CONVERSATIONAL ANALYTICS OF CLINICAL TRIAL OPERATIONS DATA

Clinical trial operations data, including site information, study protocols, adverse event reports, training and certification records, and more, is inherently complex and comprises a range of variables, including studies, phases, substances used in trials, site locations, and many more. The data may be stored in many different formats in multiple locations, including SQL databases, spreadsheets, unstructured repositories (like email systems), and third-party systems.

Traditional methods require users to exhaustively cross reference data across a myriad of dashboards or develop complex database-driven solutions. Such solutions are often opaque and getting answers usually requires tricky, time-consuming work by an analyst skilled in the domain and knowledgeable about where the data can be sourced and how to interpret it correctly, which consumes valuable time and resources.

Knowledge graphs can integrate data from multiple underlying sources and support the implementation of intuitive natural language user interfaces that understand the questions being asked and return accurate, complete answers.

Example Queries

A properly implemented conversational analytics system for clinical trial operations applications will make it easy to answer questions like this:

- How many clinical studies were initiated in the last N years?
- How many patients have been enrolled in clinical studies X, Y, and Z this year?
- Which studies have enrolled patients in Country A between Date 1 and Date 2?
- How many studies have achieved the “CTR approved” milestone in the last N years?
- What are the therapeutic areas covered by Program X?
- Who are the principal investigators for Program Y?
- How do the objectives of Program Z align with studies currently being conducted?
- What common therapeutic areas are being addressed across programs X, Y, and Z?
- Which institutions are collaborating on programs A, B, and C?
- In which countries or regions are most of our clinical studies conducted?
- Which studies has Investigator D participated in?
- Which active substance is being tested in Program Y?
- Compare the median timeframe from final protocol approval to first subject/first dose for all studies with sourcing model hybrid operating model (HOM) to studies with a different sourcing model and same study phase.
- Please summarize the differences of milestones across all active studies.

These are just a few examples of the thousands of complex queries that users can submit to a conversational analytics system. They will receive detailed, complete, and accurate responses based on all the data available to the system. The benefits in terms of operational efficiency are obvious; no one has to write any code, queries, or configure a dashboard to gather the necessary information and users need not have detailed knowledge of the source data formats and storage systems involved.

Knowledge graphs provide understandable, comprehensive organizational context to enterprise analytics solutions, especially to those employing generative AI.

Start with Enhanced Data Quality

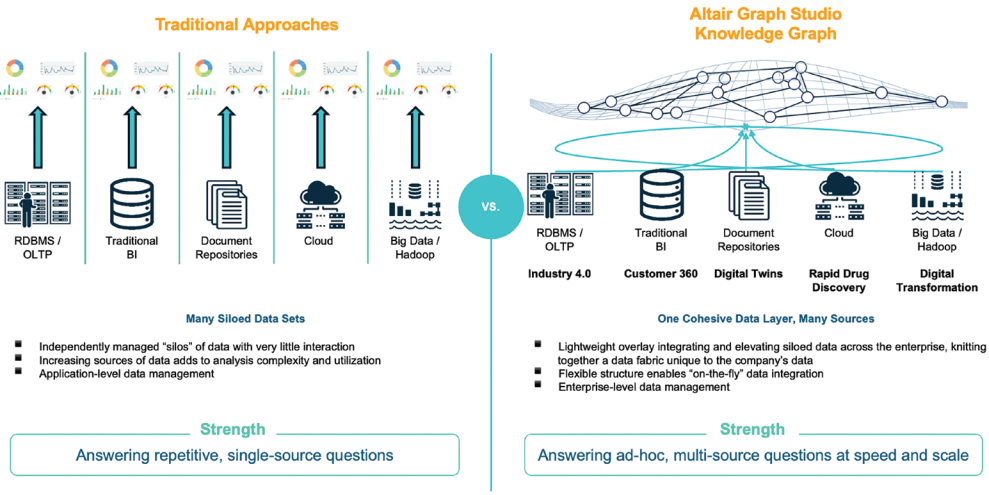
Clinical trial operations generate vast amounts of data collected from diverse sources. It is often complex and difficult to access. Users must know where all the operations data is stored and how the data relates to the questions the clinical trial managers are likely to ask.

When implementing a knowledge graph, the data engineering team uses operational data to tell a large language model (LLM) how to access the data needed to answer business users' questions. They configure the graph to represent the data using ontologies that describe the contents of the graph using the domain's natural language. High-quality data cleansing is essential to the success running of any large-scale clinical trial and allows trial administrators to trust the integrity of their reports and analyses.

There are numerous products on the market to automate data cleansing processes, including specialized software like Altair® Monarch®, part of the Altair® RapidMiner® data analytics and AI platform. Knowledge graphs, including Altair® Graph Studio™, can also cleanse data and, given their additional utility in improving the quality of responses from LLMs, are excellent fits with the requirements of large clinical trial analysis projects. Pharmaceutical data management teams can use Graph Studio to build knowledge graphs that provide extensive data profiling metadata for structured, semi-structured, and unstructured data sources and enhance the quality of that data on the fly.

Creating a Knowledge Graph

Clinical trial operations data is often siloed, with data stored in relational databases, text files, and other documents. Users must employ multiple technologies and approaches, including dashboards, spreadsheets, and email to locate and query the data. Questions that require information from multiple data silos are difficult and time-consuming to answer, and subject to incompleteness errors.



Traditional approaches to building a data fabric do a good job of answering repetitive, single-source questions. However, in clinical trial applications, this capability is rarely adequate. A data fabric built using a knowledge graph enables users to ask questions that require access to and correlation between many different data sources. Further, knowledge graph-based data fabrics respond to queries quickly and at scale.

Seamless access, sharing, and governance of data is extremely valuable in the context of clinical trial analysis, especially for large-scale trials, and a much better approach uses a knowledge graph. The knowledge graph provides a unified view of all trial operations data and enhances data accessibility, quality, and security:

- **Enhance data accessibility**: Knowledge graphs break down data silos and ensure that all users have immediate access to relevant data.
- **Improve data quality and consistency**: Their integrated data governance and metadata management capabilities enable knowledge graphs to maintain high levels of data quality and consistency.
- **Increase agility**: Knowledge graphs allow administrators to react quickly as new logistical challenges arise during trials and easily assemble overviews of current and past trials.
- **Reduce costs**: By streamlining data management processes and reducing the need for multiple data integration tools, knowledge graphs reduce the time and effort required to manage data and free up administrators to focus on analysis.
- **Enhance security and compliance**: The robust governance and security measures incorporated into knowledge graphs ensures the protection of intellectual property and sensitive medical data, including personally identifiable information.
- **Shorter time-to-insight**: By providing unified views of all accessible data, knowledge graphs enable comprehensive, accurate, and efficient analysis and immediate answers — even to questions users may not have gone to the trouble of asking using traditional methods.

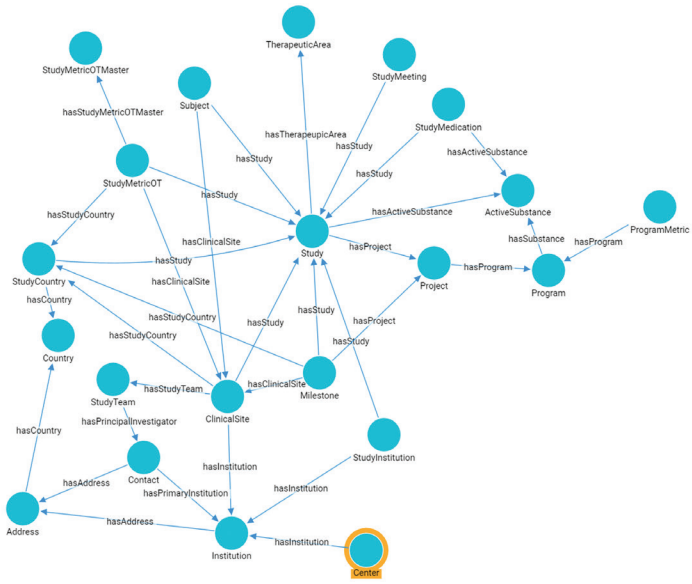
Knowledge graphs build cohesive semantic layers that describe and connect all relevant data sources and make them accessible through simple user interfaces. Further, building a knowledge graph is a great approach for supporting conversational, interactive analytics.

Knowledge graphs form flexible, transparent layers sitting on top of data stacks. They span all different types of data sources, including data lakes, relational databases, document repositories in the cloud, and so on. They create a single layer that stitches together structured and unstructured content into a semantic layer that describes the relationships between concepts contained within the data. The implementation of knowledge graphs allows users to query across all data sources at speed and at scale.

Building Context: Ontologies, Graph RAG and Competency Questions

Generative AI (genAI) models rely on context to produce hallucination-free responses to queries, in the context of text-to-graph queries that context includes the ontology, which is a description of the available data in the domain's natural language.

Put simply, an ontology is a structured framework that enables LLMs to understand and generate accurate queries by navigating the relationships between concepts and entities as well as their attributes within a specific set of data. A well-built ontology is not static and does not have to be re-structured as new use cases arise; it will grow and adapt to accommodate unforeseen queries.



The ontology for a clinical trial operations model includes data about the study itself, clinical sites, subjects, and milestones. Graph Studio automatically optimizes the ontology and prompts to improve end user navigation of the knowledge graph through natural language queries, enabling users to receive accurate responses confined by the scope of the ontology.

LLMs have made interactive natural language processing (NLP) user interfaces ("chatbots") more useful and effective than ever. However, on their own, LLMs generate responses based on probabilities of patterns in the data they were trained on; they have no real comprehension of questions, answers, or the data itself. This means these models can — and do — generate incorrect or nonsensical information; they can "hallucinate." These types of error-filled responses can occur when an LLM lacks sufficient additional factual context needed to answer the questions being asked. In addition, LLMs could not have been trained on current clinical trial operations data. These factors prohibit their use in clinical trial operations, especially since many answers must be computed in real time from a variety of confidential data sources.

Limitations to RAG and Graph RAG Approaches

Retrieval-augmented generation (RAG) is a valuable technique that limits a genAI model's frame of reference to real and vetted information. Graph RAG exploits the contextual information from knowledge graphs to reduce hallucinations. It adds context to prompts that increase the accuracy of the generated response and reduces the likelihood of hallucinations. Graph RAG enables the model to find all the relevant bits of information available to the system, capture multi-step relationships that standard RAG can easily miss, and produce more useful insights — while limiting query responses to information grounded in facts.

Incorporating knowledge graphs as part of the stack provides improved, more informed, and contextually relevant responses, since graph RAG utilizes knowledge graphs to capture and represent relationships between data points and conceptual entities. However, neither vector embedding-based RAG nor graph RAG are sufficient to overcome the issues associated with answering clinical trial operations-related questions, which must be computed in real time.

An Improved Method: Graph Query Generation

Altair has developed a new technique that overcomes the limitations of RAG and graph RAG approaches for applications like this. We integrate the data into a knowledge graph by loading, cleaning, interlinking, and describing it using an ontology. The ontology describes the contents of the knowledge graph in the domain-specific language of the system's users. Data engineers — not data scientists — normally complete this work.

Once we have a working knowledge graph, we use the ontology describing what the knowledge graph contains as context to aid the accurate creation of queries by an LLM from user questions. The user enters questions in a conversational chat client.

The ontology conveys to the LLM the contents and structure of the previously constructed knowledge graph. It provides information that allows the LLM to create graph queries that when executed against the knowledge graph provide the answer to the users question as a result. The chat client displays the result using a visual card chosen by the LLM. Card types include written text, tables, and data visualizations.

The ontology informs and constrains the LLM, allowing it to create accurate queries instead of hallucinated query elements or entirely hallucinated answers that other approaches may generate.

Use Altair Copilot to Support Ad Hoc Querying and Streamlined Analysis

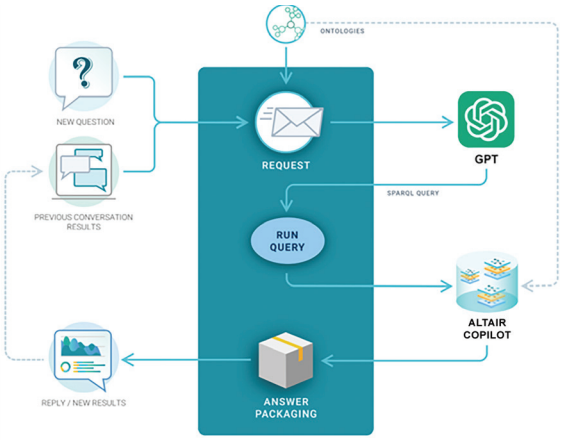
Altair® CoPilot™, a natural language query (NLQ) user interface, allows users to ask questions using everyday language to interact with all the data available to the system. This removes barriers to data access and makes analysis of clinical trial operations data much more efficient and flexible than traditional methods.

In this type of implementation, a knowledge graph and LLM work together to understand what the user is asking and quickly produce accurate, relevant answers, visualizations, and even suggestions. Users can explore data easily, ask follow-up questions, and refine queries in seconds. They can make fully informed, data-driven decisions about next steps quickly and effectively.

The responses from an Altair CoPilot implementation will be accurate, understandable, and traceable since they are simple query responses to data stored in a database. This means the system cannot produce hallucinatory responses. If a user asks a question that requires understanding of a concept not included in the ontology, Altair CoPilot will respond that the answer is not contained within the data; it will not guess or make something up. For questions that can be answered by the data, users will receive precise, accurate, and complete responses.

A properly implemented system does not send the actual data, including confidential clinical trial data and personally identifiable information, to ChatGPT or any other external system. All data remains in place and under the control of the organization. The system sends only an ontology describing the data model along with the query to the external system.

The system acts like a participant in a conversation between two people. One person asks a question, gets a response, and then asks a follow up question; the responder understands that the follow up question is based on the information in the initial response. An example: A user asks how many clinical trials utilize a particular test site, and then asks how many other sites operate under the same set of conditions.



In Graph Studio-based systems, user queries are fed to the knowledge graph on the right through Altair CoPilot, to the organization's GPT endpoint, which generates a SPARQL query based on the question asked. The system executes the SPARQL query in Altair Graph Lakehouse against the knowledge graph. It then shows those results in clear, understandable visualizations and text.

Optimize Clinical Trial Operations

Building an interactive interface like this allows anyone within the organization, even people without extensive domain knowledge or technical skills, to ask questions about the clinical trial operations and receive informative, accurate responses. Pharmaceutical research teams therefore spend no time writing code for complex queries and can concentrate on understanding and analyzing the data.

With this type of system, users can query enormous amount of clinical trial operations data using ordinary language — and they can use the human language (German, English, Japanese, and so on) they prefer — to access their data. Rather than waiting days or weeks for a report that then requires follow-up queries, they get answers in a format they can use immediately in presentations, emails, meetings, and more.

The Next Five Years

Once a system is in place, the applications go well beyond queries of trial operations data. Drug development teams may use the same system to draft new clinical trial protocols, generate reports to government bodies, prepare requirements documentation, write programming code, and more.