UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

TÍTULO: Interoperabilidad semántica en el manejo de datos normativos sobre la presencia de agroquímicos en alimentos

AUTORES: Carlos Francisco Ragout

DIRECTOR: Dr. Alejandro Fernández

CODIRECTOR: Dr. Diego Torres

ASESOR PROFESIONAL:

CARRERA: Licenciatura en Sistemas

Resumen

El objetivo de esta tesina es demostrar de qué manera las tecnologías de la web semántica ofrecen soluciones a la problemática de la publicación interoperable de datos normativos sobre la presencia de agroquímicos en alimentos. Para esto se creó una ontología que permite representar formalmente este dominio y se elaboró un pipeline de transformación de datos de sus fuentes originales a un dataset semántico. Se utilizó este dataset para luego demostrar cómo estas tecnologías facilitan la ejecución de distintas operaciones de interoperabilidad entre los datasets demostrando la diferencia entre usar la estrategia propuesta respecto de los métodos existentes.

Palabras Clave

Ontología, AGROVOC, OpenRefine, LMR, Límite máximo de residuo, Fitosanitario, Pesticida, Agroquímico, Web Semántica, Provenance, SPARQL

Conclusiones

La publicación de datos normativos sobre los residuos de productos fitosanitarios mediante el uso de tecnologías de la web semántica propicia la interoperabilidad, permitiendo que los datos puedan ser consumidos fácilmente por agentes inteligentes y se fomenta la colaboración entre los distintos actores del dominio, eliminando barreras como el idioma o el soporte en el que se encuentran los datos. Por otra parte, la información de proveniencia en los dataset semánticos brinda confianza sobre el origen de los datos y permite realizar una trazabilidad a las fuentes originales.

Trabajos Realizados

En este trabajo de grado se creó una ontología, LMR-O, que permite mediante la reutilización de ontologías existentes y vocabularios propios describir el dominio de conocimiento de la publicación de datos normativos respecto a los límites máximos de residuos de fitosanitarios en alimentos. Se desarrolló además un pipeline robusto, reproducible y trazable, y se brindaron las herramientas que sientan las bases para la transformación de los datasets tradicionales a datasets semánticos.

Trabajos Futuros

Como trabajo futuro se contempla la posibilidad de continuar desarrollando la ontología LMR-O para abarcar más áreas del conocimiento del dominio, así como la necesidad de enriquecer las ontologías utilizadas como base. De la misma manera se plantea mejorar las herramientas ad hoc y las de código abierto utilizadas para la creación de los datasets semánticos de forma tal que puedan seguir dando soporte al proceso de transformación mientras éste evoluciona y es mejorado.

Fecha de la presentación: Febrero 2021

Interoperabilidad semántica en el manejo de datos normativos sobre la presencia de agroquímicos en alimentos

Carlos Francisco Ragout

22 de febrero de 2021

Resumen

El objetivo de esta tesina es demostrar de qué manera las tecnologías de la web semántica ofrecen soluciones a la problemática de la publicación interoperable de datos normativos sobre la presencia de agroquímicos en alimentos. Para esto se creó una ontología que permite representar formalmente este dominio y se elaboró un pipeline de transformación de datos de sus fuentes originales a un dataset semántico. Se utilizó este dataset para luego demostrar cómo estas tecnologías facilitan la ejecución de distintas operaciones de interoperabilidad entre los datasets demostrando la diferencia entre usar la estrategia propuesta respecto de los métodos existentes.

Agradecimientos

A cada compañero, amigo y docente que directa o indirectamente aportó su grano de arena durante este largo camino recorrido.

A mis directores de tesis, en especial a Alejandro Fernández por su dedicación y compromiso con este trabajo.

A mi familia, al Enano y a mis amigos.

Índice general

Capítulo 1

Introducción

1.1. Fitosanitarios y sus residuos en alimentos

Llamamos productos fitosanitarios a aquellas sustancias o mezclas de sustancias destinadas a prevenir, atraer, repeler o controlar cualquier plaga de origen animal o vegetal durante la producción, almacenamiento, transporte, distribución y elaboración de productos agrícolas y sus derivados[1]. Su uso conlleva la presencia de residuos en los productos tratados, por lo que se hace necesario regular la cantidad máxima de estos residuos que pueden contener los alimentos, y de esta manera asegurar que no supongan un riesgo para la salud de los consumidores. Se entiende por residuo a las sustancias activas, los metabolitos y los productos de degradación o reacción de las sustancias activas utilizadas en productos sanitarios que están presentes en alimentos[2].

En Argentina se utilizan mas de 400 productos fitosanitarios diferentes y, según datos de 2014[3], sólo las empresas que componen la Cámara de Sanidad y Fertilizantes (CASAFE) utilizaron más de 304 millones de litros/kilos. De estos, cerca del 87 % corresponden a herbicidas (de los cuales 62 % del total son glifosatos), 5,78 % a insecticidas, 3,14 % a fungicidas y el restante 4,07 % a otros productos.

1.2. Normativa respecto a la presencia de fitosanitarios

La demanda por obtener mayores rendimientos de cultivos ha llevado al hombre a desarrollar y utilizar productos fitosanitarios cada vez mas avanzados. De acuerdo a la Organización Mundial de la Salud (o WHO por sus siglas en inglés), se utilizan en el mundo más de 1000 pesticidas diferentes[4], algunos de los cuales pueden durar años en el suelo o el agua. Por ejemplo, el diclorodifeniltricloroetano (o DDT) debido a su no degradabilidad puede viajar a lugares distantes e incluso, al ser liposoluble, puede acumularse en animales y cuerpos humanos [5]. Es por eso que existen normativas respecto a qué productos se pueden utilizar y en que cantidad.

Actualmente los límites o tolerancias máximas de residuos de productos fitosanitarios en alimentos o LMR, son regulados por varios organismos gubernamentales de casi todos

los países del mundo. Así mismo, existen diferentes organizaciones relacionadas a la salud alimenticia que publican recomendaciones sobre los LMR que muchos de estos organismos antes mencionados toman como guía. Estas normas tienen como objetivo asegurar niveles seguros de residuos de plaguicidas y contar con alimentos inocuos que no presenten peligros para la salud humana.

A nivel mundial, la Comisión del Codex Alimentarius, creada por la FAO y la OMS, fija recomendaciones para el LMR. Lo propio hace la United States Environmental Protection Agency (US EPA por sus siglas en inglés). En la Argentina, la norma 934-2010 del Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA) establece los requisitos que deben cumplir los productos y subproductos agropecuarios para consumo interno. En esta resolución se regulan aspectos sobre los límites máximos de residuos de plaguicidas, se establece un listado de productos fitosanitarios que se hallan exentos del requisito de fijación de tolerancias así como también un listado de principios activos prohibidos y restringidos.

Es importante además, para lograr que las aplicaciones de fitosanitarios no generen LMR superiores a los establecidos en las normas, respetar el tiempo que debe transcurrir entre la aplicación del producto fitosanitario y la cosecha del alimento. Esto se conoce como Tiempo de Carencia.

1.3. El desafío de la interoperabilidad en datos relativos al LMR

Cada organización posee sus métodos para construir y publicar estos datasets, los cuales difieren en su formato (pdf, xml, doc, xls, web, etc), en su contenido (tablas, dibujos, listas), en su idioma (español, inglés, chino, etc) e incluso, en algunos casos, utilizan diferente terminología. Tras una revisión de los datasets existentes podemos indicar que estos métodos de publicación de información, de los cuales se sabe muy poco, son propensos a errores que se acarrean versión tras versión pasando desapercibidos. Este abanico de formas de representación de lo que esencialmente es información de la misma naturaleza, hace que no exista una manera sencilla de poder realizar comparaciones entre dos datasets de organismos diferentes. Además, dentro de una misma organización, es muy difícil identificar diferencias entre dos versiones del mismo conjunto de información (por ejemplo, correspondiente a distintos años). Las mencionadas dificultades para combinar y comparar datos de distintas fuentes impacta negativamente en el trabajo de quienes dependen de esos datos, como es el caso de los investigadores en salud alimentaria y producción de alimentos, los productores de alimentos, y a quienes definen políticas públicas entre otros.

Parte de la problemática radica además en lo difícil que puede resultar encontrar y llevar un registro de esta información, que por otro lado suele publicarse en formatos propietarios y que no son aptos para poder ser leídos por una computadora sin un procesamiento previo de los datos. Esto es especialmente notable al utilizar estas fuentes de información en el mundo académico o en investigación, en donde la falta de información sobre el origen o la versión particular de los datos que se utilizan, pueden convertirse en un limitante al momento de poder reproducir el resultado de un experimento[6].

En ese contexto, ofrecer estrategias de publicación interoperable de datos normativos sobre la presencia de agroquímicos en alimentos, trae importantes beneficios. La web-semántica propone un conjunto de métodos y herramientas que pueden aportar en este sentido.

El objetivo de esta tesina es demostrar de qué manera las tecnologías de la web semántica (RDF/OWL y SPARQL entre otras) ofrecen soluciones a las problemáticas anteriormente mencionadas.

1.4. Organización de este documento

Esta tesina está dividida en 7 capítulos. En el primer capítulo se hizo una introducción a la problemática de la publicación de datos relativos al LMR.

En el siguiente capítulo se comentarán tecnologías de la web semántica relacionadas con esta problemática y se describirán casos actuales de metodologías de publicación de datos normativos.

En el tercer capítulo se presentará la estrategia general que se utilizará para encarar el proceso de transformación de datasets. También se hará una descripción de las herramientas que darán soporte a dicho proceso.

En el cuarto capítulo se presentaran las ontologías que representan partes del dominio de la publicación de LMR. Se definirá además, la ontología que permitirá representar los datos normativos relativos al LMR en un formato semántico que propicie la interoperabilidad y mantenga la trazabilidad a las fuentes originales.

En el quinto capítulo se definirá el pipeline de trabajo que permitirá la transformación de los datos normativos a la ontología propuesta. Además se enriquecerá la ontología ChEBI de manera de contener más conceptos y sinónimos para conceptos previamente existentes.

En el sexto capítulo se aplicará el proceso anteriormente descripto en los casos prácticos de Argentina y Brasil. Finalmente realizaremos consultas sobre los datasets resultantes utilizando la tecnología SPARQL para de esa manera demostrar las ventajas de los datasets semánticos versus los originales.

Por último, se describirán en el séptimo capítulo las conclusiones obtenidas y trabajos futuros en base al trabajo realizado y la experiencia que obtenida.

Capítulo 2

Trabajo Relacionado

2.1. La Web Semántica

La World Wide Web (o WWW por sus siglas en inglés) es un sistema de información donde los documentos y otros recursos web se identifican mediante URLs (como https://unlp.edu.ar/), los cuales pueden estar interconectados por hipertexto y son accesibles mediante internet[7]. La figura 2.1 ilustra una colección de recursos interconectados por hipertexto.

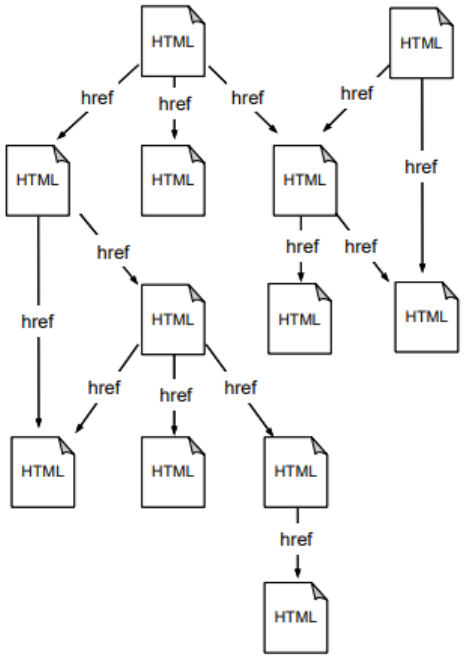

Figura 2.1: Web actual

Hoy en día una persona puede, mediante el uso de buscadores, encontrar casi cualquier cosa en la web. Si bien esto supone una gran ventaja respecto de otras maneras de consumir información, también puede volverse un problema. Se estima que hay 4.5 mil millones de usuarios y que para 2023 dos tercios de la población mundial poseerán una conexión a internet[8]. En 2008, Google ya reportaba 1 billón de URLs únicas[9]. Encontrar información en este universo de datos puede ser sumamente difícil. Desarrollar programas que realicen estas tareas, con el estado actual de la web, puede ser una tarea casi imposible, y mantener estos algoritmos al ritmo de cambio de la web es aún peor.

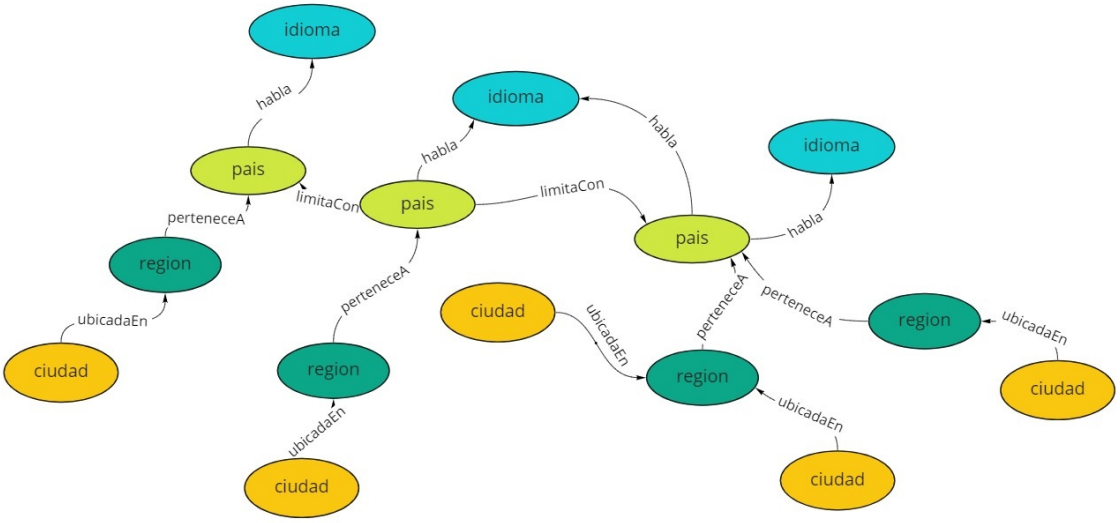

Figura 2.2: Web semántica

La web semántica[10], representada en la figura 2.2, propone superar las limitaciones de la web actual mediante el uso de tecnologías que permitan la publicación y procesamiento de datos que puedan ser legibles por aplicaciones informáticas. Estas aplicaciones, también llamadas agentes inteligentes, son automáticas, es decir, no poseen operadores humanos. Para ello, la web semántica busca clasificar, dar estructura, describir el contenido y significado de datos y sus relaciones y anotar recursos de manera explícita mejorando internet y ampliando la interoperabilidad entre los sistemas[11][12].

En este trabajo se hará uso de ontologías y vocabularios semánticos existentes de modo de poder describir formalmente en términos de web semántica el dominio de la publicación de datos normativos de LMR en alimentos. Entre estos antecedentes, que describiremos en detalle en la subsección 4.1.2, podemos destacar a AGROVOC, el cual es un vocabulario elaborado por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO[1]) publicado por primera vez al principio de los años 80.

[1] http://www.fao.org/

Otro antecedente destacable aplicado también a la web semántica, es el concepto de Provenance[13]. Es importante registrar el origen de los documentos ya que en una web en la que abunda la información, muchas veces falsa, imprecisa o de poca calidad, el hecho de poder identificar dicho origen puede ser una herramienta de suma importancia, si no la más importante, a la hora de determinar si estos datos son confiables. Si se planea enfocar el tema de la interoperabilidad de datos, haciendo transformaciones de su fuente original a la web semántica, es importante incluir información de donde fueron tomados estos datos.

2.2. Publicación de datos Normativos

La publicación de datos normativos de LMR en alimentos se lleva a cabo por variadas instituciones en todo el mundo. El nivel de interoperabilidad que estas instituciones alcanzan y la facilidad para el consumo y posterior análisis de estos datos varía considerablemente según el organismo que publique dichos datos.

En el caso de Argentina, SENASA publica en su sitio web[2] periódicamente un archivo en formato Excel el cual posee la información normativa de los LMR en alimentos. Con cada nueva publicación, una nueva versión del archivo es subida al portal web y la vieja versión deja de estar accesible para los usuarios. Esto es así, salvo que éstos sepan la url del archivo anterior al que intentan acceder, la que típicamente es un prefijo (idéntico para todas las versiones) seguido del mes y año de publicación. El nombre del archivo es la única manera de identificar el origen de los datos, ya que ni en el contenido mismo se encuentra la fecha de creación del mismo. Las operaciones que se pueden realizar sobre este dataset, son aquellas que nos permita realizar la planilla de cálculo, como por ejemplo, ordenar y buscar palabras. La información en el dataset a menudo aparece escrita de diferentes maneras y no es raro encontrar errores de tipeo u ortografía, por lo que la interoperabilidad es muy difícil.

En Brasil, ANVISA (Agência Nacional de Vigilância Sanitária) posee un sitio interactivo[3] que permite al usuario navegar y consultar el conjunto de datos. En este caso la información se encuentra bien estructurada (es decir, sin errores ni sinónimos) por lo que encontrar lo que se está buscando resulta bastante sencillo. Además posee información como el grupo y la fórmula química de las sustancias fitosanitarias. Es posible descargar la información en formato pdf o CSV.

La Unión Europea en la resolución 396/2005 del Parlamento Europeo establece los límites máximos de residuos para plaguicidas. La información es publicada, al igual que en caso de Brasil, en un sitio web interactivo[4] que permite la navegación de los datos mediante consultas en un buscador. El sitio se encuentra en varios idiomas y posee información de las normas en las que se establecieron los valores de los LMR. Es también posible descargar el dataset en formato xml, más adecuado para sistemas informáticos. Sin embargo, el dataset

[2]http://www.senasa.gob.ar/normativas/resolucion-934-2010-senasa-servicio-nacional-de-sanidad-y-calidad-agroalimentaria

[3]http://portalanalitico.anvisa.gov.br/monografias-de-agrotoxicos

[4]https://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/mrls/?event=search.pr

viene dividido en seis archivos y no es posible interpretar los datos sin algún procesamiento previo de los mismos. Esto quiere decir que, un investigador que quisiera hacer uso de esta información, debería primero dedicar tiempo y esfuerzo a encontrar la manera de consumirla antes de poder dedicarse a su análisis.

En China, la norma GB 2763—2019 establece los estándares nacionales de seguridad alimentaria y los límites máximos de residuos de pesticidas. La norma es publicada por el Ministerio de Agricultura y Asuntos Rurales, y puede obtenerse, en idioma chino, en el sitio de la Asociación China para la Nutrición y la Salud en los Alimentos[5]. La descarga es un archivo en formato pdf que a su vez se encuentra comprimido (.rar).

Otro caso interesante es el de Estados Unidos, donde el Departamento de Agricultura y la Agencia de Protección Medioambiental (USDA y EPA respectivamente) resolvieron delegar el manejo de la información de los LMR en alimentos en una empresa privada llamada Bryant Christie Inc.[6]. Esta empresa brinda de forma gratuita la información normativa de los Estados Unidos y ofrece planes de pago para acceder a la información de 130 países. Este portal web en su versión libre es similar a los que poseen Brasil o la Unión Europea: permite realizar búsquedas interactivas, exportar el dataset hasta cierto número de registros, etc. A estas opciones, la versión paga le agrega, entre otras funcionalidades, poder comparar datasets de diferentes países, el cual es uno de los objetivos de este trabajo.

Sin embargo, también se destacan algunos puntos en contra. En primer lugar los datasets que se exportan de este sitio web son en un formato privativo (Excel) y suelen ser un subconjunto del total (el sistema sólo permite exportar hasta una determinada cantidad de registros de una sola vez).

En segundo lugar no es posible determinar de donde vienen los datos, el usuario sólo puede consumirlos y deberá confiar en que provienen de fuentes autorizadas. Por último, las comparaciones u operaciones que se pueden hacer dado un conjunto de datos son limitadas. Solamente se pueden comparar principios activos y cultivos. Si bien un usuario podría responder con facilidad las preguntas del tipo "¿Qué principios activos puedo utilizar en Manzanas?" o "¿Con qué cultivos puedo utilizar Metalaxyl?", no hay una manera sencilla de responder preguntas más complejas e igualmente interesantes, como "¿Qué principios activos se utilizan en Manzanas y no en Peras?" o "¿Cuál es el principio activo que se utiliza en mayor cantidad para cada cultivo diferente?". Cabe destacar que este problema no es propio de este dataset o metodología de publicación, sino que es común en todos los casos analizados. Se buscó en este trabajo de tesis brindar una solución a estos problemas mediante el uso de tecnologías de la web semántica.

[5] http://www.cnhfa.org.cn/fagui/show.php?itemid=442
[6] https://www.bryantchristie.com/

Capítulo 3

Estrategia general

3.1. Pipeline de transformación de datos

Para abordar la problemática de la publicación interoperable de datos normativos sobre la presencia de agroquímicos en alimentos se plantea la creación de un framework o pipeline de trabajo. Este proceso tiene como objetivo la transformación de datos existentes, datos muchas veces no estructurados y con errores, a una ontología y vocabularios capaz de representar distintos datasets y que propicie la interoperabilidad, manteniendo la trazabilidad a las fuentes originales. El proceso debe ser a la vez reproducible, robusto y efectivo.

El pipeline de transformación está basado en el uso de ontologías existentes, a fin de garantizar interoperabilidad y reusabilidad de herramientas probadas. Cada una de las ontologías elegidas está enfocada en representar de manera formal un aspecto específico de la publicación de LMR en alimentos. La combinación de estas ontologías permite reusar el conocimiento existente de cada una de ellas para representar el dominio como si fuera un mosaico. Sin embargo, como ninguna de las ontologías elegidas se corresponde estrictamente con el dominio del LMR en alimentos, este mosaico se encuentra incompleto y presenta grietas. Para cubrir estas partes faltantes, la estrategia a que tomada fue, en primer lugar, crear los vocabularios necesarios dentro de las ontologías originales. De esta manera generamos ontologías "aumentadas" que con el paso del tiempo se podrán volver todavía más específicas al dominio que representan. Un ejemplo de esto, es la creación de conceptos que representan principios activos o aptitudes dentro de ChEBI, la ontología de entidades químicas.

Por otro lado, se creó también una ontología que permite realizar una representación formal y lo más completa posible del dominio de la publicación de datos normativos sobre residuos de productos fitosanitarios en alimentos. Esta ontología, la cual se describe más adelante en este trabajo, persigue el mismo objetivo de interoperabilidad y reusabilidad que persigue el proceso general, por lo que también se construye en base a términos y ontologías existentes además de los términos propios.

El proceso de transformación se puede describir, en líneas generales, como un pipeline con los siguientes pasos:

1. **Formateo de datos:** Transformaciones simples en los datos no estructurados para generar una versión depurada de los mismos.

2. **Alineación de datos:** Asociación de los datos (esencialmente cadenas de caracteres) a términos semánticos.

3. **Generación de un dataset semántico:** Generación de las tripletas RDF y agregado de información de proveniencia.

El objetivo del pipeline es transformar los datos publicados en internet por los diferentes organismos, en cualquier formato en que se encuentren, a tripletas RDF que conformen un dataset semántico que será apto para ser consumido por sistemas automáticos y personas.

Los conjuntos de datos semánticos que obtengamos como parte de esta estrategia, serán luego utilizados para realizar una evaluación de esfuerzo, y eficacia en la ejecución de distintas operaciones de interoperabilidad entre los datasets, y mantenimiento de los mismos, a fin de demostrar de qué manera las tecnologías de la web semántica (RDF/OWL y estrategias de Linked open Data) ofrecen una estrategia superadora para la publicación interoperable de datos normativos sobre la presencia de residuos de productos fitosanitarios.

3.2. Herramientas

Para poder dar soporte a las diferentes etapas del proceso de transformación de datos se hará uso de diferentes herramientas que se describirán a continuación.

3.2.1. OpenRefine

Una de las herramientas más importantes que se usará a lo largo del pipeline es OpenRefine[1] 3.4.1. OpenRefine es una aplicación de escritorio de código abierto originalmente creada por Google para limpieza de datos generalmente desordenados y transformación de los mismos a otros formatos. Esto se conoce generalmente como Data Wrangling (o arreglo de datos). Data Wrangling se conoce también como el proceso de recolección, selección y transformación de datos con el fin de agregar valor o realizar análisis de los mismos[14].

OpenRefine nos permite no sólo filtrar, buscar y ordenar la información, como se ve en la figura 3.1, sino que también es posible limpiarla, exportarla en diferentes formatos y extenderla mediante el uso de información externa o incluso utilizando servicios web. La herramienta brinda además trazabilidad en los cambios, pudiendo observar la secuencia de pasos que se llevo a cabo hasta el estado actual del dataset. Esta secuencia de pasos puede ser exportada en formato GREL (General Refine Expression Language). Este es un formato propio de OpenRefine basado en JSON utilizado para manipular información en la herramienta. El lenguaje usa algunos tipos de datos como boolean, number, string, date y array. Además provee algunas variables que representan información como las celdas y filas

[1] https://openrefine.org/

en la tabla de OpenRefine o metadata como por ejemplo el identificador del concepto con el que se reconcilió un término[15].

Poder exportar los pasos realizados permite reutilizar las secuencias de acciones en futuros datasets. De esta manera, se puede crear una biblioteca de secuencias que luego pueden ser aplicadas a diferentes conjuntos de datos, de forma de compartimentizar las transformaciones en diferentes categorías (secuencias para limpieza de datos, secuencias para alineación, secuencias para corrección, etc) a la vez que se ahorra tiempo y se reduce la posibilidad de cometer errores.

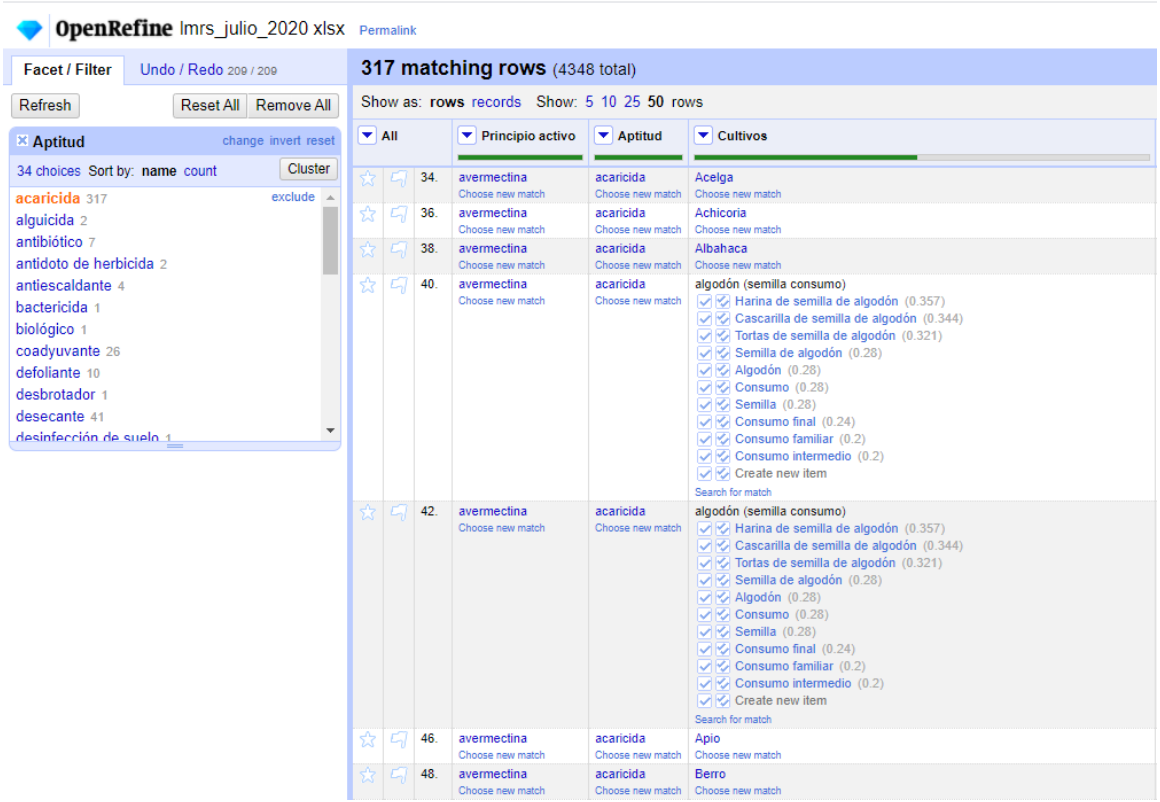

Figura 3.1: Filtrando por la columna Aptitud en OpenRefine

Junto con esta herramienta, se utilizará la extensión RDF[2] versión 1.3.0, que nos permitirá hacer uso de archivos en este formato como servicio de reconciliación (o alineación). Un servicio de reconciliación es un servicio web o archivo RDF que, dado un texto que es un nombre o etiqueta para algo y, opcionalmente, algunos detalles adicionales, devuelve una lista clasificada de entidades potenciales que coinciden con los criterios. El texto candidato no tiene que coincidir perfectamente con el nombre oficial de cada entidad, el objetivo de la reconciliación es pasar de un nombre de texto ambiguo a entidades identificadas con preci-

[2]https://github.com/stkenny/grefine-rdf-extension

sión[16]. Por ejemplo, dado el texto "sal", un servicio de conciliación probablemente debería devolver la fórmula química de la sal (NaCl), el nombre químico de la sal (cloruro de sodio) y la forma imperativa del verbo salir (sal). Se puede decir que el objetivo de un servicio de reconciliación es obtener buenos candidatos que representen los términos contenidos en una base de datos determinada. Para esto utilizan los siguientes elementos:

1. Una cadena de texto que representa el nombre o el título de una entidad
2. El tipo de la entidad, como por ejemplo Persona o Principio Activo. Este elemento es opcional.
3. Una lista de propiedades y sus valores, que se pueden usar para refinar la búsqueda. Por ejemplo, al reconciliar una base de datos de libros, el nombre del autor o la fecha de publicación son pueden ser útiles para descartar entidades similares. Esta lista es opcional.

La extensión RDF para OpenRefine nos permitirá usar como servicios de reconciliación endpoints SPARQL así como también dumps RDF de ontologías existentes. Usaremos diferentes servicios de reconciliación basados en archivos a lo largo del trabajo, entre ellos versiones de ChEBI que nos permitirán comprobar cómo es la efectividad en la alineación con y sin sinónimos, así como también el servicio de AGROVOC para la alineación con cultivos.

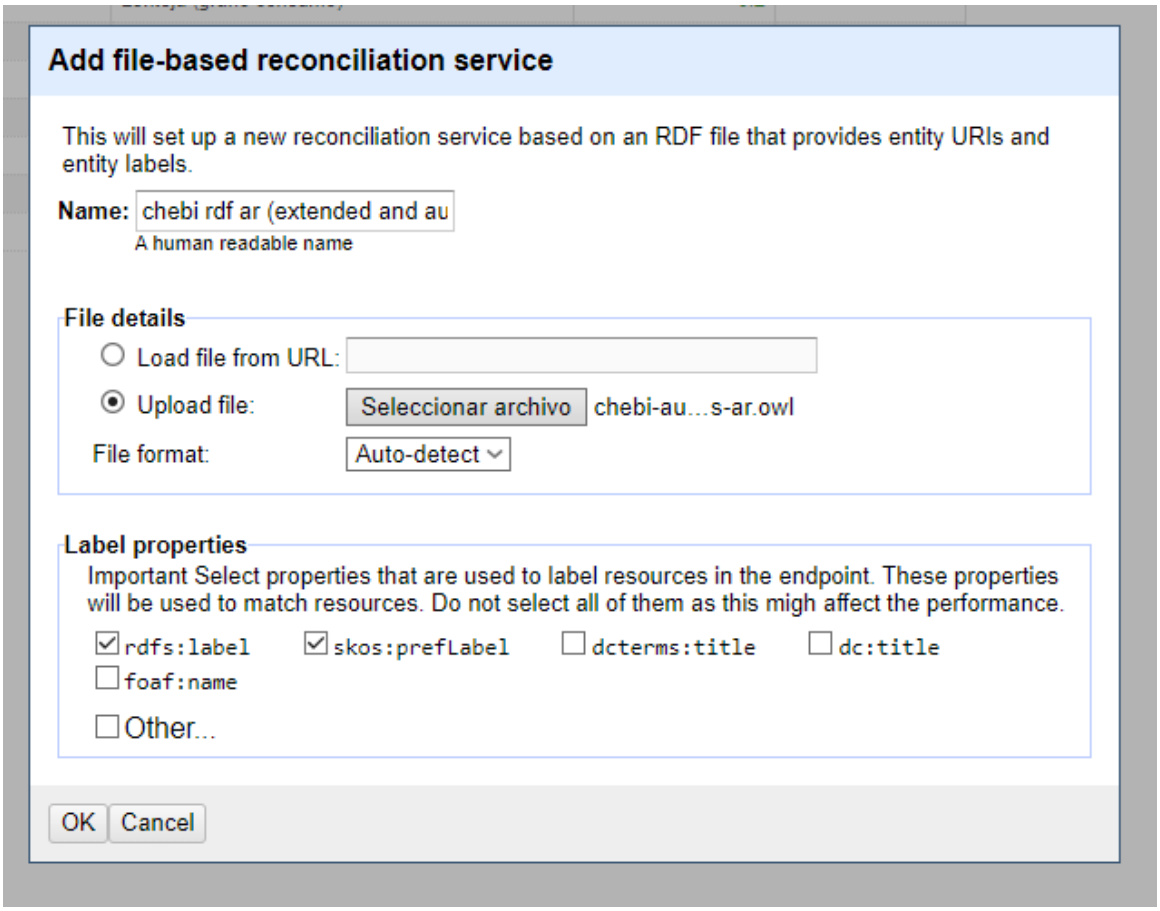

Figura 3.2: Agregando un servicio de reconciliación basado en un archivo RDF

En la Figura 3.2 se pueden apreciar las opciones de configuración que ofrece la herramienta al momento de agregar un servicio de reconciliación basado en RDF.

Finalmente, utilizaremos OpenRefine para exportar los datos transformados y alineados a un formato RDF.

3.2.2. Protégé

Protégé es un editor de ontologías y un manejador de conocimiento libre y open source creado por el Stanford Center for Biomedical Informatics Research[3], el cual provee una interfaz gráfica que, entre otras funciones, nos permiten diseñar, modificar, visualizar y exportar ontologías. La aplicación está escrita en Java y sigue un modelo de framework, para el que los miles de usuarios crean diferentes plugins.

Se utilizará Protégé 5.5.0 para la creación de la ontología LMR-O y cada uno de sus

[3]https://bmir.stanford.edu/

términos, lo cual se exportará en el archivo LMR-O.owl que podrá encontrarse en el repositorio de la tesina.

3.2.3. Apache Jena Fuseki

Una de las ventajas de convertir los datos a una representación de la web semántica es la posibilidad de realizar consultas utilizando SPARQL. SPARQL es un lenguaje de consulta de datos pensado para trabajar sobre RDF definido por la W3C[17]. Este lenguaje esta definido como un estándar de la RDF Data Access Working Group (DAWG) y es reconocido como una de las tecnologías mas importantes de la web semántica[18]. De la misma manera que con SQL un analista puede realizar consultas sobre una base de datos, utilizaremos SPARQL para obtener información de los datasets resultantes del proceso de transformación de datos que se describirá a lo largo de este trabajo. Se puede ver en el snippet 1 un ejemplo de consulta SPARQL.

```
SELECT distinct ?crop (count(distinct ?activePrinciple) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role <http://purl.obolibrary.org/obo/CHEBI_33288>.
    ?subject lmro:appliesTo ?crop.
}
group by ?crop
limit 1
```

Snippet 1: Consulta SPARQL

Si bien Protégé posee su propio plugin de consulta SPARQL integrado, el mismo presenta varias deficiencias. En principio, no posee ningún tipo de ayuda visual o highlighting de palabras claves, autocompletado o historial de consultas. Tampoco es posible copiar o pegar dentro del editor de consultas o del panel de resultados, ni deshacer cambios con atajos del teclado. Más problemático aun, es que la implementación de SPARQL que posee Protégé no es la más actual, por lo que no es raro encontrarse con funciones que no están disponibles. Todas estas dificultades vuelven a esta herramienta incómoda y sumamente impráctica. Para poder sortear estos inconvenientes se utilizará el servidor SPARQL Jena Fuseki de Apache[4] en su versión 3.17.0. Este servidor posee varios modos de ejecución, tanto por línea de comandos como mediante una interfaz gráfica. Entre otras ventajas destacables de Jena Fuseki, se encuentran la posibilidad de realizar consultas sobre varios datasets al mismo tiempo, consulta de namespaces en servicios externos, resultados formateados, facilidades para exportar resultados en varios formatos, búsqueda de texto en resultados, etc. Es posible ver en la figura 3.3 la interfaz gráfica de la herramienta con varias de estas opciones.

[4]https://jena.apache.org/documentation/fuseki2/

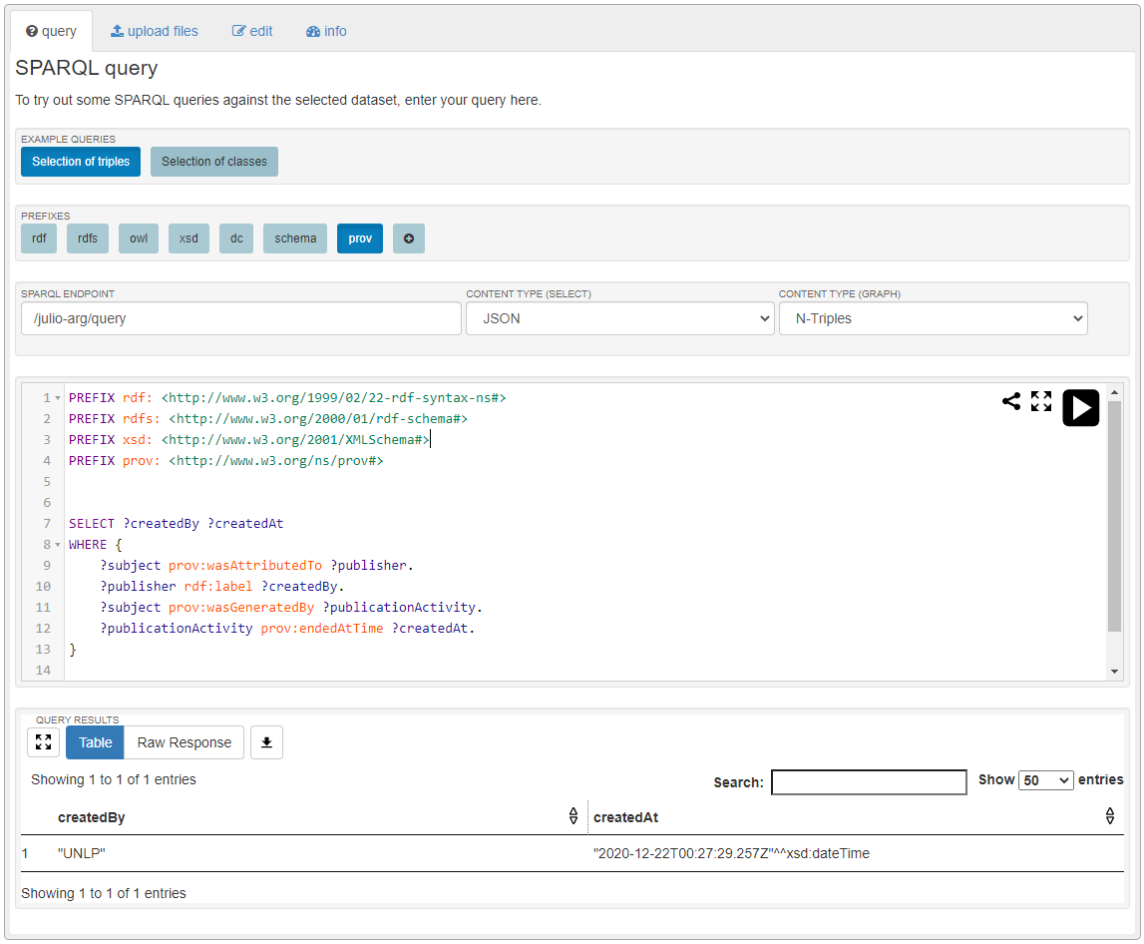

Figura 3.3: Consulta SPARQL en Jena Fuseki

Analizando casos prácticos sobre los datasets obtenidos, Fuseki, mediante SPARQL, nos permitirá demostrar cómo las nuevas representaciones semánticas presentan una mejora sustancial al momento de realizar consultas de datos respecto a la modalidad en la que los mismos se publicaron originalmente.

3.2.4. Otras herramientas

Además de las herramientas mencionadas, se utilizarán algoritmos ad hoc que darán soporte a distintas partes del pipeline. Cada uno de estos algoritmos tendrá una función bien determinada y se los describirá oportunamente a lo largo de este trabajo de tesis. Se podrá encontrar también esta descripción en el apéndice A junto con su ubicación en el repositorio de la tesina.

Capítulo 4

Ontología de datos normativos relativos al LMR

Una ontología es una representación formal de un dominio, que provee un estándar semántico que puede emplearse para anotar datos donde se definen conceptos clave, así como las relaciones que existen entre esos conceptos[19]. Las ontologías permiten que los conceptos del dominio representados sean fáciles de interpretar a la vez que facilitan la interoperabilidad.

4.1. Redes de ontologías

4.1.1. Requerimientos de una ontología de LMR

A continuación analizaremos los requerimientos que debería satisfacer una ontología para representar el dominio de la publicación de datos normativos respecto a los LMR de fitosanitarios en alimentos. Para ello utilizaremos como referencia la guía para la especificación de requerimientos de ontologías propuesta por la metodología NeOn[20].

1. **Propósito, alcance y lenguaje de implementación.**

   El propósito de la nueva ontología es brindar marco estandarizado para la publicación de datos relativos a las tolerancias máximas de residuos de productos fitosanitarios en alimentos que facilite la interoperabilidad. De esta manera se busca brindar facilidades para combinar y comparar datos de distintas fuentes. Finalmente se desea proveer información sobre el origen de los datos y de los procesos de transformación y curación de los mismos. El lenguaje de implementación seleccionado es OWL.

2. **Usuarios finales**

   Los usuarios finales de la ontología son:

   - Miembros de organismos encargados de la publicación de datos relativos a los LMR y/o de definición de políticas públicas.

- Investigadores de áreas a fines a la salud alimentaria y producción de alimentos.
- Los productores de alimentos.

3. **Usos**

   Los principales usos que motivan a esta ontología son:

   - Publicar datos relativos a los LMR de fitosanitarios en alimentos.
   - Realizar consultas sobre datos normativos.
   - Comparar y realizar consultas sobre diferentes dataset.
   - Realizar auditoría sobre el origen de los datos y los cambios que estos sufrieron.
   - Facilitar el procesamiento de los datos por parte de computadoras

4. **Requerimientos**

   Para expresar los requerimientos de la ontología utilizamos la técnica de preguntas de competencia, como también se utiliza en la metodología NeOn. La ontología debería poder responder al menos a las preguntas que podemos ver en la tabla 4.1.

Pregunta	Respuesta
¿Cuál es el límite máximo de residuo para el principio activo Procloraz en tomate?	2mg/kg
¿Qué fungicidas puedo utilizar para el cultivo Manzana?	Flutriafol, Flusilazole, Folpet, Fosetil Aluminio, ...
¿Qué principiso activos puedo utilizar para el cultivo Pimiento?	Gamacialotrina, Hidróxido de cobre, Imadacloprid, ...
¿Cuántos principios activos diferentes se publican en determinado dataset?	138
¿Existe el principio activo Tetraconazole en el dataset de Argentina?	Si
¿Qué principios activos están presentes en el dataset de Argentina y no en el de Brasil?	Terbacil, Tembotrione, Teflutrina,...
¿Qué principios activos están presentes tanto en el dataset de Argentina como en el de Brasil?	Betaciflutrin, Betacipermetrina, Tiametoxam, ...
¿Cuántos usos tiene el principio activo Metribuzin?	7
Dado los datasets de Argentina y Brasil, ¿qué país permite mas residuos de Metsulfuron Metil en Zanahoria?	Argentina
¿Qué cambió en el dataset de Argentina en su última publicación?	Se agregó un registro
¿Quién publicó el dataset de Argentina?	SENASA

Figura 4.1: Requerimientos de la ontología

4.1.2. Ontologías existentes y relevantes a LMR

Teniendo en cuenta los requerimientos mínimos planteados en el punto anterior, y en pos de garantizar la interoperabilidad y de no "reinventar la rueda", se describen a continuación qué ontologías existen hoy que cubren algún punto del dominio planteado.

Agrontology - la ontología de AGROVOC

Agrontology es un vocabulario OWL que proporciona un conjunto de propiedades específicas del dominio al tesauro de AGROVOC[21]. AGROVOC es un vocabulario controlado que abarca todos los ámbitos de interés de la Organización de las Naciones Unidas para la

Alimentación y la Agricultura (FAO). Este vocabulario no sólo modela conceptos relacionados a la agricultura sino que también incluye otros aspectos como la pesca, la nutrición, la alimentación, las ciencias forestales y el medio ambiente y cuenta con miles de conceptos disponibles en diferentes idiomas[22].

AGROVOC es un esquema de concepto (del inglés, concept scheme) RDF / SKOS-XL. SKOS, o Sistema de Organización de Conocimiento Simple, es un modelo de datos común para compartir y vincular sistemas de organización de conocimiento a través de la Web. Muchos sistemas de organización del conocimiento, como tesauros, taxonomías o esquemas de clasificación, comparten una estructura similar y se utilizan en aplicaciones parecidas. SKOS captura gran parte de esta similitud y la hace explícita, para permitir el intercambio de datos y tecnología entre diversas aplicaciones[23]. En tanto SKOS-XL define una extensión que proporciona soporte adicional para describir y vincular entidades léxicas[24]. En el esquema de AGROVOC las relaciones se expresan mediante los predicados SKOS "skos:broader" y "skos:narrower". Por ejemplo, dado el concepto `http://aims.fao.org/aos/agrovoc/c_13551` que representa a "Papa", podemos encontrar relacionado mediante "skos:broader" al concepto `http://aims.fao.org/aos/agrovoc/c_8174` utilizado para representar a "Hortaliza". De manera inversa, conceptos como "Papa", "Mandioca" (`http://aims.fao.org/aos/agrovoc/c_9649`) o "Pimiento" (`http://aims.fao.org/aos/agrovoc/c_14727`), entre otros, se relacionan mediante el predicado "skos:narrower" con "Hortaliza".

Para las relaciones no jerárquicas entre conceptos y términos, AGROVOC utiliza la propiedad de SKOS "skos:related", y una serie de subpropiedades de ésta (que agrupadas forman el vocabulario Agrontology)[25]. Por ejemplo, los términos para representar "Derecho de autor" (`http://aims.fao.org/aos/agrovoc/c_a282c214`) y "Regalías" (`http://aims.fao.org/aos/agrovoc/c_0102e3f7`) se encuentran relacionados por la propiedad "skos:related" mencionada.

Agrontology, mediante AGROVOC, puede brindar a nuestra ontología los conceptos y los textos en varios idiomas diferentes de los cultivos que pueden aparecer en una publicación de datos relativos al LMR.

PROV-O - La ontología de Provenance

PROV-O es una ontología que forma parte de la familia de especificaciones PROV, producida por el World Wide Web Consortium o W3C. Esta ontología expresa el modelo de datos PROV (o PROV-DM) usando el lenguaje de ontologías web OWL2[26]. PROV-DM distingue las estructuras centrales, que forman la base de la información de Provenance. PROV-DM está organizado en seis componentes, que tratan respectivamente de:

1. entidades y actividades, y el momento en que fueron creadas, utilizadas o terminadas
2. entidades derivadas de otras entidades
3. agentes responsables de las entidades que se generaron y las actividades que ocurrieron

4. una noción de paquete, un mecanismo para apoyar la procedencia de la procedencia
5. propiedades para vincular entidades que se refieren a lo mismo
6. colecciones que forman una estructura lógica para sus miembros[27]

La familia de especificaciones PROV, está diseñada para promover la publicación de información de Provenance en la Web y ofrece una base para la interoperabilidad entre diversos sistemas de gestión de procedencia. Es además deliberadamente genérico y agnóstico del dominio, y brinda mecanismos de extensión que pueden utilizarse para modelar dominios específicos[28].

PROV-O forma parte de varios sistemas de información de Provenance y ontologías, como la Human Computation Ontology, así como por proveedores de Linked Data, como DBPedia. Es posible encontrar una lista detallada de implementaciones y uso de PROV-O, y del resto de las especificaciones normativas de la familia de documentos PROV, en el Reporte de Implementación de PROV de la W3C.

En PROV-O las clases y propiedades se definen de manera que no sólo se puedan utilizar directamente para representar información de Provenance, sino que también se pueden especializar para modelar detalles de Provenance específicos para una variedad de dominios. Por lo tanto, la ontología puede ser directamente utilizada en aplicaciones y mientras que al mismo tiempo puede servir como modelo de referencia para crear ontologías de Provenance específicas para un dominio.

Es posible que los consumidores de PROV-O solo necesiten utilizar partes de la ontología, según sus necesidades y según la cantidad de detalles que deseen incluir para describir la información de procedencia. Por esta razón, los términos PROV-O (clases y propiedades) se agrupan en tres categorías que proporcionan una introducción incremental a la ontología:

- Términos de punto de partida: proporcionan la base para el resto de la ontología. Estos términos, que podemos ver en la figura 4.2, se utilizan para crear descripciones de Provenance simples que se pueden elaborar utilizando términos de otras categorías.

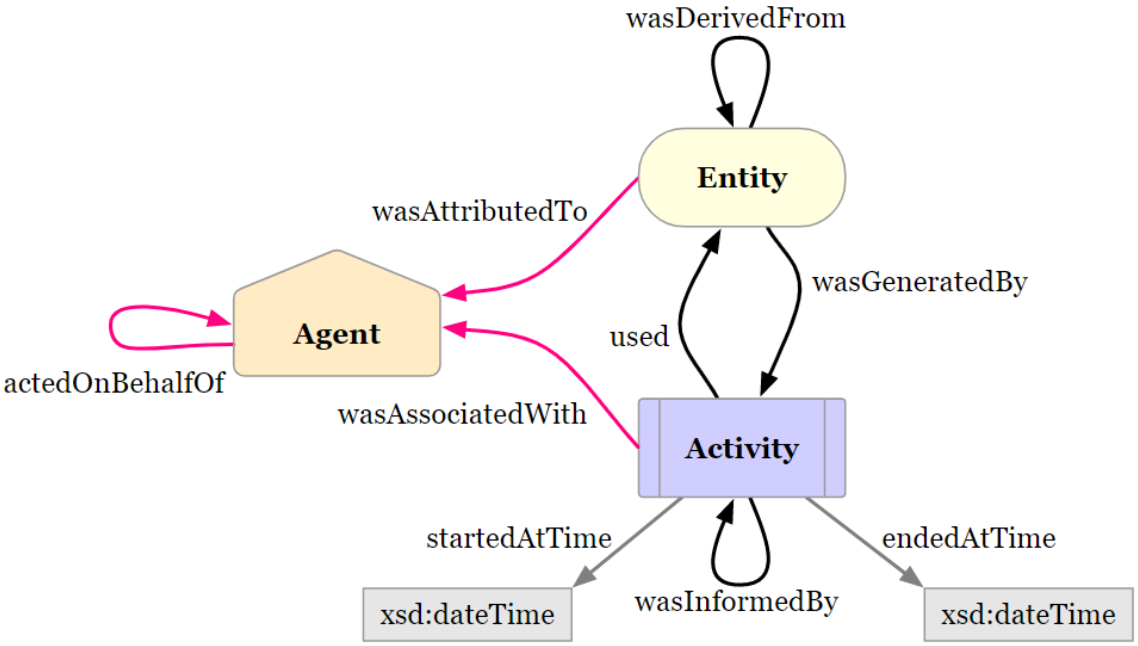

Figura 4.2: Términos de la categoría Punto de Partida

- Términos ampliados: proporcionan términos adicionales que se pueden utilizar para relacionar clases en la categoría Punto de partida. Muchos de los términos de esta categoría son subclases o subpropiedades de los de la categoría Punto de partida.
- Términos para relaciones de calificación: proporcionan información detallada sobre las relaciones binarias afirmadas mediante las propiedades de las categorías Punto de partida y Expansión. Si bien las relaciones de estas últimas categorías se aplican como aserciones binarias directas, los términos de esta categoría se utilizan para proporcionar atributos adicionales de las relaciones binarias.

Es posible partir de Prov-o para registrar el origen de los datos y la información de proveniencia de la ontología propuesta.

ChEBI - Entidades químicas de interés biológico

ChEBI (Chemical Entities of Biological Interest) es una base de datos y ontología de pequeñas entidades moleculares desarrollada por el Instituto de Bioinformática Europeo[1] (o EBI por sus siglas en inglés). ChEBI define como "pequeña entidad molecular" a cualquier átomo, molécula, ión, par iónico, radical, ión radical o entidad que sea constitucional o isotópicamente distinta y que pueda ser distinguible por sí misma. Estas entidades pueden

[1] https://www.ebi.ac.uk/

ser productos de la naturaleza o productos sintéticos utilizados para intervenir en los procesos de los organismos vivos (ya sea deliberadamente, como una droga o un insecticida, o no intencionalmente, como cualquier producto químico que pueda ser encontrado naturalmente en el medio ambiente)[29].

ChEBI es ampliamente usado como referencia para una amplia gama de bases de datos de bioinformática, como por ejemplo UniProt (una base de datos libre de información de proteínas). También se utiliza como base de conocimiento para fines de minería de datos y texto, y como componente químico de varias ontologías, incluida la Ontología Genética o la anteriormente mencionada AgrO. En el contexto de la web semántica, clases ChEBI se utilizan para representar la propiedad tipo (rdf:type) de los químicos en la representación RDF de la base de datos PubChem (una de las más grandes bases de datos abiertas de información química)[30].

La información contenida en ChEBI proviene de múltiples fuentes de datos, algunas incorporadas manualmente y otras mediante referencias cruzadas automáticas. Estas referencias automáticas son añadidas mediante la comparación de nombres y/o estructuras contenidas en las bases de datos de referencia. Algunas de estas bases de datos que se utilizan son:

- Secuencias de proteínas
  - UniProtKB: información respecto a proteínas, su clasificación y función.
- Enzimas
  - BRENDA: Sistema de información de enzimas.
  - IntEnz: Base de datos integral de enzimas
- Pequeñas moléculas
  - NMRShiftDB: base de datos de estructuras orgánicas.
  - PubChem: información de actividades biológicas.
- Interacciones moleculares
  - IntAct Interactions: base de datos de evidencia de interacciones moleculares.
  - CompTox: registro público de toxicología.

Es importante mencionar que toda la información contenida en ChEBI es no propietaria o derivada de una fuente no propietaria, por lo que está disponible para cualquiera.

La ontología ChEBI está dividida en tres sub-ontologías separadas:

- Estructura molecular: en donde entidades moleculares o partes de ellas son clasificadas de acuerdo a composición y estructura. Por ejemplo hidrocarburos, ácidos, etc.
- Rol: se divide en tres subcategorías:

    - Rol químico: clasifica entidades en base al rol que cumplen en un contexto químico. Por ejemplo un inhibidor.
    - Rol biológico: clasifica las entidades en base al rol biológico que estas cumplen. Por ejemplo un antiviral o una hormona.
    - Aplicación: clásica de acuerdo al uso que hace un humano de la entidad. Por ejemplo un herbicida o un insecticida.
- Partícula subatómica: clasifica partículas que son más pequeñas que un átomo. Por ejemplo un electrón o un fotón.

Un ejemplo concreto de roles que puede tener una substancia se puede ver en la imagen 4.3.

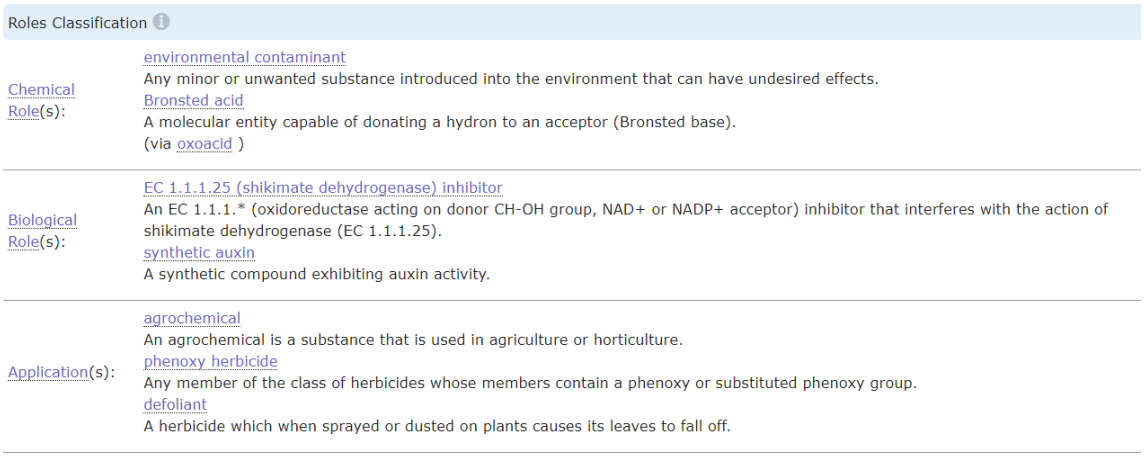

Figura 4.3: Roles de 2,4'-dichlorobiphenyl

La ontología ChEBI posee relaciones que fueron definidas ad hoc, como por ejemplo la relación "is tautometer of" (en español "es tautómetro de") o "is conjugate acid of" ("es ácido conjugado de"). Además de estas relaciones, incorpora dos relaciones definidas en la Ontología de Relaciones, "is a" "y is part of" (en español "es un" y "es parte de")[31]. La Ontología de Relaciones es una colección de relaciones creadas con la intención de formar una base y estandarizar las relaciones entre entidades para futuras ontologías que la incorporen[32]. A continuación se muestran las relaciones en la ontología ChEBI:

- is a
- is part of
- has role
- is conjugate acid of

- is conjugate base of
- is tautometer of
- is enantiometer of
- has functional parent
- has parent hydride
- is substituent group from

Esta ontología puede ser consultada desde su sitio web oficial[2], donde es posible realizar búsquedas y acceder a un sinnúmero de otros recursos gracias a las anteriormente mencionadas referencias cruzadas automáticas, así como también descargada en formato OWL.

ChEBI puede ser de suma utilidad en nuestra futura ontología ya que posee una enorme cantidad de compuestos químicos que se utilizan como principios activos en el ámbito del agro, por lo que se pueden utilizar dichos elementos ya existentes para representar a los fitosanitarios a los que se hace referencia en los datos normativos.

AgrO - Ontología Agronómica

AgrO es una ontología que busca describir prácticas agronómicas, técnicas y variables usadas en experimentos agronómicos[33]. Es desarrollada por un equipo de expertos de centros de investigación del CGIAR (acrónimo para Consultative Group for International Agricultural Research). Esta organización es una asociación global que reúne a varias organizaciones internacionales comprometidas en la reducción de la pobreza rural, el incremento en la seguridad alimentaria, la mejora en la salud y la nutrición y el manejo sustentable de los recursos naturales.

La ontología se construye a partir de rasgos y parámetros identificados por agrónomos, el Diccionario de Datos ICASA (International Consortium for Antimicrobial Stewardship in Agriculture) y otras ontologías existentes como la Ontología del Medio Ambiente, que busca formalizar y estandarizar conceptos relacionados al medio ambiente[34], la Ontología de Unidades, que provee conceptos relacionados a unidades métricas, etc. AgrO se enriquece además con el apoyo de varios científicos que aportan su conocimiento del dominio[35][36].

AgrO es actualmente utilizada por AgroFims, un sistema de gestión de información de campo desarrollado por CGIAR que consiste en una aplicación web la cual se usa para diseñar plantillas de recolección de datos para experimentos agrícolas mediante la selección de variables anotadas con AgrO y otras ontologías relevantes. AgrO ademas es utilizado por la Universidad de Florida (UF) e investigadores asociados con el Proyecto de Mejora e Intercomparación de Modelos Agrícolas (AgMIP) como una terminología de referencia estándar para permitir la generación y reutilización de datos model-ready.

[2]https://www.ebi.ac.uk/chebi/init.do

AgrO se basa en las categorías formales de la Ontología Formal Básica (o BFO por sus siglas en inglés). Esta ontología define conceptos de alto nivel como qué es una entidad, un proceso o una función y es usada como base no solo por AgrO sino por muchas otras ontologías desarrolladas por el colectivo denominado Fundición OBO (Open Biological and Biomedical Ontology). Este colectivo es un conjunto de desarrolladores de ontologías comprometidos con la colaboración y el cumplimiento de principios compartidos y cuya misión es desarrollar una familia de ontologías interoperables que estén lógicamente bien formadas y sean científicamente precisas[37]. Es en pos de la interoperabilidad con otras ontologías que alguno de los términos que se pueden encontrar en AgrO son algo genéricos. Además en Agro se reutilizan términos para evitar la proliferación y duplicación de los mismos entre las diferentes ontologías que la componen. Se muestran en la figura 4.4 las ontologías que forman parte de AgrO.

Ontología	Contenido usado en AgrO	Ejemplo de términos
Ontología Formal Básica (BFO)	Bases de la ontología	“entidad”, “contínuo”, “ocurrencia”
Entidades Químicas de Interés Biológico (ChEBI)	Entidades químicas y sus roles bioquímicos	“rol fertilizante”, “entidad química”
Ontología del Medioambiente (ENVO)	Biomas y entidades	“suelo”, “campo”, “bosque”
Ontología de Artefactos de Información (IAO)	Entidades de información	“tiene timestamp”, “tiene medida”
Ontología de Investigaciones Biomédicas (OBI)	Protocolos e instrumentación	“objetivo logrado por”, “logra objetivo planeado”
Ontología de Características y Fenotipos (PATO)	Cualidades de entidades	“Adyacente a” “contribuye a”, “concentración de”
Ontología de Condición Experimental de Plantas (PECO)	Estudio de tipos, condiciones de crecimiento, tratamientos abióticos	“exposición al contenido de arcilla”, “exposición al contenido de limo”
Ontología de Plantas (PO)	Anatomía de la planta, morfología y crecimiento	“hoja”, “semilla”, “brote”
Ontología de Características de Plantas (TO)	Rasgos fenotípicos en plantas	“rendimiento”
Ontología de Unidades (UO)	Unidades	“m”, “kg”, “l”

Figura 4.4: Ontologías que forman parte de AgrO

AgrO puede ser consultada de manera online desde el repositorio de ontologías biomédicas[3] (OLS) o descargada en formato OWL.

4.2. LMR-O

Uno de los resultados de esta tesina es la ontología de datos relativos al LMR (desde ahora, LMR-O). LMR-O se propone, como su nombre lo indica, representar el dominio de los datos normativos respecto a tolerancias máximas de residuos de productos fitosanitarios en alimentos cuya publicación es generalmente llevada a cabo por organizaciones gubernamentales. Se busca además, poder representar la información de proveniencia (o Provenance) de estos datos así como también los procesos de transformación y curación de los mismos, si es que los hubo.

LMR-O reutiliza y se basa, en gran parte, en términos existentes de otras ontologías, esto evita tener que definir términos ya existentes, proporcionando interoperabilidad a la vez que la vuelven mas sencilla de entender si el usuario esta familiarizado con estas ontologías. Las ontologías utilizadas por LMR-O se pueden observar en la figura 4.5.

Adicionalmente LMR-O brinda un pequeño conjunto de clases y propiedades propios que ayudan a representar el dominio mencionado. Se utilizará el prefijo "lmro:" para indicar entidades que pertenecen a esta ontología.

Ontología	Prefijo	Contenido usado	Términos de ejemplo
Prov-o	prov:	Entidades relativas a Provenance	Entity , Activity, Organization
CHeBI	chebi:	Principios activos y sus aplicaciones	ChemicalEntity, Application, 2,4-D
AGRONTOLOGY	-	Tesauro de cultivos	Tomatoe, Carrot
AGRO	agro:	Entidad cultivo	Crop
Unit Ontology	uo:	Unidades	m, kg, l

Figura 4.5: Ontologías que forman parte de LMR-O

La principal clase de LMR-O es la clase lmro:Record. Esta clase representa un único registro de información normativa sobre LMR. Un lmro:Record establece para un alimento, el límite de residuo máximo para un principio activo en un rol específico. Por ejemplo, un lmro:Record podría representar la información "El límite de residuo máximo para el principio activo Triadimefon si se lo utiliza como fungicida en tomates, es de 0,2mg/kg". Puede suceder en la práctica, que un mismo principio activo se utilice para varios cultivos, o que que tenga varios roles diferentes, en ese caso se utilizará un registro para cada cultivo y rol diferentes.

[3]https://www.ebi.ac.uk/ols/index

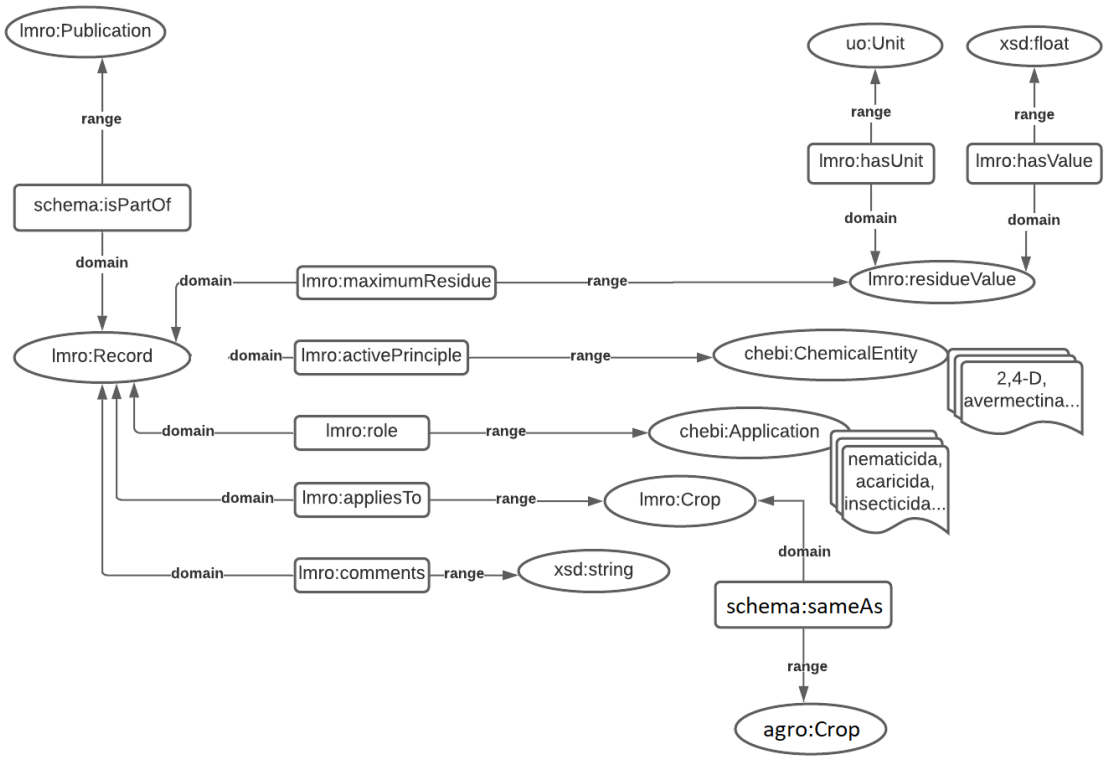

Figura 4.6: Entidades involucradas para representar un registro de información normativa en LMR-O

Como se puede apreciar en la figura 4.6, se utiliza la clase lmro:ResidueValue para determinar el valor y la unidad del LMR. La clase uo:Unit, proveniente de la Ontología de las Unidades de Medida (o, del inglés, Unit Ontology) posee como subclases todas las unidades que podemos llegar a necesitar en nuestro dominio, como miligramos por kilogramo, gramos por kilogramo, partes por millón, etc.

Cada lmro:Record, y por consiguiente cada lmro:ResidueValue, pertenecen a una publicación. Esto está representado por el concepto lmro:Publication, el cual tendrá asociado la información de provenance.

Para establecer el principio activo del registro se utilizan las subclases de chebi:ChemicalEntity, la cual representa entidades moleculares (o sus partes) y sustancias químicas. ChEBI también aporta la clase chebi:Application, la cual determina cuál es el rol del principio activo en el registro (fungicida, insecticida, rodenticida, etc).

LMR-O también hace uso del vocabulario de AGROVOC, el cual proporciona nombres de cultivos y sus traducciones a varios idiomas diferentes.

```
<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11>
  a <http://www.lifia.info.unlp.edu.ar/lmro#Record> ;
  ns0:activePrinciple <http://purl.obolibrary.org/obo/CHEBI_28854> ;
  ns0:role <http://purl.obolibrary.org/obo/CHEBI_24527> ;
  ns0:appliesTo <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11/Crop> ;
  ns0:maximumResidue <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11/ResidueValue> .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11/Crop>
  a ns0:Crop ;
  ns0:a <http://aims.fao.org/aos/agrovoc/c_6599> ;
  rdf:label <http://aims.fao.org/aos/agrovoc/xl_es_1299486832833> .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11/ResidueValue>
  a ns0:ResidueValue ;
  ns0:hasValue "0.05"^^xsd:float ;
  ns0:hasUnit <http://purl.obolibrary.org/obo/UO_0000308> .
```

Figura 4.7: Un registro LMR y la información asociada

Se puede observar en la figura 4.7 como quedaría representada la información de un registro de LMR en formato turtle. Específicamente este registro nos indica que el principio activo "2,4-D" (representado por la uri de ChEBI `https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI_28854`) puede utilizarse como Herbicida (otro recurso de ChEBI `http://purl.obolibrary.org/obo/CHEBI_24527`) en el cultivo "arroz" (recurso AGROVOC `http://aims.fao.org/aos/agrovoc/c_6599.html` cuyo label en español es `http://aims.fao.org/aos/agrovoc/xl_es_1299486832833.html`). A su vez el límite de residuo máximo, indicado por el resurso `http://www.lifia.info.unlp.edu.ar/lmro/data/ar/11/ResidueValue`, es de 0.05 miligramos por kilogramo (recurso de la Ontología de las Unidades `http://purl.obolibrary.org/obo/UO_0000308`). Este registro pertenece a la publicación identificada por la uri `http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publication`.

Otro ejemplo interesante ocurre cuando el principio activo está exento, es decir, no hay un valor máximo de residuo fijado. La figura 4.8 muestra precisamente este caso, donde la sustancia "Avermectina" se encuentra exenta para el cultivo "Semilla de Arroz". Esto es representado con el concepto lmro:Exempt.

```
<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/41>
  a <http://www.lifia.info.unlp.edu.ar/lmro#Record> ;
  ns0:activePrinciple <http://purl.obolibrary.org/obo/CHEBI_50344> ;
  ns0:role <http://purl.obolibrary.org/obo/CHEBI_22153> ;
  ns0:appliesTo <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/39/Crop> ;
  ns0:maximumResidue ns0:Exempt .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/39/Crop>
  a ns0:Crop ;
  ns0:a <http://aims.fao.org/aos/agrovoc/c_25473> ;
  rdf:label <http://aims.fao.org/aos/agrovoc/xl_es_1299487280829> .
```

Figura 4.8: Registro LMR con principio activo exento

Existen casos donde los datasets no poseen información sobre el rol que cumple un prin-

cipio activo en su aplicación sobre un cultivo. A fin de que el dataset sea lo más consistente posible, en estos casos el rol será representado por el concepto lmro:Any. Un ejemplo de esto se puede ver en la figura 4.9.

```
<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/5>
  a <http://www.lifia.info.unlp.edu.ar/lmro#Record> ;
  ns0:activePrinciple <http://purl.obolibrary.org/obo/CHEBI_73173> ;
  ns0:appliesTo <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/5/Crop> ;
  ns0:maximumResidue <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/5/ResidueValue> ;
  ns0:role ns0:Any .
```

Figura 4.9: Registro LMR con rol Any

Como se puede observar en la figura 4.10, LMR-O posee información relativa al origen de los datos mediante el uso de clases y propiedades de PROV-O. Específicamente se utilizan las tres clases primarias de PROV-O: prov:Entity, prov:Activity y prov:Agent. Regularmente las organizaciones gubernamentales, tales como SENASA en Argentina o ANVISA en Brasil, ponen a disposición en sus portales web la información relativa al LMR en alimentos. Esto puede ser una tabla en excel, una página web interactiva, un archivo en formato CSV etc. Este documento, que luego utilizaremos como input para nuestro proceso de transformación a un dataset semántico, es de suma importancia desde el punto de vista de la Provenance ya que nos indica de dónde provienen los datos.

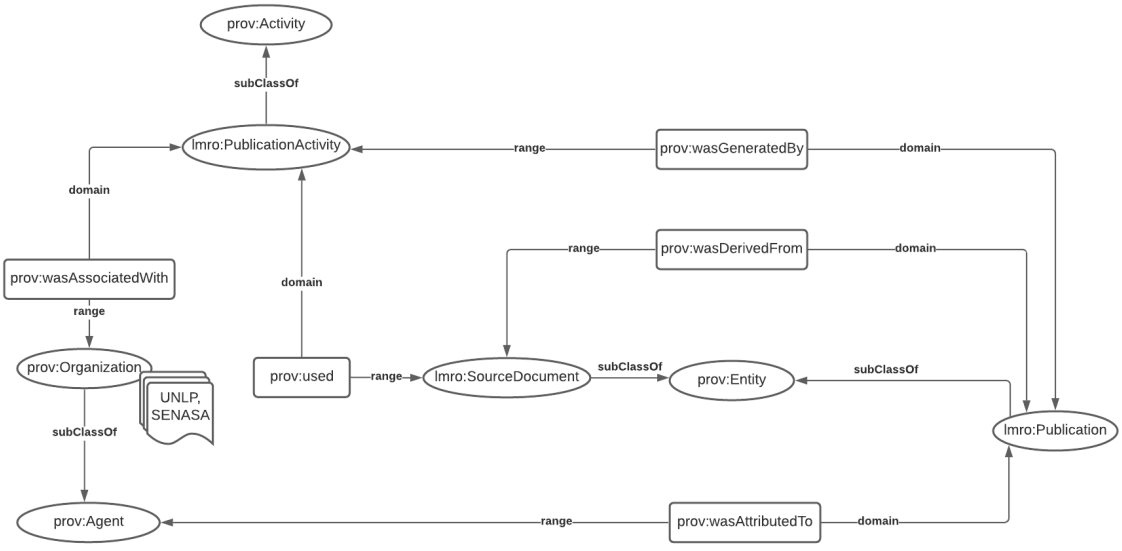

Figura 4.10: Entidades involucradas para representar el origen de los datos en LMR-O

En LMR-O utilizaremos el concepto lmro:SourceDocument para registrar esta información. SourceDocument utiliza términos que provee el Conjunto de Metadata de DublinCore,

un vocabulario de conceptos utilizados para describir recursos. Utilizaremos de dicho volabulario las propiedades descriptas en la figura 4.11.

Término	Uri	Descripción
creator	`http://purl.org/dc/elements/1.1/creator`	La entidad primaria responsable de crear el recurso.
date	`http://purl.org/dc/elements/1.1/date`	Un punto o período en el tiempo asociado a un evento en la vida del recurso.
format	`http://purl.org/dc/elements/1.1/format`	El formato, medio físico o dimensiones del recurso. Una buena práctica es utilizar como formato un tipo MIME (ej. xlsx, doc, etc).
language	`http://purl.org/dc/elements/1.1/language`	Idioma del recurso.
description	`http://purl.org/dc/elements/1.1/description`	Una descripción libre del recurso.
title	`http://purl.org/dc/elements/1.1/title`	Es el nombre del recurso.
identifier	`http://purl.org/dc/elements/1.1/identifier`	Una referencia inequívoca del recurso. Utilizaremos un hash sha256.

Figura 4.11: Términos de Dublin Core utilizados en LMR-O

Otro aspecto interesante a modelar es la tarea de generación de nuestro dataset semántico. Utilizaremos para esto el concepto lmro:PublicationActivity, el cual es subclase de prov:Activity. De la misma manera que un lmro:SourceDocument está asociado a la organización que lo generó, la actividad de publicación utiliza la propiedad prov:wasAssociatedWith para describir dicha organización (prov:Organization). Es importante aclarar que la entidad que genera el dataset original puede ser diferente a la que genera el dataset semántico. En los casos de estudio de esta tesis se utilizarán conjuntos de datos generados por SENASA y ANVISA, mientras que el resultado final es un dataset generado por la Universidad Nacional de La Plata. Utilizaremos la propiedad schema:sameAs para asociar las organizaciones con los recursos correspondientes en wikidata.

La actividad de publicación hará uso del documento fuente (indicado con la propiedad prov:used) para generar la publicación (lmro:Publication) que contiene todos los registros LMR.

```
<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/PublicationActivity>
  a ns0:PublicationActivity ;
  prov:wasAssociatedWith <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publisher> ;
  prov:startedAtTime "2020-12-22T00:27:29.236Z"^^xsd:dateTime ;
  prov:used <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/SourceDocument> ;
  prov:endedAtTime "2020-12-22T00:27:29.257Z"^^xsd:dateTime .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Organization>
  a prov:Organization ;
  rdf:label "SENASA" ;
  schema:sameAs <https://www.wikidata.org/wiki/Q6972721> .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publisher>
  a prov:Organization ;
  rdf:label "UNLP" ;
  schema:sameAs <https://www.wikidata.org/wiki/Q784171> .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publication>
  a ns0:Publication ;
  prov:wasDerivedFrom <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/SourceDocument> ;
  prov:wasGeneratedBy <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/PublicationActivity> ;
  prov:wasAttributedTo <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publisher> .

<http://www.lifia.info.unlp.edu.ar/lmro/data/ar/SourceDocument>
  a ns0:SourceDocument ;
  dc11:creator "http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Organization" ;
  dc11:date "2020-07-02T00:00:00.000Z" ;
  dc11:format "xlsx" ;
  dc11:language "Español" ;
  dc11:description "Informacion respecto a los LMR publicada por SENASA en el mes de Julio 2020" ;
  dc11:title "lmrs_julio_2020" ;
  dc11:identifier "802EF781D909A6B3B501F3242F924070213B8F8DC012502C66B2464C976F598C" .
```

Figura 4.12: Información de Provenance relativa a la publicación de SENASA del mes de Julio 2020

Podemos ver en 4.12 la información de Provenance de un dataset semántico.

LMR-O y sus términos están anotados utilizando las propiedades de metadata más comunes que se deberían tener en cuenta a la hora de describir una nueva ontología[38]. Para esto se reusan propiedades de vocabularios existentes, lo que aumenta la interoperabilidad y legibilidad de la ontología.

La figura 4.13 identifica a los elementos que componen la ontología LMR-O que, como se mencionó en 3.2.2, fue creada utilizando la herramienta Protégé.

Nombre	Descripción
lmro:Record	Registro LMR (límite máximo de residuo).
lmro:ResidueValue	El valor y la unidad que tiene el registro LMR.
lmro:Crop	Un cultivo. Es equivalente (schema:sameAs) a la entidad Crop de la ontología Agro (`http://purl.obolibrary.org/obo/AGRO_00000325`).
lmro:Publication	La publicación que se genera con la PublicationActivity. Contiene los registros LMR.
lmro:PublicationActivity	Actividad que genera la publicación con los registros de LMR.
lmro:SourceDocument	Documento fuente utilizado para crear el dataset semántico. Se utiliza como input en la actividad de publicación.
lmro:maximumResidue	Residuo máximo permitido para un fitosanitario.
lmro:activePrinciple	Los principios activos son las sustancias que se utilizan como bactericidas, fungicidas, insecticidas, herbicidas, etc.
lmro:role	Rol que cumple un fitosanitario. Ejemplos de roles son fungicida, herbicida, molusquicida, insecticida, acaricida, regulador de crecimiento, etc.
lmro:appliesTo	Cultivo sobre el que se utiliza el principo activo.
lmro:comments	Comentarios, información adicional.
lmro:hasUnit	Unidad en que se mide el LMR.
lmro:hasValue	Valor que posee un registro LMR.
lmro:Any	Cualquier valor. Un LMR que aplica a cualquier valor aplica a cualquier cultivo del dataset.
lmro:Exempt	No sujeto a restricciones. Que un fitosanitario posea LMR exento significa que no hay una restricción en cuanto a la cantidad de residuo que puede dejar el dicha sustancia

Figura 4.13: Elementos que componen LMR-O

Capítulo 5

Pipeline de transformación a la ontología propuesta

5.1. Pasos generales

Como mencionamos anteriormente, cada organización publica sus datos normativos respecto al LMR de la manera que más cree conveniente. Esto hace que los datasets sean muy heterogéneos entre sí. Se pueden encontrar conjuntos de datos publicados en formato web, tablas en formato Excel o CSV, listas en formato Word o archivos PDF que no sólo contienen los datos normativos sino también descripciones del principio activo y de su estructura química como es el caso del dataset que publica la Agencia Nacional de Vigilancia Sanitaria (ANVISA) de Brasil entre otros. Esto significa que los pasos que deberán llevarse a cabo para transformar un conjunto de datos específico a la ontología propuesta pueden variar radicalmente.

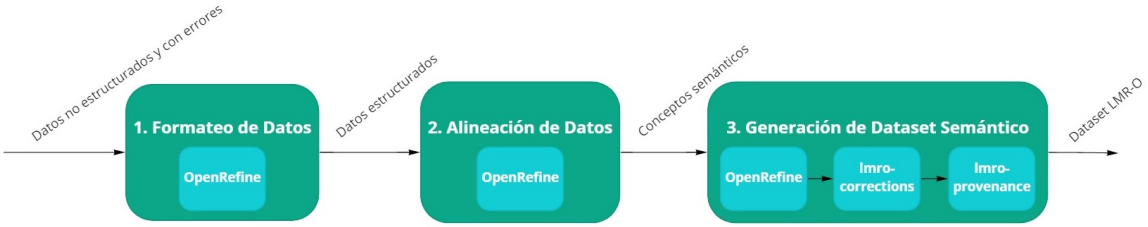

Figura 5.1: Pipeline de transformación de datos

Sin embargo, los pasos que daremos estarán enmarcados en el pipeline de transformación de datos a la ontología propuesta, cuyo espíritu describimos en la sección 3.1. Como se puede ver en la figura 5.1, cada paso tiene un objetivo específico y genera la entrada para el paso siguiente. El objetivo del pipeline es contar con un proceso reproducible, robusto y efectivo (con sus herramientas de soporte) para la tarea de transformación de los datos en los formatos originales al dataset en formato semántico.

5.1.1. Formateo del dataset

Darle formato al dataset significa realizar transformaciones simples al conjunto de datos de modo que cada valor cumpla con las siguientes condiciones:

1. El dato representa una única entidad. Por ejemplo, si el dataset consiste en una tabla que posee la columna Cultivos, entonces cada celda de dicha columna debe representar a un único cultivo. Un valor que cumpla esta condición sería "Arroz". Por el contrario "Arroz, Cebada" no la cumple.

2. La cadena de texto representa un valor y no posee agregados de ningún tipo. Si tuviéramos el valor " Tomate (1) " deberíamos expresarlo como "Tomate".

3. El dato representa a un único valor y a una única representación del mismo. En algunos casos los datos representan una entidad y alguna forma alternativa de de describirla. Por ejemplo, el valor "abamectina/avermectina". En este caso deberíamos eliminar una de las dos formas.

4. El dato no es una abreviación salvo que ésta sea ampliamente utilizada y reconocida en el dominio. Por ejemplo, la abreviación "in" no es una abreviación reconocida y debería expresarse como "insecticida" para obtener más probabilidades de alineación automática. Por otro lado, la abreviación "mcpa" es comúnmente utilizada en el ámbito químico para representar al compuesto "ácido 2-metil-4-clorofenoxiacético" por lo que su sustitución no brindaría grandes beneficios.

5. El dato no posee espacios innecesarios, como espacios al principio o final de la cadena de texto o múltiples espacios entre palabras.

6. La forma de representar los datos debe ser consistente. Por ejemplo, si representamos el valor "arroz con cáscara" luego no sería consistente utilizar el valor "arroz c/cáscara" o "arroz (con cáscara)"

Dependiendo del dataset, esto podría consistir en quitar espacios innecesarios, explotar filas que contienen varios valores, eliminar diferentes formas de representar un mismo valor y cualquier otra operación necesaria para que los datos sean consistentes entre sí y lo mas parecidos posibles a sus representaciones semánticas. En el caso de que estuviésemos construyendo un dataset nuevo, considerar las consignas mencionadas anteriormente nos ayudará no sólo a tener más efectividad a la hora de alinear los datos con sus equivalentes semánticos sino también a que el dataset sea mas fácil de interpretar y más mantenible.

5.1.2. Alineación de datos

Llamamos alinear o reconciliar los datos, a transformar las cadenas de caracteres (strings) que encontramos en los datasets en recursos de la web semántica. La eficacia de la alineación depende de la calidad y la completitud del vocabulario. Esto significa que la elección

de la ontología que se utilice es un aspecto fundamental para obtener un alto porcentaje de éxito en la alineación. En LMR-O utilizamos la ontología de entidades químicas ChEBI, la cual abarca el mismo dominio que los principios activos utilizados para la publicación de los LMR, por lo cual sería de esperarse una alta efectividad de alineación. Por otro lado, utilizamos el tesauro de AGROVOC para asociar los productos alimenticios con sus contrapartes semánticas. Este tesauro cubre el dominio de la agricultura, el cual se intersecta con el nuestro de productos alimenticios y cultivos, pero que no son exactamente lo mismo. Por esta razón, la eficacia de alineación puede variar ampliamente según cuánto más se acerque el dataset al dominio de la industria alimenticia o, por el contrario, cuánto más se acerque al dominio del agro y la botánica.

Utilizaremos un dump de la ontología ChEBI como servicio de reconciliación en la extensión RDF de OpenRefine, lo que nos permitirá alinear los datos respecto a principios activos así como también el rol que cumplen los mismos. Así la cadena de texto 2,4-D pasa a ser chebi:28854 (el término semántico que lo representa en la ontología ChEBI) o la cadena de texto para el rol "Herbicida" se convertirá en chebi:24527. Es importante mencionar que esta ontología se encuentra en inglés, y muy pocos de los conceptos que la componen tienen algún sinónimo en español. Es por esto que las alineaciones automáticas que encontremos serán de términos cuya representación es muy parecida o directamente la misma en ambos idiomas (2,4-D se representa igual tanto en español como en inglés). Esto significa que, contrario a lo esperado, muy pocos términos serán alineados correctamente de manera automática. Para lograr sortear este problema, extenderemos la ontología ChEBI agregando las traducciones al español de aquellos conceptos que estén dentro del alcance de esta tesis.

Otros valores que debemos alinear son los cultivos. Para estos valores se usará el tesauro AGROVOC, el cual posee "labels" en varios idiomas, incluyendo el español.

5.1.3. Generación del dataset semántico

El último paso consiste en generar el archivo RDF en base a los resultados obtenidos de aplicar los pasos 1 y 2. Para esto la extensión RDF de OpenRefine ofrece la posibilidad de crear una plantilla indicando como será para cada fila la representación en RDF. Esta plantilla la podremos exportar en formato GREL y podría ser reutilizada en sucesivas actualizaciones del mismo dataset e incluso, si respetamos los nombres de columnas utilizados, para diferentes conjuntos de datos.

OpenRefine genera una tripleta o conjunto de tripletas por cada una de las filas existentes en nuestro dataset. Sin embargo, algunas de las entidades que deben ser creadas para representar la información de proveniencia son únicas (por ejemplo el concepto de Actividad de Publicación mencionado en 4.10). También existen casos en que, aun no siendo únicos, hay conceptos que no deberían generarse una vez por fila. Tomemos el caso del cultivo "tomate", donde vamos a varias sustancias activas, y por ende varios registros en el dataset representando esta información. Sería incorrecto generar tantos conceptos "tomate" como filas en la tabla.

Otra limitación al generar las tripletas RDF ocurre en el paso de alineación de datos,

donde se relacionan los nombres de cultivos con los correspondientes labels de AGROVOC para un idioma en particular. Siguiendo con el tomate como ejemplo, en un dataset en español, la alineación se dará dado con el concepto cuyo URI es `http://aims.fao.org/aos/agrovoc/xl_es_1299487137447` (el label "Tomate"). Por otro lado, en un dataset en italiano este mismo término se alineará con la URI `http://aims.fao.org/aos/agrovoc/xl_it_1299487137562` (el label "Pomodori"). Esto a futuro dificultará las consultas entre datasets en diferentes idiomas, ya que obligará al analista a saber cómo se llama al cultivo en cada idioma que desea buscar. AGROVOC proporciona conceptos más amplios que sortean este tipo de dificultades. Sería deseable asociar el término "Tomate" al concepto `http://aims.fao.org/aos/agrovoc/c_7805`, que es al que se asocian cada uno de los labels de skosxl que expresan "tomate" en cada idioma disponible. Esto, sin embargo, no es posible desde OpenRefine ya que este concepto, el `http://aims.fao.org/aos/agrovoc/c_7805`, no posee labels propios que el servicio de reconciliación pudiera utilizar al momento de buscar candidatos para los términos de cada registro, sólo referencias a ellos. Vemos en la figura 5.2 la relación entre estos conceptos.

```
<!-- Concepto tomate -->
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/xl_es_1299487137447">
      <notation xmlns="http://www.w3.org/2004/02/skos/core#"
rdf:datatype="http://aims.fao.org/aos/agrovoc/AgrovocCode">7805</notation>

<!-- Label español -->
</rdf:Description><rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/xl_es_1299487137447">
     <literalForm xmlns="http://www.w3.org/2008/05/skos-xl#" xml:lang="es">Tomate</literalForm>
</rdf:Description>

<!-- Label italiano -->
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/xl_it_1299487138308">
     <literalForm xmlns="http://www.w3.org/2008/05/skos-xl#" xml:lang="it">Pomodori</literalForm>
</rdf:Description>

<!-- Union del concepto con labels -->
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_7805">
     <prefLabel xmlns="http://www.w3.org/2008/05/skos-xl#"
rdf:resource="http://aims.fao.org/aos/agrovoc/xl_es_1299487137447"/>
     <prefLabel xmlns="http://www.w3.org/2008/05/skos-xl#"
rdf:resource="http://aims.fao.org/aos/agrovoc/xl_it_1299487137562"/>
</rdf:Description>
```

Figura 5.2: Extracto de la estructura de AGROVOC

Dadas estas dificultades, el paso de generación del dataset semántico se valdrá, en primera instancia, de la extensión RDF de OpenRefine para realizar una primera versión del conjunto de datos. En segundo lugar, se utilizarán dos algoritmos ad hod que permitirán generar los conceptos relativos a cultivos y valores de LMR por un lado, y aquellos relacionados al origen de los datos por el otro.

El primer algoritmo, denominado "lmro-corrections", tiene como función principal resolver todos los inconvenientes mencionados anteriormente y realizar las correcciones necesarias para que el dataset exportado de OpenRefine, además de ser semántico, sea compatible

con la ontología LMR-O. Este sencillo script recorre cada uno de los registros lmro:Record exportados y realiza las tres operaciones siguientes:

1. **Agregar el concepto lmro:Any para el rol**

   Si el registro no posee rol, entonces se asume que el principio activo puede cumplir cualquier rol.

2. **Resolver el valor del LMR**

   Si el registro no posee un valor para el LMR válido (es decir, numérico), entonces se asume que el principio activo está exento. Esto se representa mediante el concepto lmro:Exempt. En caso contrario, cuando el valor es válido, se crea un concepto lmro:ResidueValue con el valor numérico y la unidad de medida (por ejemplo mg/Kg o partes por millón).

3. **Resolver el valor del cultivo**

   Si el registro no posee un valor para la propiedad lmro:appliesTo, entonces se asume que el principio activo es válido para cualquier cultivo. Esto se representa mediante el concepto lmro:Any. Por otro lado, si lmro:appliesTo posee un valor, este puede ser una URI, identificando un label de AGROVOC, o una cadena de caracteres indicando que el cultivo no se pudo alinear con la ontología. Para el primer caso, se buscará el concepto más amplio (denominado broader) en AGROVOC y se lo utilizará en lugar del label original. Si el valor es una cadena de caracteres, la propiedad se mantendrá sin cambios.

El segundo algoritmo será el encargado de la generación de entidades de proveniencia para LMR-O. Este script utilizará como entrada el dataset semántico generado en el paso anterior, en donde se añadirá la nueva información, así como también un pequeño archivo en formato JSON con información adicional. Esta información consiste en la URI que se utilizará para los nuevos conceptos generados, la información de la organización que publica el nuevo dataset (en nuestro caso, La Universidad Nacional de La Plata) y finalmente información relativa al archivo fuente que se utilizó para generar el conjunto de datos.

Más información de estos scripts puede verse en el apéndice A.

5.2. Extendiendo ChEBI

Gran parte del éxito que se obtiene en la generación del dataset semántico depende directamente de cuán completas son las ontologías que se utilizan para relacionar simples strings de texto con entidades existentes en la web semántica.

La ontología ChEBI se encuentra mayormente en inglés y posee pocos conceptos expresados en otros idiomas. Esto quiere decir, que muchos términos de los datasets que son utilizados en los casos de estudio, si bien se encuentran presentes en la ontología, no se alinean con ningún término. Existen también casos en los que el concepto se encuentra expresado de una manera alternativa en el dataset a como está en la ontología (es el caso de

los acrónimos por ejemplo). Finalmente hay en los datasets términos que no se encuentran en la ontología, ya sea porque son de un dominio diferente o porque aun no fueron incluídos en la misma.

Es preciso sortear estas dificultades antes de aplicar los pasos de generación del nuevo dataset semántico. Esto nos permitirá concentrarnos en la información del conjunto de datos y no tanto en las carencias que esta ontología (y cualquier otra) pudiera tener. Por este motivo, se decidió extender ChEBI basándonos nuevamente en OpenRefine y en dos scripts específicos que harán mas sencillo agregar sinónimos para diferentes representaciones de conceptos ya existentes y la creación de aquellos que no existan.

5.2.1. Asociando conceptos

Idealmente, esta es una tarea que debería hacerse automáticamente con algún servicio de traducción. En este trabajo, debido a lo específico del dominio (entidades químicas) y la escasez de servicios de traducción disponibles lo haremos de manera manual.

El proceso, si bien algo tedioso, es sencillo y es una tarea que se realizará una sola vez por idioma, al menos para tantos términos. Podemos dividir este proceso en cuatro pasos:

1. Alinear con la ontología ChEBI original las columnas Principio Activo y Aptitud.

2. Para los términos que no se hayan alineado, buscaremos en ChEBI el concepto que los representa, ya sea una traducción al inglés o una manera alternativa de expresarlo. Muchos términos serán similares en inglés, como en el caso de “acetoclor” o “herbicida” cuyas traducciones son “acetochlor” o “herbicide” respectivamente. Sin embargo habrá otros más difíciles de identificar, como “m.c.p.a.” cuya representación en ChEBI es “(4-chloro-2-methylphenoxy)acetic acid”. En estos casos será preciso investigar un poco más para dar con el término correcto.

3. Una vez alineados todos (o casi todos) los términos de la ontología, exportamos el archivo GREL que genera OpenRefine con todos los pasos que hicimos cuando alineamos manualmente los conceptos.

4. Finalmente utilizaremos un algoritmo construído para la ocasión con el fin para insertar los nuevos términos y, generar así, la versión extendida de ChEBI.

5.2.2. Creando nuevos conceptos

Los términos para los que no encontramos conceptos existentes en ChEBI deberán ser agregados a la ontología. Es posible filtrar en OpenRefine por conceptos no alineados de manera de obtener rápidamente estos elementos.

El algoritmo que se utilizará para crear los conceptos nuevos será un pequeño script javascript que correremos utilizando node. El script tendrá dos entradas, la ontología original y la lista de elementos que serán insertados en formato JSON. En cada elemento de la lista se especifican sus atributos y contenido, pudiendo agregar estructuras anidadas. El

script recorrerá la lista de elementos, agregando los nuevos términos con sus atributos, namespaces, elementos anidados y contenido en general. Cabe destacar que este algoritmo debería poder utilizarse para extender cualquier tipo de archivo en formato RDF. Las listas que se generaron para los casos de estudio se podrán encontrar en el repositorio de esta tesis en GitHub.

5.2.3. Agregando sinónimos

Este algoritmo será muy similar al mencionado en el punto anterior. Nuevamente el script tendrá dos entradas, la ubicación del archivo con la ontología a modificar y el archivo GREL donde asociamos cada término, y su salida será la ontología extendida. En este caso usaremos como ontología de entrada al archivo que generamos con el algoritmo de la subsección 5.2.2, de esta manera nuestra nueva versión de ChEBI tendrá tanto conceptos nuevos como sinónimos para conceptos ya existentes.

El algoritmo es sumamente sencillo. Iteraremos sobre cada elemento del archivo GREL, de el obtendremos el concepto alineado (por ejemplo "herbicida") y el concepto con el cual se alineó (por ejemplo "herbicide"). Buscaremos éste último en la ontología original, la cual recibimos como input en el algoritmo, y agregamos el sinónimo. Debido a una limitación de OpenRefine, la herramienta sólo toma el primer label que encuentra para cada concepto. Esto quiere decir, que la etiqueta que usemos para agregar el nuevo sinónimo no debe estar repetida. La ontología ChEBI utiliza las etiquetas rdfs:label, oboInOwl:hasExactSynonym, oboInOwl:hasRelatedSynonym de manera que, para no repirlas, utilizaremos skos:prefLabel.

Una vez ejecutado el algoritmo, tendremos nuestra versión extendida de ChEBI (a partir de ahora, ChEBI Extendido), la cual utilizaremos como servicio de reconciliación en OpenRefine.

Capítulo 6

Evaluación

Como se mencionó en la sección 1.3, una de las problemáticas actuales es la dificultad que existe al realizar operaciones básicas de consulta dentro de un dataset. Tampoco es posible para un investigador, de manera sencilla, realizar comparaciones de datos entre diferentes datasets o diferentes versiones del mismo. Ya sea que la publicación de los datos sea en portales interactivos, en archivos pdf o en planillas de cálculo, el analista posee pocas herramientas para procesar la información. En general, estas herramientas se reducen a búsqueda de cadenas de texto y ordenación por criterios limitados.

La evaluación del pipeline de transformación de datos propuesto se hará en dos partes.

En primer lugar, se evaluará la factibilidad de aplicar el pipeline en dos casos reales: la publicación de datos relacionados al LMR por parte del SENASA en Argentina y lo propio por parte de ANVISA en Brasil.

En segundo lugar, se demostrarán las ventajas que poseen los nuevos dataset semánticos respecto de sus contrapartes convencionales, haciendo hincapié en cuanto a la facilidad que proporcionan las tecnologías de la web semántica para consultar estos conjuntos de datos

Utilizando el lenguaje de consulta SPARQL, demostraremos cómo un consumidor de los datos podrá, utilizar funciones de agregación y manipulación de datos mucho más poderosas. Podrá también realizar consultas entre diferentes versiones de los datasets o incluso entre datasets de diferentes organismos, los cuales pudieron haber sido publicados en diferentes idiomas y formatos. Esta modalidad tiene como ventaja asociada, a su vez, la posibilidad de exportar el resultado de las consultas en formato RDF. Anteriormente, exportar subconjuntos de dos datasets publicados en formatos diferentes (por ejemplo, xlsx y csv) no era posible sin al menos la transformación de uno de los datasets o la copia manual por parte del usuario de los datos deseados.

Adicionalmente, con el nuevo formato de datos semánticos es posible consultar la información de proveniencia de diferentes datasets dentro de una organización.

Se describirán a continuación algunos ejemplos de consultas sobre diferentes escenarios.

6.1. Caso práctico I: El dataset de Argentina

6.1.1. Descripción del dataset

La norma 934-2010 del Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA) establece los requisitos que deben cumplir los productos y subproductos agropecuarios para consumo interno. En esta resolución se regulan aspectos sobre los límites máximos de residuos de plaguicidas, se establece un listado de productos fitosanitarios se hallan exentos del requisito de fijación de tolerancias así como también un listado de principios activos prohibidos y restringidos. La información se publica en el sitio del SENASA y, al día de la fecha (02/08/2020) se puede descargar en formato excel desde este link (la versión es de Julio 2020). También es posible encontrar acá la versión de Febrero 2020.

El dataset consiste de un libro Excel de tres hojas de las cuales sólo la primera tiene datos. En esta hoja encontramos 3545 filas organizadas en una tabla de cinco columnas:

1. **Principio activo**

   El fitosanitario en cuestión. Por ejemplo "cloropicrina" o "Propanil". En algunos casos puede haber nombres alternativos del fitosanitario, como por ejemplo "Abamectina/Avermectina".

2. **Aptitud**

   Función que puede cumplir el fitosanitario. Puede tomar uno o varios de los siguientes valores:

   - Acaricida
   - Alguicida
   - Antiescaldante
   - Coadyuvante
   - Defoliante
   - Desecante
   - Feromona
   - Fitorregulador
   - Fungicida
   - Gorgojicida
   - Herbicida
   - Hormiguicida
   - Insecticida
   - Nematicida
   - Preservador de madera

- Repelente
- Rodenticida
- Tratamiento de semillas

Algunas celdas pueden además tener una o varias abreviaciones (como “ne”para nematicida o “he” para herbicida). El valor de esta celda se encuentra siempre entre paréntesis.

3. **Cultivos**

   Cultivos a los que se puede aplicar el fitosanitario. Por ejemplo “soja” o “centeno”.

4. **Residuos (mg /Kg)**

   Residuo máximo permitido del fitosanitario (LMR). Publicado como miligramos por kilogramo. La celda toma valores positivos (sin unidad) o la palabra “Exento” cuando no hay un límite máximo.

5. **Post Cosecha**

   Se indica con la palabra “Po” cuando el principio activo puede aplicarse luego de cosechado. En caso contrario esta celda se encuentra vacía.

Nombre original	lmrs_julio_2020.xlsx
Hash (sha256)	802EF781D909A6B3B501F3242F92407021 3B8F8DC012502C66B2 464C976F598C
Formato	Excel (xlsx)
Contenido	Tabla
Idioma	Español
Fecha de descarga	09-08-2020

Figura 6.1: Descripción del dataset de Argentina

Se describe brevemente en la figura 6.1 al dataset de Argentina.

6.1.2. Paso 1: Formateo del dataset

Este paso recibe como entrada el archivo descripto en la sección 6.1.1 y su salida va a ser un archivo del mismo tipo (en este caso un archivo .xlsx) con los datos ya formateados. Al usar la herramienta OpenRefine es posible exportar además en formato GREL (Google Refine Expression Language) la secuencia de operaciones que se llevó a cabo en el proceso de formateo. De esta manera se puede automatizar este paso y reutilizar las transformaciones en datasets de similar estructura.

A continuación se describen los pasos que se llevarán a cabo para el formateo de los datos del dataset de Argentina, teniendo en cuenta las condiciones mencionadas en 5.1.1:

1. Eliminar espacios innecesarios de todas las filas. Esto consiste en eliminar espacios al principio o final de los valores de cada celda así como también espacios de más entre palabras. Para esto, se utilizarán tres funciones que provee OpenRefine:

   a) Trim: elimina espacios innecesarios al principio y final del valor de la celda.

   b) To Lowercase: cambia el valor de la celda a minúscula.

   c) Collapse consecutive whitespaces: elimina espacios innecesarios entre palabras.

2. Eliminar los sinónimos y abreviaciones que podemos encontrar tanto en la columna de Principio Activo como en la de Aptitud. Si bien OpenRefine provee la facilidad de modificar todas las celdas similares en simultáneo (es decir, bulk), esto implica tener que buscar manualmente los valores que debamos modificar.

3. Se pasan los valores de las columnas Principio Activo, Aptitud y Cultivos a minúscula. Esto ayudará a encontrar sus equivalentes semánticos de manera mas fácil en el Paso 2.

4. En la columna Aptitud se eliminan los paréntesis que aparecen en todas las celdas.

5. En los casos en que haya varias aptitudes para un mismo principio activo, se dividen en tantas filas nuevas como aptitud diferente haya. Por ejemplo: para la fila 1, principio activo “cloropicrina”, el valor de la celda de aptitud es “(Herbicida - Fungicida - Nematicida)”. Luego de la transformación, habrá tres filas de “cloropicrina” cada uno con un valor de aptitud diferente (“Herbicida” la primera, “Fungicida” la segunda y “Nematicida” para la tercera). Este mismo proceso se repetirá con la columna Cultivos. Estas operaciones incrementan la cantidad de registros en el dataset de 3545 a 4348.

6. Creación de columna Unidad. Esto nos ayudará en el siguiente paso a alinear el valor del LMR a una unidad de la ontología UO. Esta columna es útil para los datasets en los que diferentes principios activos se expresan en diferentes unidades de medidas. En el caso del dataset de Argentina, todos los principios activos utilizan la misma unidad (miligramo por kilogramo) por lo que podemos obviar esta columna y utilizar un valor estático al momento de generar el dataset semántico.

6.1.3. Paso 2: Alineación de datos

El segundo paso del proceso recibe como entrada la salida del paso anterior. Esto podría ser un nuevo archivo excel con los datos normalizados o un archivo CSV. También, como en este caso específico se utiliza la herramienta OpenRefine para ambos pasos, se podría continuar desde donde se dejó en 6.1.2.

Para poder comprobar la necesidad de formatear el dataset, vamos a realizar la alineación tanto con el dataset original como con la versión formateada.

Principio Activo

Alinearemos la columna Principio Activo con la ontología ChEBI. Para esto se utilizó como servicio de reconciliación una release de ChEBI disponible para descargar desde el sitio web del Instituto Europeo de Bioinformática[1] (o EBI por sus siglas en inglés).

Al alinear (o reconciliar) la columna de Principios Activos del dataset de la Argentina con la ontología ChEBI, sin aplicar el paso de formateo obtenemos un total de 1173 coincidencias automáticas. Esto significa que se encontraron, sin la necesidad de intervención de un operador humano, 1173 conceptos en la ontología que representan principios activos del dataset. Esto es un 34 % del total de los registros del conjunto de datos.

La falta de efectividad en la alineación de esta columna con el dataset original se debe principalmente a dos razones: en primer lugar, y como mencionamos anteriormente, los conceptos en ChEBI están en inglés. Si bien hay algunos sinónimos, no son particularmente frecuentes y menos en idioma español. Por ejemplo, una celda cuyo valor es "teflutrina" no se alinea correctamente con ningún concepto, pero "tefluthrin" lo hace automáticamente con ChEBI:9430 tefluthrin.

En segundo lugar, hay casos de celdas que poseen no sólo el principio activo, sino también información extra, como una manera alternativa de escribirlo, un nombre alternativo o alguna información de la composición química del principio activo.

Luego de aplicadas las consignas mencionadas en 5.1.1, el porcentaje de efectividad es del 33 %, lo que representa un total de 1419 registros. Si bien el porcentaje es más bajo, hay que considerar que, al explotar las filas (aquellas que poseían varios valores se dividieron en varias filas nuevas), el número total de filas es mayor.

Utilizando la versión extendida de ChEBI alcanzamos una efectividad en la alineación del 90,11 % (3918 registros) en la alineación automática, la cual aumenta al 93,99 % luego de aceptar las sugerencias que proporciona la herramienta (aquellos candidatos de los que se tiene una certeza menor al 90 %).

Se resumen en la figura 6.2 los resultados obtenidos que se describieron anteriormente.

Dataset + Ontología	Celdas alineadas	Porcentaje alineado
Dataset original + Ontología original	1173 de 3545	≈34 %
Dataset formateado + Ontología original	1419 de 4348	≈33 %
Dataset formateado + Ontología extendida	4087 de 4348	≈94 %

Figura 6.2: Resultados de la alineación de principios activos

Aptitud

La columna Aptitud se alineará con la ontología ChEBI. Este paso es análogo al anterior, pero para el rol que cumple el principio activo.

[1] ftp://ftp.ebi.ac.uk/pub/databases/chebi/

Tanto con la versión sin modificar del dataset, como con la versión formateada, no se obtiene ningún concepto alineado, ni siquiera sugerencias de posibles coincidencias.

Hay varias razones para esto. En principio, y como ya se ha comentado anteriormente, la falta de sinónimos en español.

En segundo lugar, el formato particular que tienen los valores de Aptitud en este dataset: todos los valores se encuentran entre paréntesis. Por ejemplo, el valor para "herbicida" está expresado como "(Herbicida)". Aun si el valor estuviera en inglés, es decir "(Herbicide)", no obtendríamos una coincidencia automática al alinearlo con la ontología, sino una sugerencia del 81.8 % con ChEBI:24527 herbicide. Análogamente podemos pensar que si la ontología tuviera un sinónimo en español para este valor, tampoco obtendríamos una coincidencia automática.

En tercer lugar, los valores de las celdas de la columna Aptitud pueden tener en ocasiones más de un elemento. Por ejemplo, el valor para "Bromuro de Metilo" es "(Insecticida - Fungicida - Herbicida - Rodenticida - Gorgojicida - Nematicida)". Cada aptitud, que acá vemos separada por un guión, debería estar en su propia fila como se mencionó en el punto 1 de la sección 5.1.1. De esta manera, la herramienta podrá tomar a cada uno de estos valores de manera individual y compararlo con los conceptos de la ontología.

Al alinear utilizando la versión extendida de ChEBI con sinónimos en español, obtenemos un 91,74 % de datos alineados.

Es interesante mencionar que luego de alinear esta columna descubrimos un error de ortografía bastante frecuente. Es el caso de la palabra "funguicida" que, al estar mal escrita, no se alinea automáticamente con la ontología. Solamente este error se encuentra presente en 93 filas, lo que representa un 2,62 % del total.

El 7 % de los elementos restantes, son conceptos que no se encuentran en la ontología. Es decir, que para lograr una efectividad del 100 % no basta con correcciones o sinónimos, sino que habría que crear los conceptos faltantes en la ontología. Nuevamente, se resume lo anteriormente mencionado en la figura 6.3.

Dataset + Ontología	Celdas alineadas	Porcentaje alineado
Dataset original + Ontología original	0 de 3545	0 %
Dataset formateado + Ontología original	0 de 4348	0 %
Dataset formateado + Ontología extendida	3989 de 4348	≈92 %

Figura 6.3: Resultados de la alineación de aptitudes

Cultivos

Para alinear la columna Cultivos utilizaremos AGROVOC. En este caso se utiliza una release de AGROVOC descargada de la página oficial del tesauro[2].

[2]http://aims.fao.org/agrovoc/releases

Realizando la alineación con el dataset original obtenemos un 10,86 % de coincidencias automáticas (385 registros).

El dataset de Argentina tiene muchas inconsistencias a la hora de representar la información de los cultivos. La mayor parte de estas inconsistencias la vemos a la hora de expresar información para más de un cultivo. En algunas ocasiones se utiliza una fila por cultivo, lo que como mencionamos en 1 es la forma recomendada, pero esto no siempre es así. Se pueden encontrar cultivos separados por coma, por barras y hasta por guiones.

Al alinear la columna de Cultivos luego de aplicado el paso de formateo, obtenemos 1350 coincidencias automáticas, lo que supone una efectividad 31,04 %.

El tesauro AGROVOC posee los nombres de cultivos en varios idiomas diferentes. Esto en ocasiones puede ser un problema a la hora de alinear los datos, ya que hay valores que se escriben de la misma manera en diferentes idiomas. Por ejemplo "Acelga" se escribe de la misma manera tanto en español como en portugués. En estos casos la herramienta nos muestra ambas sugerencias pero no realiza la reconciliación automáticamente. Si utilizáramos una versión reducida de AGROVOC, sólo con términos en español, tendríamos 598 coincidencias más. Esto es una mejora de alrededor del 13 %, lo que elevaría el porcentaje total a más del 44 %.

Para este caso práctico además consideraremos que todos los cultivos definidos con el sufijo "(consumo)" o "(grano consumo)" se alinearán sin tener en cuenta dichas cadenas. De esta manera el término "trigo (grano consumo)" se alineará con "trigo", "cebolla (consumo)" con "cebolla", etc.

Teniendo en cuenta las diferentes decisiones tomadas a la hora de alinear este dataset, la alineación alcanza un porcentaje aproximado de %76,50.

Al igual que con la columna Aptitud, es necesario crear los conceptos faltantes en AGROVOC para poder aumentar el porcentaje de datos alineados.

Los resultados obtenidos se resumen en la figura 6.4.

Dataset + Ontología	Celdas alineadas	Porcentaje alineado
Dataset original + Ontología original	385 de 3545	≈11 %
Dataset formateado + Ontología original	1350 de 4348	≈31 %
Dataset formateado + Ontología sólo español	3327 de 4348	≈76,50 %

Figura 6.4: Resultados de la alineación de cultivos

Unidad

La columna Unidad la alinearemos con la ontología de las Unidades de Medida o UO. Para reconciliar las unidades se utiliza una release de UO descargada de BioPortal[3], un repositorio de ontologías biomédicas.

[3]https://bioportal.bioontology.org/ontologies/UO

6.1.4. Paso 3: Generación del dataset semántico

La extensión RDF de OpenRefine permite generar una plantilla la cual se utilizará para traducir de fila en el dataset a tripleta RDF. Aquí se indicará como la herramienta deberá hacer la conversión de cada fila del conjunto de datos en tripletas. Como se puede observar en la figura 6.5, las propiedades de cada registro se mapean en propiedades de la ontología LMR-O, por lo que es necesario incluir dicha ontología en el apartado de prefijos disponibles.

Para esta plantilla, indicaremos las siguientes reglas:

1. Cada fila equivaldrá a un lmro:Record, el cual tendrá cinco propiedades: lmro:role, lmro:activePrinciple, lmro:appliesTo, lmro:comments, lmromaximumResidue.

2. La propiedad lmro:activePrinciple tendrá el valor de alineado en la columna Princio Activo. Se debe especificar que el valor será una URI.

3. De la misma manera, lmro:role tendrá el valor de alineado en la columna aptitud y se debe especificar este valor será una URI.

4. En lmro:appliesTo usaremos una expresión condicional que nos permitirá usar el valor alineado, si existiere, o el valor original en caso contrario:

   ```
   if(cell.recon.match.id != null, cell.recon.match.id, value)
   ```

   El tipo de dato a usar será texto.

5. La propiedad lmro:comments tendrá el valor, de tipo texto, de la columna Post Cosecha.

6. La propiedad lmro:maximumResidue tendrá el valor de la columna Residuos (mg /Kg) en tipo texto.

Se debe indicar además, la URI que identificará cada nuevo recurso semántico. Para el caso de Argentina se usará la URI base “http://www.lifia.info.unlp.edu.ar/lmro/data/ar/”.

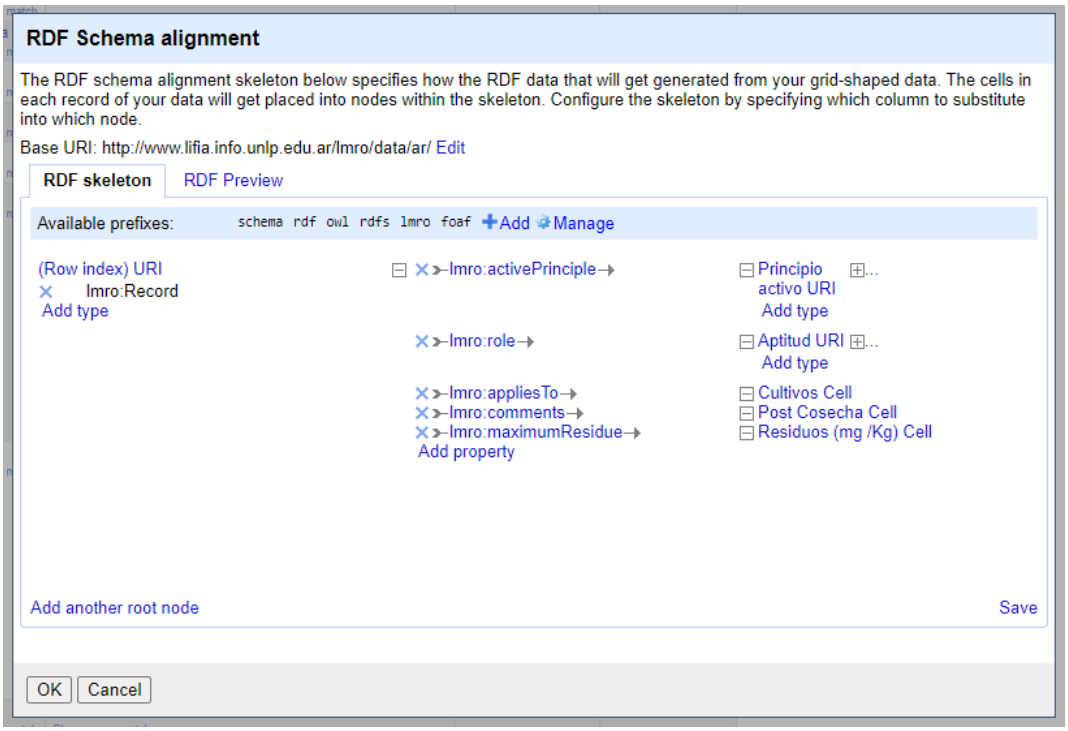

Figura 6.5: Plantilla de exportación a RDF

Es posible exportar esta plantilla y reutilizarla con futuras versiones de los datos o incluso con otros dataset de similar estructura. En el repositorio de este trabajo de tesis se encuentra disponible la plantilla que fue generada para este caso de estudio.

Una vez armada la plantilla, procedemos a descargar las tripletas de nuestro nuevo dataset. Este archivo será utilizado como entrada para el algoritmo “lmro-corrections”. Este script recibirá como entrada adicional el archivo de AGROVOC de donde obtendrá los conceptos que representan a cada uno de los cultivos. Luego será el turno del algoritmo de generación de información de Provenance. La información de proveniencia propiamente dicha que se utilizará puede verse en el snippet 2.

```
{
  "namespace":  "http://www.lifia.info.unlp.edu.ar/lmro/data/ar/",
  "sourceDocument": {
    "date": "2020-07-02T00:00:00.000Z",
    "format": "xlsx",
    "language": "Español",
    "description": "Informacion de los LMR publicada por SENASA en Julio 2020",
    "title": "lmrs_julio_2020",
    "identifier": "802EF781D909A6B3B501F3242F924070213B8F8DC012502C66B2464C976F598C",
    "createdBy": {
      "name": "SENASA",
      "uri": "https://www.wikidata.org/wiki/Q6972721"
    }
  },
  "publisher": {
    "name": "UNLP",
    "uri": "https://www.wikidata.org/wiki/Q784171"
  }
}
```

Snippet 2: Información de proveniencia para el dataset argentino

Finalmente, luego de correr el algoritmo consideraremos finalizado el proceso de transformación.

6.1.5. Conclusión

Como se pudo observar a lo largo del caso práctico, este dataset presenta numerosos desafíos.

En primer lugar, el conjunto de datos resulta muy heterogéneo en cuanto a la manera de describir términos específicos. Se pueden observar casos en que un mismo concepto de la vida real se expresa de formas diferentes o se encuentra abreviado. Para establecer el LMR para el cultivo gladiolo, el dataset utiliza los términos “Gladiolo”, “Ornamentales (Clavel, gladiolo, rosal)”, “Ornamentales, Gladiolos, tulipanes, iris y narcisos”, “Gladiolo, Clavel, Crisantem, Rosa, Florales”, “ornamentales (Gladiolos y arbustos) y “Florales y ornamentales (Crisantemo, gladiolo, clavel, rosal). En casos como este se decide agrupar a todas las variantes bajo el concepto más representativo, (en el ejemplo, “Gladiolo”). Para esta tesina no se han agregado sinónimos en AGROVOC (se demostró cómo podría realizarse esto con ChEBI en la sección 5.2), por lo que aquellos términos que no se reconcilien correctamente, serán exportados como una cadena de caracteres en lugar de una URI que identifica a un recurso semántico.

El dataset también es heterogéneo en la manera de estructurar la información. Por ejem-

plo, es frecuente encontrar todas las combinaciones de un principio activo-aptitud ocupan una sola fila en la tabla mientras que en otros casos cada combinación ocupa su propia fila. Tampoco es inusual que se utilicen caracteres diferentes para separar palabras u opciones (por ejemplo, guiones o barras). Todos estos aspectos hacen más difícil el trabajo de transformar los datos a una representación semántica y exigen frecuentemente que se tomen decisiones de diseño que sin duda afectan el resultado final.

En segundo lugar, se pudo comprobar que hay gran cantidad de conceptos que no se han podido encontrar en las ontologías seleccionadas. En algunos casos la decisión de agregar estos conceptos o sus sinónimos a la ontología correspondiente es sencilla y no deja lugar a dudas. En otros casos, es un poco más complicado determinar si el concepto realmente falta dentro de la ontología o si el mismo está expresado de alguna manera diferente (por ejemplo, un elemento químico podría estar con un nombre coloquial o con uno científico). En el caso específico de los cultivos, es muy frecuente encontrar en el dataset varias presentaciones diferentes de un mismo cultivo. Tomando el caso del algodón, por ejemplo, encontramos en los datos que publica SENASA siete versiones diferentes de este producto: algodón, algodón (aceite), algodón (con cáscara), algodón (fibra), algodón (pellets), algodón (semilla consumo) y algodón (torta). La mayoría de estas variantes no tienen su contraparte en AGROVOC y es una decisión de diseño si concentramos todas estas variantes bajo el término "Algodón" y si creamos conceptos diferentes para cada una.

Finalmente y pese a estas dificultades se logró crear un dataset semántico, que si bien no contiene la totalidad de la información original, es perfectamente adecuado para evaluar el esfuerzo, y eficacia en la ejecución de distintas operaciones de interoperabilidad entre diferentes conjuntos de datos.

6.2. Caso práctico II: El dataset de Brasil

6.2.1. Descripción del dataset

La Agencia Nacional de Vigilancia Sanitaria de Brasil (ANVISA) publica en su sitio[4] informes sobre los requisitos que deben cumplir los fitosanitarios. El sitio está en portugués y al día de la fecha (02/08/2020) es posible descargar en formato PDF un informe por cada principio activo. Cada informe posee una breve descripción del fitosanitario, de su composición química y métodos de uso. Posee además una tabla conteniendo los cultivos a los que aplica el fitosanitario, la modalidad de aplicación, el límite máximo del residuo y el tiempo de espera.

También está disponible una web interactiva que permite diferentes modalidades de búsqueda y la posibilidad de descargar el conjunto de datos en formato CSV (Comma Separated Value). El archivo CSV posee 4822 filas organizadas en 14 columnas de las cuales nos interesan las siguientes cuatro:

1. **DS_INGREDIENTE_ATIVO**

[4] https://www.gov.br/anvisa/pt-br

El fitosanitario en cuestión, en idioma portugués. Por ejemplo “Cloretos de benzalcônio” o “alfa-cipermetrina”.

2. **NU_LMR**

   Residuo máximo permitido del fitosanitario (LMR). En este dataset todos los LMR están publicados como miligramos por kilogramo. La celda toma valores positivos (sin unidad).

3. **NO_CULTURA**

   Cultivos a los que se puede aplicar el fitosanitario, en idioma portugués. Por ejemplo “Algodão” o “Tomate”.

4. **DT_FIM_VIGENCIA**

   Fecha en la que dejó de ser válida la fila.

Se describe en la figura 6.6 al dataset de Brasil.

Nombre original	TA_MONOGRAFIA_AGROTOXICO.csv
Hash (sha256)	83208CBB4C39A6273209C5A4DCE7CF52 7292681748933013 9E0 8EF1D63058B73
Formato	Comma Separated Value (csv)
Contenido	Tabla
Idioma	Portugués
Fecha de descarga	29-11-2020

Figura 6.6: Descripción del dataset de Brasil

Es importante mencionar que a diferencia del caso de estudio Argentino, este dataset no contempla el concepto de rol. Si bien no es un problema a la hora de generar nuestro nuevo conjunto de datos semántico, es algo a tener en cuenta a la hora de realizar comparaciones entre datasets que sí poseen esta información. Por ejemplo, una consulta válida para el dataset argentino sería “¿Qué fungicidas se pueden utilizar en el cultivo Tomate?”, mientras que para el de Brasil no habría manera de diferenciar entre principios activos cuáles son fungicidas y cuáles no. Para estos casos, y como decisión de diseño, utilizaremos el concepto lmro:Any, que en esta circunstancia representa a “cualquier” rol. De esta manera la consulta planteada anteriormente devolvería a todos los principios activos utilizados en el tomate según ANVISA, ya que todos ellos cumplen el rol “cualquiera”.

Utilizaremos para el caso de estudio de Brasil, el archivo CSV ya que nos permite usarlo como entrada en la herramienta OpenRefine, la cual usaremos nuevamente para realizar las alineaciones con las ontologías elegidas y la exportación a RDF.

6.2.2. Paso 1: Formateo del dataset

A diferencia de lo que ocurría con el caso Argentino, el dataset que publica el gobierno de Brasil ya se encuentra formateado según las pautas que describimos en 5.1.1. Este dataset ya se encuentra informatizado y ANVISA cuenta con una web que permite al usuario realizar consultas por cultivo o fitosanitario. Es posible que durante el proceso de informatización de los datos, éstos hayan pasado por un etapa de formateo similar a la que describimos en este trabajo.

El cambio más significativo que se realizará en este paso será eliminar los datos que no estén mas en vigencia, o dicho con otras palabras, aquellas filas del dataset que posean un valor para la columna de Fecha de Vigencia. Esta operación reducirá las filas a 4190.

Se renombrarán, además, las columnas "DS_INGREDIENTE_ATIVO", "NU_LMR" y "NO_CULTURA" por "Principio activo", "Residuos (mg /Kg)" y "Cultivos" respectivamente. Esto no sólo hará más legible la información sino que nos permitirá utilizar la plantilla RDF que exportamos en el caso práctico anterior sin realizar mayores cambios. Por una cuestión de legibilidad, se removerán el resto de las columnas con las cuales no vamos a trabajar.

6.2.3. Paso 2: Alineación de datos

A continuación se describen los resultados de alinear las dos columnas que poseen texto en este dataset: Principio Activo y Cultivos (originalmente "DS_INGREDIENTE_ATIVO" y "NO_CULTURA" respectivamente).

Principio Activo

Alinearemos la columna Principio Activo con la ontología ChEBI. Para esto se utilizó como servicio de reconciliación una release de ChEBI disponible para descargar desde el sitio web del Instituto Europeo de Bioinformática[5] (o EBI por sus siglas en inglés).

Al alinear la columna de Principios Activos del dataset de Brasil con la ontología original de ChEBI, obtenemos un total de 8 % coincidencias automáticas, lo que equivale a 307 términos alineados y 3882 sin alinear.

Utilizando la versión extendida de ChEBI (sólo sinónimos) alcanzamos una efectividad en la alineación del 99 % (4167 registros). Esto quiere decir que sólo 22 filas del dataset de Brasil no fueron encontrados en ChEBI y requerirán nuevos conceptos en dicha ontología. Este número es realmente pequeño comparado al caso de estudio argentino, donde el numero de filas sin alinear para la misma versión de ChEBI fue superior a 250.

Los resultados de la alineación de la columna Principio Activo se pueden ver en la figura 6.7.

[5] ftp://ftp.ebi.ac.uk/pub/databases/chebi/

Dataset + Ontología	Celdas alineadas	Porcentaje alineado
Dataset original + Ontología original	307 de 4189	≈8 %
Dataset original + Ontología extendida	4167 de 4189	≈99 %

Figura 6.7: Resultados de la alineación de principios activos

Cultivos

Utilizando la release anteriormente mencionada de AGROVOC, se alineará la columna Cultivos.

Realizando la alineación con el dataset original obtenemos aproximadamente un 54 % de coincidencias automáticas (2246 registros). Sin embargo hay gran cantidad de sugerencias de alineación debido al ya mencionado problema de aquellas palabras que se escriben igual en diferentes idiomas, por lo que el número total aumenta sin mucha dificultad a hasta aproximadamente 96 % dejando de esa manera 4019 registros alineados y sólo 170 sin alinear. El porcentaje restante requerirá nuevos términos en AGROVOC que representen estos conceptos. La figura 6.8 resume los resultados obtenidos.

Dataset + Ontología	Celdas alineadas	Porcentaje alineado
Dataset original + Ontología original	2246 de 4189	≈54 %
Dataset original + Ontología sólo portugués	4019 de 4189	≈96 %

Figura 6.8: Resultados de la alineación de cultivos

6.2.4. Paso 3: Generación del dataset semántico

Utilizando el esqueleto RDF que generamos y exportamos para el caso argentino, procedemos a generar las tripletas para el dataset de Brasil. La única diferencia con la plantilla argentina, será la URI que se utilizará para identificar a las entidades semánticas. En este caso, la URI base será "http://www.lifia.info.unlp.edu.ar/lmro/br/data/". Es posible reutilizar la misma plantilla ya que renombramos las columnas de modo que posean el mismo nombre que en el caso práctico argentino. Si esto no fuera así, solo habría que indicar el nombre actual de cada columna y la plantilla funcionará correctamente.

El archivo RDF generado luego de exportar los datos de OpenRefine será luego utilizado como input para el script "lmro-corrections". Una vez corrido este algoritmo, se usará su salida en el script de generación de información de Provenance.

```
{
  "namespace":  "http://www.lifia.info.unlp.edu.ar/lmro/data/br/",
  "sourceDocument": {
    "date": "2020-10-11T00:00:00.000Z",
    "format": "csv",
    "language": "Portugués",
    "description": "Informacionde LMR obtenida de ANVISA en Octubre 2020",
    "title": "TA_MONOGRAFIA_AGROTOXICO",
    "identifier": "83208CBB4C39A6273209C5A4DCE7CF5272926817489330139E08EF1D63058B73",
    "createdBy": {
      "name": "ANVISA",
      "uri": "https://www.wikidata.org/wiki/Q295815"
    }
  },
  "publisher": {
    "name": "UNLP",
    "uri": "https://www.wikidata.org/wiki/Q784171"
  }
}
```

Snippet 3: Información de proveniencia para el dataset de Brasil

Podemos ver en el snippet 3 la información de proveniencia que se utilizará junto con las tripletas generadas en el paso anterior.

Una vez mas, luego de correr el algoritmo daremos por finalizado el proceso de transformación.

6.2.5. Conclusión

A diferencia de lo ocurrido en el primer caso práctico, el dataset que confecciona ANVISA resulta bastante sencillo de abordar. Este dataset ya viene formateado según las pautas descriptas en la sección 5.1.1. Esto probablemente es así, ya que el dataset proviene de un sistema informatizado el cual inherentemente requiere de una información bien estructurada.

Otro aspecto positivo de este dataset es que la gran mayoría de los términos pudieron ser encontrados en las ontologías elegidas, por lo que fueron muy pocos los conceptos nuevos que debieron ser agregados.

El mayor y quizás único inconveniente con el dataset de Brasil, es que no posee información respecto del rol (también llamado aptitud) que cumplen los principios activos en cada cultivo. Como se mencionó anteriormente, utilizaremos el concepto lmro:Any y se asumirá que todos los productos fitosanitarios cumplen cualquier rol. De esta manera, ante una hipotética consulta del tipo "¿Qué herbicidas permite ANVISA en el Café?" la respuesta incluirá a la totalidad de las sustancias permitidas para dicho cultivo.

Nuevamente se logró generar un dataset semántico el cual se utilizará, en conjunto con el dataset del caso práctico uno, para evaluar el impacto del uso de tecnologías de la web semántica en la publicación de datos normativos de LMR en alimentos.

6.3. Escenario I: Consultas dentro de un mismo dataset

En este caso utilizaremos el dataset de Argentina como caso testigo, para lo cual importaremos el archivo RDF que fue generado en la sección 6.1 a Jena Fuseki.

1. **Todos los herbicidas**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?activePrinciple
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role <http://purl.obolibrary.org/obo/CHEBI_24527>.
}
```

activePrinciple
http://purl.obolibrary.org/obo/CHEBI_64163
http://purl.obolibrary.org/obo/CHEBI_5008
http://purl.obolibrary.org/obo/CHEBI_133237
http://purl.obolibrary.org/obo/CHEBI_83569
http://purl.obolibrary.org/obo/CHEBI_27744

Figura 6.9: Todos los herbicidas

Esta consulta nos permite obtener todos los herbicidas del dataset. En total, se han exportado 130 herbicidas diferentes. Por cuestiones de espacio, sólo se muestran los primeros 5 conceptos. Las URI que se pueden ver en la figura 6.9 son las que identifican en ChEBI a “diquat dibromuro”, “fenoxaprop etil”, “indaziflam”, “pendimetalin” y “glifosato” respectivamente.

2. **El principio activo con el límite de residuo permitido más alto**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?activePrinciple ?lmr ?unit ?crop
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:maximumResidue ?residue.
  ?residue lmro:hasValue ?lmr.
  ?residue lmro:hasUnit ?unit.
  ?subject lmro:appliesTo ?crop.
}
order by desc(?lmr)
limit 1
```

activePrinciple	lmr	unit	crop
http://purl.obolibrary.org/obo/ CHEBI_27744	500.0	http://purl.obolibrary.org/obo/ UO_0000308	alfalfa (forraje)

Figura 6.10: Principio con LMR más alto

Podemos ver en 6.10 que la sustancia con mayor residuo permitido es el glifosato, utilizado en el cultivo alfalfa (forraje), el cual permite hasta 500mg/Kg. La URI `http://purl.obolibrary.org/obo/CHEBI_27744` identifica en ChEBI al concepto "glyphosate", mientras que `http://purl.obolibrary.org/obo/UO_0000308` hace referencia en la ontología de las unidades de medida a "milligram per kilogram". Vale aclarar que "alfalfa (forraje)" es un string y no una URI ya que no se alineó correctamente con ningún concepto de AGROVOC.

3. **Las 5 aptitudes con más principios activos**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?role (count(?activePrinciple) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role ?role.
}
group by ?role
order by DESC(?count)
limit 5
```

role	count
http://purl.obolibrary.org/obo/CHEBI_24127	1275
http://purl.obolibrary.org/obo/CHEBI_24852	1192
http://purl.obolibrary.org/obo/CHEBI_24527	956
http://purl.obolibrary.org/obo/CHEBI_22153	317
www.lifia.info.unlp.edu.ar/data/lmro/84	173

Figura 6.11: Las 5 aptitudes con más principios activos

Como se puede ver 6.11, los cinco roles con más principios activos en el dataset de Argentina son "fungicida", "insecticida", "herbicida", "acaricida" y "tratamiento de semillas" respectivamente. Estas cinco aptitudes concentran un 89,99 % del total de las sustancias. Es interesante destacar que "tratamiento de semillas" está representado por una URI propia, lo que quiere decir que es un concepto creado en la versión extendida de ChEBI que se generó en la sección 5.2.

4. **El cultivo para el que se aceptan más principios activos**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>
PREFIX schema: <http://schema.org/>


SELECT distinct ?aCrop (count(distinct ?activePrinciple) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:appliesTo ?crop.
    ?crop schema:sameAs ?aCrop
}
group by ?aCrop
order by DESC(?count)
limit 1
```

aCrop	count
http://aims.fao.org/aos/agrovoc/c_14477	139

Figura 6.12: El cultivo que acepta más principios activos

El cultivo que más principos activos diferentes permite utilizar en este dataset es la "Soja", aquí representado por la URI de AGROVOC `http://aims.fao.org/aos/agrovoc/c_14477`. En este cultivo se pueden utilizar 139 sustancias que cumplen 12 roles diferentes.

5. **El cultivo con mas acaricidas**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>
PREFIX schema: <http://schema.org/>

SELECT distinct ?aCrop (count(distinct ?activePrinciple) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role <http://purl.obolibrary.org/obo/CHEBI_22153>.
    ?subject lmro:appliesTo ?crop.
    ?crop schema:sameAs ?aCrop
}
group by ?aCrop
order by DESC(?count)
limit 1
```

aCrop	count
http://aims.fao.org/aos/agrovoc/c_541	25

Figura 6.13: El cultivo con más acaricidas

EL cultivo con más acaricidas (ChEBI URI `http://purl.obolibrary.org/obo/CHEBI_22153`) es la "Manzana", representada por el concepto AGROVOC `http://aims.fao.org/aos/agrovoc/c_541`. Como se puede ver en la figura 6.14 existen en este dataset 25 acaricidas diferentes que se pueden utilizar en este cultivo.

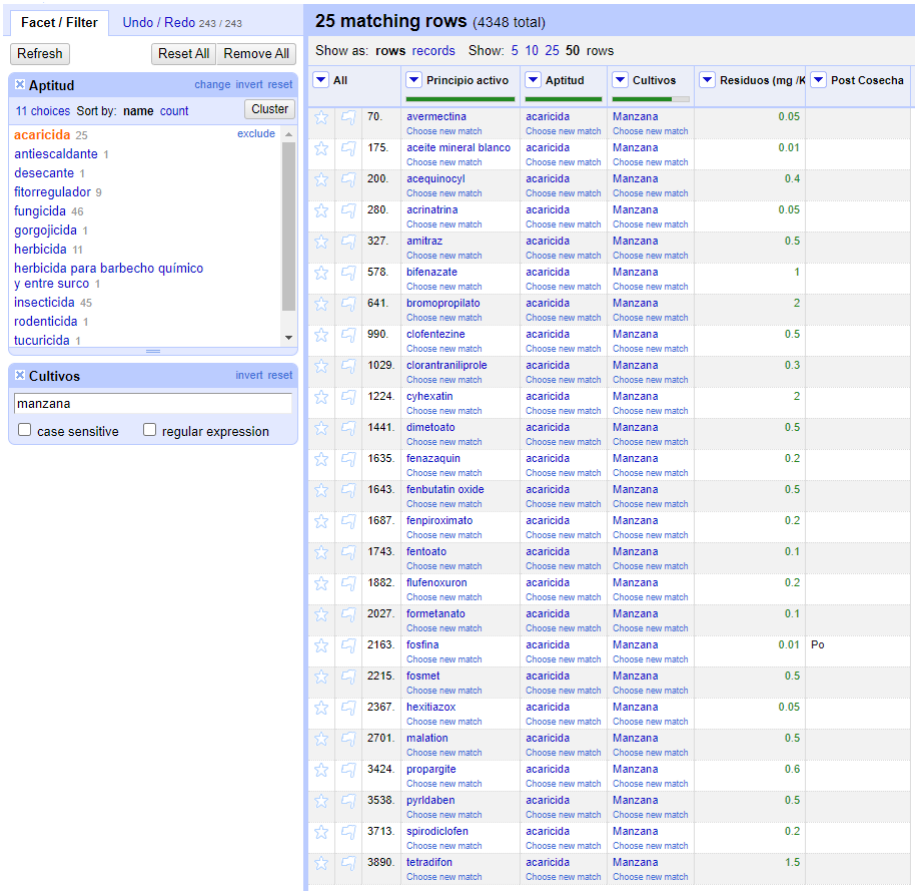

Figura 6.14: Acaricidas para manzanas en el dataset de Argentina

6. **El principio activo con más roles**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?activePrinciple (count(distinct ?role) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role ?role.
}
group by ?activePrinciple
order by DESC(?count)
limit 1
```

activePrinciple	count
http://purl.obolibrary.org/obo/CHEBI_39275	6

Figura 6.15: El principio activo con más roles

El principio activo con más roles es el concepto ChEBI `http://purl.obolibrary.org/obo/CHEBI_39275` "bromomethane", el cual posee un sinónimo agregado en la sección 5.2 para "bromuro de metilo", nombre con el cual figura en el dataset argentino. Según este dataset, esta sustancia cumple simultáneamente los roles de fungicida, gorgojicida, herbicida, insecticida y nematicida.

7. **El principio activo usado en más cultivos diferentes**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?activePrinciple (count(distinct ?crop) as ?count)
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:appliesTo ?crop.
}
group by ?activePrinciple
order by DESC(?count)
limit 1
```

activePrinciple	count
http://purl.obolibrary.org/obo/CHEBI_40909	60

Figura 6.16: El principio activo usado en más cultivos diferentes

Podemos encontrar en el dataset 60 cultivos diferentes para los que se puede utilizar "azoxistrobina" (ChEBI URI `http://purl.obolibrary.org/obo/CHEBI_40909`).

8. **Todos los principios activos, roles y limites máximos para durazno, siempre y cuando el LMR sea mayor a 7. Ordenado por valor máximo de residuo**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>
PREFIX schema: <http://schema.org/>


SELECT  ?activePrinciple ?role ?lmr
WHERE {
  ?crop schema:sameAs <http://aims.fao.org/aos/agrovoc/c_5638>.
    ?subject lmro:appliesTo ?crop.
    ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:role ?role.
  ?subject lmro:maximumResidue ?residue.
  ?residue lmro:hasValue ?lmr.
  FILTER(?lmr > 7)
}
order by DESC(?lmr)
```

activePrinciple	role	lmr
http://purl.obolibrary.org/obo/ CHEBI_34608	http://purl.obolibrary.org/obo/ CHEBI_24127	15.0
http://purl.obolibrary.org/obo/ CHEBI_23414	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
http://purl.obolibrary.org/obo/ CHEBI_28909	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
http://purl.obolibrary.org/obo/ CHEBI_31440	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
http://purl.obolibrary.org/obo/ CHEBI_81907	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
http://purl.obolibrary.org/obo/ CHEBI_81908	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
http://purl.obolibrary.org/obo/ CHEBI_82019	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
www.lifia.info.unlp.edu.ar/data/ lmro/10	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0
www.lifia.info.unlp.edu.ar/data/ lmro/14	http://purl.obolibrary.org/obo/ CHEBI_24127	10.0

Figura 6.17: Sustancias con LMR superior a 7mg/Kg en Durazno

Los principios activos con LMR superior a 7mg/Kg permitidos en Durazno (URI AGROVOC `http://aims.fao.org/aos/agrovoc/c_5638`) para el dataset de Argentina son las sustancias fungicidas “chlorothalonil”, “captan”, “sulfato tribásico de cobre”, “sulfato cúprico pentahidratado”, “hidróxido de cobre”, “óxido cuproso”, “folpet”, “iprodione”, “sulfato tetracúpico tricálcico” y “oxicloruro de cobre”.

9. **Principios activos exentos (limitado a los primeros 5)**

```
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>
PREFIX schema: <http://schema.org/>


SELECT distinct ?activePrinciple ?aCrop
WHERE {
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:maximumResidue <http://www.lifia.info.unlp.edu.ar/lmro#Exempt>.
  ?subject lmro:appliesTo ?crop.
  ?crop schema:sameAs ?aCrop.
}
limit 5
```

activePrinciple	aCrop
http://purl.obolibrary.org/obo/ CHEBI_81760	http://aims.fao.org/aos/agrovoc/ c_21298
http://purl.obolibrary.org/obo/ CHEBI_60607	http://aims.fao.org/aos/agrovoc/ c_12140
http://purl.obolibrary.org/obo/ CHEBI_9495	http://aims.fao.org/aos/agrovoc/ c_12487
http://purl.obolibrary.org/obo/ CHEBI_81760	http://aims.fao.org/aos/agrovoc/ c_728
http://purl.obolibrary.org/obo/ CHEBI_9495	http://aims.fao.org/aos/agrovoc/ c_25473

Figura 6.18: Sustancias con LMR exento

En esta consulta se pueden ver aquellos principios activos que no poseen un valor de LMR específico, sino que se encuentran exentos para determinado cultivo. Podemos ver en la figura 6.18, que por ejemplo la sustancia "metalaxyl-M" se encuentra exenta para el cultivo "Lenteja".

10. **Información del documento fuente**

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>


SELECT ?title ?format ?language ?id ?date ?description
WHERE {
  ?subject dc:date ?date.
  ?subject dc:description ?description.
  ?subject dc:format ?format.
  ?subject dc:identifier ?id.
  ?subject dc:language ?language.
  ?subject dc:title ?title.
}
```

Esta consulta SPARQL nos permite ver la información de proveniencia relativa al documento que se utilizó como fuente para la generación del dataset semántico. Este archivo, como se puede ver en la figura 6.19, es la planilla de excel que publicó SENASA en el mes de Julio de 2020.

11. **La institución que publicó el dataset semántico**

title	format	language	id	date	description
lmrs_julio_2020	xlsx	Español	802EF781D909A6B3B501F3242F924070213B8F8DC012502C66B2464C976F598C	2020-07-02T00:00:00.000Z	Informacion respecto a los LMR publicada por SENASA en el mes de Julio 2020

Figura 6.19: Información de proveniencia

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX schema: <http://schema.org/>


SELECT ?createdBy ?sameAs
WHERE {
  ?subject prov:wasAssociatedWith ?publisher.
   ?publisher rdf:label ?createdBy.
   ?publisher schema:sameAs ?sameAs.
}
```

createdBy	sameAs
UNLP	https://www.wikidata.org/wiki/Q784171

Figura 6.20: Institución que publica el dataset semántico

En la figura 6.20 se puede ver la institución que publica el dataset semántico. Se muestra el idenificador único de recurso de Wikidata que se especificó oportunamente en el algoritmo de proveniencia, descripto en la subsección 5.1.3.

6.4. Escenario II: Consultas entre diferentes versiones de un mismo dataset

En este escenario se utilizará el dataset generado en la sección 6.1. En el sitio web del SENASA es posible descargar también la versión del dataset del mes de Febrero. Aprovechando las operaciones exportadas de OpenRefine y siguiendo los mismos pasos del proceso de transformación descriptos en 3.1, se generó el dataset semántico para esta versión. Importando ambos datasets al servidor SPARQL Fuseki es posible realizar consultas como las que se describen a continuación.

1. **Fecha de creación de cada dataset semántico**

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>


SELECT ?activity ?createdAt ?date
WHERE {
  ?activity rdf:type <http://www.lifia.info.unlp.edu.ar/lmro#PublicationActivity>.
  ?activity prov:endedAtTime ?createdAt.
  ?activity prov:used ?source.
  ?source dc:date ?date.
}
```

activity	createdAt	date
http://www.lifia.info.unlp.edu.ar/lmro/data/ar/feb/ PublicationActivity	2021-01-14T16:06:58.531Z	2020-02-02T00:00:00.000Z
http://www.lifia.info.unlp.edu.ar/lmro/data/ar/ PublicationActivity	2021-01-14T16:08:05.174Z	2020-07-02T00:00:00.000Z

Figura 6.21: Fecha de creación de los datasets

La fecha de generación de los dataset semánticos y de publicación de los documentos fuentes que se usaron de base para su elaboración.

2. **Nombres y hashes de los documentos fuentes de cada dataset**

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>


SELECT ?title ?id
WHERE {
  ?subject rdf:type <http://www.lifia.info.unlp.edu.ar/lmro#SourceDocument>.
  ?subject dc:title ?title.
  ?subject dc:identifier ?id.
}
```

title	id
lmrs_julio_2020	802EF781D909A6B3B5 01F3242F924070213B 8F8DC012502C66B2464C976F598C
lmrs_febrero_2020	7E3B4B6B2C65D2032339C E4E9B5B2681E68B782E21 2D0466A3142CED2AA9D14D

Figura 6.22: Nombres y hash de los archivos fuentes

Esta consulta nos permite conocer el nombre original de los archivos fuentes y el hash sha256 que se utiliza como identificador.

3. Cantidad de registros LMR de cada dataset

```
1  PREFIX prov: <http://www.w3.org/ns/prov#>
2  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3  PREFIX dc: <http://purl.org/dc/elements/1.1/>
4  PREFIX schema: <http://schema.org/>
5
6
7  SELECT ?title
8         (count(distinct ?subject) as ?records)
9  WHERE {
10   ?subject rdf:type <http://www.lifia.info.unlp.edu.ar/lmro#Record>.
11   ?subject schema:isPartOf ?publication.
12   ?publication prov:wasDerivedFrom ?source.
13   ?source dc:title ?title.
14 }
15 group by ?title
```

title	records
lmrs_julio_2020	4348
lmrs_febrero_2020	4347

Figura 6.23: Cantidad de registros LMR en cada dataset

Esta consulta nos permite descubrir que el dataset que SENASA publica en Julio de 2020 posee un registro más que la versión de febrero del mismo año.

4. **Cultivos, principios activos y valor de LMR con el nombre del archivo fuente (limitado a los primeros 5)**

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX schema: <http://schema.org/>
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT ?activePrinciple ?crop ?lmr ?dataset
WHERE {
  ?subject rdf:type <http://www.lifia.info.unlp.edu.ar/lmro#Record>.
  ?subject lmro:activePrinciple ?activePrinciple.
  ?subject lmro:appliesTo ?crop.
  ?subject lmro:maximumResidue ?residue.
  ?residue lmro:hasValue ?lmr.
  ?subject schema:isPartOf ?publication.
  ?publication prov:wasDerivedFrom ?source.
  ?source dc:title ?dataset.
}
limit 5
```

activePrinciple	crop	lmr	dataset
http://purl.obolibrary .org/obo/CHEBI_35883	http://www.lifia.info. unlp.edu.ar/lmro/data /ar/feb/feb/156/Crop	0.01	lmrs_febrero_2020
http://purl.obolibrary .org/obo/CHEBI_27744	http://www.lifia.info. unlp.edu.ar/lmro/data /ar/feb/feb/144/Crop	0.1	lmrs_febrero_2020
http://purl.obolibrary .org/obo/CHEBI_64163	http://www.lifia.info. unlp.edu.ar/lmro/data /ar/15/Crop	2.0	lmrs_julio_2020
http://purl.obolibrary .org/obo/CHEBI_39314	http://www.lifia.info. unlp.edu.ar/lmro/data /ar/feb/feb/370/Crop	0.3	lmrs_febrero_2020
http://purl.obolibrary .org/obo/CHEBI_81970	http://www.lifia.info. unlp.edu.ar/lmro/data /ar/417/Crop	0.1	lmrs_julio_2020

Figura 6.24: Cultivos, roles y sustancias de cada dataset (limitado a los primeros 5)

Es posible con esta consulta ver los registros de cada dataset en simultáneo. En este caso se pueden ver de la publicación de febrero y de la publicada en Julio.

6.5. Escenario III: Consultas entre diferentes datasets

Se utilizará para este escenario los casos prácticos de Argentina y Brasil, generados en las secciones 6.1 y 6.2 respectivamente.

1. **Los organismos que publican cada dataset original**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>


SELECT ?title ?creatorName
WHERE {
  ?source rdf:type <http://www.lifia.info.unlp.edu.ar/lmro#SourceDocument>.
  ?source dc:title ?title.
  ?source dc:creator ?creator.
  ?creator rdf:label ?creatorName.
}
```

title	creatorName
TA_MONOGRAFIA_AGROTOXICO	ANVISA
lmrs_julio_2020	SENASA

Figura 6.25: Organismos que generaron los datasets

Esta consulta permite ver de dónde proviene la información relativa a los LMR disponible en el servidor SPARQL. Cómo se puede observar en la figura 6.25, los datasets fueron originalmente generados por ANVISA y SENASA.

2. **Principios activos que se utilizan en Durazno en cada dataset (limitado a los primeros 10)**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX schema: <http://schema.org/>
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>
PREFIX prov: <http://www.w3.org/ns/prov#>


SELECT ?activePrinciple ?lmr ?organizationName
WHERE {
  ?crop schema:sameAs <http://aims.fao.org/aos/agrovoc/c_5638>.
  ?record lmro:appliesTo ?crop.
  ?record lmro:activePrinciple ?activePrinciple.
  ?record lmro:maximumResidue ?residue.
  ?record schema:isPartOf ?publication.
  ?publication prov:wasDerivedFrom ?source.
  ?source dc:creator ?organization.
  ?organization rdf:label ?organizationName.
  ?residue lmro:hasValue ?lmr.
}
order by ?activePrinciple
```

activePrinciple	lmr	organizationName
http://purl.obolibrary.org/obo /CHEBI_133237	0.01	SENASA
http://purl.obolibrary.org/obo /CHEBI_133305	0.4	SENASA
http://purl.obolibrary.org/obo /CHEBI_137425	0.08	ANVISA
http://purl.obolibrary.org/obo /CHEBI_23414	10.0	SENASA
http://purl.obolibrary.org/obo /CHEBI_27744	0.2	SENASA
http://purl.obolibrary.org/obo /CHEBI_27744	0.2	ANVISA
http://purl.obolibrary.org/obo /CHEBI_27864	20.0	ANVISA
http://purl.obolibrary.org/obo /CHEBI_28909	10.0	SENASA
http://purl.obolibrary.org/obo /CHEBI_28909	5.0	ANVISA
http://purl.obolibrary.org/obo /CHEBI_3015	0.5	SENASA

Figura 6.26: Principios activos utilizados en Durazno

Esta consulta SPARQL retorna todos los principios activos que pueden ser utilizados para duraznos según ANVISA y SENASA. Es interesante resaltar que al utilizar el término más genérico para este cultivo (URI de AGROVOC `http://aims.fao.org/aos/agrovoc/c_5638`) no es necesario especificar el nombre en cada idioma ("Durazno" en español y "Pêssego" en portugués). Por esta misma razón, esta consulta seguirá funcionando aun cuando se agreguen más datasets a futuro, siempre y cuando éstos estén en alguno de los idiomas soportados por AGROVOC (actualmente el término "Durazno" se encuentra en 24 idiomas).

Es fácil de observar en la figura 6.26 que no todas las sustancias están permitidas por ambos organismos. Por ejemplo, el elemento "indaziflam" (ChEBI URI `http://purl.obolibrary.org/obo/CHEBI_133237`) está autorizado para su uso en duraznos por SENASA y no por ANVISA, mientras que el caso contrario ocurre para "proexadiona cálcica" (ChEBI URI `http://purl.obolibrary.org/obo/CHEBI_137425`). También es posible ver que, si bien algunos principios activos son comunes en ambos datasets, el valor del LMR puede diferir. Es el caso de "iprodione" (ChEBI URI `http://purl.obolibrary.org/obo/CHEBI_28909`), que para ANVISA se permiten hasta 5mg/Kg mientras que SENASA autoriza hasta 10mg/Kg.

Obtener esta información con los documentos originales, si bien no sería imposible,

puede resultar verdaderamente tedioso. Este tipo de consultas permite ver las ventajas de las tecnologías de la web semántica (en este caso los datos en formato RDF y las consultas SPARQL) para la publicación de datos normativos frente a las modalidades tradicionales.

3. **Principios activos que se usan en el mismo cultivo en ambos dataset pero en diferente cantidad**

```
PREFIX schema: <http://schema.org/>
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT ?activePrinciple ?name
WHERE {
  {
    ?cropAR schema:sameAs ?name.
    ?recordAR lmro:appliesTo ?cropAR.
    ?recordAR lmro:activePrinciple ?activePrinciple.
    ?recordAR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publication>.
    ?recordAR lmro:maximumResidue ?maximumResiudeAR.
    ?maximumResiudeAR lmro:hasValue ?lmrAR.
  }
  FILTER EXISTS
  {
     ?cropBR schema:sameAs ?name.
    ?recordBR lmro:appliesTo ?cropBR.
     ?recordBR lmro:activePrinciple ?activePrinciple.
    ?recordBR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/br/Publication>.
     ?recordBR lmro:maximumResidue ?maximumResiudeBR.
     ?maximumResiudeBR lmro:hasValue ?lmrBR.
    FILTER(?lmrAR != ?lmrBR)
  }

}
LIMIT 5
```

activePrinciple	name
http://purl.obolibrary.org/obo /CHEBI_81833	http://aims.fao.org/aos/agrovoc /c_13551
http://purl.obolibrary.org/obo /CHEBI_34714	http://aims.fao.org/aos/agrovoc /c_541
http://purl.obolibrary.org/obo /CHEBI_81848	http://aims.fao.org/aos/agrovoc /c_13551
http://purl.obolibrary.org/obo /CHEBI_35883	http://aims.fao.org/aos/agrovoc /c_14477
http://purl.obolibrary.org/obo /CHEBI_3009	http://aims.fao.org/aos/agrovoc /c_7805

Figura 6.27: Sustancias con diferente LMR para un mismo cultivo

Esta consulta de SPARQL nos permite ver 5 sustancias habilitadas tanto por ANVISA como por SENASA, para el uso en el mismo cultivo, pero en cantidades diferentes. Estas sustancias, que se pueden ver en la figura 6.27, son: "trifloxistrobin" en papa (0.2mg/Kg en Argentina y 0.02mg/Kg en Brasil), "dimetoato" en manzana (0.5mg/Kg en Argentina y 2mg/Kg en Brasil), "dimetomorf" (0.1mg/Kg en Argentina y 0.03mg/Kg en Brasil), "benalaxil" en tomate (0.5mg/Kg en Argentina y 0.1mg/Kg en Brasil) y "fosfina" en soja (0.01mg/Kg en Argentina y 0.1mg/Kg en Brasil).

En los datasets generados hay un total de 519 principios activos en esta situación.

4. **Principios activos que se usan en el cultivo pera en Argentina pero no en Brasil (limitado a los primeros 5)**

```
PREFIX schema: <http://schema.org/>
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT ?activePrinciple
WHERE {
  {
    ?cropAR schema:sameAs <http://aims.fao.org/aos/agrovoc/c_5645>.
    ?recordAR lmro:appliesTo ?cropAR.
    ?recordAR lmro:activePrinciple ?activePrinciple.
    ?recordAR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publication>
  }
  FILTER NOT EXISTS
  {
     ?cropBR schema:sameAs <http://aims.fao.org/aos/agrovoc/c_5645>.
    ?recordBR lmro:appliesTo ?cropBR.
     ?recordBR lmro:activePrinciple ?activePrinciple.
    ?recordBR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/br/Publication>.
  }
}
LIMIT 5
```

activePrinciple
http://purl.obolibrary.org/obo/CHEBI_52136
http://purl.obolibrary.org/obo/CHEBI_39346
http://purl.obolibrary.org/obo/CHEBI_39382
http://purl.obolibrary.org/obo/CHEBI_132592
http://purl.obolibrary.org/obo/CHEBI_39211

Figura 6.28: Sustancias autorizadas para pera en Argentina

Vemos en la figura 6.28 las URIs que representan a los principios activos “glufosinato de amonio”, “esfenvalerato”, “flufenoxuron”, “metilciclopropeno” y “spinosad”, los cuales son autorizados por SENASA para ser utilizados en la pera (el cual está representado por la URI AGROVOC `http://aims.fao.org/aos/agrovoc/c_5645`) pero no por ANVISA.

Esta consulta nos permite rápidamente ver para un determinado cultivo qué principios activos están autorizados en un país pero no en el otro. De igual manera, podemos invertir las subconsultas para obtener el resultado opuesto (los principios permitidos

en Brasil pero no en Argentina) o eliminar el operador NOT para obtener las sustancias que ambos países comparten. De esta manera podemos saber que para este cultivo en particular, hay un total de 108 sustancias autorizadas para Argentina que no lo están para Brasil, 10 sustancias en Brasil no autorizadas en Argentina y 12 principios activos habilitados en ambos países.

Obtener esta información con los datasets originales es sumamente dificultoso, ya que requiere analizar ambas fuentes minuciosamente, identificando principios activos en idiomas diferentes. Este proceso se debe repetir para cada cultivo diferente que el usuario desee analizar, pudiendo llevar horas sólo para una docena de cultivos. Por otro lado, con el dataset semántico solo basta con actualizar la URI que identifica el cultivo deseado en ambas subconsultas para obtener los resultados deseados.

5. **Principios activos permitidos en Brasil pero no en Argentina (limitado a los primeros 5)**

```
PREFIX schema: <http://schema.org/>
PREFIX lmro: <http://www.lifia.info.unlp.edu.ar/lmro#>


SELECT distinct ?activePrinciple
WHERE {
  {
    ?recordAR lmro:activePrinciple ?activePrinciple.
    ?recordAR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/br/Publication>
  }
  FILTER NOT EXISTS
  {
     ?recordBR lmro:activePrinciple ?activePrinciple.
    ?recordBR schema:isPartOf
    <http://www.lifia.info.unlp.edu.ar/lmro/data/ar/Publication>.
  }
}
LIMIT 5
```

activePrinciple
http://purl.obolibrary.org/obo/CHEBI_38963
http://purl.obolibrary.org/obo/CHEBI_30473
http://purl.obolibrary.org/obo/CHEBI_133997
http://purl.obolibrary.org/obo/CHEBI_39348
http://purl.obolibrary.org/obo/CHEBI_83060

Figura 6.29: Sustancias autorizadas sólo en Brasil

En esta consulta se obtienen las primeras 5 sustancias de aquellas aprobadas por ANVISA en Brasil pero que por el contrario no fueron autorizadas en Argentina por parte del SENASA. Estas sustancias son “hidróxido de fentina”, “triazofós”, “flupiradifurone”, “etofenprox”, y “Ditiocarbamato”.

En total 70 principios activos se utilizan en Brasil que no están autorizados en Argentina.

Capítulo 7

Conclusiones y trabajo futuro

7.1. Conclusiones

A lo largo de este trabajo se obtuvo documentación que permitió interpretar y comparar datasets de diferentes organismos (ANVISA y SENASA) y versiones. Dicha documentación se encontraba en formatos e idiomas diferentes.

Posteriormente se creó una ontología, LMR-O, que permite mediante la reutilización de ontologías existentes y vocabularios propios describir el dominio de conocimiento de la publicación de datos normativos de LMR. Esta ontología permite describir de manera inequívoca datos y relaciones entre los datos que, como hemos visto en su formato original, pueden resultar ambiguos. Al mismo tiempo, poseer los conceptos que representan a las entidades del dominio permite eliminar errores de tipeo.

Junto con la ontología se desarrolló un proceso robusto, reproducible y trazable, y se brindaron las herramientas que sientan las bases para la transformación de los datasets tradicionales a datasets semánticos.

Con los datasets obtenidos nos propusimos demostrar las ventajas que se pueden obtener utilizando las tecnologías de la web semántica. La publicación de datos normativos sobre los residuos de productos fitosanitarios mediante el uso de estas tecnologías propicia la interoperabilidad ya que la información se encuentra en un formato diseñado para ser leído por máquinas. Esto permite que los datos puedan ser consumidos fácilmente por agentes inteligentes. Mediante el uso de ontologías y recursos semánticos se fomenta la colaboración entre los distintos actores del dominio, eliminando barreras como el idioma o el soporte en el que se encuentran los datos. Un investigador en Brasil, sólo necesita hacer la consulta SPARQL adecuada para obtener el valor de residuo máximo de un producto fitosanitario para un cultivo en particular, sin la necesidad de conocer el idioma original en que se publicó esta información. De la misma manera podría hacer la misma consulta en su dataset local para realizar comparaciones de datos, todo esto sin tener que realizar transformación alguna para poder ver la información.

La información de proveniencia en los dataset semánticos brinda confianza sobre el origen de los datos y permite realizar una trazabilidad a las fuentes originales. A medida que los

conjuntos de datos crecen con las subsecuentes publicaciones, es posible también seguir la evolución en la normativa de los fitosanitarios. Un investigador podría no solo consultar el valor del LMR para una sustancia actualmente, sino también en un punto específico en el tiempo. Actualmente esto no es posible en los datasets analizados.

7.2. Trabajo futuro

A pesar de los resultados obtenidos aun queda mucho camino por recorrer.

Las ontologías elegidas para formar parte de LMR-O modelan un dominio cuyo conjunto de conocimiento se intersecta con el de la publicación normativa de datos de LMR. Sin embargo muchas veces estos conjuntos no son iguales. Un claro ejemplo sucede con AGROVOC. Este tesauro modela áreas que la FAO considero de interés, incluyendo alimentación, nutrición, agricultura, pescadería, forestación y medio ambiente. A pesar de esto hay muchos conceptos de la industria alimenticia y agropecuaria que no se encuentran en el vocabulario. Para mejorar este aspecto, sería interesante ampliar las ontologías de base de LMR-O, agregando conceptos, sinónimos y traducciones, de manera que éstas vayan ganando en especificidad respecto al dominio que se pretende representar. Un ejemplo de esto se llevó a cabo en 5.2 con la ontología ChEBI. Estas tareas de enriquecimiento de las ontologías podrían incluso hacerse mediante una interfaz gráfica que permita al usuario decidir si para un término no encontrado en la ontología correspondiente, agregar un sinónimo o crear un nuevo concepto. Dicha interfaz también podría contribuir en la proveniencia de la información, agregando para cada elemento nuevo, la metainformación correspondiente, como quién agregó el término o en qué fecha entre otros.

De la misma manera que se mejoran y se especializan las ontologías que la componen, sería un trabajo continuar trabajando sobre LMR-O. Si bien esta primera iteración fue suficiente para este trabajo de tesis, esta ontología puede seguir creciendo para abarcar más áreas del conocimiento del dominio. Nuevos conceptos como tiempos de carencia o momentos de aplicación de fitosanitarios, entre otros, podrían formar parte del vocabulario.

En cuanto a las herramientas también existen limitaciones. En OpenRefine, por cada registro disponible se crea al menos una entidad semántica. Esto genera el inconveniente de que ciertas entidades que deben ser únicas no puedan ser creadas utilizando esta funcionalidad. Por ejemplo, el usuario sólo necesita representar una única vez al documento original del que se tomaron los datos (representado por el concepto lmro:SourceDocument), no tendría sentido que hubiera una representación de este elemento por cada principio activo en el dataset. Por situaciones como esta, en este trabajo se utilizan algoritmos ad hoc para realizar ciertos operaciones. Sin embargo, como mencionamos en 3.2.1, OpenRefine es una herramienta de código abierto que podría eventualmente ser mejorada para convertirse en la herramienta definitiva del pipeline de transformación a LMR-O.

Finalmente sería interesante a futuro ver crecer el conjunto de datos agregando más versiones y más datasets. Un dataset más rico y variado presentará más oportunidades de análisis de datos y podría incluso motivar el uso de los mismos en otras áreas del a informática como la minería de datos o el business intelligence.

Apéndice A

Algoritmos de soporte

Este anexo contiene información detallada de los scripts que se crearon para dar soporte al proceso de transformación a la ontología propuesta, LMR-O. Estos algoritmos podrán ser encontrados en el repositorio de la tesina (`https://github.com/cfragout/interoperabilidad-semantica-lmro`).

Todos los algoritmos fueron escritos en Javascript y utilizan NodeJS para correr (probados en la versión 10.20.1). Antes de ejecutar cada uno de estos algoritmos es necesario instalar sus dependencias corriendo el siguiente comando:

```
$ npm install
```

A.1. ontology-augmenter

A.1.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Algoritmos/ontology-augmenter`

A.1.2. Descripción

Este algoritmo se utilizó para extender ChEBI a fin de realizar pruebas respecto a la cantidad de alineaciones automáticas que se obtienen con diferentes versiones de la ontología mencionada.

El script posee dos entradas, la ontología original y la lista de elementos que serán insertados en formato JSON. En cada elemento de la lista se especifican sus atributos y contenido, pudiendo agregar estructuras anidadas. El algoritmo recorre la lista de elementos, agregando los nuevos términos con sus atributos, namespaces, elementos anidados y contenido en general. Cabe destacar que este algoritmo debería poder utilizarse para extender cualquier tipo de archivo en formato RDF.

Se encuentra en el repositorio el archivo `ejemplos/nuevos_aptitud_principios-activos-ar-br` que puede ser utilizado como entrada de este algoritmo así como también una versión reducida de ChEBI utilizada para realizar pruebas rápidas.

A.1.3. Ejecución y parámetros

```
$ node --max-old-space-size=9216 index -t chebi.owl -i principios-activos.json
-n "www.lifia.info.unlp.edu.ar/data/lmro/"
```

- -t: archivo que contiene la ontología.
- -i: el archivo GREL de OpenRefine con los conceptos que se van a agregar a la ontología.
- -o: el nombre del archivo de salida.
- -n: URI que se utilizará como namespace en los nuevos conceptos.

A.2. chebi-synonyms

A.2.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Algoritmos/chebi-synonyms`

A.2.2. Descripción

Este algoritmo itera sobre cada elemento del archivo GREL de entrada. Este archivo, obtenido de OpenRefine, posee las alineaciones manuales que se hicieron en el dataset. El objetivo de este script es que esas alineaciones sean agregadas a la ontología para que luego OpenRefine, en subsecuentes reconciliaciones de nuevos datasets, encuentre automáticamente los conceptos correctos. Por ejemplo, en ChEBI alinearemos manualmente el término "herbicida" con el concepto "herbicide". Una vez ejecutado el algoritmo, tendremos nuestra versión extendida de ChEBI, la cual utilizaremos como servicio de reconciliación en OpenRefine y donde automáticamente se alineará la cadena "herbicida" con el concepto correcto.

Se encuentra en el repositorio el archivo `ejemplos/aptitud_principio-activo-ar.json` que puede ser utilizado como entrada de este algoritmo así como también una versión reducida de ChEBI utilizada para realizar pruebas rápidas.

A.2.3. Ejecución y parámetros

```
$ node index -t chebi.owl -i aptitud_principio-activo.json
```

- -t: archivo que contiene la ontología.
- -i: el archivo GREL de OpenRefine con los sinónimos que se van a agregar a la ontología.
- -o: el nombre del archivo de salida.

A.3. lmro-corrections

A.3.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Algoritmos/lmro-corrections`

A.3.2. Descripción

La función principal de este algoritmo es resolver aquellos puntos para los que no funciona correctamente o no es suficiente la exportación a RDF de OpenRefine. Este sencillo script recorre cada uno de los registros lmro:Record exportados y realiza las tres operaciones siguientes:

1. **Agregar concepto lmro:Any para el rol**

   Si el registro no posee no posee rol, entonces se asume que puede el principio activo puede cumplir cualquier rol.

2. **Resolver valor del LMR**

   Si el registro no posee un valor válido (es decir, numérico) para el LMR, entonces se asume que el principio activo está exento. Esto se representa mediante el concepto lmro:Exempt. En caso contrario, si el valor es válido, se crea un concepto lmro:ResidueValue con el valor numérico y la unidad de medida (por ejemplo mg/Kg o partes por millón).

3. **Resolver valor del cultivo**

   Si el registro no posee un valor para la propiedad lmro:appliesTo, entonces se asume que el principio activo es válido para cualquier cultivo. Esto se representa mediante el concepto lmro:Any. Por otro lado, si lmro:appliesTo posee un valor, este puede ser una URI, identificando un label de AGROVOC, o una cadena de caracteres indicando que el cultivo no se pudo alinear con la ontología. Para el primer caso, se buscará el concepto más amplio (denominado broader) en AGROVOC y se lo utilizará en lugar del label original. Si el valor es una cadena de caracteres, la propiedad se mantendrá sin cambios.

Se encuentra en el repositorio el archivo `ejemplos/dataset-ar.json` que puede ser utilizado como entrada de este algoritmo así como también una versión reducida de AGROVOC utilizada para realizar pruebas rápidas.

A.3.3. Ejecución y parámetros

```
$ node --max-old-space-size=9216 index -a agrovoc.rdf -i dataset-ar.rdf -n
http://www.lifia.info.unlp.edu.ar/lmro/data/ar/
```

- -a: archivo de AGROVOC que se utilizará para obtener información de cultivos.
- -u: URI de la unidad de medida que se utilizará en el dataset.
- -n: namespace que se utilizará en el dataset.
- -i: el archivo RDF que contiene el dataset.
- -o: el nombre del archivo de salida.

A.4. lmro-provenance

A.4.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Algoritmos/lmro-provenance`

A.4.2. Descripción

El algoritmo de generación de entidades de proveniencia para LMR-O utiliza como entrada el dataset semántico generado por el script "lmro-corrections" y un pequeño archivo en formato JSON con información adicional. Esta información consiste en:

- namespace: URI que se utilizará como namespace en las entidades de provenance.
- sourceDocument: objeto con información de provenance para el documento fuente que se utilizó para generar el dataset.
  - date: fecha de publicación o creación del archivo.
  - format: formato del archivo (por ejemplo xlsx o csv).
  - language: idioma del archivo.
  - description: breve descripción del archivo.
  - title: titulo original del archivo.
  - identifier: hash que identifica el archivo. También se puede utilizar para determinar si el archivo original fue modificado.
  - createdBy: información del creador o publicador del archivo original.
    - name: nombre del organismo que publica el archivo original (por ejemplo SENASA o ANVISA).
    - uri: URI que puede utilizarse para identificar unívocamente a la institución. Puede utilizarse, por ejemplo, una URI de wikidata.
- publisher: información del creador del dataset semántico.

- name: Nombre del organismo que se encuentra creando el dataset semántico (por ejemplo Universidad NAcional de La Plata).
- uri: URI que puede utilizarse para identificar unívocamente a la institución. Puede utilizarse, por ejemplo, una URI de wikidata.

Se encuentra en el repositorio el archivo `ejemplos/provenance-ar.json` que puede ser utilizado como entrada de este algoritmo así como también un dataset obtenido luego de ejecutar el algoritmo "lmro-corrections".

A.4.3. Ejecución y parámetros

```
$ node --max-old-space-size=9216 index -p provenance.json -i dataset.rdf -o
dataset-with-provenance.rdf
```

- -p: el archivo con la informacion de provenance.
- -i: el archivo RDF que contiene el dataset.
- -o: el nombre del archivo de salida.

Apéndice B

Documentos adicionales

Se adjunta en el repositorio de esta tesina el archivo OWL con la definición de la ontología LMR-O creada en Protégé, así como los documentos originales utilizados para cada uno de los casos de prácticos descriptos en el capítulo 5. Además están disponibles los archivos GREL exportados de OpenRefine con las secuencias de pasos realizadas para formatear y alinear los datasets.

B.1. LMR-O

El archivo LMR-O.owl, que se encuentra ubicado en `https://github.com/cfragout/interoperabilidad-semantica-lmro/blob/master/LMR-O/LMR-O.owl`, contiene la especificación de la ontología LMR-O. La ontología posee las siguientes anotaciones:

- schema:name: LMR-O
- schema:description: LMR-O es una ontología que busca modelar el dominio de la publicación de los límites mpaximos de residuos (o LMR) de fitosanitarios en alimentos.
- schema:creator: Carlos Francisco Ragout
- schema:contributor: Alejandro Fernandez, Diego Torres, Carlos Pintor
- schema:dateCreated: 2020-12-13
- schema:schemaVersion: 1.0.0

B.2. Documentos relativos al caso práctico I

B.2.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Caso%20Practico%20I%20-%20Argentina`

B.2.2. Descripción

- 1. `formateo-ar.json`: archivo GREL con los pasos realizados para formatear el dataset de Argentina.
- 2. `alineacion-ar.json`: archivo GREL con los pasos realizados para alienar el dataset de Argentina.
- 3. `esqueleto-rdf-ar.json`: archivo GREL con la plantilla utilizada por la extensión RDF de OpenRefine para exportar el dataset.
- 4. `provenance-ar.json`: archivo JSON con la información de proveniencia que se utilizó como entrada en el algoritmo "lmro-provenance".
- `dataset-LMR-0-ar-with-provenance.rdf`: archivo RDF con el resultado final de aplicar el pipeline de transformación de datos.
- `lmrs_julio_2020.xlsx`: archivo original publicado por el SENASA en su sitio web oficial. Posee la información normativa respecto a los LMR del mes de Julio de 2020.
- `ChEBI/nuevos_aptitud_principios-activos-ar-br.json`: archivo JSON que se utilizó como entrada para el script "chebi-augmenter". Contiene los conceptos nuevos que se agregaron en la ontología ChEBI.
- `ChEBI/aptitud_principio-activo-ar.json`: archivo GREL que se utilizó como entrada para el script "chebi-synonyms". Contiene las asociaciones manuales hechas en OpenRefine y se utiliza para extender la ontología ChEBI.

B.3. Documentos relativos al caso práctico II

B.3.1. Ubicación

`https://github.com/cfragout/interoperabilidad-semantica-lmro/tree/master/Caso%20Practico%20II%20-%20Brasil`

B.3.2. Descripción

- 1. `formateo-br.json`: archivo GREL con los pasos realizados para formatear el dataset de Brasil.
- 2. `alineacion-br.json`: archivo GREL con los pasos realizados para alienar el dataset de Brasil.
- 3. `esqueleto-rdf-br.json`: archivo GREL con la plantilla utilizada por la extensión RDF de OpenRefine para exportar el dataset.

- 4. `provenance-br.json`: archivo JSON con la información de proveniencia que se utilizó como entrada en el algoritmo "lmro-provenance".
- `dataset-LMR-0-br-with-provenance.rdf`: archivo RDF con el resultado final de aplicar el pipeline de transformación de datos.
- `TA_MONOGRAFIA_AGROTOXICO.csv`: archivo original exportado del sitio oficial de ANVISA. Posee la información normativa respecto a los LMR de Brasil hasta el mes de Noviembre de 2020.
- `ChEBI/nuevos_aptitud_principios-activos-ar-br.json`: archivo JSON que se utilizó como entrada para el script "chebi-augmenter". Contiene los conceptos nuevos que se agregaron en la ontología ChEBI.
- `ChEBI/principio-activo-br.json`: archivo GREL que se utilizó como entrada para el script "chebi-synonyms". Contiene las asociaciones manuales hechas en OpenRefine y se utiliza para extender la ontología ChEBI.

Bibliografía

[1] A. D. N. Fernández, V. Viciana, "Valoración del impacto ambiental total por agroquímicos en la cuenca del rio mendoza."

[2] J. M. G.-B. G. F.-D. d. P. V. I. M. Eva Mª Martín Cruz, Jose Luis Alonso-Prados, "Límites máximos de residuos de productos fitosanitarios en alimentos y piensos," *Phytoma*, no. 199, 2008.

[3] "Estudio de mercado 2014 de productos de protección de cultivos." `https://www.casafe.org/pdf/2018/ESTADISTICAS/Informe-Mercado-Fitosanitarios-2014.pdf`. Accedido: 2021-01-11.

[4] "Pesticide residues in food." `https://www.who.int/news-room/fact-sheets/detail/pesticide-residues-in-food`. Accedido: 2021-01-11.

[5] M. R. Jan, J. Shah, M. A. Khawaja, and K. Gul, "Ddt residue in soil and water in and around abandoned ddt manufacturing factory," *Environmental monitoring and assessment*, vol. 155, no. 1-4, pp. 31–38, 2009.

[6] C. Brown, "Semantic web technologies for data curation and provenance," in *19th International Congress of Metrology (CIM2019)*, p. 26002, EDP Sciences, 2019.

[7] "World wide web." `https://en.wikipedia.org/wiki/World_Wide_Web`. Accedido: 2021-01-11.

[8] "Cisco annual internet report (2018–2023) white paper." `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`. Accedido: 2021-01-11.

[9] "We knew the web was big...." `https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`. Accedido: 2021-01-11.

[10] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[11] M. Ruiz-Casado, E. Alfonseca, and P. Castells, "Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia," *Data & Knowledge Engineering*, vol. 61, no. 3, pp. 484–499, 2007.

[12] "Web semantica." https://es.wikipedia.org/wiki/Web_sem%C3%A1ntica. Accedido: 2021-01-11.

[13] R. Q. Dividino, "Managing and using provenance in the semantic web," 2017.

[14] "What is data wrangling and why does it take so long?." https://www.elderresearch.com/blog/what-is-data-wrangling-and-why-does-it-take-so-long, Apr. 2018. Accedido: 2020-09-10.

[15] "General refine expression language." https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language. Accedido: 2021-01-11.

[16] "Reconciliation service api." https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API. Accedido: 2020-08-11.

[17] A. Seaborne and E. Prud'hommeaux, "SPARQL query language for RDF," W3C recommendation, W3C, Jan. 2008. https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[18] "Sparql." https://en.wikipedia.org/wiki/SPARQL. Accedido: 2021-01-13.

[19] T. Gruber, *Ontology*, pp. 1963–1965. Boston, MA: Springer US, 2009.

[20] M. C. Suárez-Figueroa, "Neon methodology for building ontology networks: specification, scheduling and reuse," in *NeOn methodology for building ontology networks: specification, scheduling and reuse*, 2011.

[21] "Agrontology." http://aims.fao.org/zh-hans/agrovoc/agrontology. Accedido: 2020-09-20.

[22] "Agrovoc tesauro multilingüe de agricultura." http://aims.fao.org/vest-registry/vocabularies/agrovoc. Accedido: 2020-09-20.

[23] A. Isaac and E. Summers, "SKOS simple knowledge organization system primer," W3C note, W3C, Aug. 2009. https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

[24] A. Miles and S. Bechhofer, "Skos simple knowledge organization system extension for labels (skos-xl) namespace document," W3C note, W3C, Aug. 2009. https://www.w3.org/TR/2009/REC-skos-reference-20090818/skos-xl.html#SKOS-REFERENCE.

[25] "Agrontology faq." http://aims.fao.org/standards/agrovoc/faq. Accedido: 2020-09-20.

[26] S. Sahoo, T. Lebo, and D. McGuinness, "PROV-o: The PROV ontology," W3C recommendation, W3C, Apr. 2013. https://www.w3.org/TR/2013/REC-prov-o-20130430/.

[27] L. Moreau and P. Missier, "PROV-dm: The PROV data model," W3C recommendation, W3C, Apr. 2013. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[28] P. Missier, K. Belhajjame, and J. Cheney, "The w3c prov family of specifications for modelling provenance metadata," in *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 773–776, 2013.

[29] "Manual de usuario chebi." https://docs.google.com/document/d/1_w-DwBdCCOh1gMeeP6yqGzcnkpbHYOa3AGSODe5epcg/edit#. Accedido: 2020-09-10.

[30] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "Chebi in 2016: Improved services and an expanding collection of metabolites," *Nucleic acids research*, vol. 44, no. D1, pp. D1214–D1219, 2016.

[31] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D344–D350, 2007.

[32] "The relation ontology." http://www.obofoundry.org/ontology/ro.html. Accedido: 2020-09-10.

[33] "La ontología agronómica." http://www.obofoundry.org/ontology/agro.html. Accedido: 2020-09-13.

[34] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, E. Consortium, *et al.*, "The environment ontology: contextualising biological and biomedical entities," *Journal of biomedical semantics*, vol. 4, no. 1, p. 43, 2013.

[35] D. Garijo and M. Poveda-Villalón, "Sobre agro." https://bigdata.cgiar.org/resources/agronomy-ontology/. Accedido: 2020-09-13.

[36] C. Aubert, P. Buttigieg, M. Laporte, M. Devare, and E. Arnaud, "Cgiar agronomy ontology," *CGIAR Agronomy Ontology*, 2017.

[37] "The obo foundry." http://www.obofoundry.org/l. Accedido: 2020-09-13.

[38] "A checklist for complete vocabulary metadata." https://w3id.org/widoco/bestPractices. Accedido: 2020-12-13.