Paper ET22

# The Potential Power of Data

Eva Kelty, Capish, Malmö, Sweden
Catharina Dahlbo, Capish, Malmö, Sweden

## ABSTRACT

Drug development is all about data. All clinical trial activities are geared towards generating data to accept/decline hypotheses, to support claims and to generate new knowledge and ideas.

Data needs to be available and understandable for the people consuming it. Utilizing a graph database with an intuitive interface makes it possible to structure data in a way that makes it easily available for non-programmers. Combined with relevant metadata data the usability increases.

This presentation will focus on how data can be organized and utilized in clinical trials, e.g., in the planning, execution, reporting and archiving. A repository containing clinical trial data provides opportunities to learn from past experiences, while planning ahead. Access to data as a clinical trial is ongoing could improve quality, speed up the process and support collaboration. During the reporting period data can be easily explored to understand outliers or subgroups.

## INTRODUCTION

Traditionally, drug discovery and development has been uncertain and costly. Today, data analytics has made it possible to change the process, speeding it up and reducing cost. Predictive modeling can identify promising candidates based on historical data, chemical structures, and potential interactions. However, vast amounts of data is generated every day.

Data is there, full of previously untapped potential, unexplored connections, and relevant insights. Data can provide information on the success rate of similar drugs in a particular indication, such as lung cancer. It can also provide insights useful to accelerate patient enrollment and identify previously successful sites for selected indications. Insights that would lead to the identification of unmet needs would help in better pipeline prioritization.

All one needs is to know how to gain access and use it efficiently. However, mastering this process might prove more challenging than it appears at first glance.

## GRAPH DATABASE

A property graph database can internally structure its components by incorporating properties, types, and labels into nodes and relations. These internal components, also known as attributes or fields, consist of numerous key-value pairs [1]. Nodes, when labeled, can serve to identify them as specific types of information units. The ability to represent information as properties within nodes contributes to the compactness of property graphs. An example of a property graph is given in Figure 1.

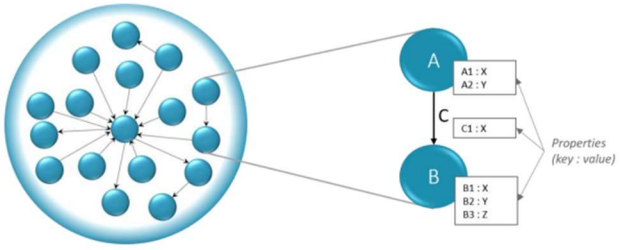


Figure 1. Simple representation of two nodes in a labelled and directed property graph.

The modeling of nodes and their relationships in a property graph is highly dependent on the database's intended purpose. In a graph database, nodes and relationships hold equal significance, but the chosen modeling approach may prioritize either nodes or relationships. The decision on which model to adopt relies on the specific questions one aims to address. It also hinges on whether the goal is to analyze the graph comprehensively for analytical purposes or if the focus is primarily on local transactions [2].

For the storage of extensive data, like those arising from clinical trials, property graph databases outshine other graph database options. This is mainly because of the faster and easier ways to access and query the data. In Figure 2 the structure and relations in a relational database vs a property graph database is illustrated.


Figure 2. Relational vs property graph database

## REPOSITORY

### FAIR PRINCIPLES – A FOUNDATION

The 'FAIR Guiding Principles for scientific data management and stewardship,' introduced in 2016, aim to enhance the Findability, Accessibility, Interoperability, and Reusability of digital assets [3]. A FAIR repository adheres to these principles, promoting the usability and impact of research data.

| | |
|---|---|
| **Findability** | Data is easily discoverable by humans and computers through efficient search and retrieval facilitated by metadata and unique identifiers. |
| **Accessibility** | Data and metadata are accessible for different users under clearly defined conditions. |
| **Interoperability** | Data and metadata are structured using standards, enabling integration with other datasets. |
| **Reusability** | Well-described data with rich metadata and provenance allows effective reuse by both humans and machines. |

FAIR repositories can be important for advancing data-driven research by ensuring that data is not only stored correctly but also maximally usable and shareable. While the principles above provide a foundation, achieving data reusability requires a practical approach and solution.

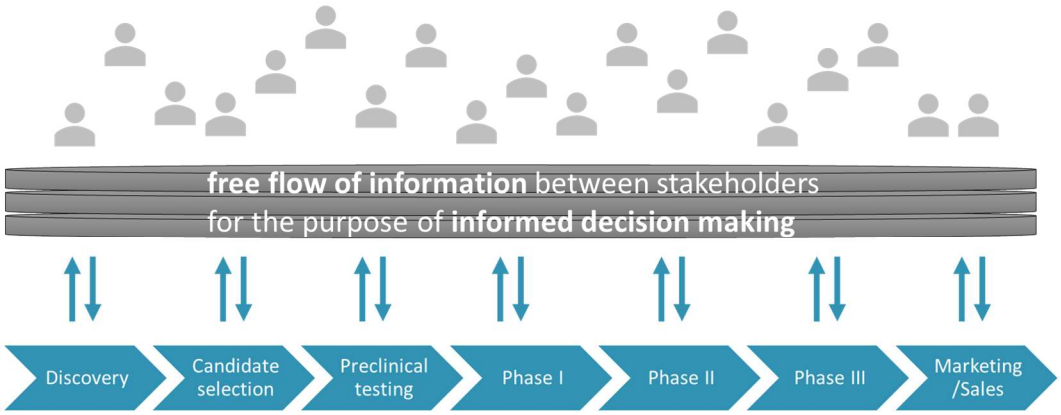


Figure 3. Information flow throughout the drug development process [4].

**METADATA MAKES A DIFFERENCE**

The repository's efficiency in data access is maximized through thoughtful organization and modeling, particularly when integrating metadata with the data. Metadata provides essential information for understanding the data, for example how it was collected and measured, by whom, where and when. Adopting an information model structured as a graph with a focus on nodes, here called Holons, opens new possibilities compared to a conventional relational database.

**ONTOLOGY**

An ontology serves as a structured representation of a knowledge domain, outlining entities (such as events and processes) and their interconnected relationships. Frequently used to give meaning and understanding to graph data. A standalone graph database lacks the inherent ability to provide meaning to its components. Instead, meaning must be explicitly provided, usually using an ontology that outlines the components of the graph.

| What? | How? | Why? |
|---|---|---|
| Describing data | Terminology | Standardization and consistency |
| Enhancing data | Metadata | Understanding data |
| Provide structure | Information model | Efficient data use |

Figure 4. What, how and why an Ontology contributes to a data repository.

An ontology comprises both of an information model, specifying how the graph's nodes and relations should be structured, and a terminology that defines all data components. Grounded in a network of information units known as "Holons" and their relationships, this model facilitates the storage of closely related data within the Holon. For instance, variables linked to an individual event, process, or entity, such as a measurement result, its unit, and the date it was measured, are stored together [5].

## KNOWLEDGE GRAPH

A knowledge graph created by combining a property graph database with an ontology and data collected in clinical trials, can be a very powerful tool providing data access for searches, exploration, and visualization.

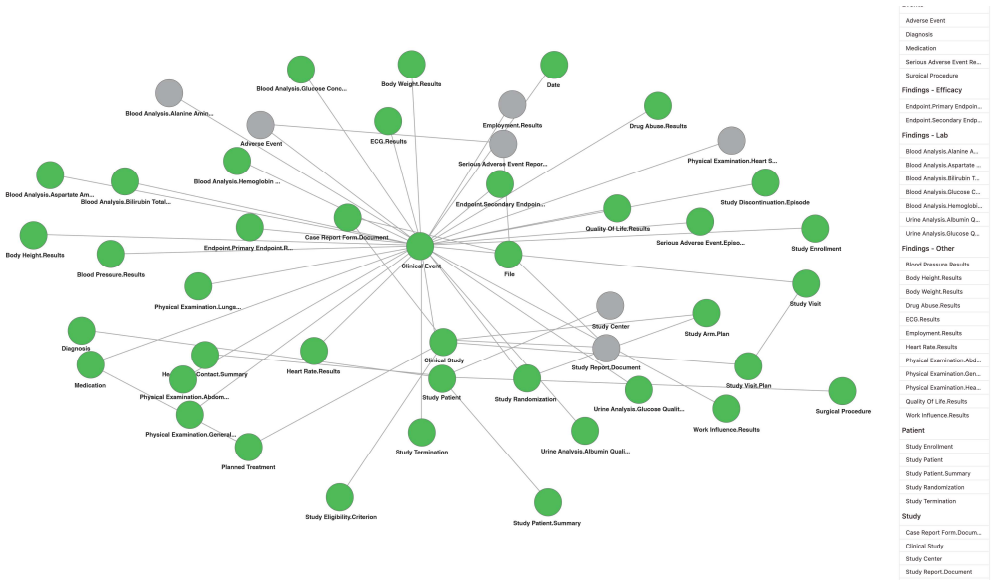


Figure 5. A knowledge graph containing nodes and relations.

## CLINICAL TRIALS

When working with clinical trials, access to information and/or data can contribute significantly to efficiencies and decision-making. A multi-trial repository contributes in different ways with knowledge from past experiences and a single-trial repository can aggregate data from multiple data sources in an ongoing trial.

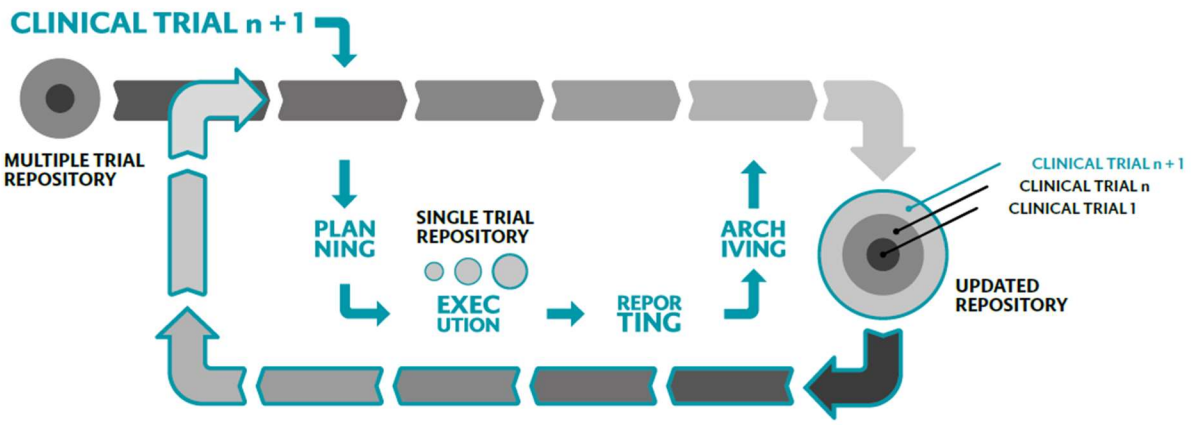


Figure 6. Clinical trial activities and data repositories.

### PLANNING

In the planning stage data, leveraging data from prior trials becomes instrumental in optimizing various aspects, including recruitment, strategies, logistics, and the selection of endpoints.

**EXECUTION**
Managing a comprehensive overview of data during an ongoing trial can be challenging, given the diverse sources and systems involved, such as eCRFs, laboratory portals, ePRO, quality of life solutions, and other devices. Much time can be saved by ensuring that the data in different systems is consistent, comprehensive, and not over-lapping. Access to data as a clinical trial is ongoing could improve quality, speed up the process and support collaboration.

**REPORTING**
As the data is aggregated in a clinical study report it may be feasible to draw on conclusions from previous trials but also to easily identify and understand outliers or subgroups.

**ARCHIVING**
Clinical trials are costly, and the generated data should be treated as the valuable asset it is. Once a clinical trial has been completed the data should be obtained from any service provider and added to a dedicated program, project, or drug repository to ensure that the knowledge gained is available for the next stage of the drug development.

## CONCLUSION

Every clinical trial is a substantial investment for companies, underscoring the critical need to safeguard access and ensure data availability. These trials hold a wealth of invaluable information beyond their financial commitment. To extract optimal value, meticulous organization is essential, facilitating versatile utilization and accessibility through advanced technology. In a well-structured and collaborative environment, data's significance goes beyond security - it becomes a dynamic force propelling transformative potential. The synergy between a systematic framework and collaboration not only strengthens data security but also unleashes its full potential as the cornerstone of innovation in clinical trials. It is within this collaborative space that data evolves from a repository to a catalyst, steering progress and innovation in the ever-evolving landscape of clinical research.

## REFERENCES


1. “RDF Triple Stores vs. Labeled Property Graphs: What’s the Difference?”, J. Barrasa, Neo4j blog.

2. “Unique Technology for the Future”, C.Dahlbo, A.Workneh, H.Drews, PP04 PHUSE EU Connect 2018

3. “The FAIR Guiding Principles for scientific data management and stewardship.” Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. Sci Data 3, 160018 (2016).

4. “Graph Data Representation – The FAIRest of Them All”, E.Kelty, P. Tormay, DS01 PHUSE EU Connect 2020

5. “Harness the power of ontologies to build better information platforms”, P.Tormay, A.Berg, DH07 PHUSE US Connect 2019


## CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:


Author Name: Eva Kelty
Company: Capish
Address: Lokgatan 8, Malmö, Sweden
Work Phone: +46 40 10 88 80
Email: eva.kelty@capish.com
Website: www.capish.com

Author Name: Catharina Dahlbo
Company: Capish
Address: Lokgatan 8, Malmö, Sweden
Work Phone: +46 40 10 88 80
Email: catharina.dahlbo@capish.com
Website: www.capish.com