PhUSE EU Connect 2018

Paper DH08

Data Transparency – Breaking Down Data Silos for Improved Insight

Dr Peter Tormay, Capish, London, UK

Anna Berg, Capish, Malmö, Sweden

ABSTRACT

Data transparency is mainly discussed in the context of external accountability and patient security. At the same time data transparency is a major opportunity for companies to share their data between departments as well as collaborators for further evaluation either in the same context but more importantly in different contexts. The main premise of such an open attitude is that the value of available data is greatly increased thus improving the return on the initial investment.

Opening up the various data silos creates a rich environment of potentially interconnected data. One key element in such an "open access" strategy is the management of the available data. Data needs to be collected, curated and stored. It also needs to be possible for people to query and extract relevant data for further analysis. This also requires that any data point can be correlated with available metadata to ensure that data can be reused and evaluated appropriately in context.

This paper will look at how graph databases and ontological data modelling can support the drive to greater data transparency and ultimately better use of data.

INTRODUCTION

With the advent of a data driven economy, it is important to change the mindset of end users and implement an appropriate infrastructure. Knowledge is power and competitive advantage was deeply rooted in secrecy and security. Unfortunately, this was not only true in the broader environment of a company but also on a more local departmental level.

The negative impact of this approach meant that data could not be evaluated to its full potential or in the worst case could not be evaluated at all as it was hidden from view in some remote data silo.

As a result, companies need to change the way they look at how they manage, integrate, curate and ultimately use data. Data transparency is a key factor in these considerations.

DATA TRANSPARENCY

ORGANISATIONAL TRANSPARENCY

Data transparency in the clinical development domain is primarily associated with Policy 70 of the European Medicines Agency (EMA), that relates to the publication of clinical data for medicinal products for human use [1]. EMA defined clinical data as clinical reports, consisting of clinical overviews, clinical summaries and clinical trial reports, as well as individual patient data. Since January 2015 all clinical reports, relating to a new marketing authorisation application through the centralised procedure need to be published. Publication of individual patient data is currently not included but will become part of a second implementation phase.

The policy provides some key pointers to the purpose of data transparency. The main objectives of making clinical data available are:

- enable public scrutiny
- improve accountability while also
- disseminate new knowledge for future research

EMA believes that the policy "has the potential to make medicine development more efficient by establishing a level playing field that allows all medicine developers to learn from past successes and failures" [1].

At the same time EMA acknowledges some of the challenges such as maintaining patient privacy and the need for protection of commercially sensitive data.

Pharmaceutical companies are still concerned about the policy. It puts a heavy onus on them as they must:

- anonymize patient data to protect patient privacy
- redact commercially sensitive information to protect their assets

While this work is certainly cumbersome, it provides companies with an additional benefit. Transparency establishes trust and enhances a company's reputation with its external stakeholders – patients and the public at large. This is something that is currently missing in the industry [2].

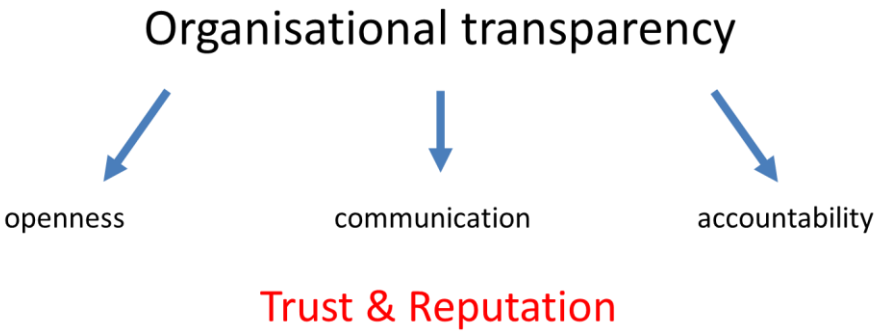

Figure 1 Organisational transparency - Trust and Reputation

Companies should take this as an opportunity to establish their transparency capabilities and make it a core value within the organisation. In addition to preparing companies better for sharing data externally, it will keep employees engaged and improve communication and discussion within the company for better decision making [3]. It needs to be clear, that transparency is first and foremost a cultural challenge. Without the right cultural mindset, no transparency initiative aimed at making data more transparent will be successful.

DATA TRANSPARENCY FRAMEWORK

Data transparency is not just about making data available and accessible to relevant stakeholders. It is concerned with the free flow of data for better decision making. Free data flow ensures better dissemination of associated information. This will result in the conversion of tacit knowledge into explicit knowledge, that is available throughout the company. There are several dimensions that need to be considered in order to achieve this goal [4].

STAKEHOLDERS

First there are the stakeholders involved. In addition to the data provider and the data consumer there is also the data source, in this case the patient. These stakeholders are connected through the information system that is used to share the information on a technical level. Looking at different stakeholders it becomes evident that one can distinguish between supply side transparency and demand-side transparency. In the first scenario data is made available by the data provider, either voluntarily or due to regulatory requirements – such as the EMA policy. In the second scenario data is requested by the data consumer often as a result of a breakdown in trust. While the proactive dissemination of data is aimed at improving trust and accountability, the reactive release of data is mainly associated with damage control.

DEPTH OF TRANSPARENCY

Transparency exists on multiple levels. In addition to the data, there is also process information and policy information. Data can be further divided into summary data and raw data, with the first being an aggregated view of the latter. Process information answers the how, it provides information about how the data was collected and how it has been analysed. Policy information answers the why. It puts the data into context. In the setting of a clinical trial, the why answers the relevance of the clinical trial in the wider context of the clinical program, the how would be related to all the procedures carried out during the study as well as the statistical analysis plan outlining the analysis, and the what would be the summary data in combination with the actual raw data.

Moving from raw data to process information provides more depth of transparency. It improves context and makes it more meaningful to the relevant stakeholders. While dissemination of raw data without process

information and policy information is possible it makes the evaluation of the data cumbersome and might even prevent it altogether. Policy and process information can often be provided independently and in many cases might be sufficient to satisfy the transparency needs of the stakeholder, particularly when it is combined with summary information relating to the outcome of the research. In other words, increased granularity of raw data results in increased usefulness of transparency.

Some transparency constraints are also connected to the depth of transparency. As pointed out earlier these constraints relate to patient privacy and the protection of commercial information. The more granular the data is, the more challenging it is to maintain patient privacy and retain control over commercially sensitive information. Looking at the EMA policy this is the current situation.

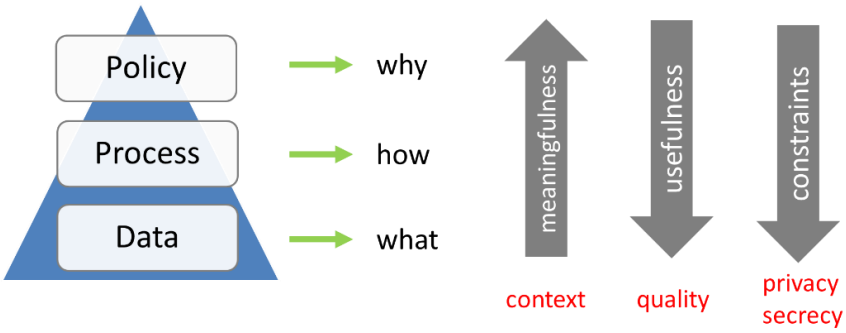

Figure 2 Depth of transparency

USEFULNESS OF TRANSPARENCY

In the previous paragraph it was already established that the usefulness of transparency increases with the granularity of the data. The more detailed the data, the more information can be gained and the deeper the insight. Usefulness is integrally linked with data quality. While the basic requirement of data quality is data accuracy there are many other data quality dimensions [5]. Data needs to be relevant. As pointed out earlier, data context is provided by including information as to why and how data was collected. This context might not necessarily be aligned with the context in which the data is consumed. Only if it is relevant and can be evaluated in the context the data is needed, it becomes useful and adds value. Data also needs to be provided in a concise and consistent manner and it needs to be interpretable and understandable. This is often referred to as interoperability. Most of the time interoperability focuses on 'semantic interoperability' or terminology to ensure that data is interpretable. Data format or 'structural interoperability' is often ignored but is equally important. Finally, data needs to be available and accessible. Availability and accessibility are two sides of the same coin. While data is made available by the data provider it also needs to be made accessible to the data consumer. In reality it is impossible for data consumers to differentiate between the two.

These various data quality challenges have recently resulted in the FAIR data principles [6] that aim to make data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

TRANSPARENCY BENEFITS AND CHALLENGES

So far, this paper has provided a framework within which data transparency can be evaluated and viewed. In this section the main benefits and challenges as derived from the framework will be considered. The key benefit on an individual level is autonomy. The higher the data transparency – the higher the autonomy. It is easy to see that complete, meaningful and understandable data that is proactively provided by the data provider enables data consumer to act on the data directly. There is no need to further engage with the data provider thus wasting valuable resources. At the same time properly implemented data transparency will improve communication as various stakeholders can engage in meaningful discussions since data is explicitly available. This positive environment will also have a profound positive effect on organisational culture in a positive feedback loop.

The main challenges as described earlier are in relationship to patient privacy and keeping commercially confidential data just that – confidential. Privacy is ensured through the anonymisation of data, while commercially sensitive data is redacted. Both practices can have a profound impact on the usefulness of the data. These challenges might also be mitigated by restricting accessibility, either by disseminating data to a small manageable group or only allowing access through a portal, that prevents dissemination of the most granular data in the first place. The ability to establish any necessary restrictions on the fly would make this process easier and would require less resource intensive processes.

THE CHANGING PARADIGM IN CLINICAL DEVELOPMENT

Drug development used to be a linear workflow moving through the various stages of drug discovery and drug development. At the end of each successful stage, the results were passed on to the next stage in order to progress the drug along the clinical development value chain. At the same time actual data was stored on departmental databases, not for future use but rather as a regulatory requirement. This resulted in data silos which completely prevented the free flow of data, hampering decision making and prohibiting the reuse of data.

With the realisation that data has a value beyond the initial context it is collected, full data transparency is needed along the clinical development value chain. Data transparency will also ensure that data doesn't only flow freely up the value chain but also down. Discovery and early development can gain much from understanding later development activities. Free flow of information will also improve communication and discussion bringing different people, functions and departments together.

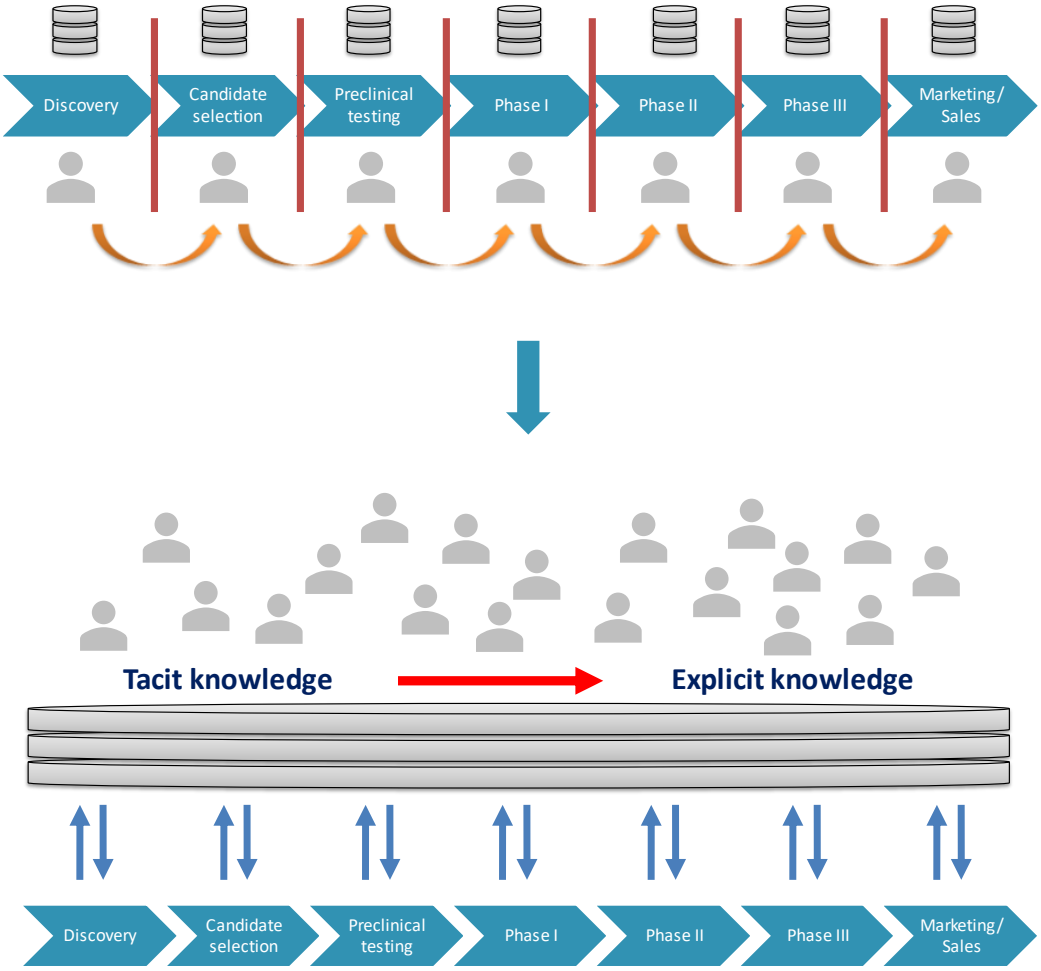

Figure 3 Paradigm shift in clinical development

Any attempt in achieving this paradigm shift, puts a tremendous onus on the information system used to store the data. It effectively functions as the main communication channel between data provider and data consumer. As such it can be looked at as a representation of reality, i.e. all the entities, events and processes as they are observed by the data provider in context, which in turn is used by the data consumer to interpret these events and processes to infer his or her view of reality. The closer reality and inferred reality are, the deeper the data transparency and thus its usefulness.

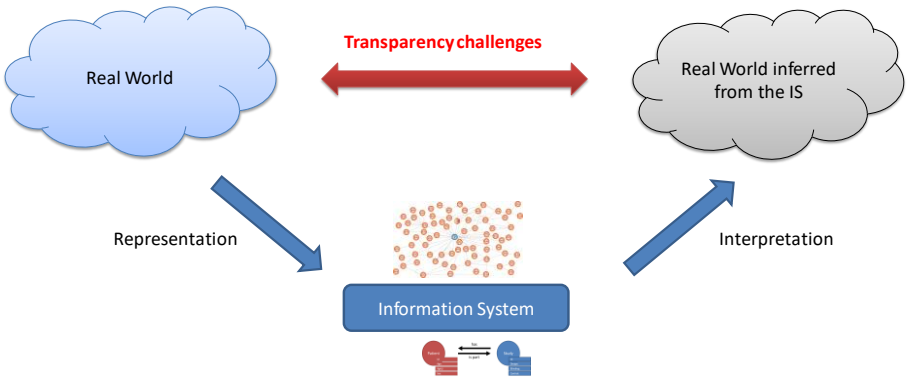

Figure 4 Information systems as a representation of the real world

For an information system to be able to represent reality as it happened it needs to be able to store all the things, events and processes as they are (collectively data nodes). In order to provide the necessary context, it also needs to be able to track how these entities are related. This is very much in the realm of ontology research.

THE ONTOLOGICAL DATA MODEL

In the context of knowledge management, ontologies are formal representations of how we view things – our current understanding of a particular knowledge domain. 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

Ontologies are also extremely flexible as they are unbounded. As our understanding of a knowledge domain expands so does the ontology. New nodes or information units, can be added as and when the need arises. In this way ontologies can be considered to be networks of information or a conceptual mind map of related concepts.

This fits very well with our data transparency requirements.

TERMINOLOGY (METADATA)

Most ontologies only deal with the generalisation and definition of terms. In this sense we are very familiar with ontologies in the form of terminologies, vocabularies and taxonomies such as MedDRA, Snomed CT and all the current metadata repositories. These simple ontologies lend themselves to annotate databases to provide context but cannot be used as the underlying data model per se.

INSTANCES (DATA)

Designed in the right way, ontologies can be extended so that they not only deal with generalisation and definitions but also with individual instances. As an example, such an ontology could define a clinical patient and a clinical study in general but would also contain data about a particular patient in a particular study.

Capish has coined the term Holon for these information units and they can be characterised as follows:

- 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 relationships to other Holons

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.

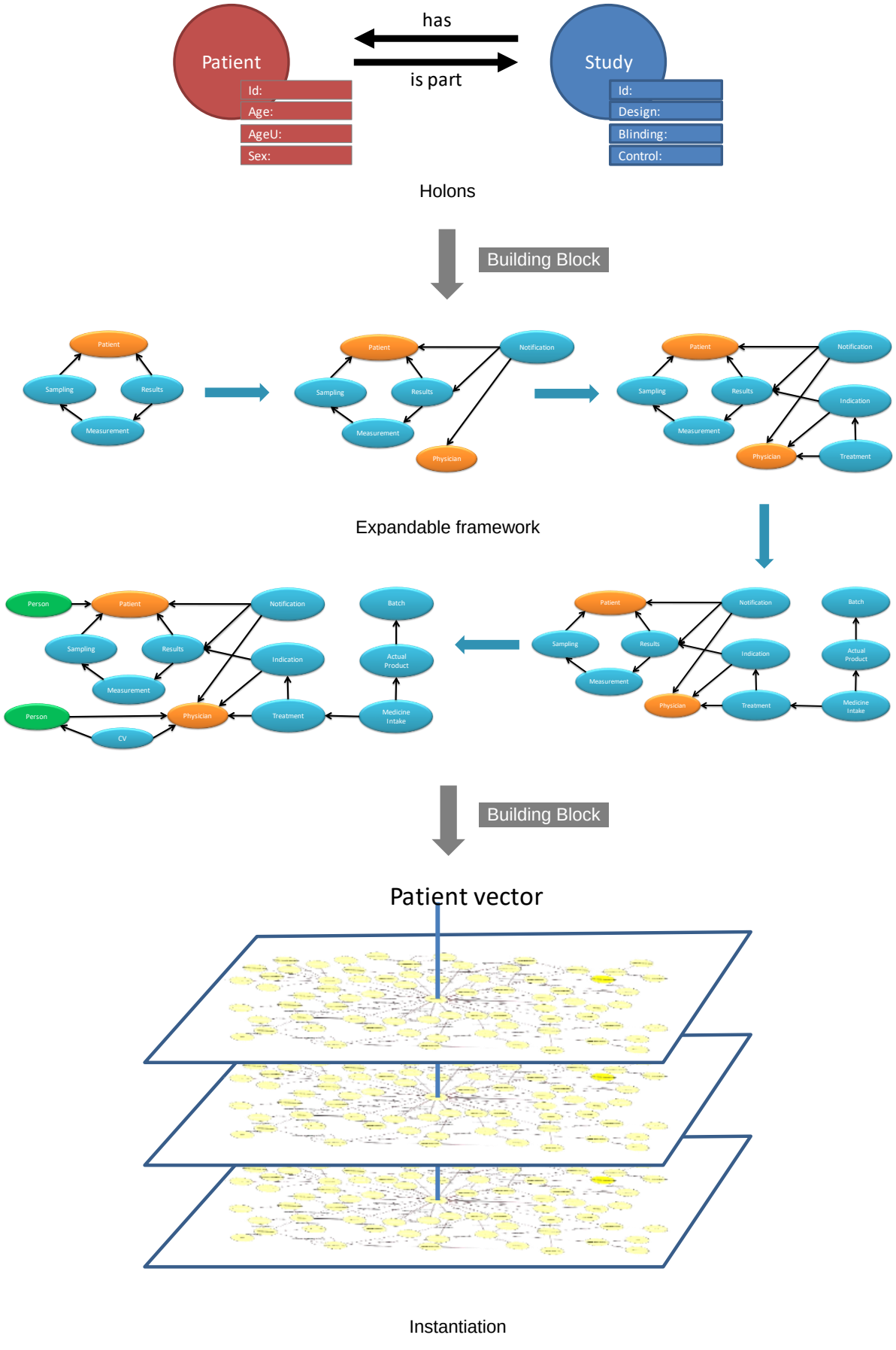

Figure 5 From Holon to Database

DATABASE IMPLEMENTATION

While the ontological data model provides the perfect fit for any data transparency requirements, it is not supported by commonly used relational databases. Any benefit this data model provides would get lost in the translation of this conceptual data model into the required physical data model.

Over the last decade a new breed of database has been developed, that provides the perfect complement for an ontological data model described in the previous section. Graph databases - in particular property graph databases - are comprised of nodes and edges. Each node is a named data object containing the main data while the edges provide the relationships between these data objects. The concept of Holons described earlier can therefore be mapped into these data objects which in turn are connected through the relationships described in the ontology. It should be noted that these property graph databases are very distinct from the commonly used RDF triple stores, that are currently used to map metadata repositories and other terminologies for simple annotation purposes (see poster PP05 for more details [7]).

As with the ontological data model, property graph databases are equally flexible and expandable, which means the database can grow not only vertically but also horizontally (i.e new types of data objects can be implemented). On a practical level this means that data can be evaluated before a final data model has been explicitly implemented.

Storing data in a database makes it generally available but not necessarily easily accessible. Building the database around an ontology that is built around how we view the world makes the navigation of the database more intuitive [8].

CONCLUSION

Data transparency provides an excellent framework to establish comprehensive data capabilities within a company that allows to share and communicate data freely for better insight. Data transparency principles inform the creation of suitable information systems and highlight the benefits of new database technologies such as graph databases underpinned by an ontological data model:

- Key drivers of data transparency are data access, data quality and data reuse.
- Easy data access is determined by a common information model based on a common terminology and the ability for the end user to interact with the data in an intuitive manner. Since clinical and healthcare data is ultimately unbounded any such model needs to be flexible in nature for future expansion and adaptation.
- Data quality is determined by the ability of the end user to infer the actual real world scenario when the data was collected. In other words, the aim of any information system is to represent reality as it was encountered at the time of data collection.
- Data reuse is concerned with the ability to use data over and over again, either in the same context or in a different context. In order to make data valuable six months, one year or five years after it was collected it is important to capture all the data associated with the actual data point of interest. This metadata provides the rich environment that is necessary to assess data again and again over time.
- Ontologies as data frameworks provide the consistent terminology required to build up knowledge bases that provide insight to the end user, while being flexible and adaptable as our understanding of a particular knowledge domain changes. They also enable the simultaneous storage of all relevant metadata to keep data relevant in the future. Ontologies can also be considered to be a representation of reality thus providing the backbone for data quality.
- Coupling these ontologies with graph database technology streamlines the process of database creation and enables the establishment of information systems that can be navigated intuitively by end users.

While data transparency and these new technologies are primarily discussed in the context of clinical development, the same conclusions can be drawn for healthcare data and information systems in the healthcare environment.

REFERENCES

[1] EMA, European Medicines Agency policy on publication of clinical data for medicinal products for human use. 2014.
[2] “The People vs Big Pharma: tackling the industry’s trust issues - Pharma Technology Focus | Issue 74 | September 2018.” [Online]. Available: https://pharma.h5mag.com/pharma_sep18/the_people_vs_big_pharma_tackling_the_industry_s_trust_issues. [Accessed: 24-Sep-2018].
[3] J. G. Bruhn, “The Decline in Organizational Transparency and the Loss of its Partner, Trust: A Restorative Agenda,” Soc. Sci. Today, vol. 4, no. 1, p. 11, 2018.
[4] M. Hosseini, A. Shahri, K. Phalp, and R. Ali, “Four reference models for transparency requirements in information systems,” Requir. Eng., vol. 23, no. 2, pp. 251–275, Jun. 2018.
[5] 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.
[6] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, p. 160018, 2016.
[7] A. Berg, H. Drews, and C. Dahlbo, “PP05-New Approach to Graph Databases,” presented at the PhUSE EU Connect, Frankfurt, 2018.
[8] P. Tormay and H. Drews, “TT06 - Reflect on your data,” presented at the annual PhUSE conference, 2016.

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