressed a strong preference for **Option 3 (Partial Move)**. This approach allows for an extended, more powerful core by promoting elements that have proven universally useful, such as `UsageExample` and `FormRestriction`, while keeping the core lean by leaving niche lexicographic structures (like complex entry nesting) within a dedicated module. This ensures that the model remains accessible to general users while providing the necessary depth for professional lexicographers.

5. Implementation

The implementation of the new OntoLex Core Module involves an iterative improvement of the model in a fully backwards-compatible manner to ensure that the model satisfies the requirements discovered in over ten years of deployment of the model across a wide range of applications. These changes are depicted in Figure 1 and are proposed changes subject to approval by the community.

[9] https://www.w3.org/2016/05/ontolex/#forms

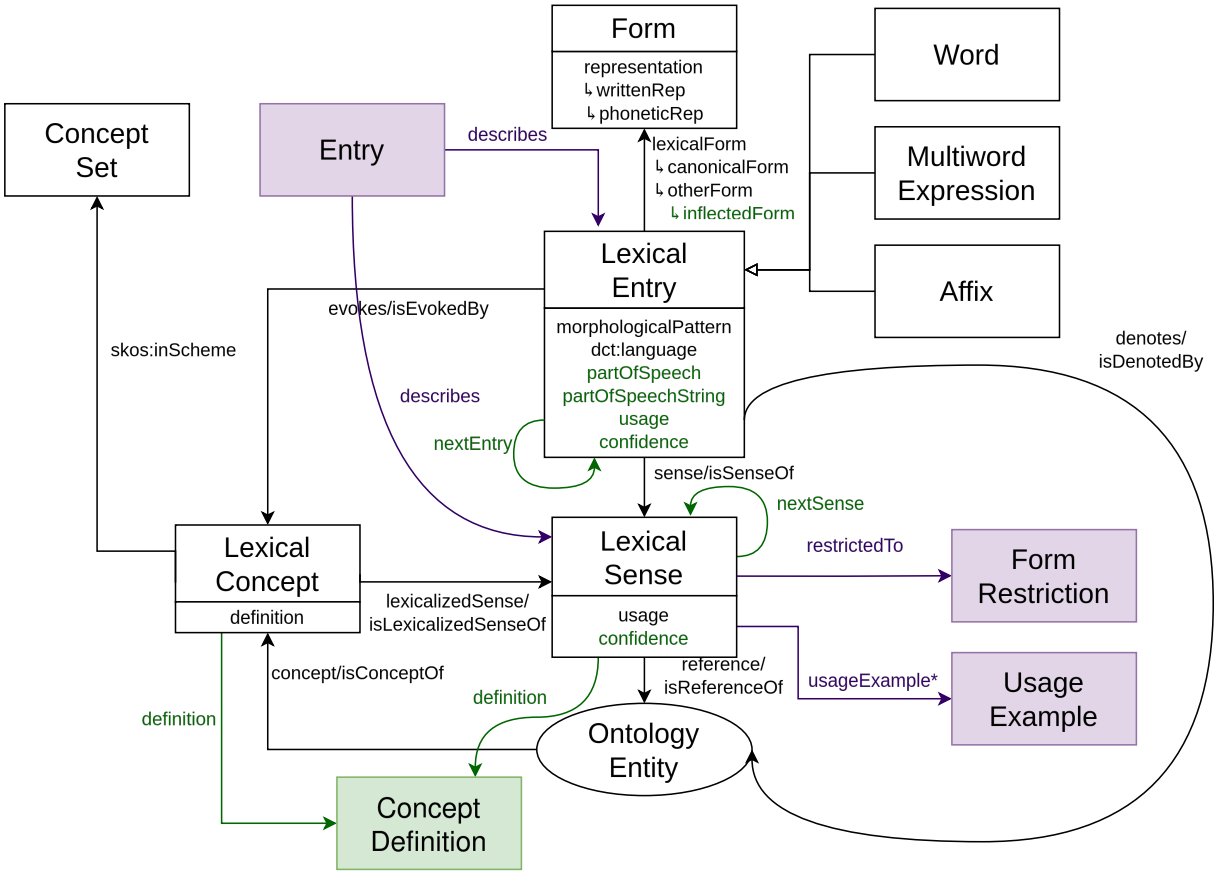

Figure 1: The proposed updated core diagram of OntoLex-Lemon. The green elements indicate new modelling to be added to the core model. The purple elements indicate modelling from modules to be promoted into the core.

5.1. Integration of Lexicographic Elements

Following the decision to adopt a hybrid integration strategy, several key classes and properties have been promoted from the Lexicography module into the OntoLex Core Module. This move enables the representation of complex dictionary structures without requiring the overhead of external modules.

Entry The introduction of the `Entry` class as a broader concept alongside the existing `LexicalEntry` allows for a more flexible grouping of lexical data, particularly for retrodigitised dictionaries where a single headword may describe multiple `LexicalEntry` instances (e.g., across different parts of speech). The `LexicalEntry` class is kept with a narrower interpretation in order to maintain interoperability as well as backwards compatibility with other modules.

Form Restrictions These elements allow a `LexicalSense` to be explicitly linked to a specific `Form`. This is essential for modelling cases where a meaning is only valid for a specific inflection (e.g., the plural “airs”).

Usage Examples Formerly part of the lexicography module, this class has been moved into the core in order to allow for the reified representation of usage examples and citations. The original usage example property was called `usageExample`, however, we will rename this to just `example` to avoid confusion between property names and support the specification in line with W3C guidelines.

To ensure backwards compatibility, previous URLs will still exist, but will have axioms such as `owl:sameAs` or `owl:equivalentProperty` to their new URLs.

5.2. New Properties and Structural Refinements

To enhance the granularity of the model and its alignment with standards like DMLex, the following properties and classes have been proposed:

Linguistic Metadata: New properties, including `partOfSpeech` and `partOfSpeechString` allow for both URI-based (LexInfo) and literal-based grammatical tagging. The

`inflectedForm` property provides a direct way to identify non-canonical forms.

Ordering: To support the sequential nature of dictionaries, `nextEntry` and `nextSense` have been introduced.

Data Quality: A `confidence` property allows for the representation of uncertainty, which is frequent in automatically generated or OCR-derived lexicons.

Definitions: A new class, `ConceptDefinition`, has been introduced to allow definitions to be treated as first-class objects, enabling them to be shared across multiple senses or concepts.

As these changes are all additive, they will not affect the backwards compatibility of the model.

5.3. Axiomatic and Domain Changes

In addition to new entities, the new core relaxes several constraints of the original model. The domain of the `usage` property has been extended to include `LexicalEntry`, allowing for broader labels (e.g., register or dialect) to be applied at the entry level. Most significantly, the strict axiom on `LexicalSense`, which previously required a reference to an external ontology entity, has been removed. This allows for "non-ontological" senses, enabling the publication of dictionaries where meanings are expressed solely through definitions or examples without the need for a formal URI reference. As these changes only relax existing constraints, this is a non-breaking change. Existing data that does use references remains perfectly valid under the new, more permissive logic.

6. Discussion

6.1. Comparison with Existing Standards

The development of the new OntoLex core has been heavily informed by a comparative analysis with other prominent lexical and lexicographic standards, ensuring that the model remains a robust bridge between the Semantic Web and traditional linguistics.

- **DMLex (Data Model for Lexicography):** As an OASIS standard, *DMLex* (Filip et al., 2024) provides a functional, abstract model for dictionary data. While DMLex is focused on the data structures required for dictionary management systems, the new OntoLex core acts as its realization in the Linked Data space. The introduction of more general entries, specific part-of-speech properties and definitions brings OntoLex closer to the DMLex core model, allowing resources to be more easily converted between these two formats. Noticeably, DMLex has a module for the modelling of etymology and this spurs the development of a potential new module for OntoLex to enable further integration of these two models.
- **TEI-Lex0:** Traditionally, the TEI (Text Encoding Initiative) guidelines, specifically the *TEI-Lex0* customization, have been the standard for the digital representation of dictionaries as documents. While TEI-Lex0 excels at capturing the layout and textual details of a source, OntoLex-Lemon focuses on the data's semantic interoperability. The new core module will reduce the friction involved in converting TEI-encoded resources into RDF by promoting features like `UsageExample` and `partOfSpeechString`, which are more directly compatible with the string-heavy metadata found in TEI headers.
- **LMF (Lexical Markup Framework):** Compared to the closed ISO (LMF) standard[10] (Romary et al., 2019), OntoLex maintains an open web-centric approach. While LMF utilises a data-category-based model that often relies on complex XML structures[11], OntoLex leverages the inherent graph nature of RDF to provide more flexible linking between senses and specific forms, a requirement that has historically been difficult to model succinctly in LMF without the need for ad hoc extensions.

6.2. Evolution of the Specification

The development of this new core module reflects a shift in the community's priorities from ontology-based modelling, as exemplified by the "semantics by reference" principle, toward "lexicographer-friendly" modelling. The original specification was highly successful in linking lexicons to formal ontologies, but it created a high entry barrier for publishers of legacy dictionaries who did not have (or need) a corresponding OWL ontology for every lexical sense. The process of evolving this specification

[10] Originally published as a single standard, ISO 24613, LMF was subsequently released as a multipart standard, including a core model ISO 24613-1:2024; a machine-readable dictionary module ISO 24613-2:2020; an etymological extension, ISO 24613-3:2021, a TEI serialisation, ISO 24613-4:2021; another serialisation in the obscure Lexical base exchange (LBX) format ISO 24613-5:2022; and finally a Syntax-Semantics module, ISO 24613-6:2024. While the original LMF standard was adopted in a number of projects and for encoding several lexicons, we are not aware of any lexicons that have been developed using the new version of LMF.

[11] The 'official' XML serialisation of LMF is in TEI-XML.

through the formation of specialised subgroups allowed us to collect feedback from a wide range of users of the models. By identifying which elements were widely used, such as reified definitions, and integrating the lexicography module more tightly with the core, we have created a model that is both more powerful and easier to adopt. This evolution demonstrates how a standard can remain viable and balance the formal world of ontologies with the needs of lexicographers. The community is discussing using SHACL (Knublauch and Kontokostas, 2017) to further improve the validation of implementations of the OntoLex-Lemon model.

7. Conclusion

The development of the new core module will further extend on the success of OntoLex as a bridge between formal ontologies and the practical needs of lexicographers. Our community-driven approach allows us to address long-standing limitations such as the handling of complex dictionary structures, multi-functional headwords, and the necessity of preserving original source metadata. Key refinements in this module promote highly useful modelling into the core, while more flexible semantics will further increase the number of use cases that OntoLex can satisfy. However, mostly, new modelling, such as for part-of-speech strings, inflected forms, and reified definitions ensure that OntoLex can capture the nuanced data required for authoritative terminological records and retrodigitised resources. Ultimately, the new OntoLex core balances the formal rigour of the Semantic Web with a "lexicographer-friendly" architecture.

This new core module will be published in accordance with the public review procedures of the OntoLex W3C Community Group and is subject to approval and comments by the full community. This acts as a basis to allow OntoLex to develop further, in particular through the development of new modules and the completion of modules to enable OntoLex to be a flexible and forward-looking standard.

Acknowledgements

This publication is based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

Bibliographical References

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1352–1362. The Association for Computer Linguistics.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference 2016*.

Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. Towards a module for lexicography in OntoLex. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017*, volume 1899 of *CEUR Workshop Proceedings*, pages 74–84. CEUR-WS.org.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Unifying morphology resources with OntoLex-morph. a case study in German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille, France. European Language Resources Association.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational morphology with OntoLex-morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86, Marseille, France. European Language Resources Association.

P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. 2007. LexOnto: A model for ontology lexicons for ontology-based NLP. In *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07*.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 401–408. European Language Resources Association (ELRA).

David Filip, Miloš Jakubíček, Simon Krek, John McCrae, and Michal Měchura. 2024. Data Model for Lexicography (DMLex) Version 1.0. OASIS committee specification draft 02, OASIS.

Manuel Fiorelli, Armando Stellato, John P. McCrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. LIME: the metadata module for OntoLex. In *Proceedings of 12th Extended Semantic Web Conference*.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Paula Diez Ibarbia, Patricia Martín-Chozas, and Elena Montiel-Ponsoda. 2025. Bringing IATE into the semantic web family. In *Proceedings of the 5th Conference on Language, Data and Knowledge: The 5th OntoLex Workshop*, pages 12–17.

Bettina Klimek, John McCrae, Maxim Ionov, James K. Tauber, Christian Chiarcos, Julia Bosque-Gil, and Paul Buitelaar. 2019. The OntoLex-lemon morphology module. In *Proceedings of the Sixth Biennial Conference on Electronic Lexicography (eLex 2019)*, Sintra, Portugal.

Holger Knublauch and Dimitris Kontokostas. 2017. Shapes Constraint Language (SHACL). W3C recommendation, W3C.

David Lindemann, Sina Ahmadi, Anas Fahad Khan, Francesco Mambrini, Federica Iurescia, and Marco Carlo Passarotti. 2023. When OntoLex meets Wikibase: Remodeling use cases. In *Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 13, 2023*, volume 3640 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Patricia Martín-Chozas, Thierry Declerck, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. 2025. Representing terminological data in the semantic web: A proposal based on OntoLex-lemon. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 31(2):171–207.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The GlobalWordNet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

John P. McCrae, Ewa Rudnicka, and Francis Bond. 2020a. English WordNet: A new open-source WordNet for English. *K Lexical News*, (28):37–44.

John P. McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020b. English WordNet 2020: Improving and extending a wordnet for english using an open-source methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*, pages 14–19.

Elena Montiel-Ponsoda, Guadalupe Aguado de Cea, Asunción Gómez-Pérez, and Wim Peters. 2011. Enriching ontologies with multilingual information. *Nat. Lang. Eng.*, 17(3):283–309.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. LMF reloaded. In *Proceedings of the 13th Conference of the Asian Association for Lexicography*.