**Abstract**

This report summarizes the progress made in developing a knowledge graph for orthopedic injuries up to Week 5. The project involves curating a knowledge graph from manually created files, supplemented by data gathered through APIs and web crawling. The graph is intended to aid in information retrieval and power recommender systems for patients and healthcare providers. This report will also outline the ongoing efforts in expanding data collection sources into the knowledge graph, this will ensure that relevant and accurate information is integrated.

**Introduction**

The motivation behind constructing this knowledge graph stems from the need to enhance the accessibility and personalization of information related to orthopedic injuries. By integrating diverse data sources into a coherent graph, the project aims to facilitate better diagnostic tools, treatment options, and patient education, thereby improving healthcare outcomes. Leveraging advanced data management techniques will ensure that a comprehensive resource base for orthopedic injury management is created properly.

**Literature Review**

The integration of knowledge graphs in medical informatics has emerged as a transformative approach to managing vast and varied data sources crucial for enhancing healthcare delivery. Knowledge graphs bring structured and semantic coherence to disparate data, facilitating enhanced data interoperability and accessibility. This capability is particularly critical in fields like orthopedics, where treatment outcomes can significantly benefit from holistic and interconnected data analysis.

**Advancements in Knowledge Graphs for Healthcare**: Recent studies underscore the adoption of knowledge graphs to synthesize information across the healthcare spectrum—from patient records and clinical studies to drug information and treatment outcomes. For example, Hänsel et al. (2023) discussed how knowledge graphs could help integrate patient diagnostic data with relevant medical research to optimize treatment pathways, thereby supporting clinical decision-making processes. Similarly, Bonner et al. (2022) utilized knowledge graphs to link genetic information with clinical symptoms to advance personalized medicine strategies.

**Entity-Centric Approach in Knowledge Graphs**: According to Nickel et al. (2015), the Entity-Centric approach to constructing knowledge graphs involves organizing information around key entities and their interrelations. This methodology is particularly well-suited to medical informatics, where entities (e.g., diseases, treatments, symptoms, and medical devices) are inherently interconnected. By focusing on entities and their relationships, this approach allows for a more intuitive representation of complex medical information, making it easier for healthcare providers to navigate and utilize this data effectively.

In the context of orthopedic injuries, where the correlation between symptoms, the efficacy of treatments, and patient outcomes must be meticulously understood and accessed, the Entity-Centric method offers substantial advantages:

* **Enhanced Data Coherence**: By centering the knowledge graph around entities such as orthopedic injuries and treatments, the graph naturally aligns with clinical workflows which typically follow a diagnosis-treatment-outcome trajectory.
* **Improved Query Performance**: Structuring data around entities simplifies querying complex relationships, enabling faster and more accurate retrieval of information about treatment options and their effectiveness.
* **Facilitation of Advanced Analytics**: Entity-centric knowledge graphs support more sophisticated analytics, such as predictive modeling and trend analysis, by providing a structured framework that can be readily analyzed.

**Methods**

**Data Sources Identified:**

1. **PubMed, Google Scholar, and American Academy of Orthopedic Surgeons**  for peer-reviewed medical articles.
2. **FDA and manufacturer websites** for information on medical devices.
3. **Patient forums and social media platforms** like HealthUnlocked, providing user-generated content on patient experiences.

These sources provide both unstructured and structured data. Specifically taking the structured data from medical research databases, and gathering unstructured data from forums.

**Data Collection:**

So far, the collection efforts have focused on extracting structured data from medical research databases and unstructured data from forums and blogs. Examples include:

* JSON documents containing metadata of research papers.
* Tables summarizing device specifications and approvals.

Within the knowledge graph, the structured data that was extracted from medical research databases (JSON documents), would be represented as nodes corresponding to the research papers. These nodes would then be linked to other nodes representing relevant entities, such as medical conditions, treatments, devices, etc. From there the attributes that would be included and associated with these nodes would be associated symptoms, side effects, and regulatory approvals respectively. Conversely, when collecting unstructured data a new approach can be taken where data is parsed and structured in order to extract relevant information. With a relational database approach, these sources would then translate into tables with links being defined by common attributes such as device names or similar symptoms. For example, some common orthopedic injuries may utilize the same device for therapy which can be linked and interconnected through these tables.

**Schema Proposal:**

The proposed schema for the knowledge graph is outlined in EdgeDB as follows:

module default {

type Injury {

required property name -> str;

multi link treatments -> Treatment;

multi link symptoms -> Symptom;

}

type Treatment {

required property name -> str;

multi link used\_for -> Injury;

multi link requires\_devices -> MedicalDevice;

}

type MedicalDevice {

required property name -> str;

multi link required\_for -> Treatment;

}

type Symptom {

required property description -> str;

multi link indicative\_of -> Injury;

}

}

A visualization of the graph structure includes nodes such as **Injury**, **Treatment**, **MedicalDevice**, and **Symptom**, with edges representing relationships like **treats**, **requires**, and **indicative\_of**.

**Completion Plans:**

* Expand the crawler to cover more sources.
* Enhance the classifier to improve data relevance and accuracy.
* Begin integrating data into the EdgeDB schema.

**Storage:**

Leveraging EdgeDB and its versatile capabilities in handling different data types will help ensure that the complex relationships between the data points are represented. An emphasis will be placed on utilizing its graph data capabilities in order to implement the knowledge base. Because of the inherent graph-like structure of the data, this implementation will seamlessly represent the complex relationships between a variety of entities. Relational storage will also be used but to a minimal extent in order to focus more on graph data capabilities. Additionally with EdgeDB’s document-based storage features exploring the unstructured data, such as JSON Documents, will be possible. While the current dataset may not include geometry data types, potential incorporation of such data will be considered for future relevance.

**Results**

Initial results indicate that while there is a wealth of available information, significant challenges remain in data normalization and relevance classification. Early tests with the schema have shown promising flexibility in representing complex medical relationships. Moving forward, further refinement of the schema and iterative testing will need to happen in order to address these challenges and optimize the knowledge graph.

**Conclusions**

The knowledge graph has the potential to significantly impact the management of orthopedic injuries by providing structured and easily accessible information. The challenges of data integration and maintaining up-to-date information are substantial but manageable with continued refinement of the crawling and classification processes.

The report will continue with detailed plans for data storage, using PostgreSQL for structured data and considering EdgeDB for its graph capabilities to effectively manage relationships and enhance query performance. The combination of relational, document-based, and graph data structures will be crucial to accommodate the diverse data types involved in this project.

In order to complete the knowledge base further expansion of data collection is required to encompass additional relevant sources. As stated previously the scope of the crawler will be broadened to include many different types of sources. From there, refinement of data classification algorithms will continue to ensure that only relevant and accurate information is integrated into the knowledge base. Finally, the creation of a UI and optimization of query performance will be prioritized in order to enhance the usability and effectiveness of the knowledge graph.

**EdgeDB**

<https://cloud.edgedb.com/org/FarisRaza1/instance/testInstance>

**References**

Bonner, Stephen, Ian P. Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L. Hamilton. "A review of biomedical datasets relating to drug discovery: a knowledge graph perspective." Briefings in Bioinformatics 23, no. 6 (2022): bbac404.

Hänsel, Katrin, Sarah N. Dudgeon, Kei-Hoi Cheung, Thomas JS Durant, and Wade L. Schulz. "From data to wisdom: biomedical knowledge graphs for real-world data insights." Journal of Medical Systems 47, no. 1 (2023): 65.

Nickel, Maximillian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. "A Review of Relational Machine Learning for Knowledge Graphs." Proceedings of the IEEE, 104(1): 11–33.