DV07

Effective Data Modelling for Effective Data Visualization

Eva Kelty, Capish, Malmö, Sweden
Dr Peter Tormay, Capish, Malmö, Sweden

ABSTRACT

"A picture says a thousand words" or so the adage goes. While true, any effective data visualization requires that it is supported by appropriate and relevant data. Without the right data, you might not be able to create the appropriate visualizations or even worse they provide the wrong picture resulting in wrong actions.

This presentation will draw from a collaboration project with a Swedish Quality Registry (CPUP) where the available data of approximately 5000 patients, collected over several years, was curated and integrated to enable the creation of a visualization application to improve data exploration not only in a cross sectional but also a longitudinal manner. It will highlight how modern graph database technology, coupled with an ontology-driven information model provides a scalable and flexible framework to facilitate effective data visualizations for the evaluation of data from various perspectives, without the need for continuous data processing and restructuring.

INTRODUCTION

Life science and health care continually struggle with complex and time-consuming processes to make data accessible, understandable and useful for end-users. Data, often comprising of a huge number of variables from different sources, can be very cumbersome for the end user to grasp and get a comprehensive view of all available data. Also ensuring that data is interoperable and reusable is challenging.

The purpose of this presentation is to demonstrate a new way of modelling and indexing data for effective and efficient visualization of life science and health care data.

EFFECTIVE VISUALIZATION

When is a visualization effective? Well, when you get the desired effect. That means you must understand the purpose you are trying to achieve and work your way from there. An effective solution for multiple purposes needs objective data as well as flexibility in the exploration of the data.

Data comes in many shapes and forms. Often clinical trial data is very structured and well-defined as it is collected to evaluate a pre-determined hypothesis. Other patient data, e.g. from patient journals, is often text-based and thus more difficult to summarize without any processing of the data. Free-text searches allows you to find content but of course it is limited to accessing information that is actually included in the text. An important aspect to consider in the data collection process.

Relating events in time is often very informative. Sometimes it is important to understand when events occur in real time, that is in absolute dates. For example, pollen allergies are more common in the spring than in the winter. Other times you want to be able to relate events to each other in relative time. E.g. did the nausea start before or after taking the first dose of a study drug. Also does the data represent a single event or is it an episode, which is ongoing during a period of time?

Another challenge that needs to be considered is what information is included in the field vs what information is included in the metadata as well as the access to metadata.

Any visualization is dependent on access to relevant data. Relevancy can be assessed in many ways. Of course, data quality must be satisfactory for the use of the data, but there are many aspects to data quality, e.g. intrinsic, contextual, representational and accessibility (Wang et al 1996).

The purpose is of course also important when it comes to how a visualization is made available, what level of flexibility is needed. Here user maturity and training requirements should be considered. A fully flexible solution where the user can create their own content using self-service components obviously requires more from the user than a static solution, perhaps only offering views of aggregate data with limited interactive features such as filtering or highlighting of data. A compromise may be guided analytics with limitations imposed on the self-service functionality.

Additional aspects to take into consideration is what level of detail of visualization is feasible. Aggregated data summarized in a chart provides a user with information regarding e.g. correlations in the full population or a selected subgroup. On the other hand detailed views of all data collected for a single patient can be very important to understand drug reactions.

INFORMATION MODELLING

Many information/data models are designed based on how data is collected (e.g. Case Report Form), or the intended type of analysis. Commonly data is stored, accessed and analyzed using traditional relational or statistical databases based on a table structure. Therefore, data from one patient is often found in multiple tables, requiring joining or merging of tables of interest using a complex query language.

An alternative way is to work with an information model constituted of small entities of information, which are linked together by relations. These small information entities, termed Holons in the Capish Information Model, can be seen as small messages representing the minimum amount of information a person with knowledge within that subject matter field needs to understand and use for decision-making. Each entity is a whole in and of itself, but is at the same time also a part of another whole, represented by the entire information model.

This modern information model is designed to be readable and understandable by humans, as well as computers. It is built on how humans perceive the world and is thus intuitive. The model allows for expansion of the model without requiring restructuring of data relationships as is commonly seen with other models. Adopting this method of information modelling removes the need for a user to have knowledge about the database structure in order to access and analyze data.

Adding a well-defined terminology, with the possibility to include existing standards and terminologies, further enhances the power of the information model by coupling metadata with the actual data. This ensures that the information is self-described and not just implicitly defined. The information coupled with a terminology in this way constitutes an ontology which is a formal representation of a knowledge domain. By utilizing an ontology, data will be easier to access, easier to explore, easier to share, and above all easier to understand.

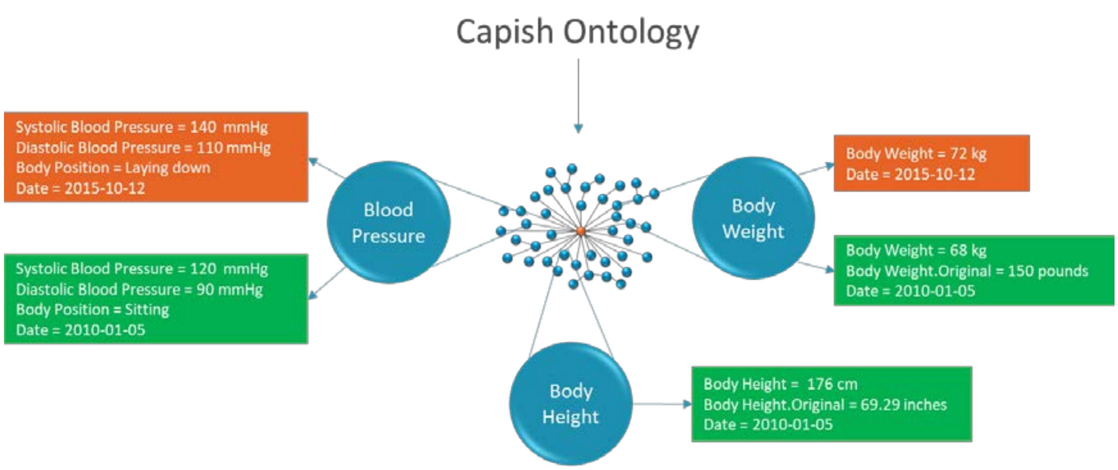

Figure 1: *The Capish Ontology and samples of Holon content*

A REAL CASE

A collaborative project has been conducted with a quality registry in Sweden where data from 5212 patients has been curated and adapted to the Capish Ontology. Thereby, the data was made available for researchers to explore correlations, identify outliers and find subgroups of interest. The patients are followed over a number of years, thus data is longitudinal, but cross-sectional views of the data was also of interest.

DATA CURATION AND INFORMATION MODELLING

Data curation spans over a range of activities; harmonization, standardization, mapping and modelling are examples of activities needed for an effective exploration.

The first data curation step in the collaboration project was to ensure consistency. Duplicate entries were handled, conflicting answers were identified and resolved in a transparent manner, text written in fields supposed to only contain numeric values were moved to a new field, misspellings were mapped to a common spelling and dates were corrected where relevant.

The second curation step was focused on standardization and aimed to make data comparable. Here data was mapped to ISO-standards and established dictionaries, it was translated to a common language, common units and value lists were applied and also reference comparisons were added.

The third step was to set a structure for the effective modelling. Data was grouped into well-known concepts in the ontology, Holons. Also, metadata was appended, time was related to events for each patient, aggregate measures were calculated, e.g. durations and number of events. In addition, episodes of events were defined based on point measurements in the collected data, to describe the continuity of e.g. use of an assistive device.

While all original data remained in the database, the curated data became the source for all aggregations, visualizations and comparisons. The curation efforts are key activities in the creation of effective data visualizations as without consistent and comparable data visualizations can be completely misleading.

VISUALIZATION AND DATA EXPLORATION

A software, Reflect, has been developed where a property graph database with a novel indexing scheme utilizing the above described information model (Dahlbo et al 2018) allows end-users to rapidly access and search data using an interactive and configurable interface.

Reflect is built on a new paradigm on how to effectively query and retrieve information from a dataset. It exploits the existence of direct and indirect relationships in the information model. The applied indexing allows for an un-reflected or reflected response to queries. As an example, a search for a value could potentially not only retrieve all the information units containing that value, but could also be configured to retrieve all directly or indirectly related information. The reflection points can vary dependent on the needs, e.g. patient, site or study. Reflection points therefore further extend the power of the modelling.

A real case example is how the registry data, after being through curation and indexing, could generate un-reflected or reflected responses to relevant queries. Depending on how a selection/search was conducted the user could obtain answers to either or both of the following questions.

Question 1. Patients with pain in thigh, what assistive devices had they been using at the same point in time?

Question 2. Patients with pain in thigh, what assistive devices had those patients ever used? And what other pain had they experienced?

Below is an example of how these questions were answered.

QUESTION 1

An un-reflected approach returns the answers to the first question, “Patients with pain in thigh, what assistive devices have they been using at the same point in time?”.

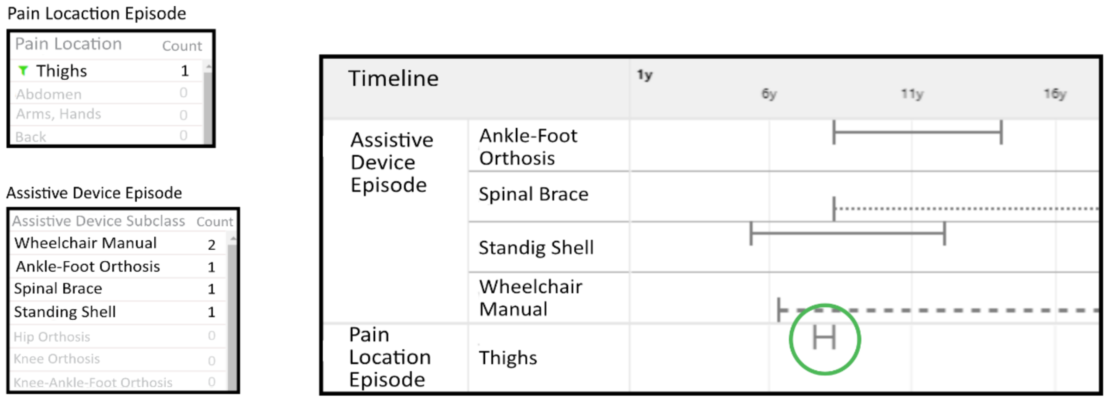

***Figure 2**: Without reflection, the answer of the question shows events that occurred at the time the patient had pain. The timelines show events as episodes, to describe the existence of pain between two dates.*

QUESTION 2

Using patient as the reflection point the solution returns the answers to the second question, “Patients with pain in thigh, what assistive devices have those patients ever used? And what other pain have they experienced?”

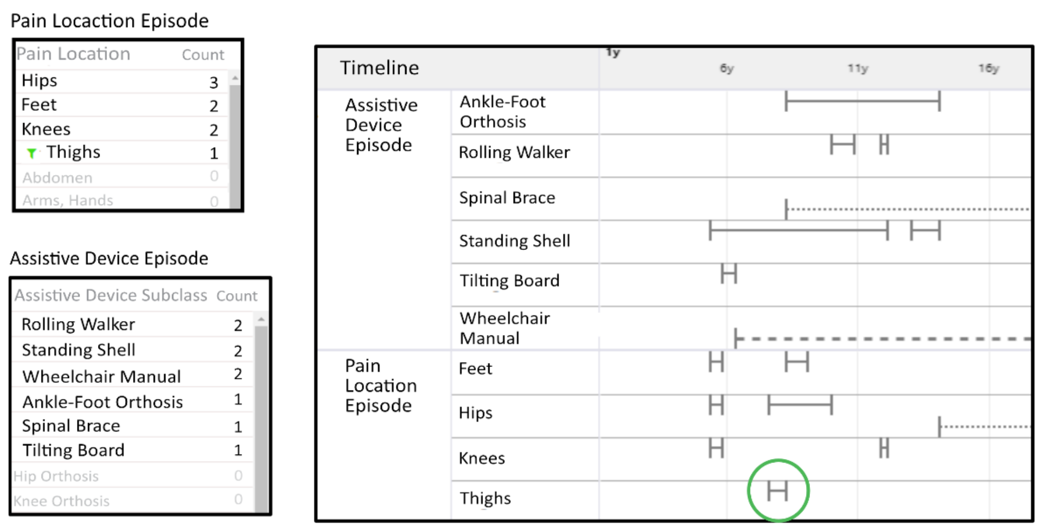

***Figure 3:** With reflection, the answer of the question finds all events that ever had occurred, independent of the time for the patient.*

Actually, utilizing the patient as a reflection point will provide the end-user with a full patient profile including original and curated data. As mentioned above, the reflection point can be changed depending on the context the data is evaluated, to e.g. site or study.

FURTHER EXAMPLES

The information model in combination with the technology allows for various ways to easily visualize data. With a few mouse clicks, among others the following can be explored;

- Patients reporting headache as Adverse Events, what other Adverse Events are they reporting?
- Patients on study drug, experiencing nausea, what other medications are they taking? When are they nauseous, before or after?

- Identify a subgroup of e.g. patients that at the age of four had a specific level of disease and experienced pain.
- Highlight the data for patients in a subgroup in the full population or a filtered subset of the population.
- Find patients with similar data for a vast number of parameters and explore their safety profile

CONCLUSION

Applying an ontology-based information model together with new graph database technology, creates new possibilities to explore and visualize data. Questions, that with traditional technologies require end-users to have assistance from programmers or savvy data analysts to get the proper response, causing delays and risk misunderstandings, can now be resolved with a limited number of clicks on a mouse.

Taking an approach to information that it is modelled for man and machine, available not retrieved, self-described not implicit as well as understandable rather than coded and applying these principles in practice creates a complete, powerful data repository containing data and metadata for reuse.

REFERENCES

R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” J Manage Inf Syst, vol. 12, no. 4, pp. 5–33, Mar. 1996.

C. Dahlbo, E. Kelty and M. Rubison, “PP20 - Understanding the Power of Information Modeling for Efficient and Effectve Visualization,” presented at the PhUSE Connect, 2018.

ACKNOWLEDGEMENTS

I would like to thank Gunnar Hägglund and Ann Alriksson-Schmidt at Skåne University Hospital for the collaboration, giving us the experiences of a collaboration project.

CONTACT INFORMATION

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

Eva Kelty
Capish Nordic AB
Carlsgatan 3
SE-211 20 Malmö, Sweden
Work Phone: +46 40 10 88 80
Email: eva.kelty@capish.com
Web: capish.com