

Harness the power of ontologies to build better information platforms

Dr Peter Tormay, Capish, London, UK

Anna Berg, Capish, Malmö, Sweden

ABSTRACT

Ontologies provide a formal representation of a knowledge domain, describing its entities, events and processes and the relationships connecting these. One key advantage of ontologies is their ability to store general knowledge constructs as well as specific ones – i.e. instances of the generalized knowledge constructs.

As such ontologies support the development of modern information platforms that aim to store a representation of the status quo for subsequent interpretation. The closer the ontology resembles the status quo, the closer the inferred reality resembles the original situation for interpretation.

Ontologies also enable the integration and combination of data from multiple heterogeneous data sources, where data can be represented in different formats, stored using different infrastructures or different terms are being used to represent the same data.

All these challenges can be addressed using an ontological approach to data modelling, which will be discussed in this presentation

INTRODUCTION

In a previous PhUSE paper [1] we pointed out that there is a clear paradigm shift in clinical development. Instead of a linear, rather siloed workflow where data is moving from one stage to the next, the industry needs an approach where data and the resulting information and knowledge is available across all functions throughout the whole value chain from discovery to market.

This has resulted in the realisation that the clinical development process is data driven. Having said that, it is not the data per se that drives this process but rather the information and the knowledge we derive from this data. In order to achieve this goal, we need to be able to understand the data in context and interpret it accordingly.

Because of the paradigm shift we eluded to earlier, we don't only have to be able to evaluate data at the point of collection but also at a later date, when additional information about how, when and in which context the data was collected is normally not readily available. In this sense modern databases are not so much archival systems but rather information platforms that in one way or another need to be representations of the data and its contextual constraints (i.e. meta data) at the time of collection so that we can understand and interpret the data at a later time. As a result, we need to create a factual representation for further interpretation so that if our view of the status quo changes our representation of the current situation can change as well.

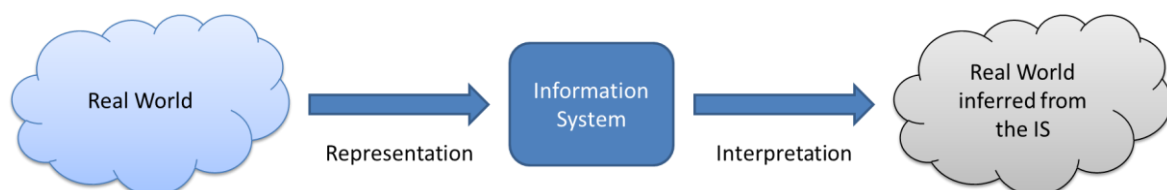


Figure 1 - Information systems as a representation of the real world

ONTOLOGICAL FRAMEWORK

In order to understand how ontologies can support the development of such a representational information model it is important to determine what an ontology actually is.

The term ontology has its origin in philosophy, where it denotes the study of what entities exist and how they relate to each other (metaphysics).

Unfortunately, the term ontology has been used to mean a variety of things within the realm of information technology, ranging from simple lists, vocabularies, terminologies, taxonomies to more complex representational frameworks. Combining these different viewpoints, we define an ontology as follows:

ONTOLOGY

A structured representation of a knowledge domain, describing its entities including properties, qualities and attributes as well as events and processes and the relationships connecting these entities, events and processes, that can be instantiated.

To start with, an ontology is a representation of a knowledge domain. An ontology can be looked at as a photograph. In the same way a photograph is a faithful representation of reality an ontology should be as close as possible to reality or our understanding of reality.

In addition, an ontology does not only cover entities or things in the way a taxonomy, a hierarchical classification of things, does. An ontology also includes their properties, qualities and attributes as well as events and processes. An ontology is capable of capturing change over time.

Ontologies also need to be well structured. A consistent terminology needs to be available to define all the different entities, events and processes. The terminology tells you what terms are used to describe a particular entity while the ontology is concerned with the entity itself. Take the term patient for example. A terminology would define what the term patient means, whereas an ontology would look at a patient as part of what exists in reality. Since the terminology is part of the ontology, an ontology is self-describing.

The main purpose of any ontology is to provide a better understanding of the existing knowledge by putting all available entities in context. By putting entities in context, you connect them to each other, creating relationships between them. These relationships are rooted in first order (or predicate) logic exemplified by the simple statement “there exists an entity X such that X is a patient and X has Asthma”. The relationships in this statement are denoted by “is a” and “has”. Relationships like this are often referred to as primitive. Primitive relationships effectively do not carry any meaning as such other than denoting that two entities are connected. Other types of relationship might provide a specific meaning – compare the two statements “Jane lives in New York” to “Jane works in New York”. While both statements connect Jane and New York, the context in which these two entities are connected is clearly different.

Another aspect about an ontology, relevant in the context of a representational information platform, is the ability to instantiate the entities, events and processes described in the ontology. As such the descriptions in an ontology are based on a generalisation, i.e. the “encoding of general features of things” [2]. All these generalisations are based on individual instances of these things. As an example, what characterises a person in general is ultimately determined by the characteristics of an individual person. Therefore, it needs to be possible to use the generalisation of a person to describe an individual person.

Looking at different ontologies there are several types of ontologies that can be distinguished. Upper or Formal Ontologies are domain-independent ontologies. In many ways these Formal Ontologies describe other ontologies and how they should be organised. You could consider them ontologies of ontologies. Domain-specific or Material Ontologies on the other hand describe specific knowledge domains using domain specific vocabulary.

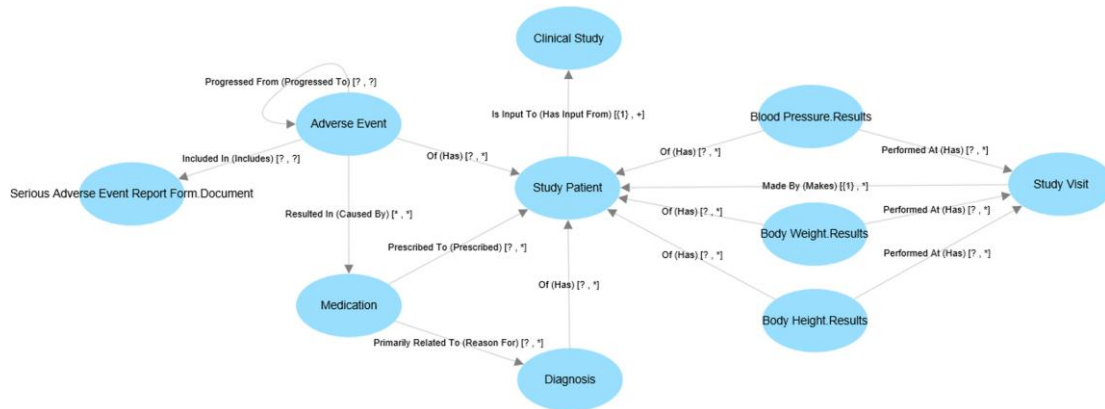


Figure 2 – A simple ontology

PURPOSE OF AN ONTOLOGY

We have already mentioned that one of the main purposes of an ontology is to provide a better understanding of a knowledge domain. By describing all the elements of a knowledge domain and its relationship, all the knowledge is made explicit and therefore usable by everybody or at least domain experts. The explicit nature of an ontology also makes it possible to evaluate the information at a later date, thus making it reusable. The terminological component of an ontology makes it also possible to integrate data from highly disparate data sources.

Ontologies are also mentioned in the context of annotating data [3], mainly in the form of assignment of entries from controlled vocabularies to particular data points. In the clinical development fields this is currently the domain of metadata repositories, which are used to annotate data in existing relational database systems.

REPRESENTATIONAL (ONTOLOGICAL) INFORMATION MODEL

Relational databases have been the “swiss army knife of the industry for decades” [4]. Relational databases have proven to be very effective as transactional systems but as analytical systems, they have many shortcomings. In particular the ever-increasing complexity of data is making the development of appropriate data models increasingly difficult or rather impossible.

People are becoming increasingly aware of the connected and linked nature of data requiring new approaches to store and evaluate complex, highly disparate data. With the advent of modern database technologies, in particular graph databases, it is becoming more and more common to represent data as knowledge graphs (not to be confused with Google's Knowledge Graph) or linked data. We described different graph database technologies as well as their benefits over conventional relational databases in previous PhUSE papers [4], [5]

BENEFITS OF GRAPH DATABASES

The conceptual data model translates directly into the graph database model. There is no need for transformations. Most data and how the data is connected can be described as a graph. In reality this is how our brain works making it rather easy to understand. There are no constraints on individual joins as with relational databases.

Graph databases are flexible and easily expandable. New nodes and relationships can be added without disturbing the existing graph. As a consequence, the data model can evolve with our understanding of the respective domain. This means that data can already be exploited or evaluated before the final data model has emerged. Repurposing of the data is also much easier as the perspective and focus can be adjusted.

Metadata can be stored directly as part of the data. Graph databases make it easy to store any data about the data as part of the model. This makes it much easier for users to understand the model and work with it.

Data can be evaluated in different contexts. Data can be explored starting from any node within the network of data points.

These benefits can be greatly enhanced through the use of an ontology. After a knowledge domain has been adequately described in an ontology, this can be directly implemented as the conceptual data model for the respective database. While the ontology provides the framework and the terminology for the database, the database provides the means and technology for the instantiation of the data. The availability of a terminology through the ontology also functions directly as the metadata for the database.

PRACTICAL MODELLING/IMPLEMENTATION CONSIDERATIONS

When you combine an ontology with a graph database to create a representational information model, there are principal modelling considerations.

The first consideration in the modelling process is dependent on the type of graph database used. Obviously, an ontology that will be paired with an RDF graph database will look different to an ontology paired with a property graph database.

In the case of an information system that is aimed at analytical processing, an intuitive data querying process and an ability to analyse data from different perspectives the starting point should be based on a property graph database with labels and fields.

In this case the ontological entities and relationships can be mapped directly into the database objects (nodes) and the edges respectively. In this case entities and nodes as well as edges and relationships are synonym respectively

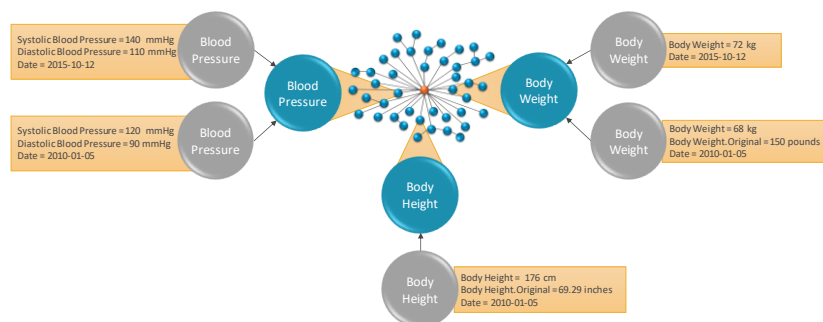


Figure 3 - Combining an ontology and graph database

The entities within the ontology should be designed in such a way that they can act as the blueprint for the nodes within the graph database during the instantiation. We believe that these entities should be

- built around concepts that can be interpreted by themselves
- classified according to content and well-established concepts
- contain information in the form of fields, groups and attributes
- contain information about related entities (nodes)

Information
Concept: Body Height Domain: 616.0754--dc23 Holon: Body Height.Results
Description Results of a Measurement of the Body Height of a
Fields Body Height Date Planned Day, In Study Study Patient Id Date.Note
Relations Study Patient Study Visit

Figure 4 - General Holon structure

We have coined the term Holon for entities modelled along these principles. The application of these principles has several implications. Different Holons might contain overlapping information in order to make them self-describing. For instance a “Patient Holon” might contain some of the information of a “Person Holon”, since the “Patient Holon” is obviously a person and therefore some of the information might be duplicated. In order to create an ontology that can be widely used, Capish for instance has decided to build its ontology around the Dewey Decimal Classification System [6] . In addition to the actual fields containing field values each Holon contains all necessary information for the interpretation of the Holon as a whole as well as each field value.

Current graph efforts focus on local graph traversals with a particular focus on relationships and the meaning of these relationships. While this allows for the evaluation of local graph patterns it makes evaluation of global graph queries more difficult. Another approach to modelling could be called a role based approach. Capish has developed its own ontology on that basis. In this approach only primitive relationships are used to indicate that two nodes are linked. Any additional meaning that might normally be conferred with the relationship is provided in separate nodes or roles. This approach allows for much better access to your data. While still some knowledge of the data model is required, the intricacies of the relationships become irrelevant, streamlining the querying process.

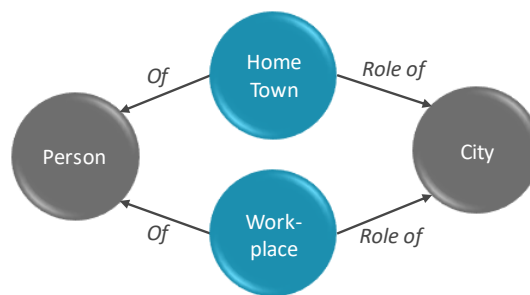
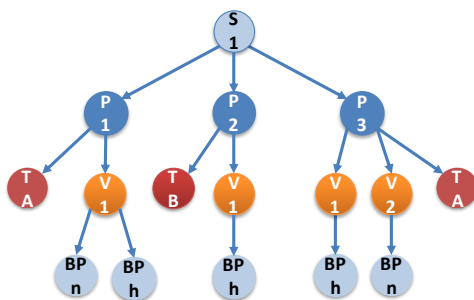


Figure 5 - Role based approach to modelling

While graph databases do not require indices, the implementation of indices clearly improves performance but also can help with querying the database. Capish for instance has introduced a relation index, which provides information of indirectly linked nodes, enabling the creation of a query language that does not require knowledge of the full data model, in particular the relationships that lead from one node to another [4].



Property Graph

- individual nodes (identity index)
- node types (node type identity index)
- property values (property index)
- existence of indirect relationships (relation index)

Select data:

Type of Node: “Patient”

With (Type of Node: “Liver Value”, Property: “Value > 5”)

Fetch:

Type of Node: “Adverse Event”, property: “Name”

Figure 6 – Database implementation

CONCLUSION

There is a clear consensus that the future of drug development and healthcare provision is driven by data – data that will help us to better understand the underlying disease, data to better evaluate the efficacy and safety of new drugs, data to make sure that patients receive the most effective treatment available.

As an industry we are still struggling to bring all the available data together in an efficient and effective way. Data is collected in many different contexts – clinical trials, healthcare records, payer records, patient reported outcomes and many more.

Due to a shifting paradigm, information systems need to become factual representation of the real world for future interpretation. As our understanding of the status quo changes we need to be able to adapt our representation of the current situation without major remodelling.

Ontologies provide a formal representation of a knowledge domain, describing its entities, events and processes and the relationships connecting these. The aim of an ontology is to provide a common understanding of the current state of particular domain knowledge, to make this knowledge and any assumptions about this knowledge explicit in order to enable reuse and analysis.

As such, we are all familiar with ontologies in the context of vocabularies, thesauri and terminologies all of which can be considered to be simple forms of ontologies.

While ontologies played an important role in the development of conceptual data models for the development of relational database systems in the past, the advent of graph databases and other NOSQL databases means that ontologies can function directly as data models in their own rights.

Ontologies designed to be used as the underlying information model for a database should follow some basic modelling rules, whereby each ontological entity should be a self described unit and links between entities should be only primitive, while specific role based entities bridge the gap of meaningful relationships.

Ontologies combined with modern graph database systems provide the flexible systems that are necessary to cope with the ever changing landscape of data complexity and our resulting understanding of how the real world works.

REFERENCES

- [1] P. Tormay and A. Berg, "DH08 Data Transparency - Breaking Down Data Silos for Improved Insight," presented at the PhUSE EU Connect, Frankfurt, 2018.
- [2] R. Arp, B. Smith, and A. D. Spear, *Building ontologies with Basic Formal Ontology*. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2015.
- [3] O. Bodenreider, "Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support," *Yearb. Med. Inform.*, vol. 17, no. 01, pp. 67–79, Aug. 2008.
- [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 the PhUSE EU Connect, Frankfurt, 2018.
- [6] M. Dewey, J. S. Mitchell, J. Beall, R. Green, G. Martin, and M. Panzer, *Dewey decimal classification and relative index*, Ed. 23. Dublin, Ohio: OCLC Online Computer Library Center, Inc, 2011.

CONTACT INFORMATION

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

Dr Peter Tormay

Anna Berg

Capish Nordic AB

Carlskgatan 3

211 20 Malmö

Work Phone: +46 40 10 88 80

Email: peter.tormay@capish.com, anna.berg@capish.com

Web: www.capish.com

Brand and product names are trademarks of their respective companies.