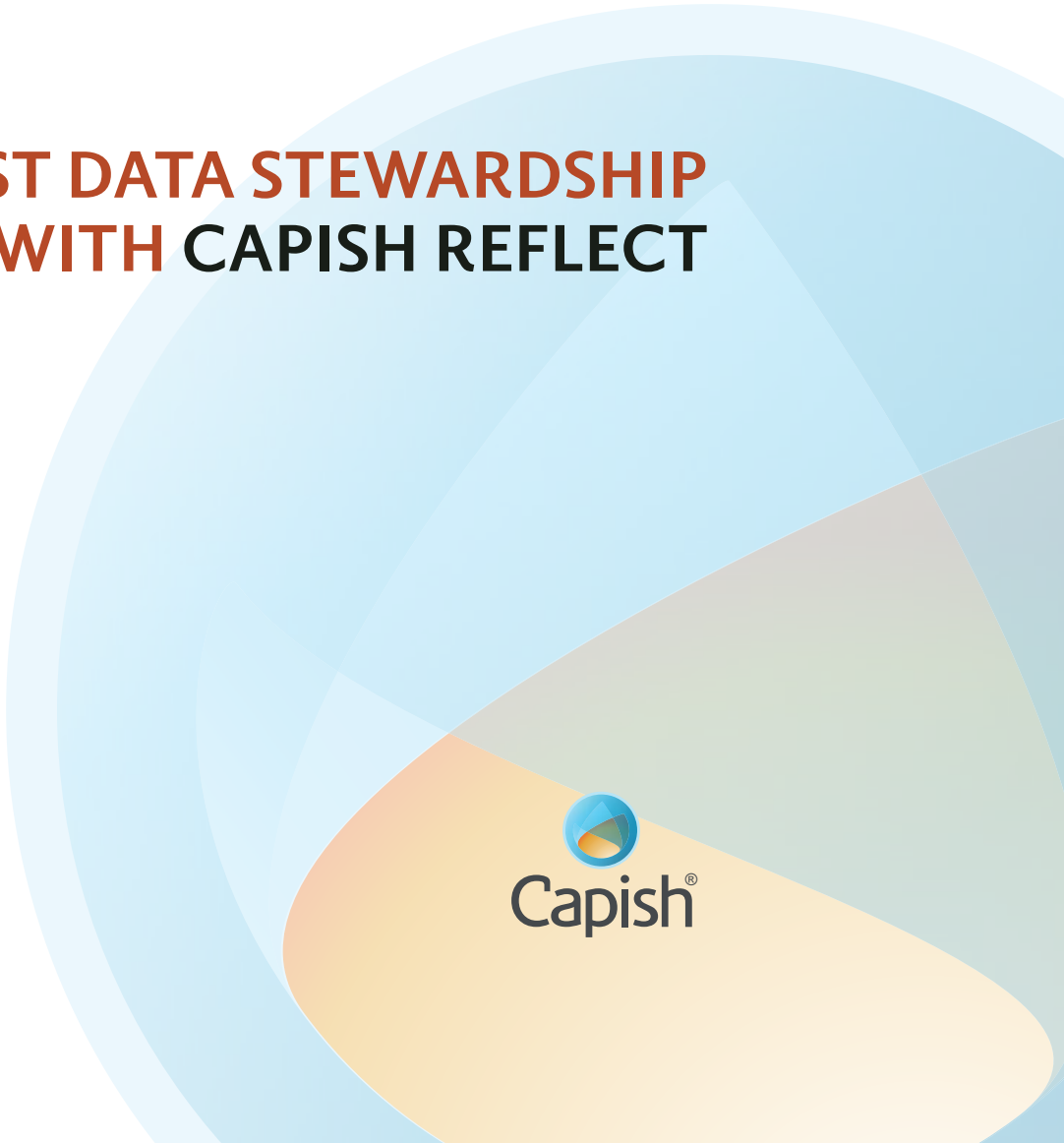
A White Paper from Capish

ROBUST DATA STEWARDSHIP WITH CAPISH REFLECT



Data is the road to success, and it has been for a long time. While this paper draws on the experience Capish has had within the pharmaceutical and healthcare industry this is true for all industries – from financial markets to retail enterprises.

Today we talk about a data driven approach to drug development. But what has changed? When it came to drug development we always relied on experiments and the data they produced – from discovery to clinical trials to market authorization. What has changed is our attitude towards data. While in the early days data was simply a means to an end, it is now center stage.

There are many reasons why this is the case. To start with, we have picked all low hanging fruits. We more or less have identified all the easy targets and we have developed the drugs aimed at these easy targets.

Luckily our understanding of the world – or more narrowly human biology, disease progression and factors is constantly evolving. Unfortunately, as a result, it is getting more and more complex. The more we understand these things the more we recognize that we are not on a linear path but that there is a multitude of interconnections that can influence any particular outcome. As a result, the data we use and need to further our understanding is getting more complex as well. And guess what, so does the analysis we are carrying out using this data.

This increasing complexity makes it necessary to bring data together from different sources to provide a holistic view of the current status quo. It also means that we need to make data explicitly available wherever possible to those people that might benefit from what the data could be telling them. As a result, data has a value beyond the original context it was collected in and can be useful in other contexts at a later stage.

Data stewardship is an organizational function based on the assertion that data is a strategic asset, that needs to be available to users in a timely and consistent manner.
As such data stewardship is an extension to classical data management that primarily looks at data acquisition, accuracy and storage to also include meaning and future accessibility.

FAIR Data Principles

In short, data is a strategic resource that enables us to make informed decisions, ultimately accelerating innovation and allowing us to bring better drugs to market.

With this we have started to appreciate that we need to treat data differently. In the past, data was simply stored for archival purposes within a particular department, creating inaccessible data silos. Today, data needs to be stored in such a manner, that it can easily be accessed by whoever might need it. It also needs to be evaluable to establish its relevance and usefulness for the current purpose.

In accessing this data it needs to be possible to establish the original context the data was collected in, in order to be able to ascertain that it is relevant for whatever it is intended to be used for now. Moving from looking at data as a simple asset to embrace it as a strategic resource requires us to move from simple databases to fully integrated information platforms.

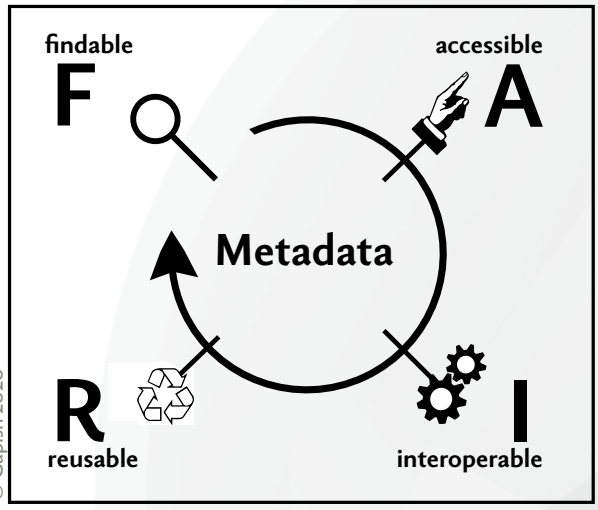
Now that we know data is important and we need to look after it, how can we establish that any solution we implement is up to scratch. This is where the FAIR data principles come into play. Several years back scientists met up to discuss exactly this question. Admittedly they were primarily concerned about published data – but the principles can be applied wherever data plays a central role.

The FAIR Guiding Principles for scientific data management and stewardship. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

It is important to realize that these principles are just that – principles. They do not provide information on a particular solution or approach. They simply provide the litmus test against which you can evaluate your particular implementation.

As the acronym FAIR indicates there are four corner stones. Data needs to be **findable**, **accessible**, **interoperable** as well as **reusable**.

FAIR Principles. Go FAIR Initiative. https://www.go-fair.org/go-fair-initiative [last accessed 03/11/2020]



What do we mean by *findable*? First, data needs to be uniquely identifiable and it needs to be possible to retrieve the data based on this unique identifier. It also needs to be possible to identify the origin of the data in order to establish the value, accuracy and relevance of the

data. This means data needs to be described by rich metadata and it needs to be possible to attribute any metadata to that data unequivocally.

Accessibility as the second corner stone refers to the ability to access the data. Data access is relevant on several levels. On the most basic level it needs to be possible to connect to the server on which the data resides. Also end users need to have the permission to log onto the data source where the data is located. Finally, we need to know exactly how the data is organized in order to be able to retrieve the data. Database models and their implementations need to be well documented, or in other words described by appropriate metadata.

Interoperability relates to our ability to understand the actual data. Data will need to be integrated, mapped to relevant data standards and so forth. This is where the use of terminologies as well as controlled vocabularies comes into play – again content needs to be described by appropriate metadata.

The ultimate goal of FAIR data is to be reusable. *Reusability* represents more of an outcome reliant on the other corner stones. If data can be identified, retrieved as well as understood, it is by definition reusable.

THE IMPORTANCE OF METADATA

We have all heard about the cliché in real estate that there are three main factors affecting a sale - location, location, location. In that vein the three most important factors affecting robust data stewardship are metadata, metadata, metadata.

Metadata is another way of saying "data about data". Metadata is all the other information that puts the actual data into context.

The metadata tells us everything else that we might need to know about the actual data, where it was collected, why it was collected, the circumstances under which it was collected as well as what exactly was collected in the first place. All this additional data about the data allows us to make the appropriate decisions about whether the data is relevant or not.

Metadata encapsulates the data, creating a data object that provides all the information necessary to understand the data.

Property Graph Databases

What are property graph databases and why do we need to talk about them? These databases are built around data objects that are connected through meaningful relationships. In graph database parlance we would talk about nodes or vertices and edges. Each node can have a label that identifies the entity they refer to.

New Approach to Graph Database. Catharina Dahlbo, Henrik Drews, Anna Berg. Poster Presentation at EU Connect 2018 http://www.phusewiki.org/docs/Frankfut%20Connect%202018/SA/PO/Papers/PP07-pp05-19569.pdf [last accessed 03/11/2020]

Each node also has a set of properties which are made up of key value pairs. These key value pairs could be considered as data fields in the conventional sense. The key aspect is that these properties are grouped together and form a defined entity within the property graph represented by the node they belong to.

All of these nodes are connected through directed, meaningful relationships. When we talk about meaningful we refer to the fact that each relationship identifies how one node is connected to another node. The result is connected data where every data point has a specific identifiable position in the overall data network.

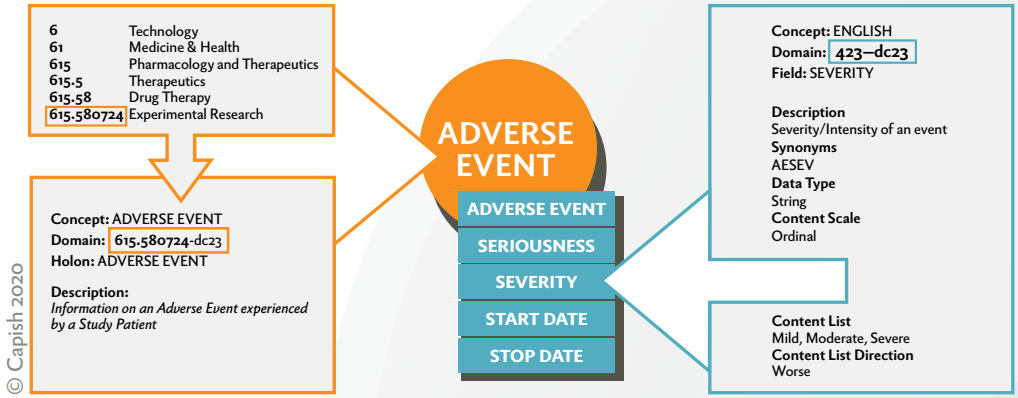
Once you understand what the FAIR data principles try to deliver and how property graph databases work, you realize that property graph databases provide the perfect tool for implementing these principles on a technical level.

CONCEPTS AND HOLONS – FINDABILITY AND INTEROPERABILITY

Nodes within a property graph database are labeled entities comprised of a group of properties. As such we can use nodes to refer to common concepts. They can effectively refer to any event, process, person or really anything that we can conceptualize and is relevant in a particular context.

Looking again at the pharmaceutical industry, these underlying concepts might refer to a clinical study, the patient or an adverse event for instance that the patient suffered during the clinical trial. The important aspect here is that the use of concepts roots these data objects in the body of available knowledge. Since these concepts are based on our understanding of a particular knowledge domain, the terminology used is therefore drawn from this body of knowledge making these data objects understandable for anybody who is knowledgeable in this particular field. While this already goes a long way to making the data stored more FAIR, these data objects can also allow us to store additional data alongside the data.

Holon – What, Why and How. Catharina Dahlbo. Poster Presentation at EU Connect 2019. https://www.phusewiki.org/docs/2019%20Amsterdam/Papers_presentations/PP/PP%20Final%20Papers/PP20.pdf [last accessed 03/11/2020]



The graph gives an overview of how a holon can look and what information it contains. At the concept level these data objects provide additional information about the concept e.g. a description.

Since metadata encapsulates the data to create a data object, why not use a database that is built on data objects to include the metadata - in the same way FAIR data principles propose. The data object in a property graph data base allows us to encapsulate the actual data directly with all the relevant metadata.

At Capish the term Holon is used to refer to these data objects. If we look up the definition for Holon, we will find that it simultaneously refers to something whole as well as a part of and this

is exactly what we are seeing here. The data stored in a single Holon provides everything we need to understand this particular event or whatever is described by the Holon, making it whole. At the same time each Holon is part of a larger network of connected data objects that tell a broader story. They are designed in such a way that they are big enough to be read and understood by themselves but small enough to serve as the building blocks of a larger data pool.

A Holon contains a reference to a particular knowledge domain. This reference is based on the DEWEY DECIMAL SYSTEM, a library classification system, that allows us to associate Holons with a particular knowledge domain.

In addition to this metadata on the Holon level extensive detailed information on each field is provided. This information includes data types, synonyms, code lists etc. again establishing the necessary information to evaluate the Holon and its content by itself. In this way Holons are completely self-describing thus removing the need for additional external documentation.

Advantages of an ontological data model

- The conceptual data model translates directly into the graph database model. There is no need for transformations. Data and how the data is connected can be described as a graph.
- 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.

THE ONTOLOGICAL DATA MODEL – ACCESSIBILITY AND REUSABILITY

So far we have only talked about how we can convert the nodes of a property graph database into Holons but not the fact that they are part of something bigger. This is where the relationships come in. As such each Holon knows exactly to which other Holon it is connected, building out the network of information.

Once we build out the whole network of Holons and their connections we end up with a Capish Ontology. Ontologies are aimed at sharing a common understanding of the structure of information among people or software agents, enable the reuse of domain knowledge, make domain assumptions explicit and enable the analysis of domain knowledge.

Ontologies are also extremely flexible as they are unbounded.

Effective Data Modeling for Effective Data Visualization. Eva Kelty, Peter Tormay. Presentation at EU Connect 2018. http://www.phusewiki.org/docs/Frankfut%20Connect%202018/DV/Papers/DV09-dv07-19253.pdf [last accessed 03/11/2020]

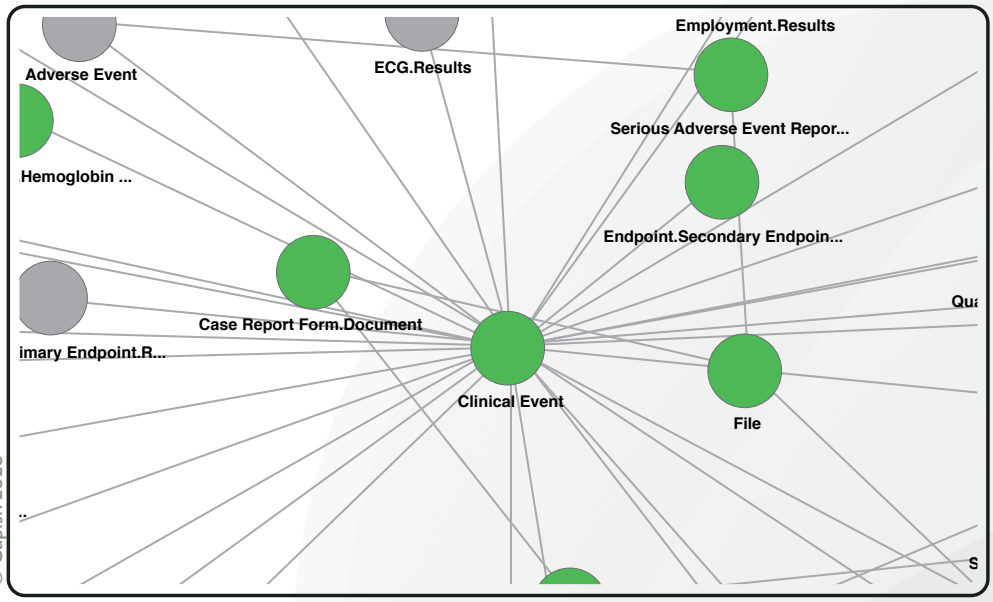
As our understanding of a knowledge domain expands so does the ontology. New Holons, can be added as and when the need arises.

Most ontologies only deal with the generalization and definition of terms. In this sense we are very familiar with ontologies in the form of terminologies, vocabularies and taxonomies, good examples being MedDRA and Snomed CT.

Metadata repositories that are currently developed to support existing databases fall into the same category. These simple ontologies lend themselves to annotate databases to provide context but cannot be used as the underlying data model *per se*. For that reason Holons are designed in such a way that they work both on the conceptual level – after all they are built around concepts but at the same time can be instantiated, in other words they can hold specific data that can be uniquely identified and attributed to a particular event or process.

The ontological data model provides the navigational map of the property database. The end result is a data structure, that is self-describing. It enables end users and other software agents to access data at any given point and explore the data from this point without the need to fully understand the underlying model. Therefore, any information can be accessed at a later date as all the information required is contained within the data model.

Capish Reflect – an information platform powered by a property graph database engine



The Infograph visualizes avaialble Holons and their relations. This is the backbone for any Capish Reflect application.

Up to this point we have discussed how property graph databases can support the implementation of FAIR data, which brings us closer to being FAIR but still leaves the final hurdle of actually implementing data fairness in a real world solution.

Capish® Reflect is a modern web based application, that is powered by a property graph database. The main distinction from a "plain" property graph database is the added user experience that

Capish Reflect provides to the end user to interact with the property graph database in the background.

Rather than being just a front-end query tool Capish Reflect offers visualization and tabulation capabilities that enable the end user to view data in context.

CAPISH REFLECT MAKES DATA ACCESSIBLE

The user interface that connects the end user to the database is built around modern web standards. This makes the need for proprietary software obsolete and allows users to access the database irrespective of their location or the device they are using.
While access is available from everywhere and anywhere, strong user authentication protects the data from prying eyes and provides role-based access to the data, restricting what individual users belonging to a particular role can see and can do with the data.

An obvious advantage of Capish Reflect, that draws directly on the property graph database backbone is the ability to view the data structure in an intuitive manner and evaluate the metadata either on a Holon level or a field level.

The organization of the data in a property graph database also makes it possible to view data from an aggregate data perspective, providing an overview while being able to "zoom" in giving a detailed view of e.g. a patient. Capish Reflect is also very flexible in the way data can be represented to the end user. Dependent on the purpose, data can be shown as part of complex visualizations, tables or in simple list boxes and panels.

A free text search box can be used to search for data values in any Holon and more specific data can be accessed by clicking on a data value inside a visualization, table or list box.

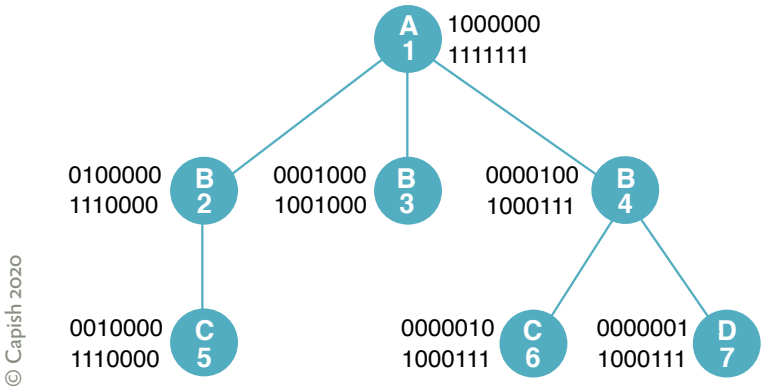
Capish Reflect offers multiples ways of taking action on selected data. While we are all familiar with such an action resulting in a set of filtered data, Capish Reflect also offers the possibility to simply highlight the search result as part of the full data set. This allows the end user to either focus on the results or view the data and metadata in context, opening up a whole new world of insights.

CAPISH REFLECT MAKES DATA FINDABLE, INTEROPERABLE AND REUSABLE

The property graph database backend matched with several indexes, makes it possible to provide the end user with multiple ways to query and interact with the data, all through a "point and click" interface or simple search boxes. Capish Reflect also offers a unique way to combine search terms for more complex searches. End users can combine search terms either by "OR", "AND" or "AND NOT" for more complex queries without ever writing a single line of code.

In line with the FAIR data principles data needs to be interoperable both for people and machines. Since Capish Reflect is backed

by a property graph database this can be easily achieved. The availability of additional metadata, providing descriptions and additional information about data points gives end users the ability to "look up" the meaning of the data directly within the application.



The patented technology behind Capish Reflect combines the property graph database with a set of binary indexes for the individual Holon, the Holon type, individual field values as well as the existing relationships for an efficient and fast way to navigate and query the data.

The fact that Holons enable the use of synonyms makes it possible to use a more verbose expressive language for end users and a more concise machine friendly language for software agents, bridging the problem that most applications have, when trying to cater for both humans and machines. The solution also caters for additional interoperability as Capish Holons can be stored as XML files, a format readable by the human eye, suitable for archiving and as a transport format.

With data stored in an ontological data model reusability is easy in Capish Reflect. The ability to find data easily and to make sense of it due to the additional information provided, makes data stored within the application truly reusable.

Capish Reflect beyond FAIR data principles

Having a FAIR data repository is obviously the first step to good data stewardship but not necessarily the only step. It is obviously important to be able to find relevant data by having a system that can be easily queried and the returned data is fully annotated for the end user making it understandable at any given point in time.

Meanwhile, data integration and retrieval are currently still the bottleneck of most data analysis projects. For that reason, the search capabilities in Capish Reflect go beyond the FAIR data principles by e.g. utilizing reflective logic that exploits the existence of indirect relations between information units using the existence of a common information unit as a reflection point.

A search for a particular patient characteristic would not only retrieve all the information units fulfilling this search criterion but in effect retrieve all patients with that characteristic and on reflection all information relating to these patients.

Hopefully this article has given you some insight into what FAIR data principles are and why we need them. In a nutshell, the FAIR data principles provide a framework to ensure data is reusable, ensur-

Reflect on your data. Peter Tormay. Presentation at the annual PHUSE conference 2016. http://www.phusewiki.org/docs/Conference%202016%20TT%20Papers/TT06.pdf [last accessed 03/11/2020]

ing continuous value from the available data - a central paradigm of good data stewardship. Capish Reflect is a solution that has been built around a property graph database with data fairness in mind thus providing the technical implementation to support your data stewardship efforts.

Capish Reflect is the ideal solution when it comes to data integration and exploration due to its underlying data structure, its query and visualization capabilities. If you want to find out more, visit our website or even better contact us for a demo.

On behalf of Capish,
Peter Tormay, London, November 2020

Contact Capish:

Capish
Lokgatan 8
211 20 MALMÖ
info@capish.com

Eva Kelty, CEO
eva.kelty@capish.com