IDC

SPOTLIGHT

ZS Associates

**Knowledge graphs are flexible and scalable platforms capable of exploring a multitude of relationships across millions of data points, providing a deep lens into systems biology, examining intricate drug-target interactions, and fueling innovation.**

# Resolving the Data Conundrum: Leveraging Knowledge Graphs to Power a Multi-Omics-Led Precision Medicine Strategy

*December 2022*

**Written by:** Dr. Nimita Limaye, Research Vice President, Life Science R&D Strategy and Technology

## Garnering Intelligence from Real-World Data

IDC estimated that on average approximately 270GB of healthcare and life sciences data was created for every person in the world in 2020. Healthcare and life sciences represented 7.3% (2.1ZB) of the total Global DataSphere of 29.1ZB in 2020. "Omics" data (e.g., genomics, proteomics, metabolomics, transcriptomics, and epigenomics data) is one of the fastest-growing segments in life sciences.

The high velocity at which data is being generated, the multidimensionality of data, the lack of standardization, and the inability to share data seamlessly across databases not only are adding to costs, fueling complexity, and scaling cybersecurity risks but also are resulting in frustrated end users (biologists and clinicians) who are spending more time grappling with accessing critical data and threading it together instead of developing valuable insights from it. Intelligence is achieved only when data is contextualized. Knowledge graphs can provide an integrated view of structured and unstructured data, linking it with key business and scientific concepts.

The industry is moving away from blockbuster medicines that serve the masses and toward precision medicine, developing therapies that are customized to serve special cohorts of patients. These therapies deliver better clinical outcomes and minimize adverse effects, but there is an urgent need to be able to map the data, which is all over the place. Life sciences organizations must connect the dots between a multitude of bioentities sitting out there to derive meaningful insights from them. Connecting these dots becomes critically important when researching disease biology and targeted therapeutics. One needs to understand the pathophysiology of a disease and how a potential molecule could modulate the actions of the target to produce the desired biological and clinical outcomes. Depending on the specific case, one needs to build out multidimensional constructs, transitioning from static models to dynamic in silico models that can examine the dynamics of disease phenotype and play out different scenarios.

## AT A GLANCE

### KEY STATS

- IDC predicts that the market for intelligent knowledge discovery software will continue to grow and will reach $11.1 billion by 2025 at a CAGR of 26.5% (source: *Worldwide Intelligent Knowledge Discovery Software Forecast, 2022–2025*).
- IDC refers to intelligent knowledge discovery software as software that uses modern methods and technologies such as deep learning-based NLP, document AI, semantic vector search, and knowledge graphs, combined with taxonomies and ontologies, to make it easier to discover contextual, accurate, and relevant knowledge.

### WHAT'S IMPORTANT

Knowledge graphs will weave a fine web, contextualizing millions of data points and threading together valuable pieces of information to derive powerful insights to scale discovery.

Knowledge graphs do just that. They play a key role in connecting the dots and establishing relationships between data from different sources, providing traceability, and they map the data, building semantic or functional relationships between data assets. The semantic layer provides real-time access to the most up-to-date version of the data. A knowledge graph provides a bird's-eye view of the data in the semantic layer and can dig deep as well. Data semantics is the "meaning" of entities and their important relationships. These meanings and relationships are established in a knowledge graph. Knowledge graphs link disparate and complex data using semantic data integration, based on ontology matching. The ontology models represent data with unambiguous, shared meaning in a human-readable, machine-understandable, and interoperable format.

A data catalog can help life sciences organizations understand where the data resides, minimizing the dependency on reshaping the data. A data fabric approach that creates an opportunity for life sciences organizations to access and learn from the vast and diverse data that they have accumulated over the years will be the game changer, fueling innovation and empowering end users.

## Definitions

- **Knowledge graphs:** A knowledge graph is a knowledge base, or a semantic network, that uses a graph-structured data model or topology to integrate data. The knowledge graph provides a graphical visualization of the relationships between these entities, namely objects, events, situations, or concepts. It includes nodes, edges, and labels. An edge defines the relationship between the nodes (or the objects). The knowledge graph uses natural language processing (NLP) to create a comprehensive view of nodes, edges, and labels through a process called semantic enrichment. This process allows the knowledge graph to understand the relationships between objects. This working knowledge is applied to other data sets, where similar objects reside. Upon completion, the knowledge graph can be queried to gain comprehensive insights.
- **Ontology:** An ontology defines a common vocabulary for researchers who need to share information in a domain. It is a model for organizing structured and unstructured information using entities and properties and the relationships between them.
- **FAIR data:** FAIR data is data based on the principles of findability, accessibility, interoperability, and reusability, with the goal of optimizing the reuse of data.

## Knowledge Graphs Are Showing the Way

Knowledge graphs are enabling life sciences organizations to understand disease biology faster and improve target and/or biomarker identification and prioritization in the following ways:

- Knowledge graphs can drive target and hit discovery as well as lead identification and optimization, thereby shortening drug discovery cycle times and enabling faster time to market.
- Attrition of drug candidates can be reduced by utilizing knowledge graph constructs that incorporate portfolio-level data input that leverages institutional learnings from both successful and failed programs and assets.
- Biobanks are dealing with big data, bringing together molecular and deep phenotypic data of hundreds of thousands of individuals to develop a holistic patient profile. Various biobanks are amassing huge volumes of data, including UK Biobank; FinnGen (Finland), and the National Center Biobank Network (Japan).

- Various large-scale genome sequencing initiatives have been implemented across the globe, including the GenomeAsia 100K initiative, the National Institute of Health's All of Us initiative to sequence the genomes of 1 million people in the United States, and the Three Million African Genomes (3MAG) project. Pharmas are also going down this path. AstraZeneca, for example, is in the process of sequencing 2 million genomes in the hunt for rare genetic sequences associated with disease.

As the volume of data continues to explode, knowledge graphs that can weave the data together to generate sophisticated insights will serve as game changers for developing a deep understanding not only of disease biology, at a precision medicine level, but also of defining factors impacting population health.

## Considerations

Biologists don't think in terms of graphs; they think in terms of pathways, systems, cells, and tissues. It becomes challenging for them to extract concepts and relationships from a knowledge graph. Unless the knowledge graph has been designed with a deep understanding of the domain and the needs of the customer, it often overwhelms the user with information. Complementing the knowledge graph with analytics and visualizations that enable biologists and clinicians to explore the knowledge using familiar constructs is critical to successful adoption and impact.

Most knowledge graphs currently address only high-level concepts and relationships (such as protein, gene, pathway, disease). However, insights that are valuable to scientists require the curation of concepts and relationships to at least two more levels of granularity (such as transcript, protein expression, protein complex).

Knowledge graphs can drive semantic interoperability, but data in the foundational layer should be standardized and meet FAIR data principles, the metadata and labeling should be complete, and the provenance of the data should be established.

The adoption of knowledge graphs requires a cultural change in the organization. The leadership must recognize and champion this cause and move beyond old ways of working. In addition, the strategy for designing the knowledge graph needs to be flipped on its ear. Rather than designing the knowledge graph based on high-level concepts and relationships (protein, gene, pathway, disease), one needs to first define the problems that need to be solved, model the biological and other scientific knowledge required to solve those problems, and then build the graph at that granularity of knowledge with domain-informed user interfaces that are meaningful to the scientists.

The transition to the use of knowledge graphs requires the development of new talent — biologists with an understanding of data sciences and data scientists who understand the biology. In this era of transformation and innovation, the molecular natives shall lead the way. Commercial strategies need to evolve to adapt to new models that focus on narrow cohorts of patient populations, changing the financial dynamics.

While traditional pharma drug discovery has been driven by bench scientists who work on identifying a hypothesis and then have their bioinformatics team validate it, this process needs to change; de novo discovery needs to be driven by in silico modeling, slashing costs and time.

System biologists need to learn a new language, a new vocabulary, and a new set of haptics. Organizations need to arrive at an alignment between what scientists want to see and what data scientists want to see.

## Considering ZS Associates

ZS Associates is a management consulting and professional services firm focusing on consulting, software, and technology. It provides services for clients in the healthcare, private equity, and technology industries.

Founded in 1983, ZS is headquartered in Evanston, Illinois, with offices in the Asia/Pacific, Europe, and Latin America regions. It has over 12,000 employees dedicated to life sciences. The ability to work at the intersection of technology, data science, and business serves as the firm's differentiator. ZS has over 400 certified data scientists, big data and cloud technologists, and clinical and real-world data practitioners; over 100 MDs and life sciences PhDs; and R&D experience in over 120 disease areas.

In 2019, ZS established a Biomedical Research practice, and in May 2022, the company acquired Intomics, a bioinformatics and systems biology company founded in Copenhagen, Denmark, in 2008. The combined organizations now accelerate and optimize pharmaceutical drug discovery and development by enabling complex analysis of biomedical data, strengthening ZS's focus on the discovery of new medicines. The team utilizes a highly curated protein-protein interaction network that supports tailored biomedical analysis spanning multiple diseases. At ZS, scientists and bioinformaticians use the network to understand disease biology at the systems biology and molecular levels.

The ZS team also builds customized knowledge graphs to solve bespoke scientific problems and "FAIRify" enterprise scientific data assets. This blend of deep data, academic and scientific expertise, technology and artificial intelligence, and expertise in designing "fit-for-purpose" knowledge graphs will accelerate drug discovery and fuel innovation.

### Challenges

ZS Associates faces the following market challenges:

- **Developing prediction models:** Building out accurate prediction models is a complex and challenging task in the highly regulated industry of life sciences.
- **Adopting knowledge graphs:** The life sciences industry is still recognizing the worth of knowledge graphs. Though this concept has been around for some time, when the implementation is more tactical, rather than coupled with deep scientific expertise and informed by specific scientific questions and explorations, it does not necessarily yield the desired results. This has created a lack of appreciation of the true value of knowledge graphs.
- **Designing knowledge graphs that address the specific use case and deliver value:** Designing a knowledge graph is not just about uploading curated data — that does not necessarily solve the problem. Knowledge graphs must have a "fit-for-purpose" design such that the data relationships that have been mapped are meaningful and the end user can find relevant information quickly and can envision relationships between data points that provide pointed and meaningful insights.
- **Defining the right use cases:** There is a need to work with customers to help them define the right use cases and then find the right data sources. The life sciences industry often struggles with identifying the right use cases, finding the right data sources to match these use cases, and integrating machine learning (ML) models into workflows.
- **Driving data interoperability:** Biomedical data usually resides in multiple databases, may exist in different sizes and shapes, and may use different ontologies. The data needs to be transformed, and ontologies need to be mapped to drive semantic interoperability.

## Conclusion

Knowledge graphs can capture and democratize institutional scientific knowledge to differentiate research organizations based on their ability to access and exploit their collective knowledge. Knowledge graphs provide a framework for building actionable insights. Unlike data lakes, which pool vast amounts of data, knowledge graphs envision relationships between data points, powering evidence-based precision medicine and fueling drug repurposing.

Knowledge graphs can also be used to contextualize social determinants of health (SDoH) data points to drive diversity and inclusion in clinical trials, and they can track evolving clinical outcome data points across quality-of-life and biomarker data to build value-based pricing models for the therapies developed. Knowledge graphs offer a creative knowledge landscape that scientists can leverage to build biological models, understand correlations, and drive decisions.

A knowledge graph-enabled data fabric can provide that magic carpet that scales innovation and elevates the user experience for R&D scientists. A "fit-for-purpose" knowledge graph is built on the foundations of deep domain expertise and powerful technology.

The increasing shift toward a multi-omics strategy, deep phenotyping, and integrating multiple sources of data to derive a truly holistic view of a patient's health profile is only going to accelerate the adoption of knowledge graphs. If ZS Associates can effectively address the challenges described in this paper, then by leveraging its knowledge graph capability and its "omics" and clinical expertise, the company can open up significant opportunities for innovation for the life sciences industry.

# About the Analyst

***Dr. Nimita Limaye,*** *Research Vice President, Life Science R&D Strategy and Technology*

Dr. Nimita Limaye is a Research Vice President with IDC Health Insights and provides research-based advisory and consulting services as well as market analysis on key topics related to R&D strategy and technology in the life sciences industry. She addresses aspects such as the role of digital transformation in discovery research and eclinical ecosystems.

## MESSAGE FROM THE SPONSOR

### More About ZS Associates

ZS's Biomedical Research team helps research organizations in the life sciences industry accelerate and improve their capabilities for in-silico research to create breakthrough science, accelerate discovery and improve the probability of success. By combining a talent pool of molecular natives with deep capabilities in omics, network biology, data science and semantic technologies, and by exploring the biological and chemical whitespace that others ignore, ZS helps discover novel targets, biomarkers and therapeutics. ZS's science first approach ensures that problems and questions are framed properly at the start to ensure answers and results that deliver the desired impact. ZS offers a unique perspective on how to bring business and patient impact through the science of R&D. ZS professionals work with business executives to align their R&D vision, benchmark capabilities, develop their R&D strategy, design organizational capabilities and lead strategic initiatives. Click here to know how ZS delivers transformative outcomes for patients through a path led by data, analytics, technology and end-to-end expertise.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

**IDC Research, Inc.**
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com