

# Generación de un grafo de conocimiento de enfermedades y síntomas en base a sus ontologías

Joaquín Sarro  
Paula Cianelli

*Facultad de Ingeniería, Universidad de la República*  
Montevideo, Uruguay  
joaquin.sarro2200@gmail.com  
paulacianelli21@gmail.com

## Resumen

Este trabajo aborda la integración de las ontologías médicas de enfermedades y síntomas para identificar relaciones entre conceptos. Se propone la creación de un grafo de conocimiento que combine la Ontología de Enfermedades Humanas (DO) y la Ontología de Síntomas (SYMP), en conjunto con otros datos de interés como son los factores de riesgo, causas y prevenciones. Se utiliza el sitio web MayoClinic.org para obtener datos y se emplea Neo4j para generar la base de grafos. En conjunto, se hace uso de un modelo pre entrenado en base al sitio web PubMed para lograr hacer la correspondencia de los datos de Mayoclinic y las dos ontologías. Finalmente, se realizan casos de estudio sobre las enfermedades cancerígenas e infecciosas en base a consultas en el lenguaje Cypher, para así mostrar posibles usos del grafo creado. Todo el código utilizado en conjunto a las consultas se encuentra en el repositorio de gitlab de la fmg del trabajo [8].

## I. INTRODUCCIÓN

El objetivo principal del artículo es abordar la necesidad de integrar las ontologías médicas de las enfermedades y síntomas para encontrar relaciones entre conceptos distintos de las mismas.

Nos referimos a una ontología médica como representaciones formales y estructuradas del conocimiento médico, diseñadas para capturar y organizar el conocimiento relevante en el campo de la medicina. Una ontología médica describe las entidades médicas, sus propiedades y las relaciones entre ellas, para así luego facilitar la interoperabilidad y el intercambio de información entre sistemas de información médica. A su vez, resultan fundamentales para facilitar la comunicación entre usuarios comunes y profesionales médicos.

Los sistemas de salud inteligentes, como los sistemas de predicción de enfermedades, requieren una base de conocimiento confiable basada en estas ontologías. Es ahí donde han surgido los grafos de conocimiento como una representación dinámica y dominante de una base de conocimiento.

Los grafos de conocimiento (KG, por sus siglas en inglés) son herramientas sofisticadas utilizadas para describir entidades del mundo real y sus interrelaciones organizadas en forma de grafo. Un grafo de conocimiento se considera una red semántica en constante crecimiento de hechos acerca de conceptos específicos (en este caso de conceptos médicos).

Es así como los grafos de conocimiento presentan varias ventajas en comparación con los datos almacenados en un modelo relacional, donde se destacan:

1. Representación flexible: Los grafos de conocimiento permiten representar datos y relaciones complejas de manera más flexible que los modelos relacionales.
2. Escalabilidad y expansibilidad: Los grafos de conocimiento son escalables y se pueden expandir fácilmente a medida que se agregan nuevos datos.
3. Consultas eficientes: Los grafos de conocimiento permiten realizar consultas de manera eficiente al aprovechar la estructura de grafo y las relaciones almacenadas. Las consultas en grafos de conocimiento pueden ser más intuitivas y rápidas que las consultas en un modelo relacional, especialmente cuando se trata de buscar conexiones entre entidades.
4. Integración de datos heterogéneos: Los grafos de conocimiento son especialmente útiles para integrar datos heterogéneos y fusionar información de diversas fuentes.
5. Razonamiento y inferencia: Los grafos de conocimiento son propicios para realizar razonamientos y aplicar inferencias.

Figura 1: Representación del scraping a realizar, <https://github.com/fediazgon/mayoclinic-scrapper>

De esta forma, conseguimos por medio de este script el navegar por las páginas de la web [mayoclinic.org](https://www.mayoclinic.org), buscar enlaces a las enfermedades y recopilar nombres de enfermedades, sus síntomas, causas, factores de riesgo y medidas de prevención. Una vez completada la extracción de datos, estos fueron almacenados en un archivo JSON para su posterior procesamiento.

```
{
  "disease_id": 1085,
  "disease_name": "Thrombophlebitis",
  "symptoms": [
    {
      "symptom_id": 4952,
      "symptom_name": "Warmth, tenderness and pain in the affected are"
    },
    {
      "symptom_id": 4953,
      "symptom_name": "Redness and swellin"
    }
  ],
  "causes": [],
  "risk_factors": [],
  "preventions": [
    {
      "prevention_id": 1122,
      "prevention_name": "Take a walk"
    },
    {
      "prevention_id": 1123,
      "prevention_name": "Move your legs regularly"
    },
    {
      "prevention_id": 1124,
      "prevention_name": "Drink plent"
    }
  ]
},
```

Figura 2: Ejemplo de estructura JSON para la enfermedad Thrombophlebitis

### III-B. Crear correspondencia entre enfermedades obtenidas y las ontologías

En este punto del trabajo, contamos con los datos obtenidos de la web de MayoClinic, donde en los mismos tenemos enumeradas enfermedades en conjunto con sus síntomas, causas, factores de riesgo y prevenciones.

El siguiente paso consta de lograr una correspondencia de estos datos mencionados con la DO y la SYMP. En la práctica, esto se traduce en obtener los datos de ambas ontologías (en particular su ID y nombre correspondiente) y realizar un macheo con los diferentes objetos de nuestro set de datos en formato JSON. Sin embargo, existen los siguientes problemas al realizar lo anterior:

1. Para el caso de las enfermedades, existen numerosos casos, los cuales su nombre en la base de información de la DO no coincide exactamente con el obtenido en la web de MayoClinic. Un ejemplo de esto es la enfermedad definida en MayoClinic como *Cytomegalovirus (CMV) infection*, la cual en la DO es denominada como *human cytomegalovirus infection*.
2. Para el caso de los síntomas, no se cuenta exactamente con el nombre del síntoma, sino con un breve texto describiendo al mismo. Por ejemplo, en los datos MayoClinic contamos con la descripción de un síntoma como *Vaginal discomfort and painful intercourse (dyspareunia)*, el cual buscamos que sea macheado con un síntoma de la SYMP (como por ejemplo SYMP:0000404,dyspareunia).

Para lograr resolver estos problemas, se hará uso del algoritmo BioWordVec [7] referenciado en el paper [1]. Este algoritmo permite obtener un modelo pre entrenando creado a partir del sitio médico <https://pubmed.ncbi.nlm.nih.gov/>, que luego podrá ser utilizado para obtener un vector por cada texto/frase que se le ingrese. De esta forma, podremos emplear dicho modelo para realizar la correspondencia para enfermedades y síntomas, donde también se obtiene un umbral de similitud que va desde 0 a 1. El modelo ofrece resultados superiores en comparación a otros modelos estandar, tal como se menciona en el paper [1] y en la siguiente gráfica incluido en el mismo.

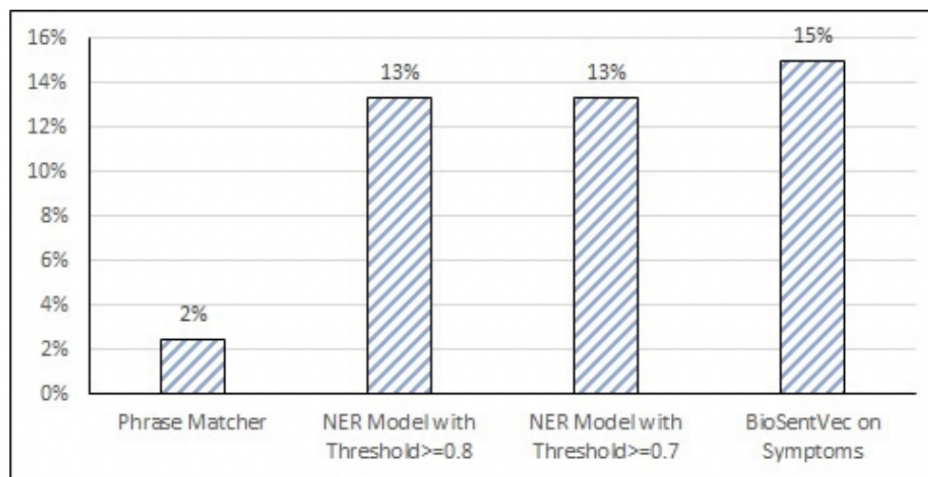


Figura 3: Comparación de BioWordVec con otros modelos

### III-C. Crear correspondencia entre enfermedades y su valor según la DO

En primer lugar, lograremos la correspondencia entre las enfermedades de MayoClinic y las de la DO. Se recuerda que esta ontología ofrece una representación estructurada y jerárquica de las enfermedades humanas, donde tiene como objetivo capturar el conocimiento sobre las mismas y sus relaciones. Esto incluye la clasificación de enfermedades, los genes relacionados, las vías biológicas afectadas y otros aspectos relevantes.

En cuanto a lo que nos compete a nuestro trabajo, nos fue requerido obtener los ids de las enfermedades en conjunto con su nombre (el cual luego es el utilizado para realizar la correspondencia). Para lograr esto, utilizamos la knowledgebase de la DO [5], donde en la misma se provee un sandbox que por medio de consultas en SPARQL se pueden obtener datos de las distintas enfermedades. Habiendo realizado una de las consultas predefinidas por la plataforma (All diseases with their parents (identified by label)), logramos obtener un csv el cual cuenta con la enfermedad, su id, y su enfermedad padre. A motivos de nuestro trabajo, por medio de un script de Python se remueve la columna correspondiente al padre (al no ser relevante para el trabajo).

Luego de contar con este csv con el par *id, enfermedad* por cada fila, creamos un programa en Python que tome este csv y los nombres de enfermedades del archivo JSON con los datos obtenidos de MayoClinic. Utilizando el modelo de BioWordVec [7], se obtiene el nombre de la enfermedad de la DO la cual tenga mayor similitud y así se realiza la correspondencia. Como dato extra, se agrega el valor del umbral en los datos.

De esta forma, se logra agregar el id según la DO a todas las enfermedades obtenidas mediante la Web de MayoClinic, en conjunto con el umbral de similitud.

```
{
  "disease_id": 5,
  "disease_name": "Soft tissue sarcoma",
  "symptoms": [],
  "causes": [],
  "risk_factors": [],
  "preventions": [],
  "DO_label": "DOID:3369",
  "DO_label_similarity": 0.9254606366157532
},
```

Figura 4: Ejemplo de la nueva estructura JSON con DO id para la enfermedad Soft tissue sarcoma

#### III-D. Crear correspondencia entre síntomas y su valor según la SYMP

Ahora resta con hallar la correspondencia entre los textos de los síntomas de MayoClinic y los nombres de la SYMP. Al igual que la DO, la SYMP ofrece una representación de los síntomas, facilitando su almacenamiento e interpretación de los datos a futuro.

Para lograr la correspondencia, se utiliza la knowledgebase de la SYMP [6], en la misma contamos con una página *symp.obo* que nos permite hacer scraping de los datos de esta ontología. De esta forma, por medio de un programa en Python logramos obtener un csv el cual cuenta con el nombre del síntoma y su id.

Luego de contar con este csv con el par *id, síntoma* por cada fila, creamos un programa en Python que tome este csv y los textos con la descripción de los síntomas para cada enfermedad del archivo JSON con los datos obtenidos de MayoClinic. Utilizando el modelo de BioWordVec [7], se obtiene el nombre del síntoma de la SYMP el cual tenga mayor similitud con el texto de MayoClinic y así se realiza la correspondencia. Como dato extra, se agrega el valor del umbral en los datos.

De esta forma, se logra agregar el id según la SYMP a todas los síntomas obtenidos mediante la Web de MayoClinic, en conjunto con el umbral de similitud.

```

{
  "disease_id": 588,
  "disease_name": "Liposarcoma",
  "symptoms": [
    {
      "symptom_id": 1862,
      "symptom_name": "A growing lump of tissue under the skin",
      "SO_label": "SYMP:0000018",
      "SO_label_similarity": 0.7173867225646973
    },
    {
      "symptom_id": 1338,
      "symptom_name": "Pain",
      "SO_label": "SYMP:0000099",
      "SO_label_similarity": 1.0000001192092896
    },
    {
      "symptom_id": 2336,
      "symptom_name": "Swelling",
      "SO_label": "SYMP:0000717",
      "SO_label_similarity": 0.9244087338447571
    },
    {
      "symptom_id": 2812,
      "symptom_name": "Weakness of the affected limb",
      "SO_label": "SYMP:0000222",
      "SO_label_similarity": 0.8390624523162842
    }
  ],
  "causes": [],
  "risk_factors": [],
  "preventions": [],
  "DO_label": "DOID:3382",
  "DO_label_similarity": 1.0
},

```

Figura 5: Ejemplo de la nueva estructura JSON con SYMP id

### III-E. Creación de la base de datos

Una vez obtenidos los datos y almacenados en el JSON anteriormente nombrado, podemos comenzar a construir nuestra base de datos en neo4j para así luego poder cargarla con los mismos.

Para esto, hicimos uso de la aplicación de escritorio que ofrece neo4j, la cual ofrece la posibilidad de crear una base de datos de grafos en un entorno local y diferentes interfaces para la gestión y visualización de la misma.

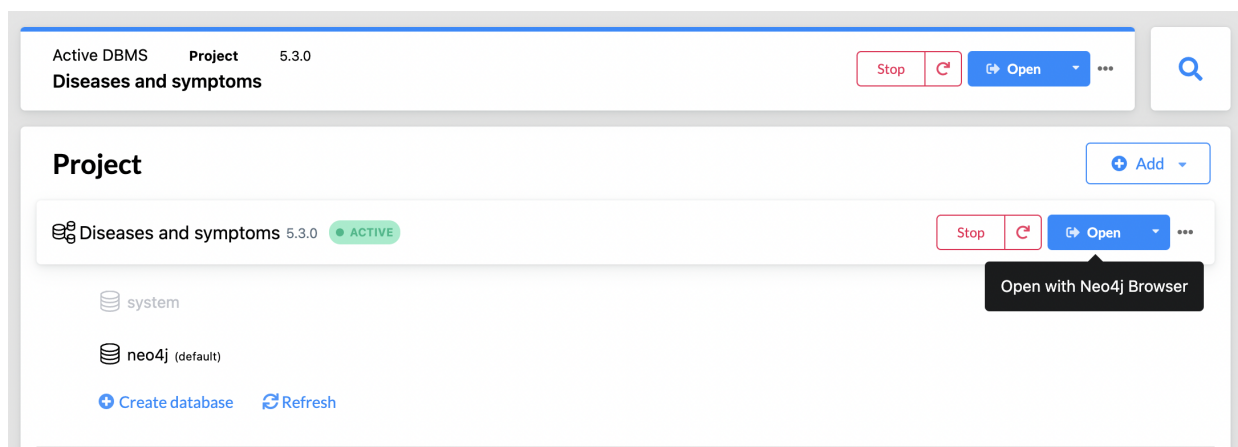


Figura 6: Base de datos corriendo en la aplicación de escritorio de neo4j

A su vez, se presenta un diagrama descriptivo de las distintas actividades realizadas, la cual fue tomada del paper de referencia [1]. En nuestro caso se omite el manejo manual de los datos para los casos de umbrales de similitud menores a 0.7, donde se incluyen indistintamente:

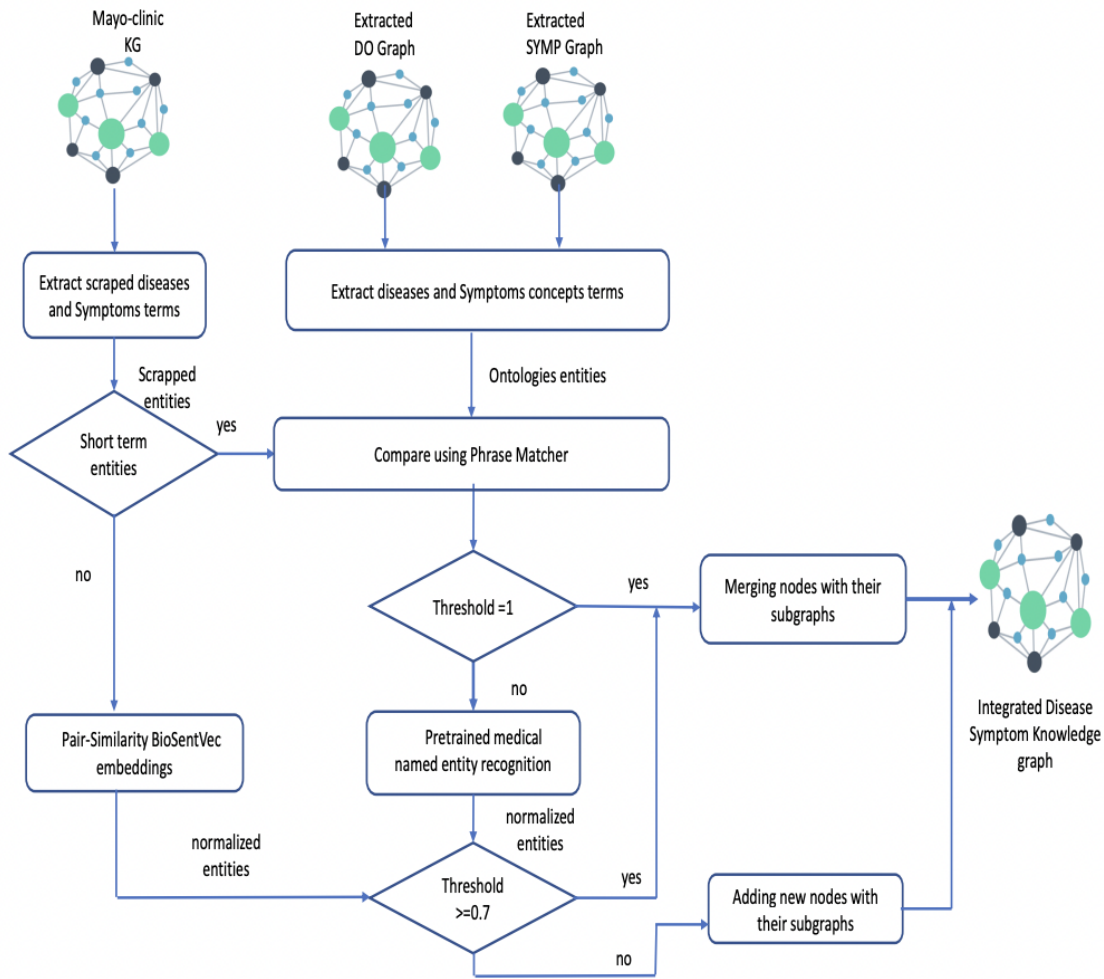


Figura 8: Diagrama de actividades realizadas

#### IV. EXPERIMENTACIÓN

En esta sección se buscará estudiar sobre el grafo distintos casos de interés por medio de consultas Cypher sobre el grafo creado, destacando el valor del contar con estos datos en un grafo y tener el etiquetado de los síntomas y enfermedades según la SYMP y la DO. Tomando como referencia el paper [1], para el caso de las enfermedades y los síntomas se utilizarán aquellos nodos que tengan un valor de umbral de similitud mayor a 0.7. Todas la consultas son incluidas en el repositorio en conjunto al código.

##### IV-A. Estudio del sub grafo de las enfermedades cancerígenas

A continuación, se realizarán casos de estudio sobre el grafo creado tomando foco en las enfermedades cancerígenas (buscando aquellas que contengan el término *cancer* en su nombre). El objetivo es simular un uso realista del grafo creado, donde el mismo pueda ser utilizado en por ejemplo sistemas médicos relacionados con el cáncer. Este sub grafo cuenta con un total de 39 enfermedades y 174 síntomas (donde estos tienen un umbral de similitud mayor a 0.7).

*IV-A1. Los 10 síntomas más comunes en enfermedades cancerígenas:* Por medio de una consulta Cypher, se logran obtener los 10 síntomas más comunes en este tipo de enfermedades, donde se incluye la cantidad de ocurrencias y su respectiva etiqueta según la SYMP. Estas etiquetas obtenidas permiten que luego estos síntomas puedan ser utilizados en otros sistemas que adopten la ontología para sus datos, donde por ejemplo se podría estudiar en un sistema con enfoque en la genética si existe una relación entre estos síntomas.

| Síntoma                        | Etiqueta según SYMP | Ocurrencias |
|--------------------------------|---------------------|-------------|
| "Losing weight without trying" | "SYMP:0000616"      | 8           |
| "Fatigue"                      | "SYMP:0019177"      | 6           |
| "Weight loss"                  | "SYMP:0000178"      | 5           |
| "Back pain"                    | "SYMP:0000632"      | 5           |
| "Ear pain"                     | "SYMP:0000632"      | 4           |
| "Bone pain"                    | "SYMP:0020023"      | 4           |
| "Nausea"                       | "SYMP:0000458"      | 4           |
| "Difficulty swallowing"        | "SYMP:0000308"      | 4           |
| "Vomiting"                     | "SYMP:0019145"      | 4           |
| "Fever"                        | "SYMP:0000613"      | 4           |

IV-A2. De los 10 síntomas más comunes en enfermedades cancerígenas, cuál es su grafo incluyendo los factores de riesgo: Para el resultado de IV-A1, se obtiene el grafo que también incluye los factores de riesgo. Se puede observar como mucho de los factores de riesgo (nodos celestes) se encuentran relacionados a varias de las enfermedades (nodos violetas). Esta visualización resulta valiosa para casos de estudio relacionados a lo mismo, así como también los resulta el tener las etiquetas de la DO y la SYMP para el uso de esta información en otros sistemas.

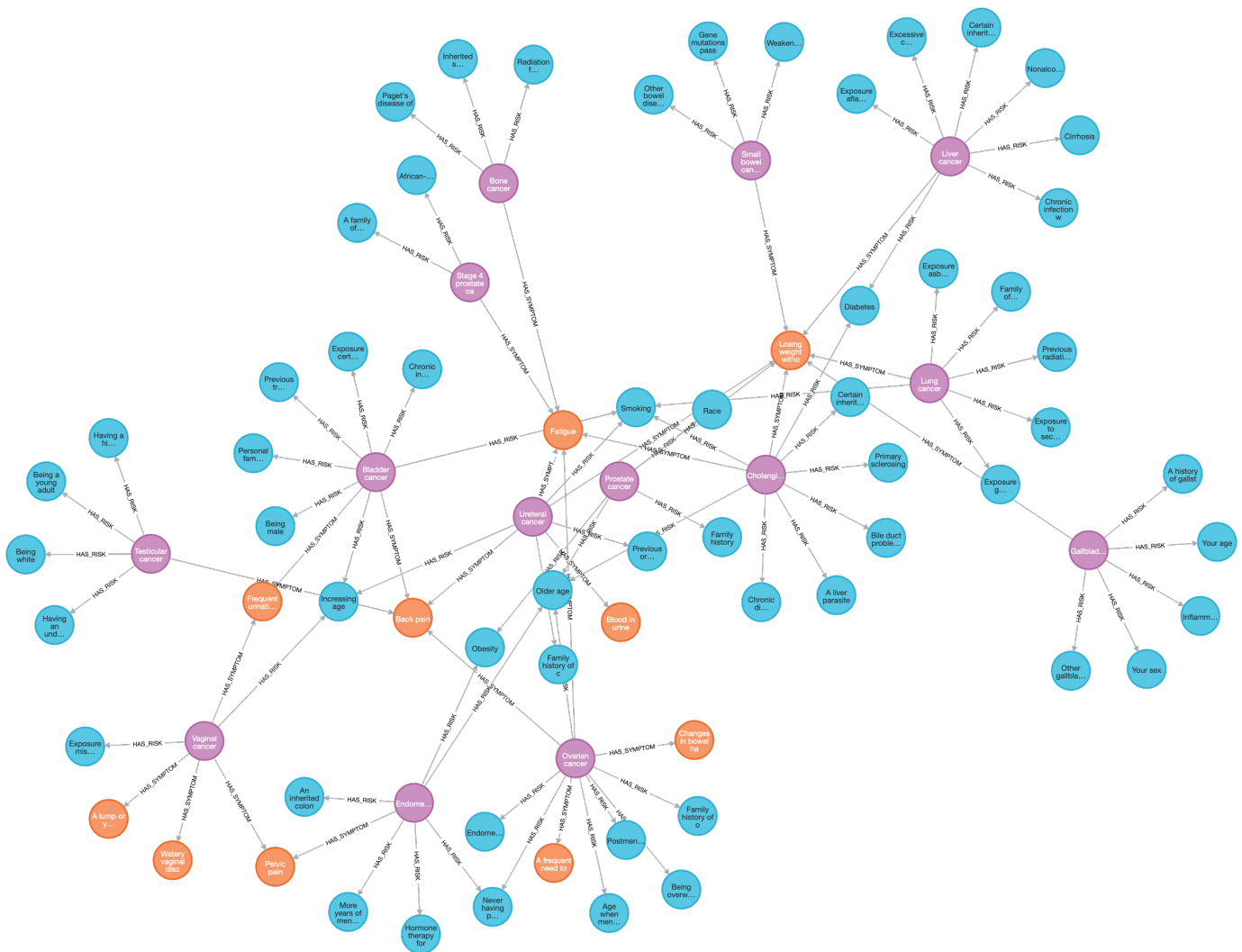


Figura 9: Grafo obtenido

IV-A3. *Las 5 causas menos frecuentes de las enfermedades cancerígenas:* Resulta interesante analizar cuales son las causas menos frecuentes de estos tipos de enfermedades según nuestro grafo, al ser estas poco frecuentes y eventualmente de interés para los expertos del campo.

| Cause                       | causeCount |
|-----------------------------|------------|
| "Epithelial ovarian cancer" | 1          |
| "Germ cell tumors"          | 1          |
| "Stromal tumors"            | 1          |
| "Lymphoma"                  | 1          |
| "Carcinoid tumors"          | 1          |

IV-A4. *Camino más corto entre el cáncer de próstata y cáncer de mama:* Al ser estos dos de los tipos más comunes, resulta natural ver cual es el camino más corto tomando en cuenta las relaciones definidas en el grafo. El resultado es que el camino más corto se da por un camino de largo dos, por medio del riesgo *obesity*.



Figura 10: Camino más corto

#### IV-B. Estudio del sub grafo de enfermedades infecciosas

Ahora se realizará lo mismo que en IV-A pero para el caso de las infecciones. Se utilizará para este estudio las enfermedades que contengan *infection* en su nombre.

IV-B1. *Los 10 síntomas menos frecuentes en enfermedades infecciosas*: Similar al otro caso, buscaremos ahora los de menor ocurrencia. Esto resulta pertinente no solo por poder estudiar estos casos excepcionales a nivel del rubro, sino que al igual que el anterior permite integrarlo con sistemas que utilizan la SYMP para sus datos.

| Síntoma                                                                                                                                                                                                                                                                                                        | SYMP[_label    | Ocurrencias |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|-------------|
| "Occasional stomach pain and nausea"                                                                                                                                                                                                                                                                           | "SYMP:0000458" | 1           |
| "Itching of the anal or vaginal area"                                                                                                                                                                                                                                                                          | "SYMP:0000398" | 1           |
| "Cellulitis is an infection of the deeper layers of skin. It causes redness and swelling on the surface of your skin. Sores or areas of oozing discharge may develop, too."                                                                                                                                    | "SYMP:0020030" | 1           |
| "Blood in the stool"                                                                                                                                                                                                                                                                                           | "SYMP:0000229" | 1           |
| "Nausea or vomiting"                                                                                                                                                                                                                                                                                           | "SYMP:0019145" | 1           |
| "Toxins produced by the staph bacteria may cause staphylococcal scalded skin syndrome. Affecting mostly babies and children, this condition includes a fever, a rash and sometimes blisters. When the blisters break, the top layer of skin comes off. This leaves a red, raw surface that looks like a burn." | "SYMP:0000671" | 1           |
| "Stomach (abdominal) cramps"                                                                                                                                                                                                                                                                                   | "SYMP:0000802" | 1           |
| "Diarrhea (often containing blood or mucus)"                                                                                                                                                                                                                                                                   | "SYMP:0020011" | 1           |
| "This contagious, often painful rash can be caused by staph bacteria. Impetigo usually has large blisters that may ooze fluid and develop a honey-colored crust."                                                                                                                                              | "SYMP:0000752" | 1           |
| "Insomnia, irritability, teeth grinding and restlessness"                                                                                                                                                                                                                                                      | "SYMP:0000076" | 1           |

IV-B2. *Grafo de prevenciones de las enfermedades infecciosas*: Las enfermedades infecciosas son muy comunes y en su mayoría prevenibles. Resulta interesante visualizar el grafo para observar como se relacionan las diferentes enfermedades infecciosas (**nodo violetas**), sus prevenciones (**nodos rojos**) y síntomas (**nodos naranjas**). En particular, se observa una alta relación de la prevención *Wash your hands*.

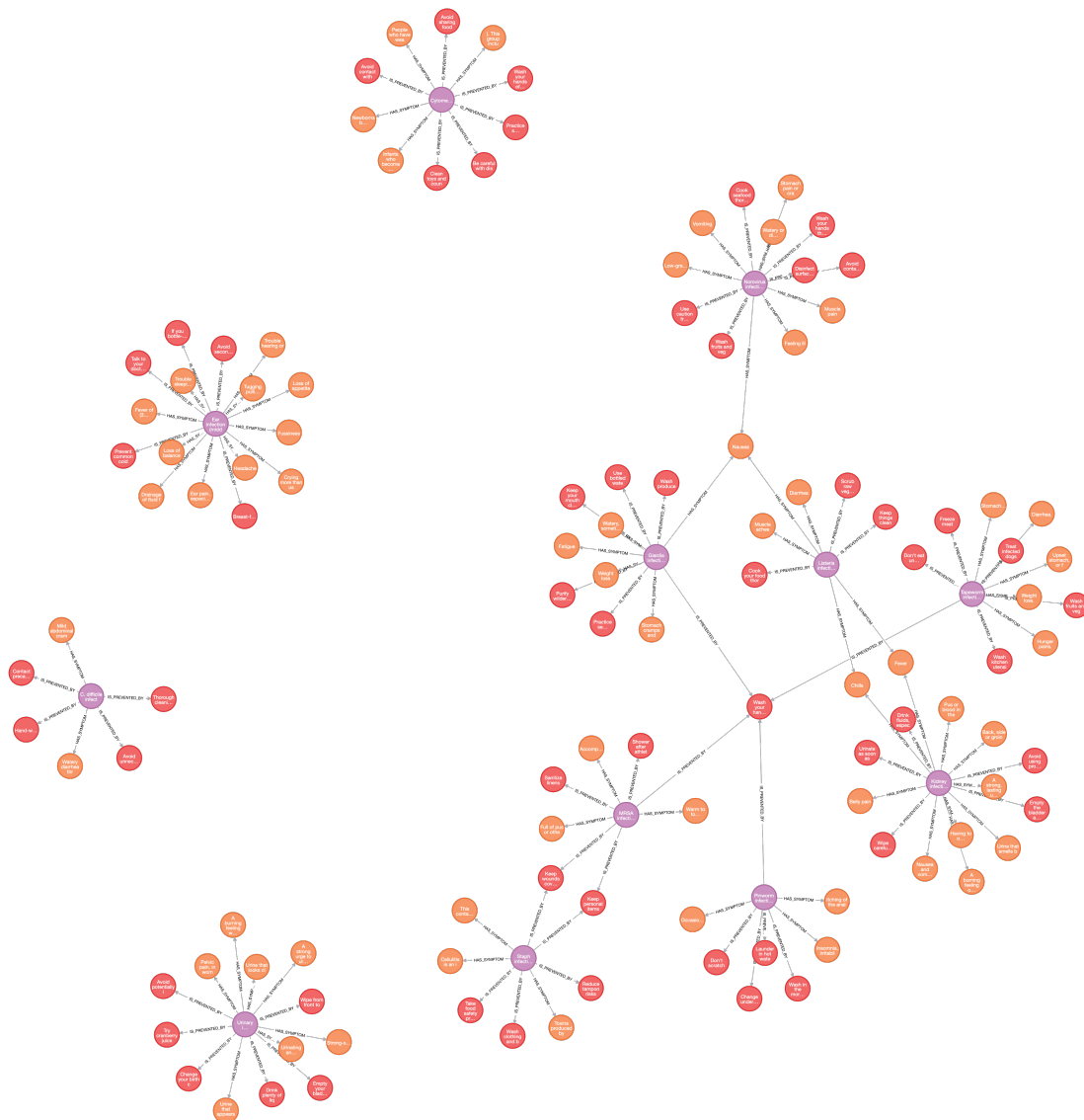


Figura 11: Grafo obtenido

## V. CONCLUSIONES Y TRABAJO FUTURO

Durante este trabajo, se logró generar un grafo de conocimiento que integra exitosamente la ontología de enfermedades (DO) y la ontología de síntomas (SYMP) utilizando la enciclopedia médica de MayoClinic como fuente de datos. Este grafo no solo incluye la integración de las ontologías, sino que también proporciona información relevante sobre factores de riesgo, causas y prevenciones de enfermedades.

El grafo de conocimiento integrado sirve como base para sistemas de predicción de enfermedades o alertas para usuarios comunes sobre posibles dolencias, siendo a su vez altamente dinámico, escalable, fácilmente reproducible, confiable, actualizado y eficiente. A su vez, este grafo enlazado a ontologías estandarizadas es fundamental para sistemas de salud inteligentes, permitiendo la posibilidad de crear flujos de datos entre diferentes sistemas que adopten las ontologías para el manejo de sus datos.

En la sección de experimentación [IV](#), se llevaron a cabo pruebas para demostrar cómo es posible utilizar este grafo para extraer sub grafos especializados, los cuales resultan de gran utilidad en sistemas de atención médica enfocados en dominios específicos, como el de las enfermedades cancerígenas o las enfermedades infecciosas. Estos casos de estudio ejemplifican cómo la representación de datos en forma de grafo enriquece el análisis y simplifica las consultas realizadas sobre la base de datos, facilitando así la obtención de información precisa y relevante.

Como perspectiva de trabajo futuro, se plantea la posibilidad de considerar la inclusión de conjuntos de datos adicionales provenientes de diversas fuentes médicas. Este enfoque permitiría enriquecer aún más el grafo de conocimiento, mejorando la correspondencia y la fiabilidad de los datos en relación con las diferentes ontologías utilizadas. Asimismo, es importante abordar posibles desafíos y limitaciones relacionados con la calidad y la actualidad de los datos provenientes de fuentes externas, y desarrollar estrategias efectivas para mantener el grafo actualizado a medida que se generen nuevos conocimientos médicos.

#### ABREVIACIONES

**KG** Grafo de conocimiento - Knowledge graph

**DO** Ontología de las enfermedades humanas - The Human Disease ontology

**SYMP** Ontología de los síntomas - The Symptoms Ontology

#### REFERENCIAS

- [1] An Automatic Generation of Heterogeneous Knowledge Graph for Global Disease Support: A Demonstration of a Cancer Use Case, <https://www.mdpi.com/2504-2289/7/1/21>.
- [2] Términos de uso de mayoclinic.org, <https://connect.mayoclinic.org/terms-of-use/>.
- [3] Librería de python scrapy, <https://scrapy.org/>.
- [4] Proyecto de GitHub para scraping y subida a neo4j, <https://github.com/fediazgon/mayoclinic-scraper>.
- [5] DO Knowledgebase, <https://disease-ontology.org/>.
- [6] SYMP Knowledgebase,
- [7] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. [BioWordVec, improving biomedical word embeddings with subword information and MeSH] <https://www.nature.com/articles/s41597-019-0055-0>. Scientific Data. 2019.
- [8] Repositorio con el código del trabajo, <https://gitlab.fing.edu.uy/joaquin.sarro/entregagrupa22bases>.