# TermiGraph

Terminología en RDF

# DOCUMENTACIÓN TERMIGRAPH:

# CONVERSOR DE GLOSARIOS MONOLINGÜES

**29 de octubre de 2025**

TeresIA

Ontology Engineering Group

# Índice

# 1. INTRODUCCIÓN

El presente documento se centra en explicar el uso y el funcionamiento de uno de los conversores de TermiGraph. En particular, este conversor se ha desarrollado para los glosarios monolingües. Este servicio se ha desarrollado en el contexto del proyecto nacional TeresIA que tiene como objetivo ofrecer un punto centralizado de acceso a terminologías.

El conversor TermiGraph transforma los archivos de entrada a RDF con el modelo de Ontolex-lemon. El RDF y los datos enlazados proporcionan un marco común para describir y vincular datos en la Web, haciendo posible la interoperabilidad y el intercambio significativo de información entre diferentes sistemas.

En el contexto de TeresIA, esta conversión es necesaria para poder incorporar los datos al portal de TeresIA así como para poder acceder al servicio de desambiguación y enlazado. Por tanto, en las próximas secciones del documento se explicará cómo usar el conversor con glosarios monolingües. En concreto, la sección 2 se centra en cómo deben estar los glosarios estructurados para su correcta conversión. A continuación, la sección 3 describe, paso a paso, cómo usar TermiGraph para convertir los glosarios a RDF y añadir los datos al portal de TeresIA. Finalmente, la sección 4 explica cómo se transforman los datos según la estructura de Ontolex-lemon.

# 2. ESTRUCTURA DEL GLOSARIO

Este conversor requiere como input un **archivo de texto plano**, es decir, un archivo con **extensión *.txt***. En cuanto a la estructuración de los datos, se debe expresar un término por línea. Tal y como se muestra en la Figura 1, **no** se debe **añadir ningún tipo de puntuación al final** de las líneas dado que el conversor divide los términos con los saltos de línea. Es decir, cualquier signo de puntuación que se añada al final de la línea, el conversor asumirá que es parte del término. Del mismo modo, no se debe añadir información adicional al término.

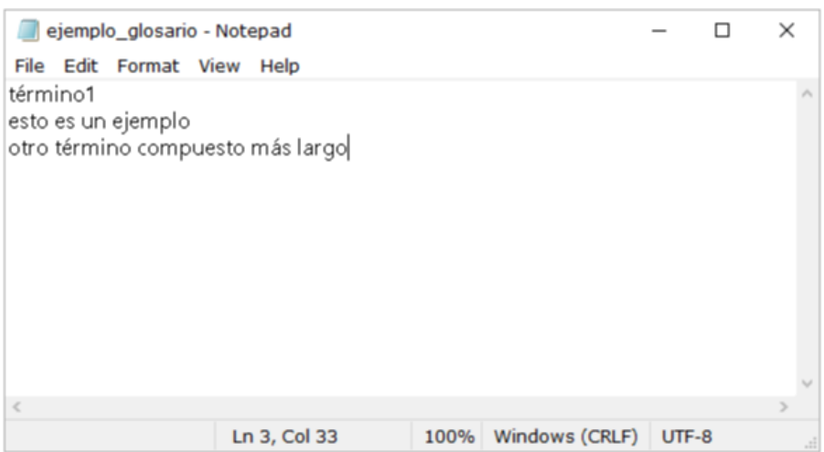


Figura 1: Ejemplo de glosario monolingüe

# 3. TRANSFORMAR UN GLOSARIO MONOLINGÜE CON TERMIGRAPH

A continuación, se explica paso a paso cómo convertir un glosario monolingüe con TermiGraph[1].

## 3.1. Paso 1: Cargar el glosario

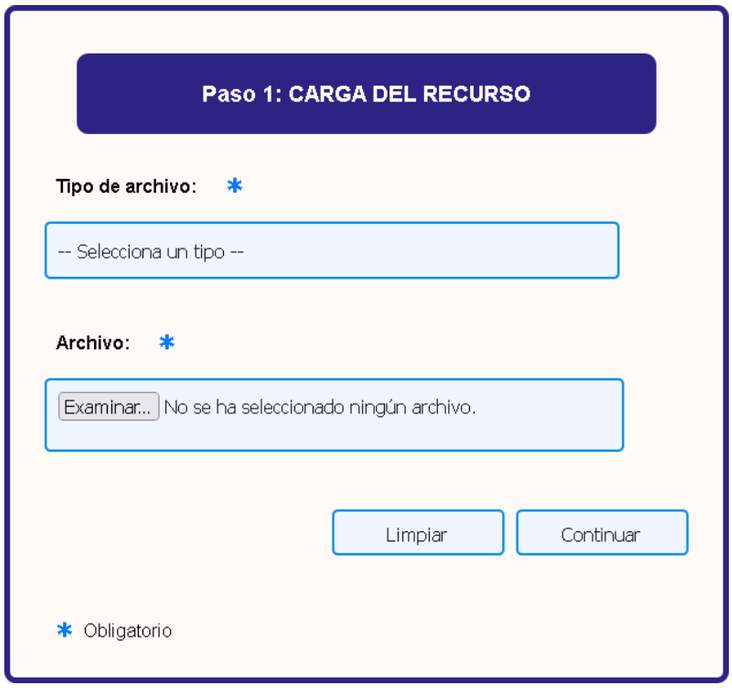


Figura 2: TermiGraph paso 1 (carga del recurso)

Tal y como se muestra en la Figura 2, en la primera pantalla del conversor se ha de cargar el recurso que se quiera convertir, indicando su tipo.

Para ello, primero se debe indicar el *Tipo de archivo*. Esta selección condicionará el conversor que se quiere utilizar. En el caso de los glosarios monolingües, se debe seleccionar **'Glosario monolingüe (lista)'**.

A continuación, en *Archivo*, se debe cargar **cargar el fichero** a convertir. Para ello, pulse 'Examinar'. Es importante tener en cuenta que dicho fichero debe estar en formato de texto plano, es decir, debe tener la **extensión *.txt***. Asimismo, tal y como se explica en la Sección 2, **cada línea** debe contener **un** sólo **término, sin signos de puntación** (salvo que sean parte del término).

Una vez se haya cumplimentado la primera página, pulse 'Continuar'. En caso de querer borrar los datos, puede pulsar 'Limpiar'.

## 3.2. Paso 2: Indicar los metadatos del recurso

A través de la segunda página del conversor se registran los metadatos del glosario (ver Figura 3).

En primer lugar, es obligatorio introducir el nombre del recurso en el campo *Título del recurso*.

[1] `https://termigraph.teresia.linkeddata.es`

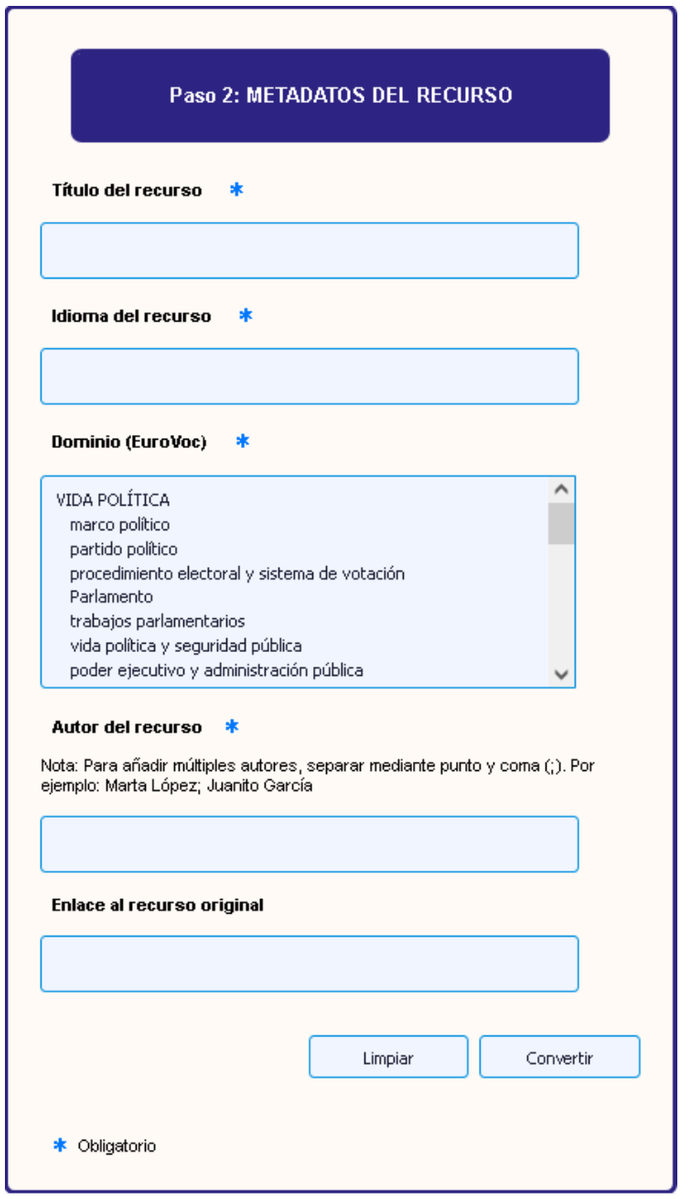


Figura 3: TermiGraph paso 2 (metadatos del recurso)

A continuación, debe especificarse el idioma del glosario en el campo *Idioma del recurso*. Dado que se trata de un recurso monolingüe, solo se admite un idioma. El idioma introducido por el usuario será buscado en Lexvo[2], una ontología que identifica un gran número de lenguas y sus respectivas etiquetas en distintos idiomas. Para asegurar una conversión correcta, **se recomienda indicar el idioma mediante su código ISO o su nombre oficial en español o en inglés**, ya que estos valores presentan mayor cobertura en Lexvo. En caso de que el idioma no se encuentre en dicha ontología, se conservará tal y como lo haya introducido el usuario.

Seguidamente, en el campo *Dominio (EuroVoc)*, se deben seleccionar uno o varios dominios temáticos relacionados con el glosario. Dado que TermiGraph ha sido desarrollada en coherencia con los demás servicios de TeresIA, las opciones disponibles se han limitado para facilitar la posterior desambiguación de términos. En particular, se emplea el esquema establecido por EuroVoc[3],el tesauro multidisciplinario de la Unión Europea.

En el campo *Autor del recurso*, de carácter obligatorio, se debe indicar el autor o los autores del

[2] `http://www.lexvo.org/`
[3] `https://eur-lex.europa.eu/browse/eurovoc.html?locale=es`

glosario. Tal como se señala en la nota, **si existen varios autores, deben separarse mediante punto y coma (;)**. Por ejemplo: Marcela Sánchez Giménez; Pedro López Rosas. Cabe destacar que, al convertirse los datos en un grafo interconectado, no se conserva ningún orden jerárquico entre los autores.

De forma opcional, puede añadirse un enlace al recurso original en el campo *Enlace al recurso original*. Aunque no es obligatorio, se recomienda completarlo para aumentar la visibilidad del glosario, especialmente si se prevé su incorporación al portal de TeresIA.

Una vez completados todos los metadatos, se puede proceder a la conversión del recurso pulsando el botón *Convertir*. Si se desea eliminar la información introducida, puede utilizarse la opción *Limpiar*.

## 3.3. Paso 3: Esperar

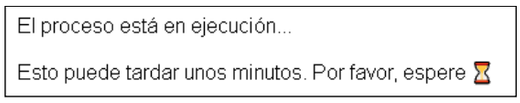


Figura 4: TermiGraph paso 3 (espera)

Mientras el glosario se convierte, un mensaje de espera aparecerá en pantalla. El conversor puede tardar varios minutos en convertir el recurso, especialmente si el recurso es grande.

## 3.4. Paso 4: Descarga del archivo

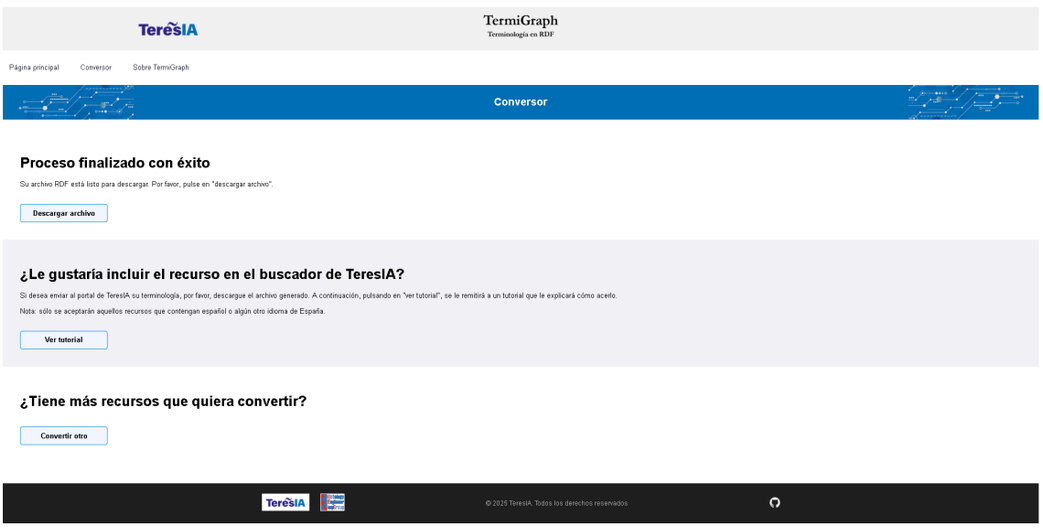


Figura 5: TermiGraph paso 4 (descarga)

Finalmente, cuando el proceso de conversión se completa correctamente, el archivo resultante puede descargarse pulsando en *Descargar archivo*.

Además, **si se desea subir los datos al portal de TeresIA, se ofrece un tutorial explicativo**. Se puede acceder a él mediante el botón *Ver tutorial*. Antes de hacerlo, se recomienda **descargar previamente el archivo** convertido, ya que será el que deberá subirse posteriormente al portal. Asimismo, se recuerda que, debido a la naturaleza del proyecto, **solo se aceptarán recursos que incluyan alguno de los idiomas oficiales de España**.

En caso de querer convertir un nuevo recurso, es posible volver a la primera página del conversor pulsando en *Convertir otro*.

# 4. REPRESENTACIÓN DE LOS DATOS CON ONTOLEX

## 4.1. Web Semántica y RDF

El RDF (Resource Description Framework)[4] es un formato estándar propuesto por World Wide Web Consortium (W3C) y sirve como base de los datos enlazados. Este formato organiza la información en tripletas compuestas por un sujeto, un predicado y un objeto; y el conjunto de tripletas genera un grafo. El sujeto, que es el primer elemento de la tripleta, representa el recurso sobre el cual se proporciona información y se identifica mediante un Identificador Uniforme de Recursos (URI, por sus siglas en inglés). El predicado, que ocupa la segunda posición, describe la relación o propiedad que conecta al sujeto con el objeto, siendo también representado por un URI. Por último, el objeto, tercer elemento de la tripleta, define el valor asociado a la propiedad del sujeto, pudiendo ser un literal (como texto) o un URI.(Chaves-Fraga et al., 2022)

Para estandarizar la representación de datos, se utilizan las ontologías, que son modelos o esquemas semánticos diseñados para describir formalmente un dominio. Las ontologías incluyen diversos componentes, como clases, propiedades e instancias. Las clases definen categorías o tipos de elementos dentro de un ámbito de conocimiento. Las propiedades representan las relaciones o atributos que vinculan clases o instancias entre sí. Por último, las instancias son ejemplos concretos o específicos que forman parte de una clase.

1. Sujeto: es aquello de lo que se habla, el recurso principal de la afirmación.
2. Predicado: es la propiedad o relación que conecta al sujeto con algo más.
3. Objeto: es el valor o recurso que completa el significado de la relación establecida por el predicado.

Aunque no hay una sola forma de representar los datos lingüísticos en RDF, OntoLex-Lemon (a partir de ahora, Ontolex) (McCrae et al., 2017) se ha convertido en la ontología o modelo más utilizado para representar recursos lexicográficos y terminológicos en RDF (Cimiano et al. 2011, Di Buono et al. 2020). Esta ontología cuenta con cuatro clases principales, tal y como se puede observar en la Figura 6: Concepto Léxico (*Lexical Concept*), Sentido Léxico (*Lexical Sense*), Entrada Léxica (*Lexical Entry*) y Forma (*Form*). La clase Entrada Léxica se utiliza para representar una palabra, frase o unidad léxica en una lengua determinada. Las diferentes realizaciones morfológicas de la entrada se expresan a través de la clase Forma. En cuanto al Sentido Léxico, se utiliza para proporcionar una conexión semántica entre la Entrada Léxica y un concepto de la ontología. Por último, un Concepto Léxico representa un concepto o idea abstracta que unifica significados entre sentidos léxicos y lenguas.

[4] `https://www.w3.org/RDF/`

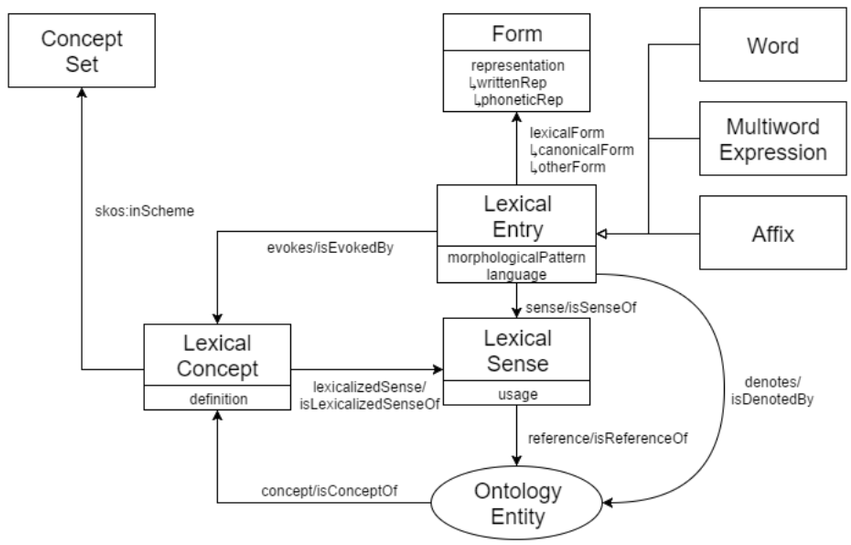


Figura 6: Diagrama del núcleo de Ontolex

Aunque Ontolex cubre la mayoría de las necesidades de representación que necesitan los glosarios monolingües, se utilizan otras ontologías para la representación de algunos aspectos como los idiomas o los datos de autoría. En concreto, se usan:

1. **Simple Knowledge Organization System (SKOS)**[5]: un estándar del W3C, utilizado habitualmente para representar datos jerárquicos en tesauros, sistemas de clasificación u otros tipos de sistemas de organización.

2. **Lexvo**[6]: Lexvo.org (de Melo, 2015) es una plataforma de datos enlazados que proporciona identificadores únicos y descripciones semánticas de lenguas y variantes lingüísticas. Su objetivo es vincular información lingüística de diferentes fuentes y ofrecerla en múltiples idiomas para apoyar la interoperabilidad y la reutilización de datos en la Web Semántica.

3. **Meta-Share**[7]: La ontología de META-SHARE es un modelo semántico formal que describe recursos lingüísticos y tecnologías de procesamiento del lenguaje, incluyendo sus propiedades, tipos, relaciones y condiciones de uso.

4. **DCMI Metadata Terms**[8]: vocabulario muy extendido para describir metadatos sobre recursos, como documentos, imágenes, conjuntos de datos o cualquier otro tipo de entidad digital o física. Por ejemplo, permite representar datos como el título o el idioma de un elemento. Esta ontología también se conoce como Dublin Core Terms (DCTerms).

5. **RDF Schema (RDFS)**[9]: una extensión de vocabulario de RDF que proporciona mecanismos para describir la estructura, la semántica y las relaciones de los datos. Este vocabulario, en este trabajo, se utiliza principalmente para describir los tipos de datos de los objetos de una tripleta.

[5] https://www.w3.org/2009/08/skos-reference/skos.html
[6] http://www.lexvo.org/
[7] http://w3id.org/meta-share/meta-share/2.0.0
[8] http://purl.org/dc/terms/
[9] http://www.w3.org/2000/01/rdf-schema#

## 4.2. Modelo de transformación de TeresIA

Siguiendo el modelo de OntoLex, en el proceso de transformación de los datos terminológicos se generan cuatro clases principales: la Forma, la Entrada Léxica, el Sentido Léxico y el Concepto Léxico, tal y como se muestra en la Figura 7.

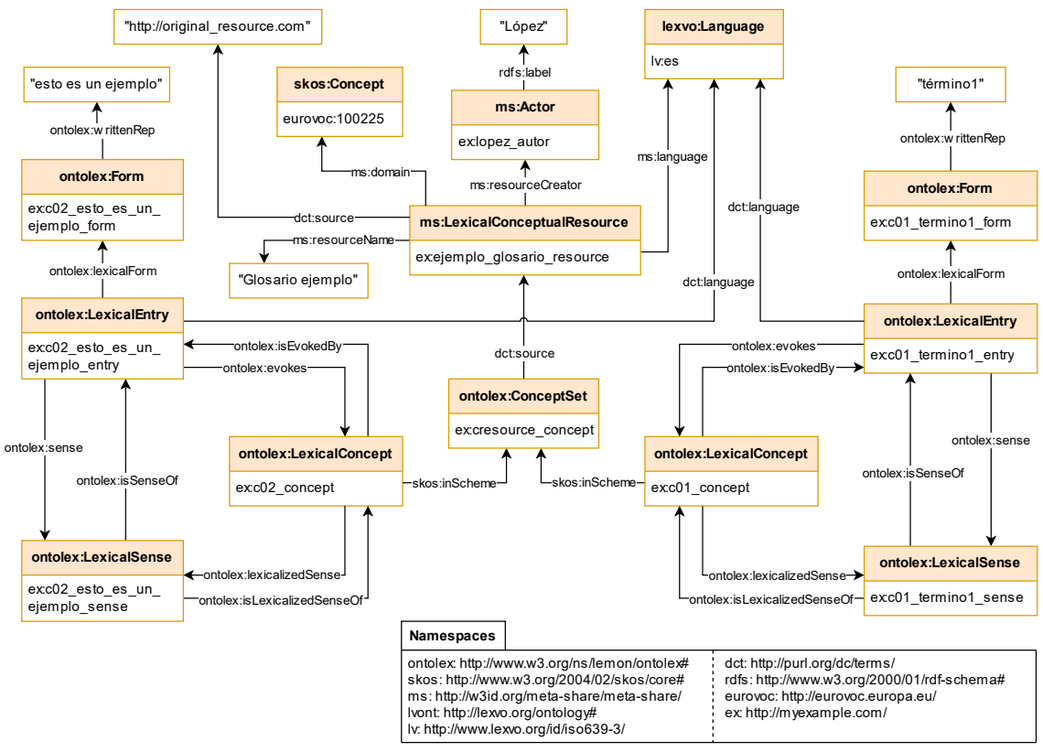


Figura 7: Diagrama ejemplo de datos convertidos a Ontolex

En primer lugar, existen tres tipos de información que se asocian con la Forma (`ontolex:Form`): la forma escrita, el género gramatical y el número gramatical. Sin embargo, dado que los glosarios no contienen información de tipo gramatical, la Forma generada únicamente incluye la propiedad `ontolex:writtenRep`, que recoge la representación escrita del término.

En cuanto a la clase Entrada Léxica (`ontolex:LexicalEntry`), en el caso de los glosarios esta clase solo contiene información relativa al idioma. Aunque el idioma no se indique explícitamente en el contenido del glosario, al tratarse de un recurso monolingüe, se asigna a todos los términos el idioma especificado en el formulario de conversión.

Para representar esta información se emplea la propiedad `dct:language`, que normalmente enlaza con una URI de Lexvo, lo que favorece la homogeneización y estandarización de los datos. Es decir, independientemente de si el usuario introduce 'es', 'spa', 'spanish' o español' en el formulario del conversor, el valor final corresponderá siempre a la URI de español en Lexvo[10].Cabe destacar que esta estandarización sólo es posible para los valores registrados en Lexvo mediante la propiedad `rdfs:label`. En caso de que el idioma introducido por el usuario no se encuentre en Lexvo, se conserva el valor original en formato texto (string).

Por cada término también se genera un Sentido Léxico (`ontolex:LexicalSense`). En esta clase suele almacenarse información sobre las relaciones entre términos (por ejemplo, sinonimia o traducción), el ciclo de vida de un término (por ejemplo, si está obsoleto) o su fiabilidad. No

[10] `http://www.lexvo.org/id/iso639-3/spa`

obstante, dado que los glosarios monolingües no incluyen este tipo de información, la clase de Sentido Léxico permanece vacía.

Asimismo, en los glosarios, por cada término se crea un Concepto Léxico (`ontolex:LexicalConcept`). Habitualmente, en esta clase se asocian definiciones y notas. Sin embargo, como los glosarios no contienen información de este tipo, la clase correspondiente al Concepto Léxico queda sin contenido terminológico adicional.

Todos los Conceptos Léxicos se agrupan dentro de un Conjunto de Conceptos (`ontolex:ConceptSet`), mediante la propiedad `skos:inScheme`. A través de este Conjunto de Conceptos se puede acceder, mediante la propiedad `dct:source`, a un Recurso Léxico/Conceptual (`ms:LexicalConceptualResource`), en el que se almacenan los metadatos del recurso.

Los metadatos del recurso se describen conforme a la ontología Meta-Share. Para representar el recurso se genera una instancia de la clase `ms:LexicalConceptualResource`. El nombre del recurso se indica mediante la propiedad `ms:resourceName`, y el idioma se especifica con `ms:language`. Como se ha mencionado anteriormente, con el fin de mejorar la estandarización y la interoperabilidad, el idioma se expresa a través de una URI de Lexvo. En caso de no encontrarse en Lexvo, se mantiene el valor textual original introducido por el usuario.

El dominio temático del recurso se indica mediante la propiedad `ms:domain`, que enlaza con una instancia de EuroVoc (de clase `skos:Concept`). Además, la propiedad `ms:resourceCreator` vincula el recurso con su autor, representado como una instancia de `ms:Actor`, cuyo nombre se especifica mediante `rdfs:label`. Finalmente, el enlace al recurso original proporcionado por el usuario se representa con la propiedad `dct:source`.

## Referencias


Chaves-Fraga, D., Óscar Corcho, and Ruckhaus, E. (2022). Guía práctica para la publicación de datos enlazados. *datos.gob.es*.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*.

de Melo, G. (2015). Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web*, 6(4):393–400.

Di Buono, M. P., Cimiano, P., Elahi, M. F., and Grimm, F. (2020). Terme-a-llod: Simplifying the conversion and hosting of terminological resources as linked data. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28–35.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.