PhUSE US Connect 2019

Paper PP01

Make the Most of Your Data – Explore Different Perspectives

Catharina Dahlbo, Capish, Malmö, Sweden
Michael Rubison, Capish, USA

ABSTRACT

In order to make information stored in datasets more accessible a new ontology-based information model, structured as a graph, has been developed. One benefit with the new approach is that it allows implementation of database indices. This makes it possible to search the data via indirect relationships and thereby get a broader result set as well as controlling the vantage point or perspective from which the data is evaluated. Since a dataset can be evaluated from different positions e.g. the patient, the study or the site, answers can be obtained for many different questions. For example, questions about site can be explored as well as questions relating to the patient.

This poster will present how it is possible to obtain the response to advanced questions from different perspectives within a single dataset and how it can be used in practice.

INTRODUCTION

Pharma companies are investing enormous amounts of time and money in clinical trials and as a result they have collected huge amounts of clinical data. Data drives the success of Pharma initiatives, but that success depends not only on the amount and quality of the data, but also how prepared it is for reuse. Data is often locked into different silos and much work is needed for integrating data so that accessibility and usability reaches a level that allows reuse to its full potential [1]. When it takes an unnecessarily long time to explore and analyze data, innovations and findings that may impact patients' lives are delayed [2].

Challenges for getting data in a reusable format include:

- Vast amounts of data lead to information overload.
- Combining data sets for cross-study analysis.
- Data structure that restricts further analysis and reuse.
- Incorporation of real-world data from disparate and nonstandard sources.

While there are excellent systems for analyzing different types of data in isolation, real value can be gained by having curated and integrated data in a unified knowledge base [3]. Furthermore, having the data in an unbiased form that could be used for analysis, not intended at the time of collection, could give unexpected opportunities [4]. Combining different datasets with real-world data increases the usage further.

In order to obtain the most from the data, it is important to collect, curate and integrate data in an unbiased form, while maintaining the context in which the data was collected. Involved parties in a trial are interested in different aspects and perspectives of the information. Therefore, the industry should explore methods utilizing the data in the most flexible, scalable and sustainable way. An unbiased data repository enables instant access and reuse of all available data without time consuming data extraction, transformation and load procedures.

PIONEERING TECHNOLOGY

In recent times, NoSQL (not only SQL) solutions and especially *graph databases* have become more common when it comes to managing large amounts of data.

As a group, graph databases provide a clear benefit over relational databases for the following reasons [4]:

- The conceptual data model translates directly into the graph database model.
- Graph databases are flexible and easily expandable.
- Metadata can be stored directly as part of the data.
- Data can be evaluated in different contexts.

All graph databases have one thing in common; they consist of nodes connected by edges. However, how the nodes and their edges are modeled and stored depends very much on the purpose of the database. In the clinical world much emphasis has been on RDF (Resource Description Framework) and triple store databases where every triplet consists of subject, predicate and object (e.g. Alan − has age – 38). It is obvious that this extremely general approach can house most data models, but at the expense of producing an enormous number of triplets, complex query language, and slow

performance.

Another more practical approach is the use of *property graphs* in which the nodes are bigger and contains much more content than a triplet in a triple store. Capish has advanced the property graph database and developed an approach to index familiar medical concepts like “Patient” or “Medication” as nodes and named relations as edges. Every specified concept serves as a building block in the resulting information graph. We call this standardized information carrier a Holon and the metadata that describes the Holons and their relations an Ontology. The strength of this approach is that the search engine only needs to keep track of a finite amount of Holons and their relations which enables the extremely high search speeds necessary for presenting an interactive user-friendly interface, while still leaving lots of information at the fingertips of the users.

The prerequisites for obtaining an information graph that can be effectively processed are outlined below [5]:

1. Develop an **ontology** that describes the data: The Capish ontology consists of an *information model,* defining how the nodes and relations of the graph should be modeled, as well as a *terminology,* describing and defining all the components of the data.
2. Establish certain **modeling principles**: Modeling principles are necessary to guarantee the consistency of the model, both in different contexts and over time. A consistent way of modeling is crucial when integrating data from different domains and when there is a need to extend an existing graph with new data. The modeling principle is a necessity for the indexing process.
3. Implement several **indices**: The fast response depends on the creation of several indices [6], represented by:
    - **Identity index** – used to identify a Holon in the database
    - **Relation index** – used to identify all Holons related to a specific Holon (e.g. a Patient) in a single step
    - **Holon definition identity index** – used to identify the Holon definition (Holon Type), thereby finding which Holons of a specific Holon Type that are present in the database

The indexing procedure facilitates analytical processing by avoiding costly node traversals. They also provide a more intuitive data querying process, since the user does not need to know the chain of relations. In particular, the relation index provides a new way to efficiently search using indirect relationships, or search for specific criteria and retrieve a broader result set.

LOOK AT DATA FROM DIFFERENT PERSPECTIVES

The implementation of various indices improves the way data can be queried and makes the data querying process faster and more intuitive for the end user. Reflective Logic is a new method to retrieve responses to queries formulated about a specific Holon Type (Reflection Point). Reflective logic [7] is a two-step method used to retrieve responses to queries by:

1. Finding all Holons of a defined Reflection Point
2. Reversing the search by finding all Holons related to the previously found Holons of the Reflection Point.

Setting a Reflection Point in, for instance, the ‘Patient’ Holon will identify all data connected to a particular patient, in one single step. The possibility to add a Reflection Point enables users to explore the database from different perspectives and easily create data subsets (cohorts) depending on what issues/aspects the user is interested in.

Through a tool, optimized for searching in the property graph database, it is possible for the end user to ask complex questions directly to the database and instantly get results. An example of the flexibility of the solution is that the data can be explored in parallel from different perspectives (Reflection Points).

USE CASE

Consider ten clinical trials with the Investigational Drug A, produced in three different batches. All trials are integrated into the same database. A serious adverse event, renal failure (in the example below referred to as AE2) has been reported for several patients.

By having data already curated and modeled regardless of the question and utilizing different Reflection Points users can answer many different questions in a few clicks.

ANALYSIS FROM PATIENT PERSPECTIVE

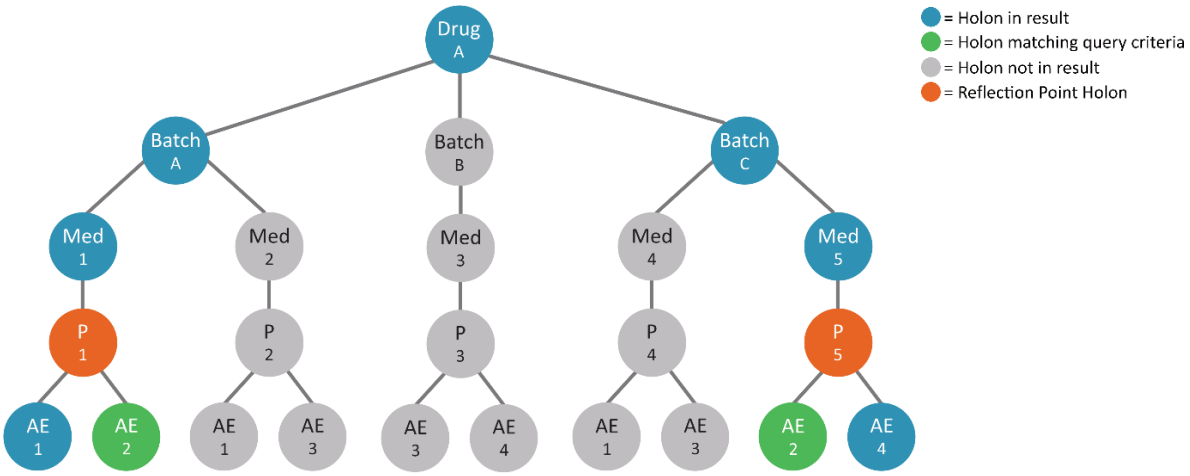

Figure 1. The AE is reflected in Patient (P). All the AEs and all other data related to every patient having experienced renal failure (AE2) is retrieved.

Figure 1 shows a simplified view of a graph to demonstrate the effect on the result set when reflecting the query in Patient. This approach will give the user an overall picture of patients who have experienced renal failure (AE2). All data that has been reported for these patients is included in the result set. The figure shows only AE and medication for patients, but all data connected to the patient such as laboratory values, treatment arm, patient reported outcomes etc. will be retrieved. Subgroups of patients matching specific query criteria can easily be created and compared to each other, or to the whole population.

Examples of questions that can be answered quickly with a reflection in Patient are:

- What other AEs have been reported for patients who have experienced renal failure?
- What do these patients have in common (e.g. batch, site, concomitant medications)?
- Is there any difference in the incidence of renal failure between the different trials?

ANALYSIS FROM BATCH PERSPECTIVE

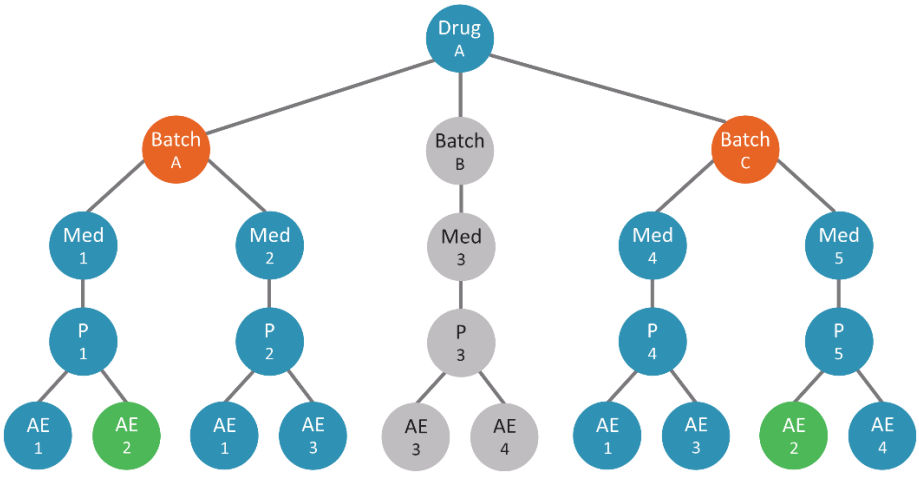

Figure 2. The AE is reflected in Batch. All the AEs and all other data related to batches where renal failure (AE2) has been reported is retrieved.

Figure 2 shows the same simplified graph as in figure 1, but the result set differs since the query is reflected in Batch. This approach will give the user an overall picture for batches where renal failure (AE2) has been reported. All data reported for these batches is retrieved in the result set including patients that had not experienced renal failure (AE2) because they have the batch in common with the patients that did. Additional information such as where the batches are manufactured can also be retrieved if available and related to Batch.

Examples of questions that can be answered quickly with a reflection in Batch are:

● What other AEs have been reported for batches where renal failure has been reported?

● Do patients, given the same batch as those with renal failure, have any evidence of elevated laboratory values that may lead to renal failure?

ANALYSIS FROM DRUG PERSPECTIVE

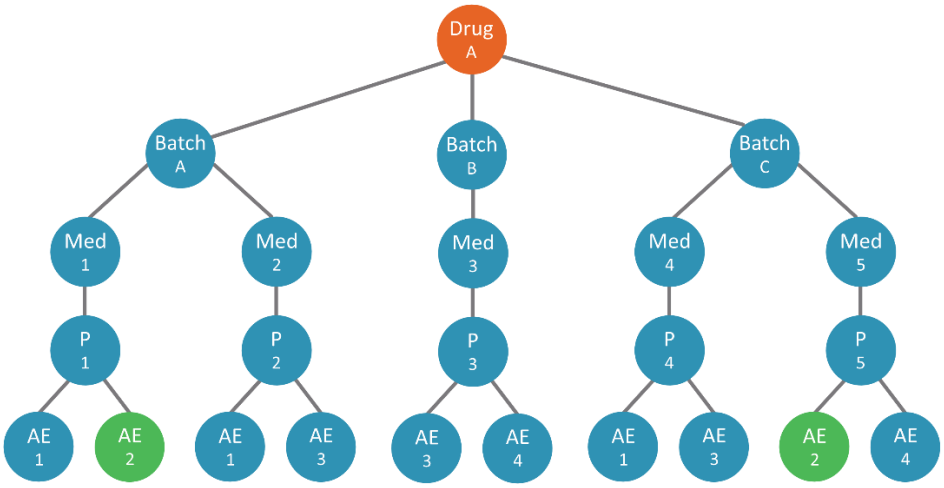

Figure 3. The AE is reflected in Investigational Drug (Drug). All the AEs and all other data related to Drug A is retrieved. This means that all batches for all patients taking Drug A are retrieved.

Figure 3 shows the same simplified graph as in figure 1 and figure 2, but the result set differs since the query is reflected in the Investigational Drug. This approach will give the user an overall picture for the investigational drug where renal failure (AE2) has been reported. All data that has been reported for patients that have taken this drug is included in the result set. All patients given Drug A, independent of batch, are retrieved as well as all medications for these patients and all related data.

Examples of questions that can be answered quickly with a reflection in Investigational Drug are:

● Is there any difference in the incidence of renal failure between the different batches?

● Is renal failure related to the drug or only to specific batches?

USE CASE SUMMARY

The figures show the results from questions asked from the three different perspectives Patient, Batch and Investigational Drug. The Reflective Logic makes it possible to quickly explore an adverse event and its related data.

In the example the focus is on the AE profile, but you can explore all the data related to the patient, batch or investigational drug and furthermore, switch between the different perspectives. On this basis new cohorts can be found and immediately be analyzed in the context of the new findings. This scenario can for example be useful when exploring whether an AE is drug related or batch related.

The advantage of this graph retrieval approach grows with the complexity of the data and the questions. By providing the end user with all available data, new hypotheses can be investigated immediately.

CONCLUSION

This approach is a breakthrough technology that allows immediate access to data and its relationships and the ability to answer complex questions in just a few clicks. Seamlessly analyzing data from different perspectives is possible using an ontology and a unique implementation of indices. The underlying database is flexible and sustainable which allows users to easily access, combine, explore and enlarge datasets.

The ability to analyze all of the data related to e.g. the patient, batch, or drug and to quickly switch between the different perspectives provides for faster and more thorough analysis. Hence allowing the data to be put to work for the greatest success. This technology solves the problem of data accessibility and usability and the advantage of this approach grows with the complexity of the data and the questions. It delivers the power of all of the data to the user, all of the time.

REFERENCES

1. P.Tormay and A. Berg, "*DH08 – Data Tranparency – Breaking Down Data Silos for Improved Insight*", presented at PhUSE EU Connect 2018
2. S. Wagers, "*Why data cleaning and curation is not something that can be handled easily*". BioSci Consulting Blog, 2018
3. P. Tormay, "*Big Data in Pharmaceutical R&D: Creating a Sustainable R&D Engine".* Pharmaceutical Medicine, 2015. 29: p. 87-92.
4. P. Tormay and H. Drews, *"TT06 – Reflect on your data"*, presented at the annual PhUSE conference, 2016.
5. A. Berg, H. Drews and C.Dahlbo, "*PP05 – New Approach to Graph Databases*", presented at PhUSE EU Connect 201**8**
6. C. Dahlbo, A. Workneh and H. Drews, *"PP04 – Unique Technology for the Future"*, presented at the PhUSE EU Connect, 2018.
7. S. Gestrelius and H. Drews, "*Patent US20130246438 – Reflective logic unlocks knowledge in datasets*", 2013

ACKNOWLEDGMENTS

I would like to thank all my colleagues at Capish for having come with valuable input in this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Catharina Dahlbo
Capish Nordic AB
Carlsgatan 3
211 20 MALMÖ, Sweden

+46 (0)40 10 88 80
catharina.dahlbo@capish.com
www.capish.com

Michael Rubison
Capish Inc.
212 Glenwood Road
Lake Forest IL 60045, USA

+1 847 917 2111
michael.rubison@capish.com