Paper TT03

Data, Metadata and the notion of Holons*

(* Information concepts)

Peter Tormay, Capish, Malmö, Sweden

ABSTRACT

There has been a fundamental paradigm shift with respect to data. The aim is to use and reuse data in many different contexts over the life of a development program or drug. This means, it is not just necessary to store the actual data value but also additional information around that value – metadata. This metadata goes well beyond simple descriptions about field labels and controlled terminology. Moreover we need to find ways to store this metadata together with the actual data - introducing the notion of Holons - in order to prevent any data loss due to a later separation of metadata and data. This presentation will look at how ontologies and graph database technology and in particular property graph databases (in contrast to RDF stores) can support the drive for an integrated information platform, where data and metadata can live together to support meaningful data exploration.

LETS TALK ABOUT DATA

What exactly is data? What do we mean when we are saying that the Pharma industry is "data driven". There is a lot of confusion when it comes to terminology. We talk about data, information, knowledge – and while these are three distinct notions that build on each other they are often used synonymously.

- **Data** is the building blocks. Data represents the raw facts, that in themselves do not provide any meaning.
- **Metadata** can be defined as data about data. This data can be looked at as additional data that qualifies other data of interest. This data effectively can be considered to provide the context in which other data of interest should be evaluated. It is noteworthy that data in one context could be considered metadata in another and vice versa (also see further down for a definition of metadata in an IT/database context).
- **Information** is derived from data by grouping data and metadata and putting it into context.
- **Knowledge** is derived from bringing information together and putting it into context. Knowledge provides a greater understanding of how things work and how they are connected.

Let's look at an example. Looking at a value of 140/90 there is very little we can say and as such it provides a simple data point. Adding the unit of millimeter of mercury (mm Hg) and designating it as blood pressure we provide more context elevating this value to providing information. Again this information does not provide much insight on its own. Only if we put this information into context of other information, such as the fact that this constitutes a high blood pressure compared to the normal values in the range of 90/60 – 120/80 mm Hg or the persons weight, smoking habits, family history etc. we start to build up knowledge. In this scenario this additional data/information could be considered to be the relevant metadata that qualifies the original blood pressure result. In conjunction this information could establish knowledge about the common causes of high blood pressure.

Taking the notion of knowledge further, one also needs to distinguish between explicit and tacit knowledge. As the terms suggest, explicit knowledge exists in a tangible form that can be easily disseminated and shared[2]. It is accessible to people and can be re-applied in various contexts. Tacit knowledge is often referred to as experience or know-how. It resides with the person who has it and is difficult to articulate, i.e make explicit. It is also important to understand that new knowledge is normally created through the internalization of explicit knowledge which is then evaluated in the context of the available tacit knowledge held by the person or other agent (i.e machine learning). Consequently, it is important to externalize available tacit knowledge, i.e. make it explicit as much as possible to benefit from its value as much as possible.

Armed with these definitions, it becomes evident that the Pharma industry is knowledge driven rather than data driven and we collect data in order to further our knowledge relating to disease pathways, mechanisms ultimately to create new treatments for those diseases.

As far as IT and the development/establishment of information systems is concerned, knowledge is treated as information, and should be seen as something that can be codified and transmitted. IT plays a pivotal role in knowledge sharing and dissemination [1].

Modern information systems need to be able to capture all available data and its context in order to create a factual representation of the real world or rather what we know about the real world for further interpretation.

In order to fulfil this goal, information systems need to store the underlying data in a consistent, easy to understand way that can be interpreted by the users of the system (people and potentially other computer agents). While not necessarily part of the system itself they also need to be accurate, objective and considered to be trustworthy in order to be used.

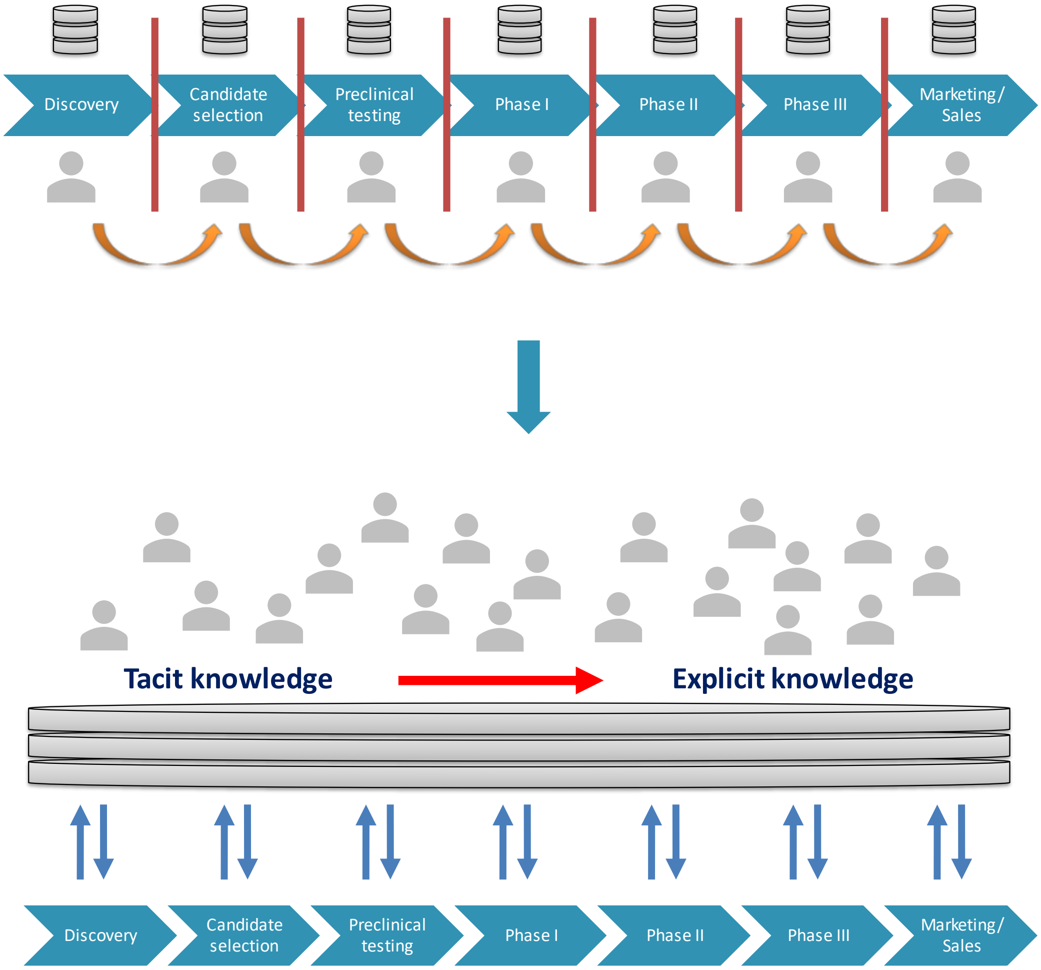

Figure 1- Tacit and Explicit knowledge across the organization

WHAT IS AN ONTOLOGY

In the context of knowledge management, ontologies are formal representations of how we view things – our current understanding of a particular knowledge domain. They are a structured representation of a knowledge domain, describing its entities, events and processes and the relationships connecting these entities, events and processes

Ontologies are aimed at

- sharing a **common understanding** of the structure of information among people or software agents
- enabling **reuse** of domain knowledge
- making domain assumptions **explicit**
- **analysing** domain knowledge

At the centre of any ontology sits the notion of concepts. While the word concept can represent many things to many people, in the context of a knowledge representation concepts can be described as the representation of the things and entities that are relevant within a particular knowledge domain (subject). Each concept tries to describe in a formal way what attributes make up a particular entity. Look at the concept of a study patient. A study patient can be described by his/her age, gender, weight etc but also the clinical study, disease etc. As far as concepts are concerned a clinical study and a disease would be concepts in their own rights. In their simplest form these different concepts are connected to other concepts in a hierarchical or taxonomical manner. A study patient for instance can be described as a type of person. At the same time more complex relationships provide the relevant context. In effect the fact what makes a study patient a study patient is its relationship to a clinical study or a disease.

These concepts are very different from simple definitions. A definition for study patient would possible be “a person that participates in a clinical study”. Ontologies are concerned with the entities and their attributes themselves.

Ontologies can be considered to be networks of information or a conceptual mind map of related concepts.

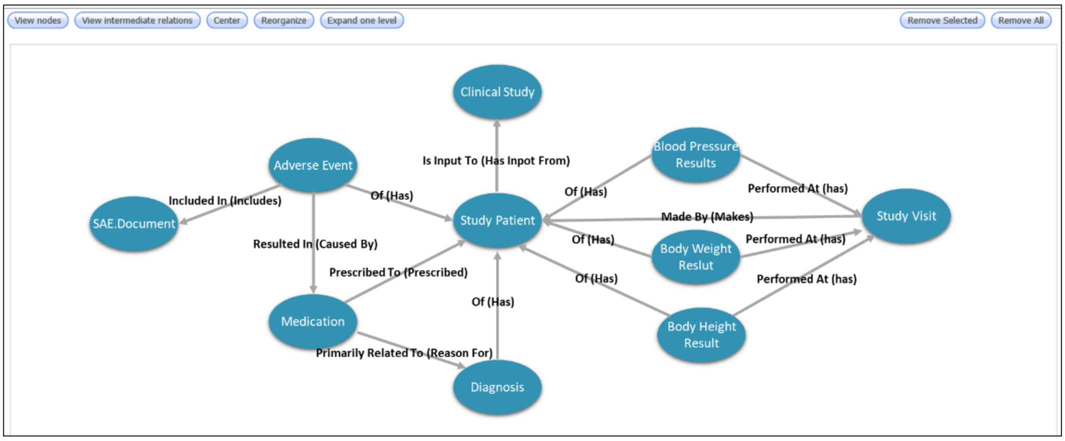

Figure 2- ontological modelling

Ontologies are also extremely flexible as they are unbounded. As our understanding of a knowledge domain expands so does the ontology. New concepts or information objects can be added as and when the need arises.

Another aspect of an ontology is the fact that it is a formal representation. In addition to the structural framework of an ontology that is defined by the entities etc, the attributes within the concepts that make up the entities as well as the relationships between those entities are described using a common terminology that ensures consistency and understanding by domain experts.

Ontologies therefore provide the structural integrity and the consistent approach required for the creation of appropriate information systems.

DATABASE SCHEMA AND GRAPH DATABASES

So far, we have discussed data, information and knowledge and a possible way to represent them using ontologies. In the end the aim is to capture this data, information and the resulting knowledge in an electronic format, so that it is available to other people for consumption.

One thing databases have in common, is the fact that they are described by a database schema. There are various types of database schemas. There is the conceptual schema, the logical as well as the physical schema. As the names suggest they deal with the underlying concepts, the rules as well as the physical implementation.

In recent years a new type of database – a so called property graph database has gain considerable traction. These are different from another graph database type referred to as RDF [3]. These databases consist of so called nodes or vertices – individual data objects, that are connected through direct relationships or edges. The interesting thing about this database type is that all three layers of the database schema are more or less the same. Graph databases are often referred to as schema less. This is merely to denote the fact that the schema can be updated without any implications for the existing database. It also means that objects, even if the represent something similar do not have to follow the same structure. While this is true in most cases a schema that describes how the database is structured will exist as it would be otherwise extremely difficult to query and make sense of the information it contains.

The database schema that is associated with a graph database is also very similar to how an ontology would be structured. While an ontology establishes the entities, processes and events that constitute the content and the structure of a knowledge domain using a common terminology, the database schema describes the structure of the database, the properties of individual nodes (vertices) and the relationships (edges) that exist as part of the database.

A graph database can therefore easily be used to represent a particular knowledge domain by mapping entities, process and events to particular nodes and define edges according to the relationships found in the ontology. In such a way the general concepts defined in an ontology can be instantiated in a database. In our example of a study patient, the ontology would provide the definitions for all the attributes that make up a particular study patient, while the database would contain data about specific study patients, where the individual attributes contain specific values associated with this particular study patient.

LET'S TALK ABOUT HOLONS

Capish has coined the term Holon for these specific nodes that are derived from an ontolohy but at the same time allow the collection/storage of data for specific instances. They can be characterised as follows:

- Concepts based on the ontology, 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 definition and description of these fields, groups and attributes (structural metadata)
- Contain relationships to other Holons establishing the context
- System independent

This means that each Holon is self-contained and encapsulates all the data – metadata and raw data - that is necessary to evaluate this particular entity, process or event without any additional information. Each Holon also "knows" how it is connected to other Holons within the ontological framework.

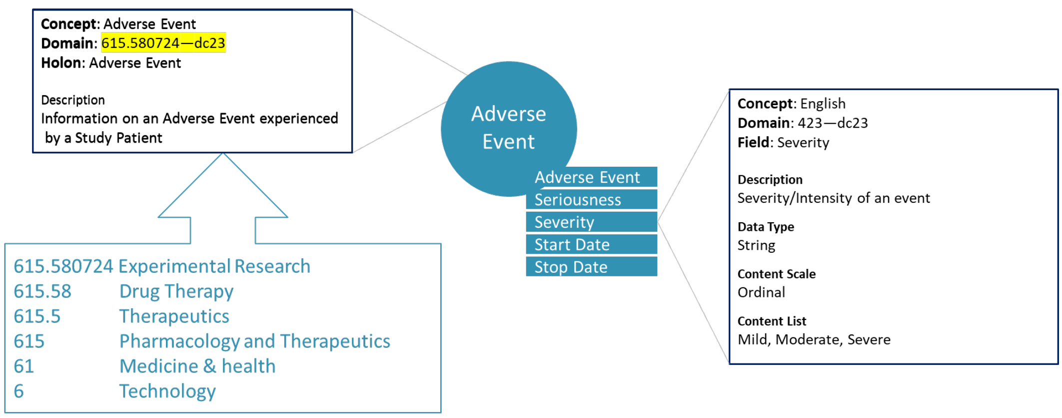

Figure 3 - The structure of a Holon

Holons can be considered as a group of data and information, that taken together provides meaningful information, from which knowledge can be derived. Apart from the actual data fields, that can hold the specific values each Holon contains information about the Scientific area it belongs to. Capish has chosen, to represent this scientific aera using the Dewey Code [4]. In addition each Holon is assigned a Concept as well as a Holon Type. In a similar way each field is also described by a field name, a concept and a corresponding Dewey code. The Dewey Classification System, which is normally used as a library classification system was chosen, as it organises knowledge domains as they are taught in universities, is widely used across the glog and is regularly updated to accommodate new scientific areas and changes.

The Dewey code together with the concept name therefore refers to a specific knowledge domain, within which domain experts will have a shared understanding of the terminology used.

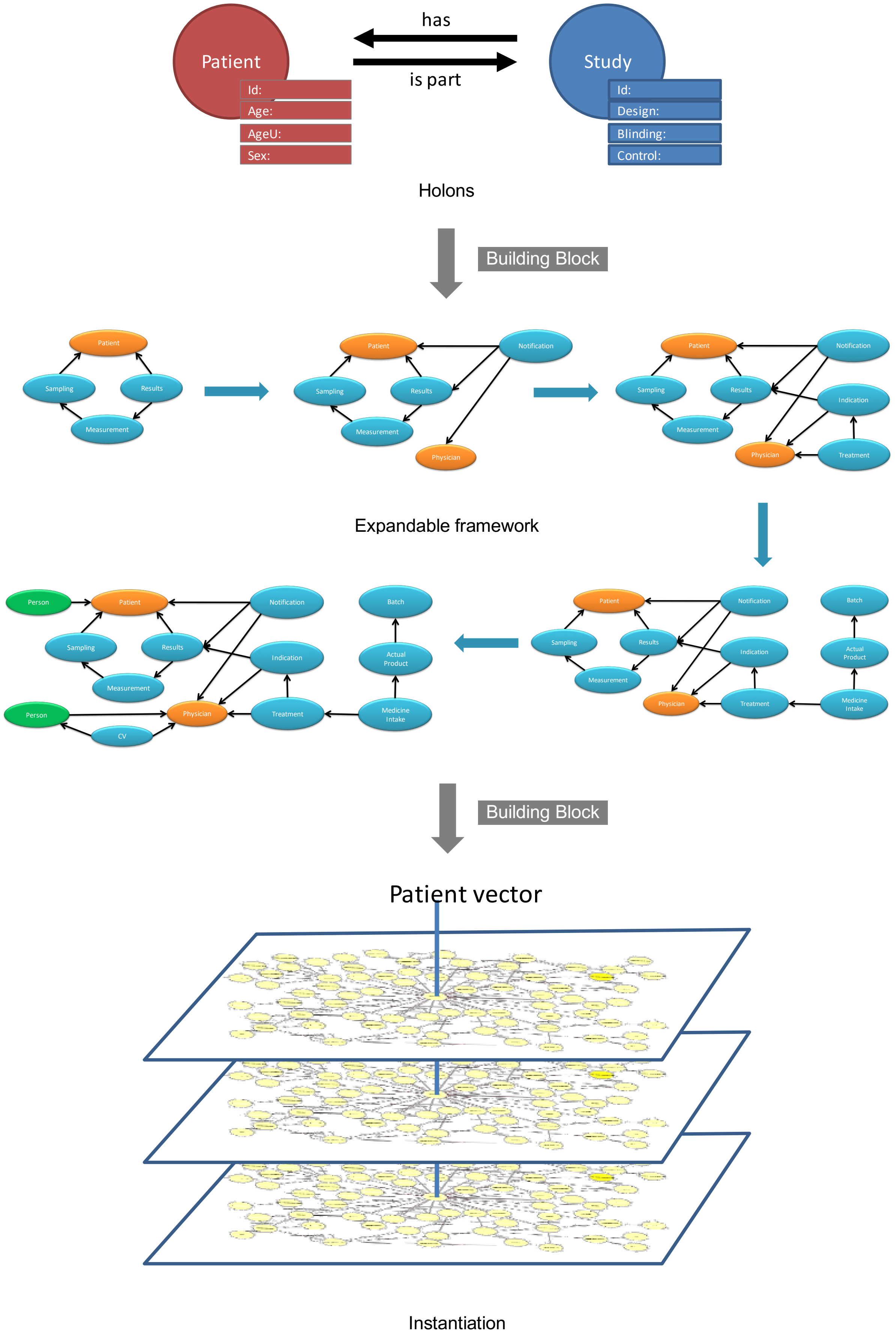

Figure 4 - From Holon to Database

SUMMARY

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.

What we need now is a way to connect all the available data in order to create a factual representation of the real world for further interpretation. As our understanding of the status quo changes our representation of the current situation needs to change as well.

In order to be able to evaluate existing data in the future it is important that we not only store the actual data value but also additional information around that value – metadata. This metadata goes well beyond simple descriptions about field labels and controlled terminology. We also need to ensure that metadata and data always stays together. The current system of independent metadata repositories and data store cannot ensure this and a separation of data and metadata is a real risk, rendering the data useless again.

This is where ontologies come in. 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 constitute simple forms of ontologies.

Ontologies also played an important role in the development of conceptual data models for the development of relational database systems. With the advent of graph databases and other so called NOSQL databases, ontologies can function directly as data models in their own rights.

Combining ontologies with graph databases provides a compelling way to store data and metadata in a single platform enabling end users to gain better insight and improve query ability of existing data.

Graph databases or the concept of linked data is not a new concept but for most people it is synonymous with RDF stores. In contrast to these RDF stores property graph databases, often referred to as native graph databases, allow the storage of data in objects, where data that pertains to the same event, incident or process is stored as a well-defined information unit. This approach to “packaging” the data provides many advantages when it comes to keeping data and metadata together introducing the notion of Holons –

Holons are self-contained information units that are sufficient in themselves to describe a particular event or process in its entirety. At the same time Holons form the building blocks of a larger network of information where Holons are connected to other Holons establishing the aforementioned factual representation of real world for further analysis.

REFERENCES

[1] Knowledge Management Tools website: http://www.knowledge-management-tools.net/knowledge-information-data.php

[2] Knowledge Management in Theory and Practice

[3] Anna Berg, Henrik Drews, Catharina Dahlbo, PP05 – New approach to graph databases, PhUSE EU Connect 2018, https://www.phusewiki.org/docs/Frankfut%20Connect%202018/SA/PO/Papers/PP07-pp05-19569.pdf

[4] https://www.oclc.org/en/dewey.html

ACKNOWLEDGMENTS

Acknowledgments go after your references. This section is not required.

RECOMMENDED READING

1. P.Tormay, "SD06 – Selection in Action", presented at PhUSE EU Connect 2019
2. P.Tormay and A. Berg, "DH08 – Data Transparency – Breaking Down Data Silos for Improved Insight", presented at PhUSE EU Connect 2018
3. E.Kelty, "DV07 - Effective Data Modelling for Effective Data Visualization", presented at PhUSE EU Connect 2018 CONTACT INFORMATION

CONTACT INFORMATION

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

Dr Peter Tormay
Capish Nordic AB
Carlsgatan 3
211 20 Malmö
Work Phone: +46 40 10 88 80
Email: peter.tormay@capish.com
Web: www.capish.com