PhUSE 2014

Paper PP35

The Capish Information Model – Simplify Access to Your Data

Anna Berg, Capish Nordic AB, Malmö, Sweden
Catharina Dahlbo, Capish Nordic AB, Malmö, Sweden

ABSTRACT

Opening up clinical data requires the ability to easily access and analyze data, irrespective of its location or the application that created it. In order to achieve this goal, data needs to be integrated from different sources and mapped to a clear and understandable common data framework. With unstructured data becoming more and more prevalent in clinical research, it becomes a prerequisite that the data framework has the ability to include all kinds of data, *e.g.* documents, comments, images etc. Capish has developed such an information model constituted of small messages of information, which are linked together by pre-defined relations. On top of this model is a well-defined terminology with the possibility to include existing standards and terminologies.

In this paper, the opportunities for creating clinical data transparency by integrating data to a well-defined, source-independent information model will be discussed. Also how challenges in protecting patient privacy and intellectual properties can be overcome.

INTRODUCTION

The European Medicines Agency has announced that it will proactively publish clinical trial data and enable access to full data sets by interested parties. In this paper, four main challenges for achieving clinical data transparency have been identified and are being discussed:

1. EASY ACCESS – Bridging the gap between user and data
    a. Integrate disparate data sources and data types
    b. Use a common information model and terminology
    c. Provide intuitive data navigation capabilities
2. DATA QUALITY - Assure the data has not been modified
3. PATIENT PRIVACY - Preserve patient integrity
4. INTELLECTUAL PROPERTIES - Protect intellectual property

The main focus of this paper is a discussion on how to bridge the gap between the different data sources and the interested user, providing the user with easy access to the data, see figure 1. Many of the existing standards and formats for storing data are designed for data from only one knowledge domain (*e.g.* CDISC) or created for a specific purpose (*e.g.* MedDRA, LOINC), which raises the need for a general data framework. To further achieve full transparency, it is also important to give interested users direct access to the data, thus minimizing the need to use database experts before being able to explore the data.

***Figure 1**. Bridging the gap between the user and the data.*

EASY ACCESS

Data transparency means that data shall be made accessible to the public. Though, simply opening up a clinical study for the public is in most cases not sufficient for a user to make sense of it, since data often resides in differently structured datasets and is coded in various standards. This means that although the data is available, an average user cannot fully use the information and explore the data. A way to provide the user with easy access is to integrate the data to a common information model and terminology, see figure 2.

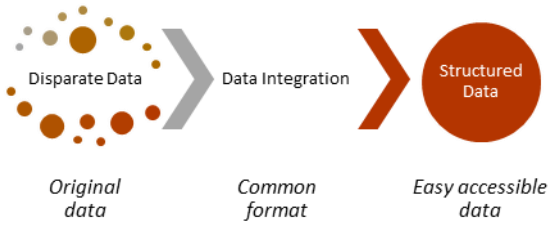

Figure 2. *Creation of structured data, which may be easily accessed.*

Patient information is unbounded in nature. The information may contain anything from sodium concentration in plasma, the date for signing an informed consent, to images from an X-ray. Putting such broad scope of information into tables requires a vast amount of tables and columns within these tables. In this aspect, medical information is completely different from normal company information that is primarily transactional in nature, like financial data, HR information or production systems, which can be well described in a relational database schema. Traditional relational databases require a database schema and handle lots of data of the same type well (new rows in a table), but when it comes to collecting information from individual patients, another data structure serves as a better solution.

In the clinical environment there is a multitude of data standards and terminologies, some of them are specific to a certain concept and some are more general. Many of the existing standards and formats for storing data are designed for data from only one knowledge domain. This means that even if each dataset is in a given standard and structure, many datasets might be difficult to combine.

DISPARATE DATA

Data can be disparate and heterogeneous in many ways, ranging from the database being used to differences in how the data is modelled within the same database. The most obvious heterogeneity is the technical one, where data is stored in completely different databases. These databases are often incompatible and differ in file format, access protocol, query language etc.

The sometimes not so obvious heterogeneity between datasets is differences in the data model being used. Even if the data is physically stored in the same database, there might have been different ways of representing and storing the same data. When looking at data structures, the database may be normalized to different degrees, resulting in data being decomposed in many ways. For example, “blood pressure” can be stored in a table together with other variables (*e.g.* “vital signs”) or in its own table. Also the design of where to put standards etc. can vary, for example whether a measurement scale shall be explicitly included in a field or be implied in a table elsewhere. Another way of data model heterogeneity is whether a generic or non-generic model has been used.

Also the nomenclature being used may differ, although it is said to be in the same standard or format. Column names (data labels) might be different, though still have the same meaning (semantics). Also the data encoding schemes (data content) may vary, with different controlled terms, value lists etc.

DATA INTEGRATION

Data integration involves combining data residing in different sources and providing users with a unified view of this data. Creating mapping-rules to combine disparate data is sometimes complicated and a time-consuming process, see figure 3. In addition to this, each standard might be updated on a regular basis, which requires a lot of maintenance to keep the mapping-rules up to date. A solution to this is to use a common information model and terminology, which can serve as a hub combining the other standards. This means that after mapping a dataset to this common information model and terminology, one might either keep it in this format for further analysis, or continue exporting the data in another standard.

Capish is using an in-house graphical data conversion tool to ease the process of harmonizing and integrating heterogeneous data. This tool is connected to the Capish® information model and terminology, assuring the output follows the format. It also provides a number of quality checks, further assuring the correctness of the integrated data.

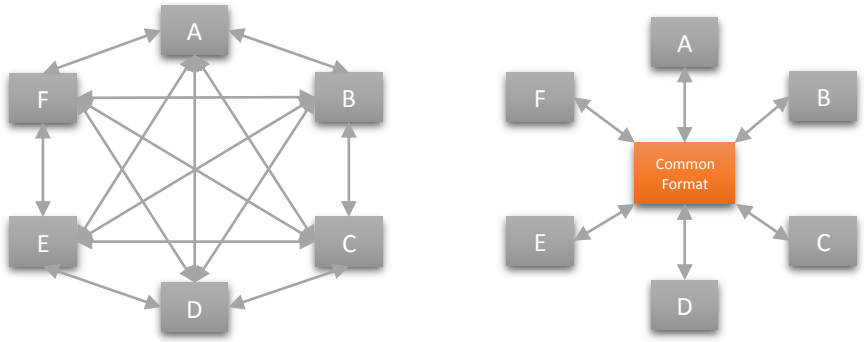

Figure 3. *A common format or data framework for all kinds of data limits the need for mapping "everything to everything" between multiple data sources. Each box in the figure symbolizes a data format or standard.*

INFORMATION MODEL

While browsing through the medical records of an individual patient, it is evident that the different entries can be seen as individual messages, using a specific medical terminology depending on the medical specialty. It is not necessary that every medical doctor will understand all of the information, such as an ECG, a blood enzyme activity, a summary of a quality of life questionnaire or an X-ray. Other important information will be the time of a specific observation, the type of observation or intervention, how it was measured or performed, who made the examination etc. The medical profession is also very semantic, using many specific terms and statements that are given both as observations and comments or opinions on observations. Although the situation looks chaotic from a database point of view, it is not so from a medical point. Medical doctors are actually interpreting this type of information daily, discussing with other doctors from different disciplines, and actively making decisions about treatment or further tests to be added to the medical records.

A way to model this complex reality is to use an information model that mimics the real world behind the collected data. Capish has developed an information model that is constituted of small units of information, which are linked together by pre-defined relations, see figure 4.

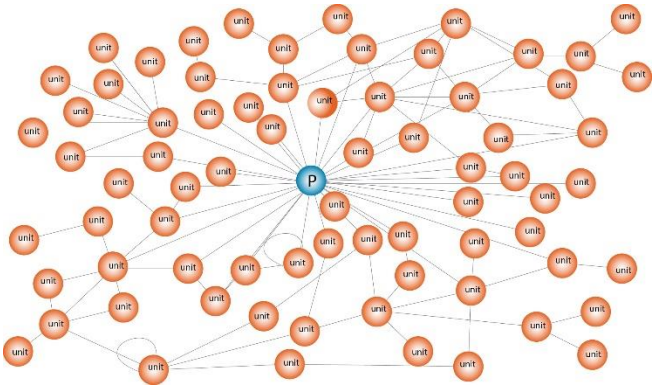

Figure 4. *The Capish information model is constituted of small units of information, which in the case of clinical data could be described as patient-centric, thereby the node called "P".*

The main idea is to model the 'reality', thus the process and environment that created the data. Many existing models are based on how data was collected (*i.e.* reflecting a CRF) or the type of analysis that was requested at that moment. Modeling information in a way that reflects the reality will create a stable, unbiased model that is hypothesis-free and ready to be used for various purposes. Further on, having the information units related to other units removes the need for knowledge of databases before accessing the data. Instead, the information can be found, understood and analyzed directly.

INFORMATION UNIT

The small information messages or units in the Capish information model are called Capish® Holons. The Capish Holon can be defined as the minimum information package a person knowledgeable within the field can understand and use for making a decision. These information units shall be large enough to be self-contained and understandable as-is, but small enough to be combined as "Lego pieces" into a complete description of, for example, a specific patient. The combinations of the Holons are realized by explicit named relations, see figure 5.

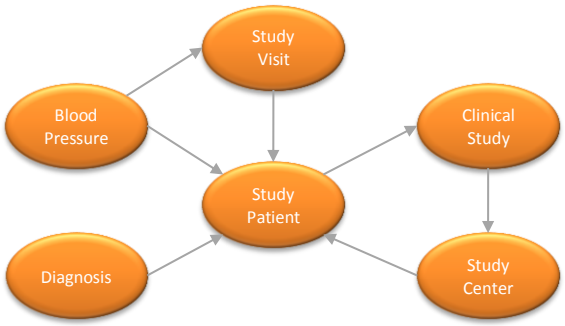

Figure 5. *An example of Capish Holons and how they can be related in the information model.*

One of the key properties of such an information unit is that it must be understandable within a given knowledge domain or discipline. This implies that the words and notions used are known within the domain, so it can be understood by another person trained within the same domain. In this way, every information unit can be said to be unbiased or hypotheses-free, as it states its message "as-is" within the knowledge domain of each and every specialty.

A Capish Holon is typically stated as a structured, schema-less document. At the top there are header information like Object Type, Domain and a globally unique identifier. The knowledge domain is specified by stating a code from the Dewey Decimal Code system (DDC). The message itself is structured into Fields, Values and Attributes. Finally there are possibilities to give named relations to other Holons. The significance of the Capish Holon is that it solves the problem of unbounded information by identifying the stable notions that we use in our everyday reasoning. Hence, it is possible to define and construct information carriers on this level of "human" interpretation and understanding that is more detailed than ordinary database tables, but yet large enough to include a number of individual data points. This limits the complexity to something that both people and computers can handle.

Data input to statistical programs and graphical interactive chart engines are mostly table based. Due to the field structure of every Capish Holon, it is possible to store and recreate any traditional relational database with them. However, tables are the result of a standardized structured report and not the information carrier itself. Because of the possibility to create and export tables from the information model, one can still use any of the many specialized programs that exist on the market today, like SAS, QlikView and Spotfire. In addition, it is in many cases possible to keep a link to the original patient data, which makes it possible to instantly jump from a specific outlier in a graph, back to the information model, where one can easily search and navigate to get a holistic view of all the reported information about that patient.

Keeping the data in small information units also makes it ideal for free or structured text searches on the entire information base. Since the individual information units are self-contained and understandable as such, one will get hits on understandable pieces of information. These hits may instantly be used for navigation via the relations, or turned into graphs directly. This opens up entirely new possibilities for text-based searches and analyses of any textual information in the patient records, which further increases the accessibility of the information.

TERMINOLOGY

Just having the information structured the same way in a common information model is not sufficient to achieve full integration of the data. Both labels and content may vary in many ways, though they might have exactly the same meaning. This implies that a common terminology also has to be used to be able to combine the different datasets. Basically, any terminology that controls both the variable names and their content might be used. However, many of the existing terminologies are not general enough to include the wide range of data that might be covered in the medical record of a patient. Further on, many terminologies are not detailed enough to control every single variable, ranging from controlling physical dimension and unit on a numerical variable, to providing value lists for variables with a limited number of content alternatives. Of course the terminology also has to handle uncontrolled variables, for example those holding free-text, where it is needed. To further give the user access to the information, it is also important to express the data clearly, without coding that forces the user to look up information elsewhere.

Capish is using a well-defined terminology on top of the information model, with the possibility to include existing standards and terminologies. Using an internally controlled terminology, with strict rules for naming of variables etc, assures that every variable is specified enough to avoid creation of duplicate variables. This also makes it easier in the mapping process, where the level of detail is known for each variable.

FILE FORMAT

Considering data transparency, it is an advantage if the resulting files are interpretable without any specific software. Keeping the data in a non-proprietary format also becomes beneficial for digital archiving. One of the key features of the Capish Holons is that they are stored as simple documents in the declarative language standard XML. The strict separation between data and database also makes the Capish information model ideal for digital archiving, since the simple structured documents used surely can be processed many decades from now.

DATA NAVIGATION

After harmonization and integration of the original data, the information is in a structured and well-defined format. To further be able to access and explore the information, an easy-to-use web-based interactive graphical interface is useful. As discussed in the section about the information model, any visualization tool might be used to analyze the information units. However, to fully take advantage of the basic ideas of the Capish information model, a visualization platform has been developed to navigate, explore and analyze the graph of related information units.

When loaded into the platform, the information for every individual patient will be matched and added to the same information of other patients, see figure 6. The results are automatically integrated on the server level and the searching and analyzing can start immediately. In other words, the user can access the information model directly, with no need to transform to and from a data model when searching. It is also possible to use the information efficiently for multiple purposes, even such purposes that are unknown at time of design.

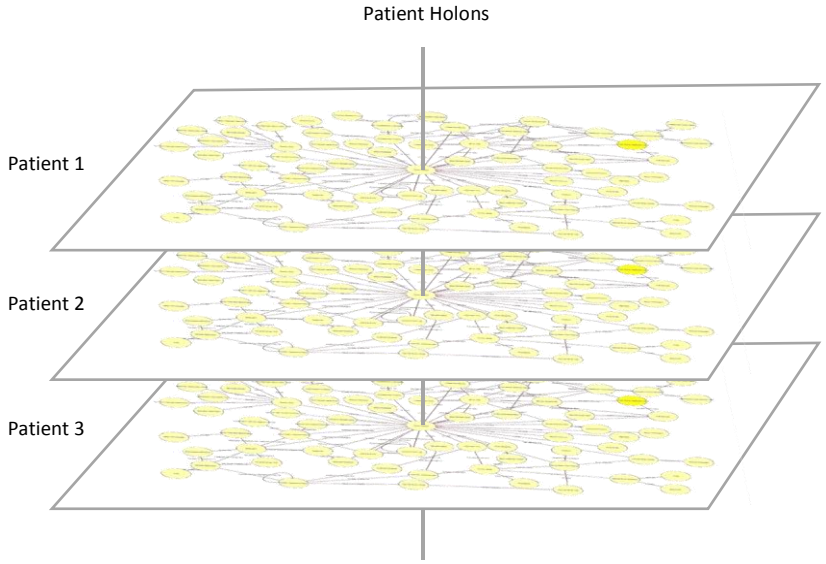

Figure 6. *The information for every individual patient is matched and added to the same information of other patients.*

The data platform is further designed to give the user advanced search capabilities, ranging from simple value searches to free text searches, which can search through the entire information content. Another advantage is the use of a reflective logic, where the user has the ability to choose centricity of the information model. The reflective logic is used to put a specific type of Capish Holon in the center of the query, making it possible to reflect the answer in any type of Holon. Using a reflective search logic will return information that would be hidden by the query if normal search logic was used.

The most useful case would probably be to reflect the search in the patient-Holon, thus retrieving the answer in distinct patients having a match of the search query somewhere in their data, see figure 7. Note that the information model as such is not centered, the centricity is chosen when looking at the data. Thus, a patient-centric view might be quickly achieved, or (if the user whishes) a “clinical study”-centric view might as well be easily created by the user itself. All dependent on the purpose and interest of the user.

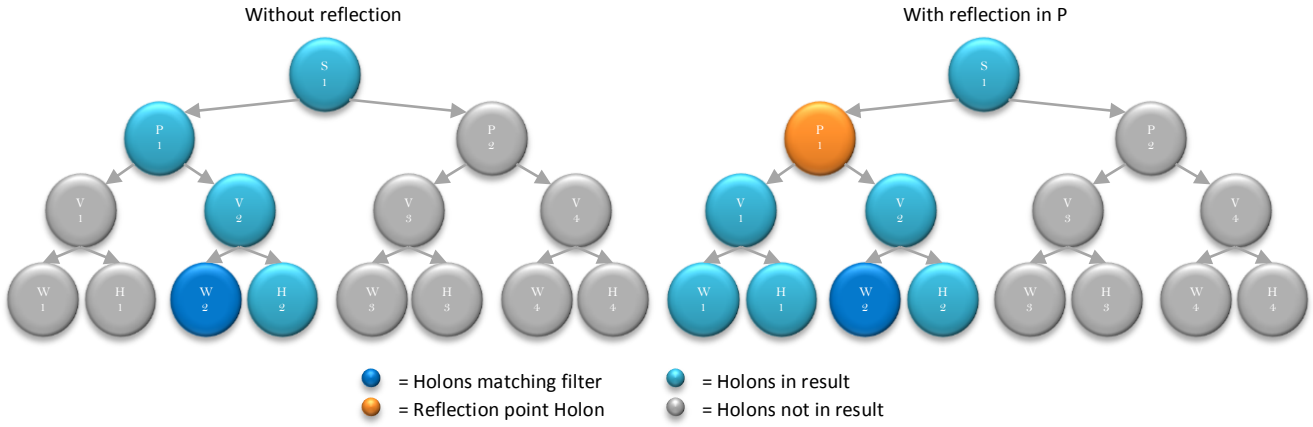

Figure 7. *An example of a search for a specific body weight (W2), performed without and with reflection in the patient-Holon. When using reflection in the patient-Holon, all Holons related to the patient will be found in the result. The Holon marked with an S is representing the Study, P is Patient, V is Visit, W is Weight and H is Height.*

On top of the data platform are applications that give the user an intuitive interface, where it is possible to move from detailed patient information to analysis of the entire population. As an example, an outlier in a population graph can be chosen and further drilled-down to view everything that has happened to that patient. From the characteristics of that patient, cohorts of similar patients can be created to further explore and investigate a specific question.

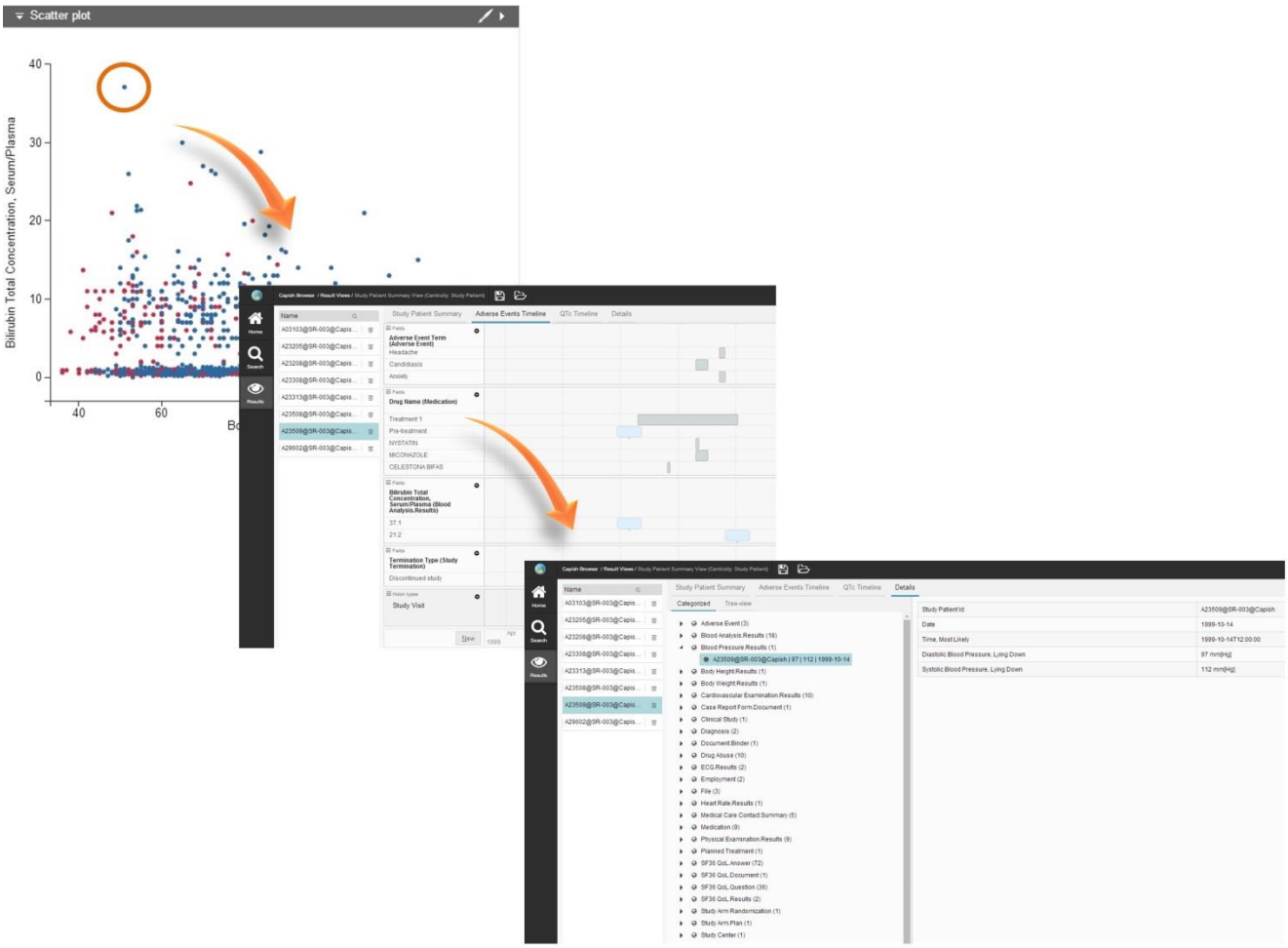

Figure 8. *An example of the drill-down principle, where it is possible to identify data points of interest in aggregate graphs and tables and drill down to the relevant patient. Screen-shots taken from the Capish applications.*

DATA QUALITY

An important challenge of data transparency is the ability to trust the quality of the data. One might want to be able to start with a result and then trace the way back to the processes that delivered the output. Or the other way around, choosing a process and trace all results. An information model constituted by small information units supports this way of free navigation between related units. Thus, it does not only add knowledge to the user, it is also congruent with explicitly finding all the information necessary to easily fulfill traceability and high quality standards.

PATIENT PRIVACY

Making clinical data accessible for the public raises the need for making sure that patient and other personal information is adequately protected. A solution to this is to handle "roles" in the information model, where a person is only represented by its role (*e.g.* a clinical study participant or a patient within health care), removing the need for displaying a personal id. While modelling the 'reality', it is of course always a real person that has taken part in a clinical study or visited the health care, but to have the possibility to easily add or remove certain data, it is convenient to divide data into a 'basic Holon' holding the personal information and 'role Holons' for the different roles a person can have, see figure 9.

The roles of one and the same person are of course related, since the same person can have different roles in different situations. When publishing the data, the role "person" might easily be hidden or removed (or never being added from the beginning at all).

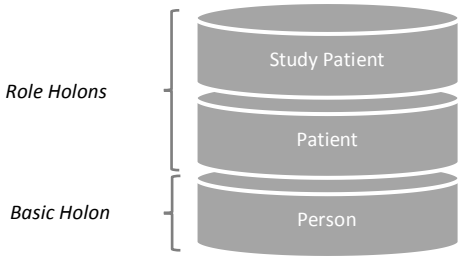

Figure 9. *A person might only be represented by a role, dependent on the circumstance.*

In the Capish information model, data from each patient can be stored in individual XML-files. While storing the information in XML-files, the information as such is completely separated from the way in which they are indexed and used. The programs used for handling the information are schema-less, meaning that they will index and present any information that is fed into the system "as is". So, instead of having complicated scripts for entering or removing a patient to a great number of tables, one just has to add or remove one more file in the system. In this way, a specific patient, or selected parts of a patient's medical record, can easily be removed from the system to ensure patient privacy.

INTELLECTUAL PROPERTY

Another issue of many pharmaceutical companies is that clinical data transparency will ruin intellectual property and innovation. By keeping the data in small information units, it is (as discussed in the section about patient privacy) easy for the owner of the data to decide which units shall be open or hidden. As an example, the data owner might want to hide the efficacy results. Then it is easy to simply hide or remove the actual information units handling that part of the data, see figure 10.

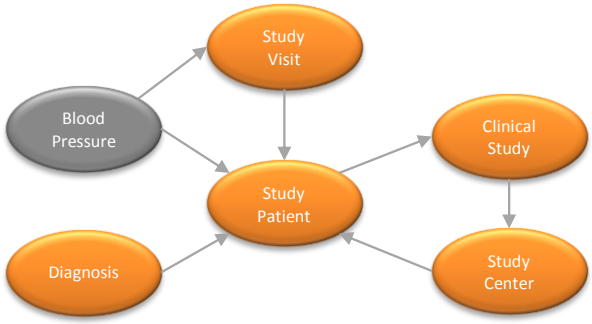

Figure 10*. The file containing the Holon holding measured blood pressures might easily be removed from the information model, with no need to modify the other Holons.*

CONCLUSION

Data transparency means that data shall be made accessible for the public. For a user to be able to access and make sense of the data, it has to be in a clear and well-defined format - and, above all, it has to be in the same format. Keeping all kinds of data in the same model is not a simple task, especially since medical data can cover a wide range of domains and originate from completely different sources. A solution that has proven to be useful in the integration of very disparate data, is to use an information model that is built on the actual reality where the data was created. Many existing models are designed based on how data was collected (i.e. reflecting a CRF) or the type of analysis that is requested at that moment.

Another important aspect while making data accessible for the public is to minimize the need of database knowledge before being able to explore the data. The information should be expressed in the domain language, making it open for any user (with access rights) who is having an interest in the data. Today many non-database experts cannot access their own data without using programmers or statisticians, this extra step limiting the way they can explore and play around with their data.

One of the advantages of keeping the data in an information model that is reflecting the reality is that it makes it hypothesis-free, only reflecting the reality behind the information. Further on, by keeping the information in small units, it is easy to choose which parts shall be open or hidden to, for example, protect patient privacy and/or intellectual property.

In addition, having all relations explicit makes it possible to search information without knowledge of the model behind the data. This will also speed up the searching process, since no joins are required to relate information. Further on, there is no need to create new applications or data marts in order to use the information for a new purpose - all information needed is already in the model.

CONTACT INFORMATION

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

Anna Berg

Capish Nordic AB

Stortorget 9

SE-211 22 Malmö

Work Phone: +46 (0)40 10 88 80

Email: anna.berg@capishknowledge.com

Web: www.capishknowledge.com