

Advancing Translational Science: Knowledge Graph Integration for Rare Disease Insights

GMG ArcData implemented a knowledge graph solution, integrating NLP workflows to enhance rare disease research and content development for the National Center for Advancing Translational Sciences.

Call for change

1. The Challenge of Rare Disease Data Management – With over 6,000 rare diseases affecting millions worldwide, the volume of medical literature and research data is vast and growing exponentially. For researchers and healthcare professionals, staying updated on new developments, treatment options, and disease classifications is a monumental challenge. Much of this information is buried in unstructured data sources such as research articles, reports, and clinical trials, making it difficult to systematically track, analyze, and utilize. The lack of efficient tools to manage, extract, and categorize this information hampers the ability to make timely, data-driven decisions regarding rare diseases.

2. Limitations of Traditional Data Systems – Traditional databases and content management systems struggle to cope with the complexity of rare disease data, often due to their inability to handle unstructured information or to connect different data points in meaningful ways. This limitation makes it difficult for curators and researchers to cross-reference findings, monitor disease trends, and update content in a timely manner. Additionally, rare disease research is highly interdisciplinary, requiring the integration of data from diverse sources such as clinical studies, genomic research, and patient registries. Without a cohesive system to unify and make sense of this information, crucial insights can be missed.

3. The Need for Semantic Understanding In order to fully leverage the wealth of data available on rare diseases, a solution is needed that goes beyond simple data storage and retrieval. Researchers and curators require a system that can understand the semantic relationships between terms, diseases, and findings. This need for semantic understanding is critical in extracting relevant insights from large datasets, enabling better categorization and annotation of rare disease information. By structuring data in a knowledge graph, researchers can uncover hidden relationships, track disease progressions, and generate new hypotheses, which can accelerate scientific discovery and improve patient outcomes.

4. Improving Collaboration and Accessibility – Another significant need is the ability for diverse teams of researchers, healthcare professionals, and content curators to access and contribute to a shared repository of knowledge. Rare disease research often involves collaboration across different fields, institutions, and countries, which can be hindered by siloed data and inconsistent information-sharing practices. A knowledge graph-based repository provides a unified platform for these stakeholders to collaborate effectively. By integrating this solution with existing workflows, the National Center for Advancing Translational Sciences (NCATS) and similar organizations can improve data accessibility, enhance content accuracy, and foster collaboration in the fight against rare diseases.





Improving Data Management & Collaboration

GMG ArcData's solution involved designing and implementing a Neo4j-based knowledge graph (KG) repository that integrates natural language processing (NLP) workflows for monitoring and classifying information on over 6,000 rare diseases. This knowledge graph serves as a central hub for aggregating and semantically annotating unstructured data, such as research articles from PubMed, allowing for more efficient extraction of relevant information. By leveraging NLP techniques, the system automatically tags and categorizes disease-related data, enabling researchers to easily navigate through complex datasets and uncover critical insights. This semantic enrichment allows for the identification of relationships between rare diseases, research findings, and evolving epidemiological trends, significantly enhancing the ability to monitor disease progression and support content curators.

The knowledge graph's integration with an existing data model ensures that it aligns with broader healthcare standards, and its deployment on a Neo4j platform facilitates advanced querying and visualization of data. GMG ArcData's approach empowers researchers and curators to efficiently explore and interact with this vast dataset in ways that were not previously possible. The solution also underwent usability and security validation to ensure it met the operational standards of the National Center for Advancing Translational Sciences (NCATS). Through this innovative platform, the research team was able to demonstrate significant scientific contributions, which have since been published, showcasing the potential for knowledge graphs to transform rare disease research and content development. The team's approach and findings were recently published in a [peer-reviewed journal](#).

About GMG ArcData LLC

Since its inception in 2019, GMG ArcData, an SBA certified service disabled veteran and minority-owned small, disadvantaged business (SDVOSB), helps government, commercial, and not-for-profit clients accomplish their goals by providing innovative solutions and evidence-based advisory services through applying software engineering, data analytics, first principles-based problem-solving, and user-centered design principles to enhance decision-making, performance, and organizational situational awareness. GMG ArcData is led by hands-on principals with MD, PhD, and PMP and over 65 years of combined experience in the private and public sectors and the military. The company is HIPAA and 42 CFR compliant.