UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CONECTANDO BIBKG CON WIKIDATA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

JORGE ALONSO CERDA VERGARA

PROFESOR GUÍA:
AIDAN HOGAN

MIEMBROS DE LA COMISIÓN:
ANDRÉS ABELIUK KIMELMAN
EDUARDO GODOY VEGA

SANTIAGO DE CHILE
2023

Resumen

BibKG es un grafo de conocimiento especializado en almacenar información relacionada con publicaciones académicas del área de las ciencias de la computación. Este proyecto está en fase temprana de desarrollo, y busca seguir aumentando su cantidad de datos almacenados. Actualmente, BibKG sólo posee 2 fuentes de datos de información, ArnetMiner y DBLP.

La idea del presente trabajo es facilitar a BibKG la incorporación de una tercera fuente de datos, Wikidata. Mientras que tanto ArnetMiner como DBLP poseen datos más orientados a información de datos académicos, Wikidata es un grafo de conocimiento más general, que posee almacenada información de todo tipo. De esta forma, permite al proyecto expandir sus posibilidades en cuanto a la creación de nuevas propiedades, en enriquecimiento de las propiedades ya existentes, y la posibilidad futura de acceder a nuevas fuentes de datos a las que tienen acceso las entidades de Wikidata, abriendo la puerta a seguir aumentando la cantidad de fuentes de información accesibles por BibKG en el futuro.

Para facilitar dicha tarea, en el presente trabajo se muestra el proceso en el cual se buscaron enlazar las entidades de BibKG con sus equivalentes en Wikidata, mediante la comparación de propiedades clave de estas que permitan definir con alta precisión y completitud si dos entidades son equivalentes o no. Para la realización de esta tarea se preprocesaron los datos de BibKG y Wikidata para poder manejarlos con mayor facilidad a la hora de realizar comparaciones, y se definieron distintos criterios de enlazamiento de entidades. Los enlazamientos realizados se guardaron en archivos en formato CSV, representando cada fila del archivo un enlazamiento realizado entre dos entidades de BibKG y Wikidata, junto con información relacionada a dicho enlazamiento.

Se lograron enlazar un 9,31 % del total de entidades de BibKG con Wikidata. De esta forma, se logró facilitar al proyecto de BibKG datos confiables para incorporarlos en su base de datos en el futuro, con información de cada enlazamiento que permita además al proyecto estimar si el enlace es seguro, mediante la asociación de cada enlazamiento con los tipos de enlazamiento con los que fue relacionado con la entidad de Wikidata, junto con algunos datos adicionales. Además, se creó un *parser* de BibKG, que permite transformar el formato de este a uno más sencillo de interpretar para la máquina, lo que puede ser de utilidad para futuros investigadores que quieran utilizar a dicho proyecto para diversos fines. De todos modos, existe un margen de mejora en cuanto a las metodologías utilizadas para enlazar datos, en particular en las que no se utilizaron identificadores externos para el enlazamiento de entidades.

Agradecimientos

Aprovecho esta instancia para agradecer a todos mis seres queridos que me han acompañado a lo largo de mi vida. Todos ellos de alguna forma han aportado de alguna forma para terminar este proceso, y para formarme como la persona que soy ahora.

Agradezco especialmente a mis padres Sergio y Verónica, que siempre han estado ahí con sus valiosas enseñanzas y apoyo incondicional, en las buenas y en las malas. No sé si algún día podré agradecerles todo lo que han hecho por mí, gran parte de lo que he logrado es gracias a la gran educación y al gran afecto que han tenido conmigo desde pequeño.

Agradezco a mi hermana Natalia por lo excelente que siempre te has portado conmigo. Buena parte de los recuerdos felices de mi infancia están relacionados contigo, estaré eternamente agradecido.

Agradezco también a mi hermanita Julieta, que quizás sin darse cuenta me ha ayudado a sobrellevar momentos complicados con su genuino cariño.

Finalmente, agradezco a mi profesor guía, Aidan. Siempre estuvo dispuesto a apoyarme con el proyecto, ha sido una persona extremadamente atenta y paciente conmigo a lo largo de todo este proceso. Creo personalmente que gran parte del progreso realizado en este proyecto no habría sido posible como tal sin él.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Las bases de datos de grafos permiten organizar la información en torno a relaciones entre entidades, objetos que pueden representar cualquier tipo de concepto definible por el ser humano. Este tipo de organización de datos tiene ciertas ventajas, como una mayor expresividad a la hora de realizar consultas en dicha base de datos, al poder expresar con mayor facilidad las propiedades de la búsqueda en cuanto al formato de las relaciones entre entidades relacionadas a lo requerido.

BibKG[1] es un proyecto creado por parte del Instituto Milenio Fundamentos de los Datos (IMFD), que consiste en un grafo de conocimiento con especial énfasis en bibliografías relacionadas a las Ciencias de la Computación, recopilando publicaciones académicas junto con sus datos asociados (autores, temáticas, referencias, etc.), y donde los usuarios pueden hacer consultas especificando las características de los enlaces de los documentos que desean encontrar. El proyecto es reciente, y la idea de los desarrolladores es la de seguir aumentando el número de publicaciones y entidades relacionadas accesibles por parte de la plataforma. Dicha base de datos funciona mediante el sistema de administración de bases de datos MillenniumDB [5].

Actualmente, si bien BibKG posee ya una cantidad razonable de asociaciones entre sus entidades con IDs externos, estos están limitados por la información proveída por sus fuentes de datos actuales, por lo que las posibilidades de enriquecer la cantidad de datos del proyecto dependen de dichos datos. Por esto, existe la necesidad de ampliar la red de conocimiento de BibKG, y de esta forma aumentar el rango de facultades del proyecto para poder ejecutar todo tipo de operaciones, y con ello aumentar las posibilidades tanto para los usuarios como para los desarrolladores.

Wikidata[2], por otra parte, consiste en un grafo de conocimiento bajo licencia de dominio público, en el que se almacenan datos más generales sobre diferentes entidades, cubriendo todo el espectro posible. A diferencia de las fuentes con las que cuenta actualmente BibKG, que son estrictamente orientadas a datos de publicaciones académicas, Wikidata es mucho más general, abarcando cualquier concepto definible por el ser humano que algún usuario

[1] `https://bibkg.imfd.cl/`

[2] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

de dicho servicio haya optado por agregar. Actualmente, si bien existen algunas entidades enlazadas de BibKG con Wikidata, están aún a una escala pequeña.

Al expandir las fuentes accesibles por BibKG, el proyecto ampliará considerablemente sus posibilidades, al tener acceso a un grafo de conocimiento tan grande como Wikidata. Por ejemplo, las entidades de publicaciones de Wikidata actualmente poseen más información en cuanto a la cantidad de referencias registradas hacia otras publicaciones que BibKG, junto con poseer más datos sobre las afiliaciones de los autores de las publicaciones. Además, Wikidata posee enlaces con otras fuentes de datos mediante sus IDs externos, por lo que enlazar BibKG con Wikidata permite utilizar a este último como "puente" para poder enlazar en el futuro a BibKG con otras fuentes, y seguir enriqueciendo su cantidad de datos.

Se necesitará del uso de herramientas que permitan analizar los datos ya existentes, y también de aplicaciones que permitan el enlazamiento masivo de datos debido a la cantidad considerable de datos que se requiere manejar, para poder crear un modelo que pueda comparar valores y relaciones entre entidades de BibKG y Wikidata. Este proceso debe ser ejecutado con la mayor completitud y eficiencia posible, y con un margen de error razonablemente bajo. Esto, para maximizar la cantidad de entidades de Wikidata accesibles por parte de BibKG, optimizar el tiempo y recursos utilizados para ejecutar los modelos creados y mantener una alta precisión en los datos obtenidos. Además, se requiere del análisis de los enlazamientos realizados para analizar la ya mencionada completitud y eficiencia del trabajo realizado por el modelo, para poder corroborar el grado de efectividad de este, y si es necesario mejorar el modelo para lograr un mejor resultado.

1.1. Objetivos

1.1.1. Objetivo General

El objetivo del proyecto es el de fortalecer la conexión entre las entidades de BibKG y Wikidata enlazando más entidades, y, de esta forma, enriquecer la cantidad de datos accesibles para BibKG. En particular, se requiere la elaboración de un sistema automático que permita encontrar entidades equivalentes entre BibKG y Wikidata, indicando cuáles autores, publicaciones, conferencias y revistas en una fuente corresponde a cuáles entidades análogas en la otra fuente. Este sistema debe ser lo más preciso y completo posible a la hora de detectar entidades equivalentes entre ambos grafos. El sistema debería quedar disponible para realizar el proceso mencionado anteriormente sobre futuras versiones de los datos de ambas plataformas.

1.1.2. Objetivos Específicos

1. Crear un analizador sintáctico (o *parser* en inglés) que permita almacenar la información del archivo *dump* de BibKG en formato JSON.
2. Crear un archivo JSON de Wikidata a partir del *dump* comprimido en formato *gz*, reduciendo el tamaño final mediante filtros de información.

3. Analizar los enlaces ya existentes entre BibKG y Wikidata para comprender su precisión y completitud.
4. Enlazar las entidades equivalentes de BibKG con Wikidata utilizando los IDs como método de comparación, analizando la completitud y precisión de los resultados del proceso.
5. Enlazar las entidades equivalentes de BibKG con Wikidata mediante distintas técnicas de comparación entre las diferentes propiedades de estas.
6. Computar el emparejamiento de entidades entre BibKG y Wikidata, junto con analizar su tamaño, el porcentaje de entidades emparejadas o no emparejadas, etc.

1.2. Evaluación

Se puede evaluar la efectividad y completitud del emparejamiento realizado estimando la cantidad de entidades existentes simultáneamente tanto en BibKG como en Wikidata que sí han sido enlazadas exitosamente, contra los que no se lograron enlazar. Como no existe un valor conocido que indique estos datos, se requiere de desarrollar alguna métrica que pueda estimar estos porcentajes. Además, se debe evaluar la precisión con la que se enlazaron las entidades, estimando de alguna forma el porcentaje de entidades enlazadas a través de la metodología utilizada que sí son equivalentes entre sí.

1.3. Estructura

La presente memoria posee la siguiente estructura:

- **Capítulo 2:** se presenta el estado del arte, donde se describen los principales conceptos relacionados con el trabajo.
- **Capítulo 3:** se presenta el proceso del preprocesamiento de los archivos de BibKG y Wikidata.
- **Capítulo 4:** se presenta el proceso de enlazamiento de entidades entre BibKG y Wikidata.
- **Capítulo 5:** se presentan los resultados del enlazamiento de entidades entre BibKG y Wikidata y la validación de dichos enlazamientos.
- **Capítulo 6:** se presentan las conclusiones del trabajo realizado, junto con el planteamiento de las limitaciones y el trabajo futuro.

Capítulo 2

Estado del Arte

En el presente capítulo se describirán los conceptos más importantes con respecto al trabajo realizado, con la finalidad de facilitar la lectura del trabajo realizado, junto con dar contexto sobre el problema y las herramientas utilizadas para la solución del mismo.

2.1. La Web Semántica

La Web Semántica[1] es un conjunto de actividades por medio de la *World Wide Web Consortium* (W3C), que tiene como objetivo el tener datos legibles por parte de máquinas, a diferencia de lo establecido por la Web tradicional.

Principalmente, los objetivos de la Web Semántica son los de tener un formato en común entre todos los datos extraídos por parte de diversas fuentes, y registrar cómo dichos datos se relacionan con los objetos de la vida real. De esta forma, una máquina tiene la posibilidad de moverse a través de la Web mediante una semántica estandarizada.

2.2. RDF

RDF[2] es un modelo de datos de la *World Wide Web Consortium*. Es un componente importante de la actividad de la Web Semántica de W3C, puesto que permite estructurar la información de tal forma que una máquina pueda comprender la información por la sintaxis presente. Consiste básicamente en datos representados mediante la estructura en tripletas de sujeto-predicado-objeto, en el que se establece una relación entre el sujeto y el objeto.

[1] `https://www.w3.org/2001/sw/`
[2] `https://www.w3.org/RDF/`

2.3. JSON

JSON[3], acrónimo del inglés *JavaScript Object Notation*, es un formato de texto de intercambio de datos, que permite organizar la información mediante una estructura definida. JSON está constituido por dos estructuras, visibles en el ejemplo de la Figura 2.1. Dichas estructuras son las siguientes:

1. Un objeto, definido como una colección de pares de nombre/valor. El nombre necesariamente debe ser una cadena de caracteres. Un objeto comienza con una llave de apertura ({) y termina con una llave de cierre (}). Cada nombre es seguido por dos puntos (:) y los pares nombre/valor están separados por una coma (,).
2. Un arreglo, definido como una lista ordenada de valores. Un objeto comienza con una llave de apertura ([) y termina con una llave de cierre (]). Los valores se separan por una coma (,).

En cuanto al formato de los valores, estos pueden ser números, cadenas de caracteres (también conocidas como *strings*), arreglos, valores de verdad (*true*, *false*), o un valor nulo (*null*).

Estas convenciones de estructuras son universales, por lo que virtualmente todos los lenguajes de programación las soportan de alguna manera. Por ejemplo, el lenguaje de programación Python posee los denominados diccionarios, un tipo de estructura de datos que posee una sintaxis similar a la de JSON.

2.4. MillenniumDB

MillenniumDB [5] es un sistema de administración de bases de datos orientados en grafos desarrollado por el IMFD, y es utilizado por BibKG para su funcionamiento. La idea de este sistema es el de mejorar la eficiencia de la representación de la información global como un grafo en comparación a los sistemas ya existentes, puesto que los desarrolladores estiman que los actuales no son lo suficientemente eficientes para dichos propósitos. El sistema está aún en desarrollo, por lo que no tiene aún todas las características esperadas por los desarrolladores.

El servicio de MillenniumDB permite cargar un archivo en forma de una base de datos de MillenniumDB, y crear un servidor para poder realizar consultas en alguna base de datos del mismo formato mediante el lenguaje de consultas del mismo. Este servicio está pensado para estar disponible para cualquier distribución de Linux x86-64, aunque se puede utilizar en Windows mediante Windows Subsystem for Linux (WSL).

El servicio prestado permite importar archivos de texto para transformarlos en una base de datos de MillenniumDB. El archivo debe tener un formato específico para poder ser leído correctamente.

[3] `https://www.json.org/json-es.html`

```json
{
    "name": "Alan Turing",
    "age": 41,
    "nationality": "British",
    "publications": [
        {
            "title": "Intelligent Machinery",
            "year": 1948,
            "journal": "National Physical Laboratory (NPL) Report"
        },
        {
            "title": "Computing Machinery and Intelligence",
            "year": 1950,
            "journal": "Mind"
        }
    ],
    "contributions": ["Turing Machine", "Cryptanalysis of the Enigma machine"],
    "occupations": ["mathematician", "logician", "cryptographer"]
}
```

Figura 2.1: Objeto JSON de ejemplo que hace referencia a Alan Turing, junto con algunas propiedades relacionadas. Se pueden observar que tanto números como strings, arreglos y otros objetos JSON son válidos como valores de una propiedad.

El lenguaje de consultas de MillenniumDB posee una sintaxis similar a la del lenguaje de consultas de Cypher [1], con ligeras diferencias en esta. Además, al estar esta característica aún en desarrollo, posee ciertas limitaciones con respecto a su símil de Cypher. Por ejemplo, por ahora existe un repertorio restringido en cuanto a la cantidad de operadores que se pueden utilizar a comparación a Cypher. Por ejemplo, el contar el número de resultados de alguna consulta con `COUNT` no funciona correctamente.

2.5. BibKG

BibKG[4], una iniciativa del Instituto Milenio Fundamentos de los Datos (IMFD), consiste en un sitio web en la que el usuario puede realizar consultas para obtener resultados de la base de datos del proyecto, que almacena datos de publicaciones académicas, junto con las entidades asociadas (autores, revistas, etc). Cada entidad puede poseer una o más propiedades, las cuales las pueden relacionar con valores (números, strings, etc.) como con otras entidades. Por ejemplo, la propiedad *author_of* relaciona a una persona con una o más publicaciones de las que haya sido autor o co-autor. Cada entidad posee una URI propia dentro de la página de BibKG (con su ID de BibKG como identificador), consistente en una vista de sus propiedades con sus respectivos valores, como se observa en la Figura 2.2.

[4] `https://bibkg.imfd.cl/`

wikidata	Q51366847
key	homepages/h/AidanHogan
mdate	2019-01-10
name	Aidan Hogan
orcid	0000-0001-9482-1982
scholar	CP-fgY4AAAAJ

author_of	• Towards a scalable search and query engine for the web. • Weaving the Pedantic Web. • Scalable Authoritative OWL Reasoning for the Web.

Figura 2.2: Perfil de BibKG de la entidad `a_Aidan_Hogan` correspondiente a Aidan Hogan, junto con algunas de sus propiedades. (Fuente: `https://bibkg.imfd.cl/identifier/a_Aidan_Hogan`)

La página web permite el acceso a las entidades presentes en la base de datos mediante los siguientes métodos de búsqueda:

1. Mediante un buscador[5] que permite buscar entidades según lo señalado en la propiedad *name* de esta. Se requiere escribir el nombre específico de lo que se quiere buscar, puesto que no se arrojan resultados con nombres similares al hacer la búsqueda ni existe un autocompletado a partir de lo escrito.
2. Mediante un servicio de consultas[6], en el que utiliza como lenguaje de consulta el propio de MillenniumDB, visible en la Figura 2.3. Este permite una alta personalización en cuanto a la realización de consultas debido a las opciones que permite el lenguaje de consultas de MillenniumDB (según las condiciones de las propiedades de las entidades), aunque requiere cierto nivel de dominio del lenguaje de consultas para ser utilizado correctamente.

En cuanto a la información almacenada en BibKG, esta está en forma de una base de datos orientada a grafos, donde se representa esta a través de entidades que se relacionan entre sí a través de enlaces, que representan algún tipo de relación entre ellas, similar a los vértices de un grafo unidos por aristas.

Actualmente, las fuentes de las que se alimenta BibKG para la obtención de enlaces son principalmente DBLP [2] y ArnetMiner [4], repositorios bibliográficos orientados a almacenar

[5] `https://bibkg.imfd.cl/browser/`
[6] `https://bibkg.imfd.cl/query/`

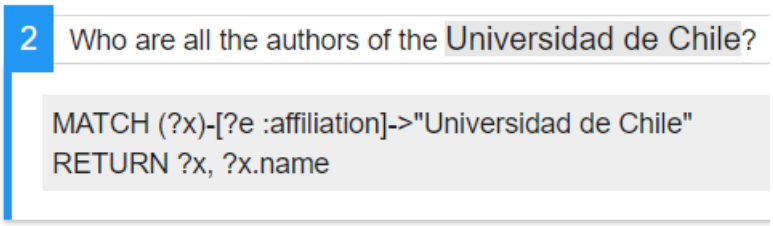

Figura 2.3: Ejemplo de consulta en el servicio de consultas de BibKG, mediante el lenguaje propio de MillenniumDB. (Fuente: `https://bibkg.imfd.cl/`)

enlaces relacionados a publicaciones científicas. En particular, DBLP se encarga de almacenar artículos relacionados a las ciencias de la computación, mientras que ArnetMiner almacena enlaces correspondientes a publicaciones académicas en general.

BibKG se encuentra aún en una etapa temprana de desarrollo, por lo que el proyecto tiene como objetivo el seguir enriqueciendo la base de datos del mismo. En particular, se tiene la intención de incorporar a Wikidata como nuevo grafo de conocimiento, de tal manera de tener acceso a sus datos. Debido a que Wikidata está diseñado para incorporar cualquier concepto definible por el ser humano como una entidad, esto permitiría expandir notablemente los posibles enlazamientos de BibKG. Actualmente existen algunos enlaces entre entidades de BibKG y Wikidata, pero son una cantidad muy pequeña considerando la magnitud de la base de datos. En particular, BibKG posee un total de 48.255 entidades enlazadas con Wikidata[7], siendo todas referentes a autores, mientras que BibKG posee más de 10 millones de entidades, con 2.741.739 de estas referentes a autores.

2.6. Wikidata

Wikidata es un proyecto administrado por la Fundación Wikimedia y editada en colaboración (esto es, donde los mismos usuarios pueden ingresar o modificar los datos almacenados en la plataforma), que consiste en una base de datos organizada en grafos. La información guardada en su base de datos no tiene ningún tipo de orientación temática, por lo que se guardan datos a partir de cualquier concepto asimilable por el ser humano. Cada entidad posee una única ID representada por un identificador con un número único y una 'Q' como prefijo, denominada QID[8], que está pensada para no favorecer a ningún idioma a la hora de su configuración. Las entidades se relacionan mediante propiedades, que describen algún tipo de conexión entre dichas entidades, y dichas propiedades también poseen identificadores únicos, representados por un número con el prefijo 'P'. Por ejemplo, Q298 es una entidad de Wikidata que hace referencia a Chile, y P30 es una propiedad que hace referencia al continente del que es parte la entidad que la posee. Al igual que BibKG, cada entidad que posee la base de datos posee una URI propia, que dirige a una vista de la página con sus propiedades y valores asociados, como se observa en la Figura 2.4.

[7] `https://bibkg.imfd.cl/docs/`
[8] `https://es.wikipedia.org/wiki/Wikidata#Conceptos`

2.6.1. Servicio de consultas de Wikidata y SPARQL

Además de poder ser editada por cualquier usuario, Wikidata puede ser accesible a cualquier usuario que quiera utilizar sus datos. Entre los medios existentes de acceso a su información, está el servicio de consultas de Wikidata[9], donde se puede obtener información del conocimiento de esta mediante consultas en el lenguaje de consultas de SPARQL [3]. Además, en la página principal de Wikidata se pueden obtener entidades mediante palabras incluidas en el nombre del idioma en el que se esté en la página. Sin embargo, el servicio de consultas permite obtener entidades mediante condiciones definidas en la consulta, por lo que de esta forma se le saca un mayor provecho a Wikidata. Sin embargo, la página web del servicio de consultas de Wikidata posee un límite en cuanto al tiempo de ejecución máximo de una consulta, por lo que no es confiable para realizar consultas de una cantidad masiva de datos.

En cuanto a SPARQL[10], este es un lenguaje de consultas que permite acceder a la información de los grafos RDF de forma selectiva, mediante el uso de tripletas que asocian entidades o conceptos con alguna propiedad u otro concepto. También posee más cualidades, como filtrar los resultados según ciertas condiciones, o construir o eliminar información.

En el caso particular del servicio de consultas de Wikidata, al escribir las tripletas que relacionan entidades, se les pone a las entidades el prefijo '`wd:`', mientras que para las propiedades va el prefijo '`wdt:`'. Se pueden determinar variables (pueden ser tanto entidades como propiedades), estableciéndose mediante un '?' como prefijo.

En el caso de las tripletas, estas definen una relación mediante la cual se acota el conjunto de entidades que se quiere manejar, funcionando en este caso como la forma de relacionar entidades y propiedades, como se observa en la Figura 2.5.

En la consulta de SPARQL realizada al servicio de consultas de Wikidata y presentada en la Figura 2.5 se puede observar que se está buscando a las entidades (con la variable asignada `?continente`) que cumplen con la condición de ser el valor en la propiedad P30 (correspondiente a 'continente') de la entidad Q298 (correspondiente a Chile). Por ende, la consulta es equivalente a buscar todas las entidades que sean los continentes en los que está situado Chile, lo que arroja como único resultado a la entidad Q18, correspondiente a América del Sur. Se observa también que se utilizó la propiedad *rdfs:label* y un filtro de idioma del valor resultante de dicha propiedad para arrojar como resultado sólo los nombres en español de las entidades encontradas en la consulta.

2.6.2. *Dumps* de Wikidata

Como alternativa, Wikidata permite la descarga del archivo *dump* con toda la información de las entidades de la plataforma[11]. Se puede descargar tanto en formato JSON como en un

[9] `https://query.wikidata.org/`

[10] `https://www.w3.org/TR/sparql11-overview/`

[11] `https://dumps.wikimedia.org/wikidatawiki/entities/`

formato RDF[12]. Ambos *dumps* están en un formato comprimido, estando tanto en formato *bz2* como en formato *gz*. En resumen, el formato *bz2* está configurado para poseer una menor cantidad de espacio en disco, mientras que el formato *gz* se descomprime más rápidamente.

2.7. Limitaciones actuales

Como se comentó anteriormente, la cantidad de enlaces ya existentes entre entidades de BibKG y Wikidata es limitada. Además, actualmente BibKG tiene enlaces a una cantidad restringida de entidades, ya que las fuentes utilizadas para la obtención de datos (DBLP y ArnetMiner) están orientadas principalmente a publicaciones académicas. Wikidata, a diferencia de los anteriores, es un grafo mucho más general, con acceso a una cantidad mucho mayor de enlaces. De generalizar más el acceso de BibKG a Wikidata, se enriquecerán de datos las entidades de BibKG, y se podrán realizar consultas al servicio de consultas con un mayor repertorio de definición de propiedades de los resultados deseados (ya que el lenguaje de consultas de MillenniumDB permite acceder a los valores de las relaciones de las entidades). Adicionalmente, Wikidata es de acceso libre, por lo que no existen problemas en cuanto a la adquisición de licencias para tener acceso a sus datos.

La realización de dicha conexión no es trivial. Se requieren de mecanismos que permitan encontrar de alguna forma coincidencias entre una cantidad masiva de datos entre las entidades de Wikidata y los de BibKG, los que no siempre presentan una simple llave en común (en especial en el caso de los autores). En estos casos, se deben considerar otros parámetros, como sus datos asociados (títulos, nombres de autores, año de publicación, etc.), y enlazar entidades en base a las similitudes entre dichos datos, minimizando el margen de error entre dichos enlazamientos.

[12] `https://www.wikidata.org/wiki/Wikidata:Database_download#RDF_dumps`

Figura 2.4: Perfil de Wikidata de la entidad Q51366847 correspondiente a Aidan Hogan, junto con algunas de sus propiedades. (Fuente: `https://www.wikidata.org/wiki/Q51366847`)

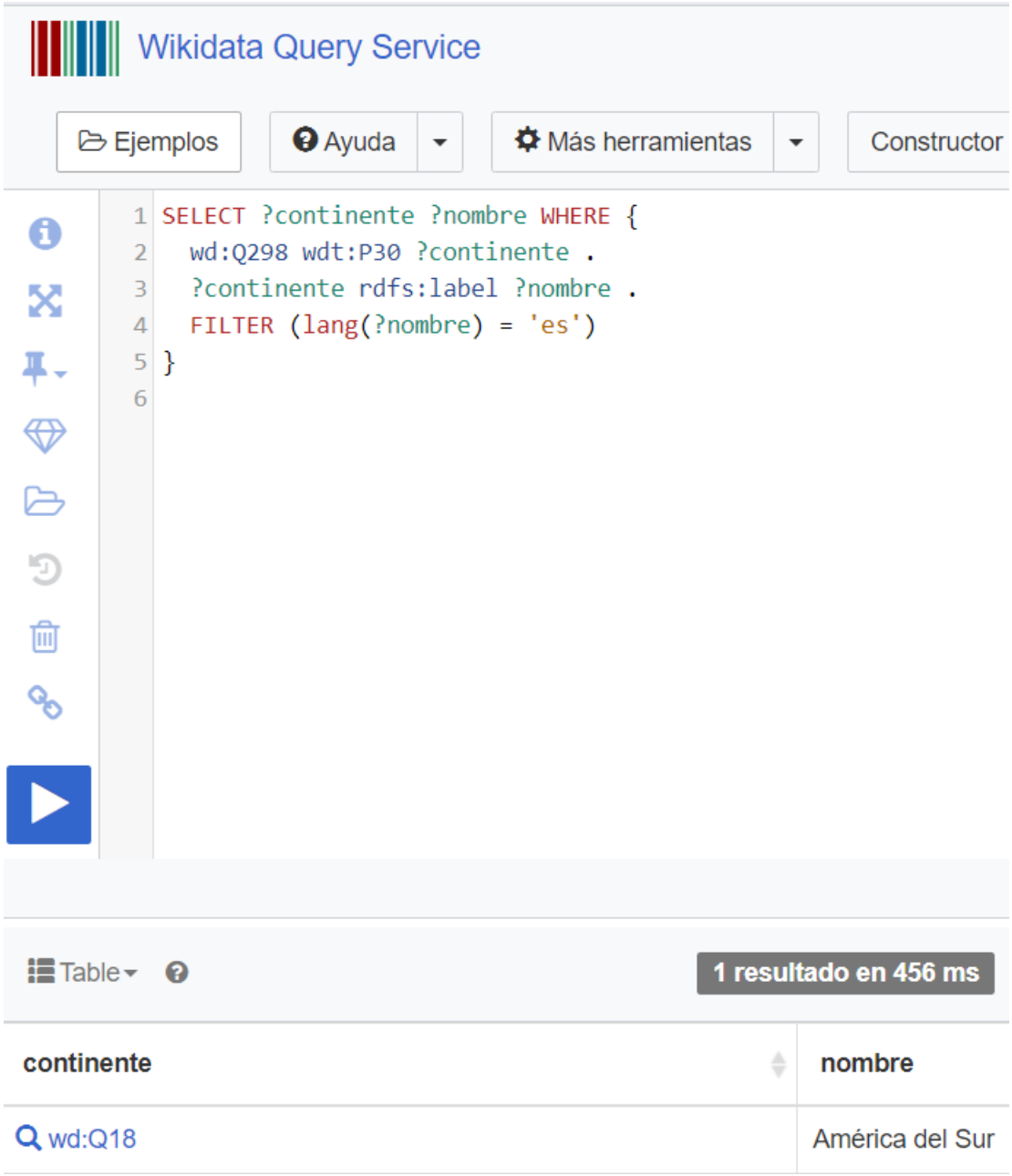

Figura 2.5: Consulta en el servicio de consultas de Wikidata junto con sus resultados, para conocer el o los continentes de los que Chile es parte, junto con aplicar un filtro para arrojar sus nombres en español.

Capítulo 3

Preprocesamiento de datos

En el presente capítulo se describe el trabajo realizado relacionado al preprocesamiento de los datos de BibKG y Wikidata, considerando tanto la planificación como la aplicación de este.

Para contextualizar los tiempos de lectura de los archivos resultantes del proceso realizado en este capítulo, es necesario mencionar las especificaciones de la computadora en la que se ejecutó todo el proyecto, que son las siguientes:

- **Sistema Operativo:** Windows 10
- **CPU:** AMD Ryzen 7 4800H 2.9 GHz
- **Almacenamiento:** disco interno SSD 512 GB, disco externo SSD 1 TB
- **Memoria RAM:** 8 GB (7.32 GB usables)

3.1. Formato del preprocesamiento de datos

Para el presente trabajo, se decidió utilizar archivos en formato JSON tanto para los datos de BibKG como para los de Wikidata, con los que se harán las comparaciones de datos y estimar posibles enlazamientos.

En cuanto a la elección del uso de *dumps* en formato JSON por sobre una base de datos convencional, se optó por este enfoque debido a su mayor simplicidad a la hora de trabajar con los datos existentes para el presente trabajo, ya que el *dump* de Wikidata permite directamente acceder a sus datos en formato JSON, tal como se explicó en el capítulo anterior. Además, el uso de bases de datos diseñadas para almacenar datos orientados a grafos (junto con la adaptación de los datos de BibKG y Wikidata a dicho formato) son algo que se escapan del alcance del presente trabajo. De todas formas, se prioriza la precisión y completitud del proceso por sobre el tiempo de ejecución de este, que sólo se ejecutará una vez (los resultados quedarán disponibles para acceder a ellos sin necesidad de ejecutar todo el proceso para

obtener la información de las entidades enlazadas, salvo que se quiera editar el algoritmo de obtención de enlaces), por lo que dicha decisión no afecta fundamentalmente a los objetivos definidos para el presente trabajo.

3.2. Preprocesando BibKG

Como fue mencionado anteriormente, los datos de BibKG están almacenados en MillenniumDB, un sistema de gestión de base de datos orientado a grafos. Dicho proyecto también posee un servicio de consultas. Existe un repositorio[1] desde donde se puede descargar el proyecto. De esta forma, es posible cargar el *dump* de BibKG en MillenniumDB, y de esta forma poder hacer consultas según el formato establecido.

Sin embargo, existen ciertos inconvenientes a la hora de utilizar este sistema para la conexión de datos. Para empezar, el programa de MillenniumDB está diseñado para realizar consultas sobre entidades específicas, no para el procesamiento masivo de sus datos. Junto con esto, MillenniumDB aún está en fase temprana de desarrollo, por lo que no están disponibles aún todas las funcionalidades que los desarrolladores consideran importantes para su completo funcionamiento.

No obstante, el inconveniente más importante es el formato mismo del *dump* de BibKG. Si bien el formato está estandarizado, no existe un intérprete como tal de los datos que permita operar las entidades y sus propiedades con facilidad, y de esta forma realizar comparaciones de datos.

3.2.1. El formato del *dump* de BibKG

En cuanto al formato de este archivo, si bien posee una sintaxis inspirada en la de Neo4J, no se ha encontrado un programa que permita poder leer y procesar los datos con facilidad, más allá de los servicios de consultas de MillenniumDB y BibKG. Estos tampoco son de utilidad para el procesamiento de los datos y la posterior comparación de datos con Wikidata, puesto que el primero requiere de la carga completa del archivo en memoria, estando además diseñado para ser utilizado para obtener entidades en particular (y no grandes cantidades de datos para procesar con algún lenguaje de programación), y el segundo sólo está disponible en la página web del proyecto. Cabe añadir que, como se señalará más adelante, no toda la información de una entidad en particular de BibKG está necesariamente en una sola línea del archivo, por lo que se requiere de leer todo el archivo en memoria para recién poder obtener con seguridad toda la información referente a dicha entidad. Por ende, es necesario un *parser* que permita comprender la sintaxis del formato particular de BibKG, junto con crear un archivo que posea una línea por cada entidad con toda la información referente a esta incluida en dicha línea y en un formato fácilmente legible para el lenguaje de programación, con el fin de economizar recursos tanto de tiempos de lectura como de espacio de memoria RAM.

[1] `https://github.com/MillenniumDB/MillenniumDB#data-model`

El *dump* de BibKG consiste en un archivo de más de 11 GB de almacenamiento sin compresión, conteniendo más de 11 millones de entidades junto con sus propiedades asociadas. Cada línea del archivo representa información respecto a una entidad en particular. Dicha línea posee dos tipos de formatos diferentes:

1. **Definición de la entidad:** Esta línea representa la definición de una entidad, donde se define su existencia, junto con el tipo de entidad que representa y sus propiedades que no hacen referencia a otras entidades. Todos estos valores se separan por espacios. Al principio de una línea se define el ID de la entidad. Después se define el tipo de la entidad, antecedido de dos puntos (':'), y finalmente van las diversas propiedades relacionadas con dicha entidad, con la particularidad mencionada anteriormente. Todo esto se puede observar a más detalle en el ejemplo de la Figura 3.1.

```
t_meltdown_m18 :Article mdate:"2020-06-25" publtype:"informal" name:"
    Meltdown" ee:"https://meltdownattack.com/meltdown.pdf" year:2018
    abstract:"The security of computer systems fundamentally relies on
    memory isolation, e.g., kernel address ranges are marked as non-
    accessible and are protected from user access." n_citation:386
```

Figura 3.1: Línea del *dump* de BibKG que hace referencia a la definición de la entidad `t_meltdown_m18` de BibKG (con la información de la propiedad 'abstract' abreviada), donde se define su ID junto con el tipo de entidad y sus propiedades que no hacen referencia a otras entidades como valor.

2. **Relaciones de entidades:** Estas líneas definen las relaciones existentes entre entidades a través de una propiedad. Esta puede ser relacionando una entidad con otra (a través de sus IDs), o entre una entidad y algún otro tipo de valor (*strings*, números). Esta línea posee la siguiente estructura:
   (a) **Entidades relacionadas:** Se establecen las entidades relacionadas a través de la escritura de los valores relacionados, separadas de un guión y un signo mayor ('->'). El orden de los valores relacionados en esta sección es importante, puesto que la propiedad está asociada a la primera entidad escrita en dicha relación.
   (b) **Propiedad:** Se establece la propiedad de la relación. Esta va antecedida por dos puntos (':').
   (c) **Característica de la propiedad (opcional):** A veces, una relación puede poseer algún valor asociado, además de la entidad relacionada. Por ejemplo, en casos en los que una entidad posea más de un valor asociado a una propiedad (por ejemplo, varios autores relacionados con una publicación), se asocia un número a la relación, en casos en los que el orden de las entidades sea importante. En particular, el orden de los autores de una publicación viene inherente a esta, por lo que cada autor debería poseer un orden fijo en todas las fuentes que posean la publicación respectiva, en caso de que la fuente de datos posea el orden como información.

   Cabe destacar que el orden en el que están situadas dichas líneas no tiene relevancia, y no necesariamente todas las relaciones correspondientes a una entidad en específico están puestas de forma consecutiva. Se puede observar un ejemplo de este tipo de líneas del archivo *dump* en el ejemplo de la Figura 3.2.

```
t_meltdown_m18->a_Moritz_Lipp :has_author orden:1

a_Moritz_Lipp->t_meltdown_m18 :author_of

t_meltdown_m18->a_Michael_Schwarz_0001 :has_author orden:2

a_Michael_Schwarz_0001->t_meltdown_m18 :author_of

t_meltdown_m18->a_Daniel_Gruss :has_author orden:3

a_Daniel_Gruss->t_meltdown_m18 :author_of

t_meltdown_m18->a_Thomas_Prescher_0002 :has_author orden:4

a_Thomas_Prescher_0002->t_meltdown_m18 :author_of

t_meltdown_m18->a_Werner_Haas_0004 :has_author orden:5

a_Werner_Haas_0004->t_meltdown_m18 :author_of

t_meltdown_m18->a_Stefan_Mangard :has_author orden:6
```

Figura 3.2: Líneas del *dump* de BibKG que hacen referencia a diversas relaciones entre entidades de BibKG. La propiedad asociada a la relación se asigna a la primera entidad a la que se hace referencia. Por ejemplo, en la primera línea, a la entidad `t_meltdown_m18` se le asigna la entidad `a_Moritz_Lipp` en la propiedad *has_author*, con orden 1. Cabe destacar que, tal como todas las propiedades a las que se les asigna valores en estas líneas, la entidad puede poseer en la propiedad asignada *has_author* más de una entidad relacionada, que se pueden ir añadiendo en otras líneas a lo largo del archivo.

3.2.2. Preprocesamiento de BibKG

La idea del *parser* del *dump* de BibKG es la de poder leer dicho archivo de forma más sencilla y eficiente en el momento de realizar las comparaciones con Wikidata.

Se determinó que los datos de BibKG se manejarán en formato JSON. Esto, ya que el *dump* escogido de Wikidata también posee JSON como formato, y debido a la facilidad con la que se puede utilizar con el lenguaje de programación Python, el escogido para realizar el proceso. Cabe destacar que los diccionarios de Python poseen una estructura muy similar a la de los objetos JSON.

El proceso realizado siguió los siguientes pasos:

1. Se lee el archivo línea por línea, y se detecta el tipo de línea de esta. El archivo es leído por partes, de tal forma que se pueda almacenar la información en la memoria RAM sin problemas. En el caso particular de este ejercicio, el archivo fue leído en 4 partes.
2. Se procesa la información de la línea según cada caso. La información se almacena en un diccionario de Python, que almacena todos los objetos relacionados a cada entidad encontrada en forma de diccionarios de Python, en los que se guardan las propieda-

des relacionadas junto con sus valores asociados. Si la entidad a la que hace referencia una línea no existe dentro del diccionario de Python, se crea dicho objeto. En el caso particular del tipo de la entidad, se almacenó en forma de la propiedad 'type', no existente previamente en el *dump*. En el caso particular en el que exista la propiedad *author_of* (correspondiente a las publicaciones de un autor), esta propiedad no se escribe. Esto, para ahorrar espacio en el archivo, puesto que esta información está ya de forma implícita en la propiedad *has_author*, que hace referencia a los autores de una publicación. Cabe añadir, para el caso de las relaciones entre entidades, se cambió el formato del contenido de estas, ya que en estos casos una propiedad puede contener más de una entidad. En estos casos, en vez de simplemente crear una lista de *strings* con los IDs relacionados, se crearon objetos JSON en su contenido, que poseen como propiedad `value` el valor de la propiedad, y pueden incluir otro valor asociado a la relación (como el orden, por ejemplo), de existir.

3. Se escribe la información de los diccionarios de Python en un archivo JSON, utilizando la librería `json` de Python para poder almacenar los diccionarios como objetos JSON.

4. Se unen todas las partes escritas previamente. Cabe destacar que como es posible que exista una referencia a una entidad en varias partes a la vez (debido a que no necesariamente las líneas referentes a una entidad están de forma consecutiva, y no siguen un orden en específico), se procesan los archivos para fusionar a los objetos repetidos de todas las partes.

Como resultado, se obtiene un archivo con un total de 11.089.213 entidades expresadas mediante objetos JSON, uno por cada línea, como la entidad ilustrada en la Figura 3.3. Cada propiedad del objeto JSON representa a la propiedad que posee el mismo nombre en BibKG, con las excepciones de las propiedades *id* y *type*, que representan al identificador de la entidad dentro de BibKG y al tipo de entidad respectivamente. En la computadora en la que fue ejecutado el proyecto, el archivo posee un tiempo de lectura de alrededor de 140 segundos.

```
{"id":"a_Aidan_Hogan",
"type": "Person",
"scholar": "CP-fgY4AAAAJ",
"orcid": "0000-0001-9482-1982",
"wikidata": "Q51366847",
"mdate": "2019-01-10",
"key": "homepages/h/AidanHogan",
 "name": "Aidan Hogan",
 "affiliation": [{"value": "Universidad de Chile"}],
 "url": [{"value": "http://aidanhogan.com/"},
 {"value": "http://sw.deri.org/~aidanh/"}]}
```

Figura 3.3: Ejemplo de la entidad `a_Aidan_Hogan` correspondiente al autor Aidan Hogan en el nuevo archivo JSON de BibKG.

3.3. Preprocesando Wikidata

Como se mencionó anteriormente, Wikidata permite el acceso a la información de sus entidades mediante *dumps* que son accesibles para todos los usuarios que lo necesiten. A diferencia de las opciones de utilizar el servicio de consultas o la página de Wikidata, esta permite un uso más personalizado de los datos de las entidades, al tener acceso de sus archivos para utilizar *scripts* personalizados para la manipulación de sus datos.

Se priorizó utilizar esta opción por sobre la API de Wikidata debido a los posibles límites de consultas al servidor, en relación a los tiempos de ejecución por cada consulta y al tamaño de la información de la respuesta de las mismas. Esto se puede observar por ejemplo con el servicio de consultas de Wikidata, que arroja un error de tiempo en caso de que la consulta enviada se ejecute por más de un minuto sin una respuesta definitiva. Si bien esto no es un problema para realizar consultas sobre ciertas entidades y/o atributos en específico, en este caso puede causar un incremento de tiempo del proceso de tal forma que sencillamente resulte inviable ejecutar dicho proceso, al ser necesario realizar una cantidad masiva de consultas a las entidades de Wikidata, debido a su gran tamaño. Por ende, guardar la información de Wikidata de forma local ayuda a evitar posibles inconvenientes resultantes de los límites de consultas y tiempos de consulta inherentes a la API de Wikidata.

3.3.1. El *dump* de Wikidata

Los *dumps* de Wikidata están disponibles tanto en formatos que implementan RDF (*nt* y *ttl*) como en formato JSON, y este último es el seleccionado para enlazar los datos con BibKG.

Como Wikidata está diseñado para contener todas las entidades que puedan ser definidas por el ser humano (y que son añadidas por la comunidad) el archivo posee un tamaño considerable. En particular, el *dump* descomprimido de Wikidata en formato JSON posee un tamaño de más de 1.5 TB de información, con más de 100 millones de entidades almacenadas. Los archivos disponibles poseen un tamaño más reducido, poseyendo el formato *gz* cerca de 120 GB de información, mientras que el formato *bz2* posee cerca de 80 GB de información. Sin embargo, como el formato está comprimido, requiere tiempo para interpretar los datos contenidos, por lo que los tiempos de lectura no son menores a los de leer el archivo mismo descomprimido.

Por lo anterior, es necesario reducir el tamaño del archivo, de tal manera de mejorar los tiempos de lectura del archivo. Para ello, es necesario quitar información redundante de los objetos JSON, y filtrar entidades que se puede asumir con alta confianza de que no pertenecen al conjunto de entidades que posee BibKG.

3.3.2. El formato de los objetos JSON en Wikidata

Los objetos JSON de Wikidata poseen un estándar en cuanto a su estructura[2]. Esta se puede resumir en los siguientes estándares:

1. **Estructuras de alto nivel:** Estas son las propiedades de más alto nivel de una entidad de Wikidata, de tal forma que son las propiedades que contienen al resto de propiedades de más bajo nivel. Entre estas se encuentran las siguientes:
   (a) **id:** El identificador de Wikidata de la entidad.
   (b) **type:** El tipo de dato que representa la entidad. Este puede ser *item* en caso de los elementos de datos, o *property* en caso de las propiedades.
   (c) **labels:** Estos datos representan a los nombres de la entidad en distintos idiomas.
   (d) **descriptions:** Contiene las descripciones de la entidad en distintos idiomas.
   (e) **aliases:** Contiene los alias de la entidad en distintos idiomas.
   (f) **claims:** Contiene cualquier número de sentencias de la entidad, agrupadas por propiedad.
   (g) **sitelinks:** Contiene enlaces a otros sitios web que describen el elemento.
2. **Estructura de *labels, descriptions* y *aliases*:** Estas propiedades comparten la misma estructura. Dentro de estas existe una propiedad por cada idioma a la que se hace referencia, y dentro de dicha propiedad todos los valores asociados al idioma respectivo. Por ejemplo, en la Figura 3.4 se observa que cada subpropiedad dentro de *aliases* representa a una lista de todos los *aliases* de cada idioma ('en' representa a los aliases en idioma inglés, mientras que 'fr' a los aliases en idioma francés), estando la información de cada *alias* dentro de un objeto JSON, con su idioma y valor.

[2] `https://doc.wikimedia.org/Wikibase/master/php/docs_topics_json.html`

```
{
"aliases": {
    "en": [
      {
        "language": "en",
        "value": "New York"
      }
    ],
    "fr": [
      {
        "language": "fr",
        "value": "New York City"
      },
      {
        "language": "fr",
        "value": "NYC"
      },
      {
        "language": "fr",
        "value": "The City"
      },
      {
        "language": "fr",
        "value": "La grosse pomme"
      }
    ]
  }
}
```

Figura 3.4: Ejemplo de valores de la propiedad *aliases* de una entidad de Wikidata. (Fuente: `https://doc.wikimedia.org/Wikibase/master/php/docs_topics_json.html`)

3. **Estructura de *claims***: En esta propiedad es donde se hace referencia a todas las propiedades que posee una entidad en Wikidata (las propiedades que poseen un identificador numérico en Wikidata antecedido por una 'P'), que se encargan de relacionar a la entidad tanto con valores inmutables (*strings*, números, etc.) como con otras entidades de Wikidata. Dentro de estas, existen las siguientes propiedades, visualizables en el objeto de ejemplo de la Figura 3.5:
   (a) **id:** Es un identificador único referente a la relación establecida, y que es único a lo largo del repositorio. No se pueden realizar suposiciones fiables con respecto a la estructura de esta.
   (b) **type:** Por convención, siempre posee como valor *statement*.
   (c) **mainsnak:** Este es un *snak*, tipo de estructura de Wikidata que se encarga de dar algún tipo de información sobre la propiedad asociada a alguna entidad. Esta es representada a través de los siguientes parámetros:
      i. **snaktype:** El tipo de valor que representa. Este puede ser *value*, *somevalue* o *novalue*.

ii. **property:** El ID de la propiedad a la que corresponde el valor.

iii. **datatype:** Este parámetro indica cómo puede ser interpretado el valor de la propiedad. Existe una amplia lista de posibles *datatypes*[3], que poseen a la vez diversos tipos de estructuras.

iv. **datavalue:** En esta estructura se encuentra el valor de la propiedad, junto con el tipo de valor que representa.

(d) **rank:** Expresa si el valor será mostrado en las consultas de Wikidata, y si se muestra de forma predeterminada en el sistema.

(e) **qualifiers:** Muestra datos para contextualizar el valor principal de la propiedad (fechas, métodos de medición, etc.). Al igual que las estructuras de los *mainsnaks* de *claims* está estructurada en *snaks*, por lo que poseen una estructura similar.

(f) **references:** Guarda referencias respecto a la información de la relación de la propiedad.

[3] `https://www.wikidata.org/wiki/Special:ListDatatypes`

```
{
    "claims": {
      "P17": [
        {
          "mainsnak": {
            "snaktype": "value",
            "property": "P17",
            "datatype": "wikibase-item",
            "datavalue": {
              "value": {
                "entity-type": "item",
                "id": "Q30",
                "numeric-id": 30
              },
              "type": "wikibase-entityid"
            }
          }
        },
        {
          "mainsnak": {
            "snaktype": "somevalue",
            "property": "P17"
          }
        }
      ],
      "P356": [
        {
          "mainsnak": {
            "snaktype": "value",
            "property": "P356",
            "datatype": "string",
            "datavalue": {
              "value": "SomePicture.jpg",
              "type": "string"
            }
          }
        }
      ]
    }
  }
```

Figura 3.5: Ejemplo de valores de la propiedad *claims* de una entidad de Wikidata, donde se puede apreciar a detalle la estructura de *mainsnak*. (Fuente: `https://doc.wikimedia.org/Wikibase/master/php/docs_topics_json.html`)

4. **Estructura de *sitelinks*:** Cada registro de esta sección representa a cada ID global de un sitio al que se hace referencia. Estas poseen a la vez los siguientes parámetros, visualizables en el objeto de ejemplo de la Figura 3.6:

(a) **site:** El ID global del sitio.

(b) **title:** El título de la página.

(c) **badges:** Posible insignia asociada con la página, que se llama mediante una lista de IDs asociados con dichas insignias.

(d) **url:** De forma opcional, se puede incluir la URL de la página.

```
{
    "sitelinks": {
      "afwiki": {
        "site": "afwiki",
        "title": "New York Stad",
        "badges": []
      },
      "frwiki": {
        "site": "frwiki",
        "title": "New York City",
        "badges": []
      },
      "nlwiki": {
        "site": "nlwiki",
        "title": "New York City",
        "badges": [
          "Q17437796"
        ]
      },
      "enwiki": {
        "site": "enwiki",
        "title": "New York City",
        "badges": []
      },
      "dewiki": {
        "site": "dewiki",
        "title": "New York City",
        "badges": [
          "Q17437798"
        ]
      }
    }
  }
```

Figura 3.6: Ejemplo de valores de la propiedad *sitelinks* de una entidad de Wikidata. (Fuente: `https://doc.wikimedia.org/Wikibase/master/php/docs_topics_json.html`)

3.3.3. Preprocesamiento de Wikidata

Como se mencionó anteriormente, el proceso consistió en reducir el tamaño del *dump* de Wikidata, con el objetivo de mejorar los tiempos de lectura, lo que permite agilizar los tiempos de ejecución de las etapas posteriores del proyecto. Los resultados fueron escritos en 3 archivos distintos, dependiendo de ciertos parámetros:

1. Un archivo JSON que contiene a todas las personas de Wikidata.
2. Un archivo JSON que contiene a todas las entidades de Wikidata que son vistas como potenciales entidades relacionadas con publicaciones y otras entidades de BibKG.
3. Un archivo JSON que posee entidades de Wikidata en las que no se pudieron encontrar relaciones con publicaciones académicas o relacionados, pero se pudieron aplicar filtros en los que se descartaron entidades que no tienen que ver con el tipo de entidades de BibKG.

Para esto, se siguió la siguiente metodología:

1. Leer el archivo *gz* línea por línea (cabe destacar que el *dump* original posee una entidad por línea), y cargar la entidad utilizando la librería *json* de Python para poder interpretarlo.
2. Aplicar filtros, para decidir si la entidad debe ser o no escrita en alguno de los archivos JSON. Los filtros fueron los siguientes:
   (a) **Archivo JSON de personas:** En este caso simplemente se buscó añadir a todas las personas de Wikidata. Para ello, se analizó la propiedad P31 de Wikidata, que posee las instancias de cada propiedad, y se consideraron las entidades que poseyeran la instancia 'ser humano', correspondiente a la entidad Q5.
   (b) **Archivo JSON de publicaciones y relacionados:** Para este caso, se consideraron todas las entidades que poseyeran como valor en la propiedad P31 alguno de los valores de la misma propiedad que poseen todas las entidades que tienen algún ID de DBLP asociado (DBLP es una de las fuentes de datos de BibKG). Para esto, previo a aplicar la reducción del *dump* de Wikidata se analizaron todas las entidades con alguno de estos IDs, y se contaron todos los valores de la propiedad P31 de cada entidad. El filtro finalmente considera a todas las entidades que posean como valor en su propiedad P31 alguno de los valores encontrados en el experimento anterior, como las entidades presentes en la Tabla 3.1.

      Existen 4 entidades relacionadas con IDs de DBLP en Wikidata:

      i. **DBLP publication ID (P8978):** Identicador de las publicaciones en la base de datos de DBLP. Existen 431.413 entidades de Wikidata que poseen esta propiedad.
      ii. **DBLP event ID (P10692):** Identicador de los eventos en la base de datos de DBLP. Existen 2.435 entidades de Wikidata que poseen esta propiedad.

iii. **DBLP venue ID (P8926):** Identicador de las conferencias en la base de datos de DBLP. Existen 4.355 entidades de Wikidata que poseen esta propiedad.

iv. **DBLP author ID (P2456):** Identicador de los autores en la base de datos de DBLP. Existen 61.695 entidades de Wikidata que poseen esta propiedad. Esta propiedad no fue utilizada para el experimento, puesto que las personas de Wikidata ya fueron consideradas para el otro archivo JSON.

Adicionalmente, se descartaron entre las propiedades más populares de Wikidata aquellas que están relacionadas con entidades que no poseen relación con las de BibKG (o sea, que no se relacionen con publicaciones académicas, autores y datos asociados), como las propiedades expuestas en la Tabla 3.2, en las que se observan propiedades relacionadas con otros tipos de objetos (por ejemplo, las propiedades P1215 y P6257 representan distintos tipos de mediciones de objetos astronómicos). Para asegurarse de que no se descartan entidades que sí son útiles (o minimizar dicha cantidad), se corroboró que estas propiedades no estén dentro de ninguna de las entidades que posean alguna propiedad de DBLP.

ID	**Nombre**	**Conteo**
Q13442814	scholarly article	415.772
Q212971	Request for Comments	9.027
Q7318358	review article	5.326
Q1143604	proceedings	3.557
Q18918145	academic journal article	2.274
Q871232	editorial	1.757
Q1348305	erratum	794
Q23927052	conference paper	693
Q815382	meta-analysis	319
Q2782326	case report	183
Q1504425	systematic review	63
Q1322187	April Fools' Day Request for Comments	63
Q5633421	scientific journal	41
Q45182324	retracted paper	41

Tabla 3.1: Tabla con las 14 instancias más recurrentes entre las entidades que poseen la propiedad P8978, correspondiente a la ID de las publicaciones de DBLP, junto con el conteo de presencias en Wikidata de entidades con dicha propiedad.

ID	Nombre	Conteo
P1215	apparent magnitude	33.122.953
P1227	astronomical filter	33.122.898
P887	based on heuristic	8.566.543
P3083	SIMBAD ID	8.152.400
P6257	right ascension	8.094.538
P6258	declination	8.094.414
P59	constelation	7.374.786
P703	found in taxon	7.168.473
P846	GBIF taxon ID	3.243.769
P351	Entrez Gene ID	3.116.063
P2326	GNS Unique Feature ID	2.894.821
P352	UniProt protein ID	2.537.311

Tabla 3.2: Tabla con las propiedades de Wikidata que se utilizan como filtro para las entidades, junto con el conteo de presencias en Wikidata de entidades con dicha propiedad.

(c) **Archivo JSON de otras entidades:** Acá se aplicó un filtro en la que se descartaron entidades que poseen valores en la propiedad P31 que indican que no pertenecen al conjunto de entidades de publicaciones y relacionados. Para ello, se observaron las estadísticas de Wikidata, y dentro de las instancias más utilizadas se descartaron las que hacían referencia a tópicos que no tienen que ver con los de BibKG, visibles en la Tabla 3.3. Por ejemplo, la propiedad Q4167836 hace referencia a las entidades que representan categorías dentro del entorno de Wikimedia.

ID	Nombre	Conteo
Q1190554	occurrence	6.109
Q16521	taxon	3.577.676
Q56061	adm territorial	8.952
Q811979	arquitectural estructure	211.166
Q11173	chemical	1.262.127
Q11424	film	285.624
Q83620	thoroughfare	2.947
Q6999	astronomical	25.762
Q4167836	wikimedia category	5.150.971
Q13406463	wikimedia article	352.836
Q11266439	wikimedia template	798.864

Tabla 3.3: Tabla de entidades de Wikidata que se utilizan como valor en la propiedad P31 (*instance of*), junto con los conteos de la cantidad de entidades que poseen dicha entidad como valor en dicha propiedad. Estas propiedades se utilizan como filtro para el archivo JSON de otras entidades, de tal forma de que las entidades que las posean en dicha propiedad no fueron incluidas en el nuevo archivo JSON.

3. Si la entidad no fue filtrada en el paso anterior, eliminar datos de esta que no tengan relevancia para el enlazamiento de datos. Los parámetros seguidos para este paso fueron los siguientes:

   (a) **Archivo JSON de personas:** La información de BibKG en cuanto a las personas es limitada. Por ende, se eliminaron todas las propiedades aparte del nombre, descripción e identificadores. Cabe destacar que todas las propiedades que hacen referencia a identificadores poseen en el parámetro *datatype* el valor *external-id*.

   (b) **Archivos JSON de publicaciones y otros:** Para los otros dos archivos, se eliminaron ciertas redundancias de datos y parámetros que no son útiles para el enlazamiento de datos. Se consideró lo siguiente:

      i. Filtrar los *labels, descriptions* y *aliases* de las entidades. Se dejarán sólo los datos en inglés de existir. De lo contrario, no se aplicarán filtros, y se añadirán los valores respectivos a todos los idiomas contenidos.

      ii. Eliminar los *sitelinks* de la entidad.

      iii. Eliminar las referencias de las propiedades de una entidad.

      iv. Eliminar los *qualifiers* de las propiedades de una entidad, siempre y cuando no hagan referencia a la propiedad P50 de Wikidata, correspondiente a los autores de una entidad, o a la propiedad P2093, correspondiente a los *strings* de autores de la entidad (o sea, los autores de una entidad sin hacer referencia a la entidad del mismo, generalmente en los casos en los que no existe o no se encontró la entidad correspondiente a dicho autor, por lo que simplemente se añade el *string* del nombre de este).

      v. Quitar ciertas redundancias de una propiedad. En esta se aplicaron varios procedimientos, como por ejemplo agregar sólo el valor de la propiedad en ciertas estructuras e ignorar el resto de parámetros de esta.

4. Escribir la entidad en el archivo correspondiente, escribiéndolo en formato JSON. A diferencia del *parser* de datos en BibKG, en este caso no es necesario cargar archivos en la memoria RAM, puesto que cada línea del *dump* de Wikidata posee toda la información respectiva a las propiedades de su entidad. De esta forma, no es necesario cargar el archivo por partes.

Se logró reducir considerablemente el tamaño del archivo *dump* de Wikidata. El archivo JSON de personas posee 10.609.909 entidades, con un tamaño de 6,32 GB descomprimidos. En tanto, el archivo de publicaciones y relacionados posee 41.232.415 entidades con un tamaño de 118 GB, y el archivo de otras entidades posee 23.509.679 entidades con 32,1 GB de memoria. En comparación, el archivo JSON de Wikidata descomprimido posee aproximadamente 101 millones de entidades, con un tamaño de aproximadamente 1,5 TB de información. Gracias a esto, se logran reducir considerablemente los tiempos de lectura del archivo, considerando que el tiempo de lectura de todo el archivo en la computadora en la que se realizó el trabajo es de alrededor de 30 minutos. Esto permite disminuir notablemente el tiempo total de ejecución del proceso de enlazamiento, lo que permite una obtención más rápida de resultados, lo que además permite trabajar con dichos datos de forma más fluida, puesto que existe un menor tiempo de espera para tomar decisiones sobre los modelos de enlazamiento y otros aspectos

que requieren de análisis de los datos. En efectos prácticos, se utilizarán los archivos obtenidos de Wikidata de publicaciones y de personas para realizar la comparación de datos de BibKG.

Capítulo 4

Enlazando BibKG con Wikidata

En el presente capítulo se describe el trabajo realizado relacionado al enlazamiento de las entidades de BibKG con su equivalente en Wikidata, incluyendo los criterios definidos para su ejecución y los archivos resultantes que incluyen la información recopilada durante este proceso.

Cabe recalcar que es importante priorizar la precisión de los enlazamientos por sobre obtener la mayor cantidad posible de enlazamientos. Por ende, los métodos utilizados para realizar las comparaciones entre entidades y determinar potenciales enlazamientos poseen un tinte más conservador, mediante el uso de ciertas técnicas que descartan enlazamientos hechos en etapas previas, en caso de determinar que existen condiciones que aumenten la posibilidad de dicha entidad de estar errónea.

4.1. Propiedades de BibKG y Wikidata

Como se explicó anteriormente, las entidades tanto de BibKG como de Wikidata poseen propiedades que relacionan a sus entidades con otras entidades o con valores fijos. Estos parámetros son los que permiten establecer equivalencias entre entidades de distintas bases de datos, al poder encontrar propiedades equivalentes de estas con valores en común que, por sí solas o en conjunto, puedan determinar con alta estima que son entidades equivalentes. Por ende, las propiedades serán la herramienta utilizada por este trabajo, puesto que son la fuente de datos más completa con la que se cuenta para el enlazamiento de entidades.

Existe más de una forma de utilizar las propiedades para realizar enlazamientos. A continuación, se detallarán los diferentes métodos utilizados para los enlazamientos:

4.2. Enlazamiento por IDs

Puesto que por definición el valor de una propiedad correspondiente a un ID es único para cada entidad, la forma más simple y precisa de poder enlazar entidades (si bien no es la única herramienta disponible) es enlazando entidades por IDs. Para esto, es necesario encontrar propiedades equivalentes entre BibKG y Wikidata, y extraer la mayor cantidad de identificadores posibles por parte de BibKG.

4.2.1. Extracción de datos de BibKG

En el caso de BibKG, esta posee propiedades que definen IDs, tanto de forma directa como indirecta, ya que cada entidad puede poseer algunas propiedades que definen directamente IDs, como también propiedades que definen estas de forma implícita. Cabe destacar que, como se utilizó DBLP para la obtención de datos de BibKG, cada entidad de BibKG obtenida a partir de esta entidad incluye a este identificador de alguna forma.

Wikidata, por su parte, posee exactamente cuatro tipos de propiedades referentes a IDs de DBLP, observables en la Tabla 4.1, notando por sus conteos de entidades con dichas propiedades presentes que estas hacen referencia principalmente a publicaciones y autores presentes como entidades en DBLP. Por otra parte, existen otras propiedades que definen explícitamente IDs externos, y que además poseen una propiedad equivalente en Wikidata, como se observa en la Tabla 4.2. En esta última, además, se observa que sólo los autores poseen directamente expresado su ID de DBLP como una propiedad de BibKG (la propiedad *key*).

Nombre Propiedad	**ID**	**Conteo**
DBLP publication ID	P8978	431.413
DBLP author ID	P2456	245.068
DBLP venue ID	P8926	4.355
DBLP event ID	P10692	2.435

Tabla 4.1: Tabla con las propiedades de Wikidata que poseen algún ID de DBLP asociado, junto con el conteo de entidades con alguna de estas propiedades asociada.

Nombre en BibKG	Conteo BibKG	ID Wikidata	Conteo Wikidata
key	2.734.446	P2456	62.393
orcid	86.633	P496	2.311.979
isbn	65.299	P957/P212	164.326
wikidata	47.170	-	-
scholar	18.026	P1960	77.433

Tabla 4.2: Tabla con el total propiedades de BibKG que indican directamente el ID de la entidad a la que pertenece en la fuente de datos relacionada con la propiedad, junto con sus conteos, las propiedades equivalentes en Wikidata y los conteos de estas en dicha plataforma.

Sin embargo, como se mencionó anteriormente, la Tabla 4.2 no representa al total de IDs potencialmente presentes en alguna entidad de BibKG, puesto que existen algunas presentes implícitamente:

1. Propiedad *ee* de BibKG: Esta propiedad se relaciona con el enlace electrónico de la entidad. El enlace puede provenir de diversas fuentes de información, por lo que no se puede asumir nada sobre el formato de este, dependiente del lugar de origen del valor de la propiedad. Algunas de estas fuentes de datos poseen un ID dentro del enlace. A su vez, existen propiedades a Wikidata asociadas a algunas de estas fuentes de datos. Se visualizan en la Tabla 4.3 los prefijos más comunes dentro de la propiedad *ee*. Se observa en esta que una gran mayoría de dichos enlaces provienen de la fuente de datos DOI, con aproximadamente un 82 % del total de enlaces electrónicos, y que gran parte de las fuentes de datos presentes en *ee* poseen un equivalente en Wikidata.
2. Propiedad *url* de BibKG: Esta propiedad se relaciona con alguna URL relacionada con la entidad. Esta propiedad puede contener URLs de DBLP, reconocibles por el prefijo de esta. En la Tabla 4.4 se observan los conteos de cada tipo de entidad con la propiedad *url* que poseen el ID de DBLP, detectando dichos casos contando la cantidad de valores con el prefijo 'db/'. Cabe destacar que las entidades del tipo 'Journal' (un total de 2.549.575) no poseen un identificador de DBLP asociado en ninguna de sus propiedades.

Prefijo	Conteo	Propiedad Wikidata
total	5.391.047	-
https://doi.org	4.440.740	P356
http://arxiv.org	245.068	P818
https://arxiv.org	95.825	P818
http://ceur-ws.org	41.177	-
http://ieeexplore.ieee.org	31.531	P6480
https://www.aclweb.org	30.845	-
http://aisel.aisnet.org	25.476	-
http://www.isca-speech.org	22.399	-
http://hdl.handle.net	19.774	P1184
http://d-nb.info	19.342	P1292
http://dl.acm.org	18.891	P2179/P3332/P3333
http://ethos.bl.uk	9.307	P4536

Tabla 4.3: Tabla con algunas de las fuentes de datos más recurrentes presentes en la propiedad *ee* de BibKG, analizados mediante las recurrencias de los prefijos de los valores referentes a la propiedad.

Tipo de entidad	Conteo
total	5.470.249
Inproceedings	2.805.365
Article	2.549.025
Incollection	66.493
Proceedings	47.046
Book	2.320

Tabla 4.4: Tabla con los tipos de entidad de BibKG que poseen un valor en la propiedad *url* que poseen como prefijo 'db/' en su valor, lo que implica contener implícitamente un ID de DBLP.

4.2.2. Enlazamiento

Para el enlazamiento, se siguió la siguiente secuencia de acciones:

1. Leer el archivo de BibKG, y cargar todas las IDs encontradas de cada entidad en memoria junto con su entidad asociada. Estas se guardan en diccionarios, uno por cada tipo de ID existente en BibKG. Procesar las IDs según sea necesario, para poder compararlos con los de Wikidata.
2. Leer el archivo de publicaciones de Wikidata, buscando si posee alguna propiedad que pueda poseer un ID comparable con alguno de Wikidata. De ser así, buscar la ID en memoria, y de encontrarse, relacionar la entidad de Wikidata con la de BibKG que

está asociada a la ID encontrada, guardándose dicha relación en memoria en caso de que el enlace no exista en ese punto. Se toman en cuenta otras consideraciones, como el análisis de casos en que una entidad de Wikidata es relacionada mediante IDs con dos o más entidades de BibKG (en estos casos, se consideran ciertos casos potenciales en los que esto ocurre, y en última instancia eliminar estos enlaces).

3. Leer el archivo de personas de Wikidata, repitiendo el proceso anterior en dicho archivo.
4. Almacenar los enlaces realizados a una lista que pueda ser procesada para ser posteriormente añadido a un archivo CSV, que posea la entidad enlazada junto con su símil de Wikidata, y el tipo de enlace realizado. Guardar los metadatos del proceso (conteos de enlazamiento, tiempo de ejecución del proceso) en otro archivo CSV.

4.3. Enlazamiento por relaciones

A partir de las entidades ya enlazadas, se pueden inferir enlaces según las propiedades dentro de estas, al acotar enormemente la cantidad de posibles candidatos a enlaces, pudiendo comparar las propiedades de las entidades enlazadas. Para este proceso se utilizarán tres submétodos que conformarán a este tipo de enlazamiento, cada uno relacionado a distintos tipos de relaciones que pueden poseer las propiedades de las entidades de publicaciones académicas y datos relacionados, tanto de forma directa como indirecta.

4.3.1. Enlazamiento de autores

Cada publicación científica que posea autores asociados posee además un orden asociado a estos autores, con cada autor poseyendo un número específico de orden dentro de este. Estos valores vienen asociados a la publicación misma, por lo que debe ser igual en toda fuente de datos que considere al orden dentro de su estructura de datos. Por ende, a través de dicho orden se pueden inferir autores a partir de publicaciones ya enlazadas con Wikidata, procurando que los nombres de las entidades relacionadas sean similares para minimizar el error del proceso.

4.3.2. Enlazamiento de publicaciones

Si bien los autores no poseen órdenes en sus publicaciones asociadas, se pueden establecer conexiones mediante los nombres de las publicaciones, verificando que estas posean un nombre similar, y no exista más de una publicación de dicho autor con el mismo nombre. Cabe destacar que en Wikidata la información de las publicaciones asociadas a cada autor están presentes implícitamente, teniendo que ingresar a la entidad respectiva a la publicación para conocer las publicaciones de cada autor. Esto también ocurre en BibKG, al ignorar la propiedad *author_of* en el *parser* de datos a la hora de crear el archivo JSON de BibKG.

4.3.3. Enlazamiento de revistas académicas

Dentro de las entidades de BibKG de tipo 'Article' (es decir, los artículos), existe la propiedad *in_journal*, que hace referencia a la entidad de la revista en la que fue publicada. Esta posee siempre una única entidad como valor, ya que por definición un artículo está presente en una única revista. Como se observa en la Tabla 4.5, la práctica totalidad de las entidades con la propiedad *in_journal* posee el tipo 'Article'.

Tipo de entidad	**Conteo**
Article	2.549.559
Journal	12
Proceedings	4

Tabla 4.5: Tabla con los tipos de entidad de BibKG que poseen un valor en la propiedad 'in_journal', junto con sus conteos.

Lo anterior permite inferir enlaces entre revistas de BibKG y Wikidata. En este último existe la propiedad P1433 (*published in*), que determina en qué trabajo mayor fue publicada la entidad a la que hace referencia. Como cada entidad de BibKG con revista asociada posee sólo una, se puede asumir con buena fiabilidad que de existir una sola revista asociada en Wikidata esta propiedad también se cumple en este último.

4.3.4. Flujo del enlazamiento

Los tres submétodos de enlazamiento integran al enlazamiento por relaciones en un mismo proceso. El flujo de enlazamiento es el siguiente:

1. Cargar los enlaces realizados en procesos anteriores mediante un diccionario con los datos asociados.
2. Leer el archivo de BibKG entidad por entidad. Se almacenan los datos relacionados con los submétodos en memoria mediante diccionarios, tales como entidades con las propiedades *has_author* o *in_journal* y nombres de entidades con autores, guardando sólo los datos necesarios para la comparación de entidades con Wikidata.
3. A partir de los datos almacenados en el paso anterior, almacenar en las entidades referentes a autores todas las entidades en las que ha participado como autor.
4. Leer el archivo de personas de Wikidata entidad por entidad, almacenando en memoria cada ID junto con su nombre asociado.
5. Leer el archivo de publicaciones de Wikidata entidad por entidad, revisando si esta posee un enlace ya existente con BibKG. De ser así, se aplican los conceptos de enlazamiento explicados anteriormente:

(a) Para el caso de enlazamiento de autores, comparar los autores de este de existir con los de su entidad equivalente de BibKG. Entre estos, si existen autores con el mismo orden asociado y con un nombre similar, se consideran enlazados y se almacena dicha relación en memoria.

(b) Para el caso del enlazamiento de publicaciones, comparar las publicaciones de los autores enlazados. Si las entidades de autores enlazadas de BibKG y Wikidata poseen una entidad con el mismo nombre y que, además, el nombre de la entidad de Wikidata sea único (o sea, que no exista otra entidad de Wikidata con el mismo nombre), y además coinciden en nombres de autores, se consideran estas entidades enlazadas, y se añaden a memoria.

(c) Para el caso del enlazamiento de revistas académicas, si la entidad leída está enlazada con BibKG y posee la propiedad P1433 (referente a la entidad donde fue publicada una entidad de Wikidata) con una única entidad como valor, enlazar directamente las entidades de la propiedad antes mencionada con la entidad de la propiedad *in_journal* de la publicación de BibKG.

En cualquier caso, si al momento de establecer una relación se verifica que la entidad de BibKG ya posee una entidad asociada de Wikidata y además es distinta a la entidad con la que se acaba de establecer una relación, el enlace se anula, se elimina el enlace previamente existente y se añade a la entidad de BibKG a una lista de entidades 'prohibidas', que nunca pueden ser enlazadas en ningún proceso.

6. Añadir las entidades enlazadas de BibKG y Wikidata a la lista de enlaces del enlazador de Wikidata, para posteriormente ser potencialmente añadidos al archivo CSV con todos los enlaces del proceso.

4.3.5. Recursión del proceso

Tanto los enlazamientos por autores como por publicaciones entregan nueva información de enlaces, por lo que cada uno puede entregar información al otro para obtener aún más relaciones. Por ende, la aplicación recursiva de dicho proceso permite la retroalimentación de datos enlazados, siguiendo el siguiente flujo de operaciones:

1. Se aplica el enlazamiento por relaciones.
2. Si en la presente iteración se realizó al menos 1 enlace, se vuelve al paso 1. De lo contrario, se acaba con la recursión.

4.4. Enlazamiento por comparación de parámetros

En casos en los que no existan IDs de ningún tipo en común entre entidades de BibKG y Wikidata ni entidades encontradas entre sus entidades asociadas en sus propiedades, se pueden buscar enlaces mediante valores en común entre sus propiedades, utilizando las que definan con mayor precisión a una entidad. En particular, los nombres, aliases, autores (en

caso de publicaciones) y URLs de entidades pueden representar a una entidad en particular, lo que permite realizar enlazamientos. Sin embargo, puesto que estas propiedades no poseen la capacidad que tiene una ID de definir entidades por sí solas (por ejemplo, pueden existir dos o más entidades con un mismo nombre o con los mismos autores y orden de estos), se deben tomar medidas tales que eviten la mayor cantidad de posibles enlazamientos erróneos.

El flujo de enlazamiento es el siguiente:

1. Cargar en memoria los valores a utilizar de las entidades de BibKG para la posterior comparación de parámetros. En los diccionarios que almacenan la información, cada valor de cada propiedad debe poseer asociada el o los IDs de todas las entidades de BibKG que poseen dicho valor en dicha propiedad (por ejemplo, en el diccionario de nombres de DBLP la llave 'Aidan Hogan' tiene asociada a todas las entidades que poseen en la propiedad *name* el valor 'Aidan Hogan').
2. Almacenar los valores a utilizar de las entidades de las personas de Wikidata.
3. Leer el archivo de publicaciones de Wikidata, entidad por entidad. Por cada una de estas, realizar la comparación de parámetros con las entidades de BibKG. Para esto, existen dos tipos de comparadores:
   (a) **Potenciales enlazadores:** Si una entidad de Wikidata posee algún valor en común de estos parámetros con BibKG, se determina que todas las entidades de BibKG asociadas a dicho valor son potencialmente enlazables con dicha entidad de Wikidata. Estos parámetros son los nombres, aliases y URLs (propiedades P856 y P953 en caso de Wikidata, que se comparan con la propiedad *ee* de BibKG). Por ejemplo, si la entidad leída de Wikidata posee el nombre 'Aidan Hogan' (o lo posee como alias), todas las entidades de BibKG que posean como nombre 'Aidan Hogan' se establecen como potencialmente enlazables con la entidad leída de Wikidata.
   (b) **Parámetros de filtración de enlaces:** A partir de las entidades de BibKG obtenidas mediante el proceso de potenciales enlazadores, se aplican filtros a través de las propiedades pertenecientes a este proceso, que son la fecha de cada entidad (utilizando exclusivamente el año) y los autores (con orden incluido). Si la propiedad de BibKG analizada no coincide con el de Wikidata con dichos parámetros, se descarta como entidad potencialmente enlazable.
4. Definir si existe un enlace o no entre la entidad de Wikidata y alguna de BibKG. Si después del proceso anterior se define que existe una única entidad de BibKG potencialmente enlazable con el de Wikidata, se realiza el enlazamiento por comparación de parámetros.
5. Añadir las entidades enlazadas de BibKG y Wikidata a la lista de enlaces del enlazador de Wikidata, para posteriormente ser potencialmente añadidos al archivo CSV con todos los enlaces del proceso.

Todo el proceso de enlazamiento de datos utilizará todos los métodos descritos anteriormente mediante el siguiente flujo de enlazamiento:

1. Aplicar el método de enlazamiento mediante IDs.
2. Escribir en un archivo CSV los datos relacionados a cada enlazamiento por IDs.
3. Aplicar el método de enlazamiento por relaciones de manera recursiva, a partir de los enlaces obtenidos en el paso anterior. Si al terminar dicho método no existieron enlaces de autores, o enlaces de publicaciones, se pasa al siguiente método. De lo contrario, se vuelve a aplicar el mismo método de enlazamiento por relaciones.
4. Aplicar el método de enlazamiento por comparaciones.
5. Aplicar el método de enlazamiento por relaciones de manera recursiva a partir de lo obtenido en el paso anterior, utilizando la misma metodología del paso 3.
6. Escribir en un archivo CSV los datos relacionados a cada enlazamiento realizado en todos los pasos anteriores.

4.5. Formato de los archivos resultantes

El producto de todos los enlazamientos realizados mediante el proceso explicado en el capítulo anterior consiste en 2 tablas de datos en formato CSV, una con el total de enlazamientos realizados, y otra con los enlazamientos obtenidos exclusivamente mediante el método de enlazamiento por IDs. Cada fila de dichas tablas representa a un enlace, relacionando el ID de BibKG de una entidad con el ID de Wikidata de su equivalente. Adicionalmente, se añade información sobre cada enlazamiento en la misma fila.

El archivo CSV de entidades totales posee las siguientes columnas:

1. **bibkg_id:** Columna con el ID de BibKG de cada entidad.
2. **wikidata_id:** Columna con el ID de Wikidata con el que fue enlazada la entidad de BibKG.
3. **previous_link:** Columna con el ID de Wikidata asociado a la entidad de BibKG, en caso de que haya estado previamente enlazada a ella en el archivo JSON de BibKG (o sea, desde antes de realizar el proceso de enlazamiento).
4. **other_wikidata_ids:** En ciertos enlaces obtenidos a partir del enlazamiento por IDs, ocurrieron casos en los que una única entidad de BibKG fue enlazada con más de una entidad de Wikidata. En dichos casos, se escriben en esta columna todas los IDs de las entidades de Wikidata enlazadas con esta, adicionales a la ID anotada en la columna *wikidata_id*. Cabe destacar que si existe más de una entidad de Wikidata añadida a esta columna, se escriben en un único *string*, separados por tres símbolos *hash* consecutivos (###).
5. **dblp_id:** En esta columna se añaden los IDs de DBLP de cada entidad de BibKG enlazada, en caso de que dicho ID pudo ser obtenido a partir de la entidad de BibKG.

6. **Columnas de tipos de enlazamiento:** A partir de este punto, todas las columnas representan un tipo en particular de enlazamiento. Cada enlace representado en el archivo presenta un '1' como valor en cada columna, de haber sido relacionado con ese tipo de enlace (y un *string* vacío, de no estarlo). Las columnas que representan cada tipo de enlace son las siguientes, en el siguiente orden:
   (a) **linked_by_id:** Entidades enlazadas a través de los IDs.
   (b) **linked_by_id_recursion_authors:** Entidades relacionadas y/o enlazadas a través de la comparación de los autores de las publicaciones previamente enlazadas a través de los IDs.
   (c) **linked_by_id_recursion_journals:** Entidades relacionadas y/o enlazadas a través de la comparación de las revistas de las publicaciones previamente enlazadas a través de los IDs.
   (d) **linked_by_id_recursion_publications:** Entidades relacionadas y/o enlazadas a través de la comparación de las publicaciones de los autores previamente enlazados a través de los IDs.
   (e) **linked_by_comparisons:** Entidades relacionadas y/o enlazadas a través de la comparación de las propiedades de las entidades de BibKG y Wikidata (sin considerar los IDs como parámetro de comparación).
   (f) **linked_by_comparisons_recursion_authors:** Entidades relacionadas y/o enlazadas a través de la comparación de los autores de las publicaciones previamente enlazadas a través del método de IDs y/o el método de comparaciones.
   (g) **linked_by_comparisons_recursion_journals:** Entidades relacionadas y/o enlazadas a través de la comparación de las revistas de las publicaciones previamente enlazadas a través del método de IDs y/o el método de comparaciones.
   (h) **linked_by_comparisons_recursion_publications:** Entidades relacionadas y/o enlazadas a través de la comparación de las publicaciones de los autores previamente enlazados a través del método de IDs y/o el método de comparaciones.

Por otra parte, el archivo CSV de las entidades enlazadas por IDs posee las siguientes columnas:

1. **bibkg_id:** Columna con el ID de BibKG de cada entidad.
2. **wikidata_id:** Columna con el ID de Wikidata con el que fue enlazada la entidad de BibKG.
3. **other_wikidata_ids:** Se escriben todos los IDs de entidades de Wikidata enlazadas con la entidad de BibKG, utilizando la misma metodología y formato que la columna homónima del archivo CSV de enlaces totales.
4. **dblp_id:** En esta columna se añaden los IDs de DBLP de cada entidad de BibKG enlazada, en caso de que dicho ID pudo ser obtenido a partir de la entidad de BibKG.

5. **Columnas de propiedades de IDs de enlazamiento:** A partir de este punto, todas las columnas representan propiedades de IDs de Wikidata con las que se realizaron las comparaciones de IDs. Si una entidad posee en su fila un valor en dicha columna, significa que la entidad de BibKG fue relacionada con la entidad de Wikidata presente en la columna *wikidata_id* mediante esa propiedad. El valor añadido en ese punto corresponde al ID con el que fueron relacionados dichas entidades (con el formato respectivo de la anotación de la ID en Wikidata, según cada propiedad). Cabe destacar que una entidad de BibKG perfectamente pudo ser enlazada con una de Wikidata mediante dos o más propiedades de IDs. Las columnas de esta sección son las siguientes, en el siguiente orden:
   (a) **DBLP publication ID:** Correspondiente a la propiedad 'P8978' de Wikidata.
   (b) **DBLP venue ID:** Correspondiente a la propiedad P8926 de Wikidata.
   (c) **DBLP event ID:** Correspondiente a la propiedad P8978 de Wikidata.
   (d) **DOI:** Correspondiente a la propiedad P356 de Wikidata.
   (e) **arXiv ID:** Correspondiente a la propiedad P818 de Wikidata.
   (f) **ieeeXplore:** Correspondiente a la propiedad P6480 de Wikidata.
   (g) **hdl.handle:** Correspondiente a la propiedad P1184 de Wikidata.
   (h) **d-nb.info:** Correspondiente a la propiedad P1292 de Wikidata.
   (i) **ACM Classification Code:** Correspondiente a la propiedad P2179 de Wikidata.
   (j) **ACM Digital Library citation ID:** Correspondiente a la propiedad P3332 de Wikidata.
   (k) **ACM Digital Library event ID:** Correspondiente a la propiedad P3333 de Wikidata.
   (l) **ethos:** Correspondiente a la propiedad P4536 de Wikidata.
   (m) **ISBN-10:** Correspondiente a la propiedad P957 de Wikidata.
   (n) **ISBN-13:** Correspondiente a la propiedad P212 de Wikidata
   (ñ) **DBLP author ID:** Correspondiente a la propiedad P4536 de Wikidata.
   (o) **ORCID ID:** Correspondiente a la propiedad P496 de Wikidata.
   (p) **Google Scholar ID:** Correspondiente a la propiedad P1960 de Wikidata.

Capítulo 5

Resultados y validación

A partir de los métodos explicados en el Capítulo 4, se obtienen los archivos CSV con los enlazamientos de cada entidad, junto con su información asociada. En el presente capítulo se mostrarán los resultados obtenidos en dichos archivos, junto con la validación de dichos resultados y un análisis a partir de lo obtenido.

Cabe destacar que no necesariamente existe una entidad en Wikidata de existir en BibKG, y viceversa, y no existe una forma conocida de determinar con gran certeza cuál es el universo total de entidades de Wikidata que existe en BibKG. De todas formas, sí se puede determinar que gran parte de los enlaces realizados poseen una alta confiabilidad, debido a que los mismos métodos de enlazamiento poseen métodos de verificación que buscan reducir el riesgo de enlazar erróneamente dos entidades. En especial, el método de enlazamiento por IDs posee una alta confiabilidad al comparar parámetros que son únicos para cada entidad, limitándose el error de esta a ingresos erróneos de valores en las propiedades referentes a dichos identificadores tanto en BibKG como en Wikidata, o múltiples IDs de un mismo tipo existentes para una misma entidad, caso último en el que el error viene de base por parte de la fuente de datos que determina el valor de dichos identificadores.

5.1. Resultados del enlazamiento

En total, se lograron enlazar, considerando los 47.170 enlaces previos, 1.009.226 entidades de BibKG de los 11.089.213 totales con entidades de Wikidata, correspondiente a aproximadamente un 9,31 % del total de entidades de BibKG presentes en el archivo JSON utilizado para el enlazamiento. De todos los métodos de enlazamiento, el método más efectivo en términos numéricos (tanto en cuando a la cantidad de enlaces producidos como de relaciones encontradas) fue el de enlazamiento mediante IDs, que enlazó a más de la mitad del total de entidades enlazadas, como se observa en la Tabla 5.1.

Método de enlazamiento	**Enlaces conseguidos**	**Referencias encontradas**
total	1.009.226	1.009.226
Enlaces previos	47.170	47.170
IDs	557.159	558.446
Autores de IDs	0	37.012
Revistas de IDs	409.287	409.287
Comparación de parámetros	15.408	219.082
Revistas de comp. de parámetros	3.779	413.061
Autores de comp. de parámetros	0	37.144
Publicaciones de IDs	0	158.593
Publicaciones comp. de parámetros	0	161.194

Tabla 5.1: Tabla con el total de enlaces conseguidos mediante todos los métodos de enlazamiento, junto con la cantidad de veces que cada método logró relacionar y/o enlazar entidades. Cabe destacar que las relaciones encontradas a partir de las entidades enlazadas por comparación de parámetros son igual a la suma de las relaciones obtenidas a partir de las entidades enlazadas mediante IDs y las entidades que fueron obtenidas mediante comparación de parámetros. Cabe destacar que los enlaces previos presentados en la tabla hacen referencia a las entidades enlazadas desde antes de la ejecución del trabajo relatado en este informe.

Se observa también en la Tabla 5.1 que muchos métodos tuvieron un resultado redundante, con una cantidad alta de referencias encontradas pero bastante más baja de enlaces conseguidos, lo que implica que se relacionaron principalmente con entidades ya enlazadas con métodos anteriores. En particular, el método de enlazamiento por relaciones de publicaciones no consiguió enlazamientos, pero sí estableció relaciones con una alta cantidad de entidades previamente enlazadas.

En cuanto a los tipos de las entidades enlazadas, se observa en la Tabla 5.2 que el proceso de enlazamiento fue más efectivo en las entidades del tipo 'Article' (artículos científicos de BibKG), tanto por cantidad como por porcentaje del total de entidades de dicho tipo. Cabe destacar que el porcentaje de personas enlazadas es particularmente bajo respecto al total, si se compara con los porcentajes de los otros tipos de entidades (aún si se sumaran las 47.170 entidades previamente enlazadas, el porcentaje no llega al 1 % del total de personas).

Tipo de entidad	Enlaces conseguidos	Total entidades BibKG	Porcentaje
Article	426.934	2.549.564	16,75 %
Journal	413.067	2.549.575	16,20 %
Inproceedings	124.456	2.805.365	4,44 %
Person	60.904	2.734.446	2,23 %
Proceedings	4.325	47.101	9,18 %
Incollection	2.345	66.493	3,53 %
Book	660	18.838	3,50 %
Phdthesis	112	80.674	0,14 %

Tabla 5.2: Tabla con la cantidad total de entidades enlazadas según el tipo de la entidad. Cabe destacar que en esta tabla se toman en cuenta las 47.170 entidades previamente enlazadas con Wikidata (todas del tipo 'Person').

El caso de las entidades enlazadas de tipo 'Journal' (las revistas académicas de BibKG) es bastante particular. A diferencia del resto de entidades de otros tipos enlazadas, (que en casi todos dichos casos no existió más de una entidad de BibKG asociada a otra de Wikidata), las 413.067 entidades de publicaciones fueron enlazadas a sólo 2.881 entidades de Wikidata. De esta forma, se determinó que existen múltiples entidades de BibKG que están repetidas de base. Para validar esta aseveración, se analizaron todas las entidades de BibKG asociadas a cada entidad de Wikidata. De las 2.881 entidades de Wikidata antes mencionadas, en 2.437 de estas se da el caso de que todas las entidades asociadas de BibKG poseen entre ellas el mismo valor en su nombre. Cabe destacar que la práctica totalidad de estos enlaces fueron obtenidos mediante las revistas que se obtuvieron verificando dónde fueron publicadas las revistas enlazadas, como señalan los datos de la Tabla 5.1.

Al ser un caso en el que se detectó un posible problema en la definición de las entidades de BibKG, para este caso en particular se creó un archivo JSON que posee la información obtenida respecto a este ítem, asociando a las entidades de Wikidata (mediante su ID) con una lista con todas las entidades de BibKG con las que fueron enlazadas, como se observa en el ejemplo de la Figura 5.1.

```
1 {"Q64432019": ["journals_make_Manzo20", "journals_make_LeakeJSM19",
    "journals_make_AraujoGSRA19",
2 "journals_make_Manzo19a", "journals_make_Manzo19", "
    journals_make_HughesFWS21", "journals_make_LinjaHNK20",
3 "journals_make_BidokiMS20", "journals_make_MusaDYE19", "
    journals_make_SehgalK19", "journals_make_LeakeM20", "
    journals_make_AhsanGISRH20"]}
```

Figura 5.1: Extracto del archivo JSON creado de enlaces de entidades de revistas de BibKG que hace referencia a los enlaces conseguidos con la entidad de Wikidata Q64432019, referente a la revista académica 'Machine Learning and Knowledge Extraction'. Esta posee un total de 12 entidades de BibKG enlazadas con esta.

En cuanto a los enlazamientos realizados mediante IDs, se observa en la Tabla 5.3 que

la propiedad que realizó la mayor cantidad de enlazamientos fue DOI. Cabe destacar que, si bien se encontraron 417.226 relaciones entre entidades de BibKG y Wikidata mediante la ID de publicaciones de DBLP, tan sólo se realizaron 194 enlaces mediante este. Esto se debe a que gran cantidad de estas entidades fueron relacionadas tanto mediante DOI como por DBLP publication ID, pero por el funcionamiento del algoritmo primero se analiza el DOI, y después la ID de DBLP.

Nombre propiedad	**Enlaces conseguidos**	**Referencias en Wikidata**
total	557.159	558.279
DOI	532.369	532.661
DBLP publication ID	194	417.226
DBLP author ID	13.093	13.093
arXiv ID	7.855	7.863
ORCID ID	321	2.249
DBLP event ID	2.055	2.138
Google Scholar ID	328	717
ISBN-13	620	786
ISBN-10	218	247
ethos	74	74
ACM Digital Library citation ID	31	47
ieeeXplore	6	6
hdl.handle	3	3

Tabla 5.3: Tabla con el total de enlaces conseguidos mediante el método de enlazamiento por IDs, junto con el número de entidades y relaciones encontradas según cada tipo de ID. Cabe destacar que existen algunas entidades que fueron enlazadas con más de una entidad de Wikidata, y que, por ejemplo, si una entidad de BibKG se enlazó con dos entidades de Wikidata, se cuantifican dos enlaces distintos a la hora de obtener los datos en esta tabla. En particular, existen 934 enlaces con la situación anteriormente descrita, correspondientes a 451 entidades de BibKG.

Cabe destacar, además, que se buscó enlazar entidades de BibKG ya previamente enlazadas con Wikidata en los casos específicos en los cuales se encuentran enlaces distintos a los originalmente almacenados. De esta forma, se enlazaron 1.120 entidades en estos casos, de los cuales en 167 casos existen discordancias entre enlaces previos y actuales. Cabe destacar que esto se puede deber tanto a enlaces antiguos que llevan a *redirects* que tienen como destino a la entidad enlazada en este proceso, o a dos o más entidades de Wikidata con un mismo ID asociado. La cantidad de entidades previas enlazadas por el proceso es baja debido a que en el proceso mismo se buscó la incorporación de información nueva en cuanto a enlazamientos, aplicando restricciones en cuanto a cuándo enlazar dichas entidades.

Se observa en la Tabla 5.4 que, salvo en el caso de las revistas académicas comentado anteriormente, casi la totalidad de enlaces fueron obtenidos mediante el método de enlazamiento por IDs. Complementando esta información con la de las tablas anteriores, se puede determinar que, debido a que casi todas las relaciones de los métodos por relaciones y por

comparaciones llevan a enlaces previos realizados en el enlazamiento por IDs, existe un relativo acercamiento al total de entidades de Wikidata que poseen un símil en BibKG, si bien no se puede definir este número con total certeza. Existen otras variables que pueden alejar al proceso de realizar un enlazamiento completo entre entidades de BibKG y Wikidata, como la creación de entidades en Wikidata en un momento posterior a la inserción de datos de BibKG. Aunque dichas creaciones de entidades posean una fuente de datos en común (como lo puede ser DBLP), esta asincronía temporal provocará que sencillamente no se pueda crear un enlazamiento entre ciertas entidades, al simplemente no existir en alguna fuente de datos. En el caso de las entidades de Wikidata que no posean algún ID externo relacionable con BibKG, además, debido a la comparación de propiedades no deterministas (como el nombre o los alias de una entidad de Wikidata con el nombre de BibKG), se confía en un preprocesamiento simple de los *strings* de los datos, que no necesariamente son similares entre las entidades equivalentes de BibKG y Wikidata.

Tipo de entidad	**Enlaces conseguidos**	**Porcentaje total de enlaces**
Article	422.746	99,02 %
Journal	1	0,00 %
Inproceedings	113.542	91,23 %
Person	13.734	22,55 %
Proceedings	4.325	100,00 %
Incollection	2.127	90,70 %
Phdthesis	79	70,54 %
Book	605	91,67 %

Tabla 5.4: Tabla con la cantidad total de entidades enlazadas mediante comparación de IDs según el tipo de la entidad, en relación a los enlaces totales realizados a lo largo del trabajo (incluyendo los enlaces iniciales).

5.2. Validación del proceso

Complementando lo argumentado en el párrafo anterior, es difícil determinar con total certeza la completitud y precisión del proceso de enlazamiento de datos, puesto que justamente el objetivo de este trabajo es lograr conectar dichos datos mediante procesos que permitan establecer equivalencias entre entidades de distintas fuentes de datos, y no existe un muestreo de entidades enlazadas más allá de las 47.170 entidades de personas previamente enlazadas con Wikidata. Sin embargo, existen algunos métodos que permiten crear una idea aproximada de la completitud de dicho proceso.

Para empezar, Wikidata posee enlaces con DBLP mediante ciertas propiedades, que fueron utilizadas para el proceso de enlazamiento de datos. Como las entidades de BibKG poseen a DBLP como una de sus principales fuentes de datos de origen, es razonable establecer que el porcentaje de entidades de Wikidata que poseen una propiedad de ID de DBLP y además existen en BibKG es cercano al 100 %. Como se observa en la Tabla 5.5, los porcentajes de entidades que se lograron enlazar mediante este método son cercanos al 100 %, siendo más

notorio en el caso de publicaciones.

Propiedad Wikidata	**Conteo Wikidata**	**Entidades enlazadas**	**Porcentaje**
Total	499.900	490.840	98,17 %
DBLP Publication ID	431.404	428.572	99,34 %
DBLP Author ID	61.709	60.028	97,28 %

Tabla 5.5: Tabla con la cantidad de enlazamientos de entidades de Wikidata con propiedades de IDs de DBLP con entidades de BibKG, con respecto al total de entidades que poseen dichas propiedades.

Otro método interesante para poder entender la completitud del enlazamiento, es observando la cantidad de autores de publicaciones enlazadas por el proceso que también están enlazados con alguna entidad de BibKG, y comparando este valor con el total de entidades en las propiedades de autor de cada publicación de Wikidata. Para este caso, también vale la pena considerar los *author strings*, es decir, los autores que no fueron registrados en la entidad con el ID de su entidad, sino que simplemente con un *string* con el nombre de dicho autor.

Se obtuvieron los datos relacionados a lo anteriormente mencionado a través de publicaciones enlazadas que poseyeran orden en sus autores, con el motivo de no considerar doblemente en la ecuación a los autores que están presentes tanto en la propiedad *author* como en la propiedad *author string* de Wikidata. Se tiene que, de un total de 146.400 entidades de autores entre todas dichas propiedades, existen 39.173 entidades enlazadas, correspondiente a alrededor de un 26,76 % del total de entidades. Además, se tiene que en promedio existe un 58,38 % de autores enlazados por entidad (esto es, promediando la relación entre autores enlazados y entidades totales de autores entre todas las entidades de Wikidata con enlaces en BibKG). Además, se procesaron un total de 753.633 distintos *string authors* entre todas las publicaciones (cabe destacar, procesando el nombre del *string* para considerar nombres parecidos como iguales). En promedio, en las entidades de publicaciones existen un 83,73 % de *string authors* con respecto al total de autores (esto es, promediando todas las proporciones entre *string authors* y total de autores de cada entidad).

Cabe destacar que el nombre de un autor en la propiedad *author string* no está estandarizado y puede variar entre propiedad y propiedad, al no poseer un ID de Wikidata y poder ser ingresado de cualquier forma por el que ingresa dichos datos, por lo que los conteos de la cantidad total de *author strings* no refleja necesariamente la cantidad total de autores que poseen una referencia en dicha propiedad entre todas las entidades de Wikidata analizadas.

Finalmente, se realizó el mismo experimento mencionado anteriormente, pero a la inversa, o sea, se contaron los porcentajes respecto a la cantidad de entidades de publicaciones de los autores enlazados de Wikidata con BibKG, Se encontró que, de las 952.616 publicaciones de Wikidata asociadas a los autores de Wikidata enlazados con BiBKG, 194.301 lograron ser enlazadas con Wikidata, representando un 20,40 % del total.

Tanto para el caso de autores de publicaciones enlazadas como para el caso inverso, hay que tener en cuenta que en Wikidata perfectamente pueden existir entidades que no poseen

un símil con alguno de BibKG, pero están relacionadas con alguna entidad de Wikidata que sí posee un enlace con BibKG. La no detección de varias de estas entidades puede tener múltiples motivos, como autores que han participado en publicaciones que no forman parte del rubro de las ciencias de la computación, entidades que no han sido creadas en BibKG o sencillamente que, si bien existen tanto en BibKG como en Wikidata, no poseen propiedades comparables (distintos nombres, no poseen propiedades de IDs, etc.).

Capítulo 6

Conclusión

A lo largo de este trabajo se abordó como objetivo principal el enlazamiento de entidades equivalentes entre BibKG y Wikidata, junto con la necesidad previa del preprocesamiento de los datos tanto de BibKG como de Wikidata para facilitar la comparación de las propiedades de dichas entidades.

Con el objetivo de facilitar la comparación de datos entre BibKG y Wikidata y evaluar posibles enlazamientos entre entidades, se preprocesaron los datos disponibles de cada fuente de datos. En el caso de BibKG, el objetivo fue el de facilitar a la máquina la lectura de los datos, debido al formato poco conocido de su archivo *dump*, y el problema de que muchas entidades poseen propiedades asignadas a lo largo de todo el archivo, dificultando la lectura de dichos datos asociados a la entidad. Se creó un nuevo archivo en formato JSON con la información de todas las entidades de BibKG, con línea del archivo representando una entidad en particular de BibKG, en forma de un objeto JSON con las propiedades asociadas a cada entidad. Por otra parte, el objetivo del preprocesamiento del archivo de Wikidata fue el de disminuir el tamaño del archivo, con la consecuente disminución de los tiempos de lectura de este.

El *parser* de datos que adapta la información de las entidades de BibKG a un archivo en formato JSON resultó exitoso. Se logró realizar un mecanismo de tal forma que a partir de cualquier archivo *dump* de BibKG con el formato de MillenniumDB se puede crear un símil en BibKG, manteniendo todas las propiedades de cada entidad en un único objeto JSON, representado en una misma línea del archivo. De esta forma, se puede trabajar fácilmente con los datos de BibKG a la hora de querer hacer uso de sus datos para cualquier trabajo académico, no sólamente para el enlazamiento de sus datos. En cuanto al preprocesamiento de los datos de Wikidata, se logró realizar este proceso de forma satisfactoria, permitiendo la comparación entre entidades de forma satisfactoria, y reduciendo considerablemente el tamaño del archivo, lo que conlleva a un menor tiempo de lectura total del archivo.

Finalmente, se consiguieron enlazar un 9.1 % del total de entidades presentes en el archivo JSON de BibKG con Wikidata. Los resultados son satisfactorios, y permitirán extender la red semántica a la que tiene acceso BibKG de forma considerable, permitiendo a las entidades el acceso a propiedades que no posee BibKG, pero sí Wikidata (instituciones, descripciones, nacionalidades, instancias, etc.), junto con la posibilidad futura de enriquecer

aún más los datos accesibles por BibKG mediante los IDs de distintas fuentes de datos que están disponibles en las entidades de Wikidata.

Cabe destacar que tanto el *parser* de BibKG como el preprocesador de datos de Wikidata y el enlazador de entidades entre BibKG y Wikidata están disponibles en el repositorio de GitHub de todo el proyecto[1], desde donde cualquier usuario puede utilizar los datos de enlazamiento, los sistemas de preprocesamiento de BibKG y Wikidata, y el enlazador de BibKG con Wikidata, tanto para mejorar su funcionamiento como para actualizar los archivos de enlazamiento con versiones futuras de los archivos de BibKG y Wikidata.

Con esto, se cumple con los objetivos principales del trabajo.

6.1. Limitaciones

En cuanto a las limitaciones del trabajo realizado, se tiene que no se ha encontrado un método para encontrar con total certidumbre todas las entidades de Wikidata que poseen un símil en BibKG, limitándose a encontrar las entidades de BibKG que poseen alguna propiedad relacionada a un ID externo y/o nombres o URLs de la entidad similares a su equivalente en BibKG. Por lo descrito anteriormente, se desconoce el porcentaje total de entidades de Wikidata que han sido enlazados con BibKG, con respecto a la cantidad total de entidades de Wikidata que poseen un símil en BibKG.

El origen principal del problema descrito anteriormente está en el criterio ambiguo de ingreso de entidades y propiedades de dichas entidades en Wikidata. Debido a la orientación que posee este de ser una base de conocimientos abierta para poseer todo tipo de entidades, no posee criterios estrictos en cuanto a qué tipo de propiedades posee cada entidad, lo que queda a criterio de las personas y/o máquinas que ingresan la información dentro de estas. Debido a esto, existe la posibilidad de crear mal un enlace debido a un error de propagación que posee como origen datos mal ingresados en alguna propiedad de alguna entidad de Wikidata (órdenes de autores de una publicación mal ingresados, poseer IDs incorrectos en alguna de sus propiedades, etc.).

También hay que considerar que Wikidata puede poseer entidades que, si bien están relacionadas con algún autor o publicación que sí logró ser enlazado con BibKG, sencillamente no existen en BibKG. Si bien puede existir una idea de la cantidad de entidades de Wikidata que poseen esta particularidad mediante la detección realizada en el proceso de validación de entidades de Wikidata que no poseen un enlace con BibKG, pero sí un autor o publicación que sí poseen un enlace con este, el número obtenido también engloba a las entidades de Wikidata que sí poseen un equivalente en BibKG, pero no fueron enlazados por el proceso.

[1] `https://github.com/JorgeCVuwu/BibKG_linker_Wikidata`

6.2. Trabajo futuro

Se proponen las siguientes ideas para mejorar la eficiencia y precisión del trabajo realizado:

- Analizar nuevos criterios de enlazamiento, con el objetivo de aumentar la cantidad total de enlazamientos. Por ejemplo, aumentar la capacidad de detectar similitudes entre nombres de autores y publicaciones de BibKG y Wikidata, con el objetivo de mejorar la efectividad de los métodos de comparación de entidades mediante comparaciones de propiedades (o sea, que no utilizan el ID como método de enlazamiento). También analizar posible uso de otras fuentes de datos externas que puedan formar puentes entre las entidades de BibKG y Wikidata.
- Considerar la aplicación de todo el proceso con los *dumps* actualizados de BibKG y Wikidata cada cierto tiempo, considerando que ambas fuentes de datos actualizan sus datos periódicamente (en el caso de Wikidata, existen actualizaciones de datos constantes). Los nuevos datos de ambas fuentes de datos potencialmente permitirán seguir aumentando la cantidad de enlaces obtenidas por el proceso de enlazamiento de datos creado en este trabajo, junto con la posible solución de casos de enlazamientos incorrectos debido a una introducción incorrecta de ciertos datos en alguna de estas fuentes de datos en el momento en el que se extrajeron dichos datos.
- Reducir aún más la cantidad de datos redundantes para los objetivos de este trabajo del archivo JSON de Wikidata creado en este trabajo, con el objetivo de seguir disminuyendo los tiempos de lectura del archivo, y con ello el tiempo de ejecución total del proceso de enlazamiento.
- Disminuir la cantidad de veces que deben leerse los archivos JSON de BibKG y Wikidata, con el objetivo de disminuir los tiempos de ejecución de los enlazamientos. Si bien esto posee como consecuencia el tener que almacenar una cantidad considerablemente mayor de datos de forma simultánea, se pueden idear nuevas metodologías que permitan disminuir la carga en memoria de datos en la computadora.

Bibliografía

[1] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. Cypher: An Evolving Query Language for Property Graphs. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, editors, *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1433–1445. ACM, 2018.

[2] Michael Ley. DBLP - some lessons learned. *Proc. VLDB Endow.*, 2(2):1493–1500, 2009. http://www.vldb.org/pvldb/vol2/vldb09-98.pdf.

[3] Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph. In Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II*, volume 11137 of *Lecture Notes in Computer Science*, pages 376–394. Springer, 2018.

[4] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 990–998. ACM, 2008. https://doi.org/10.1145/1401890.1402008.

[5] Domagoj Vrgoc, Carlos Rojas, Renzo Angles, Marcelo Arenas, Diego Arroyuelo, Carlos Buil Aranda, Aidan Hogan, Gonzalo Navarro, Cristian Riveros, and Juan Romero. Millenniumdb: A persistent, open-source, graph database. *CoRR*, abs/2111.01540, 2021.