Concept Maps: Theory, Methodology, Technology
Proc. of the Fifth Int. Conference on Concept Mapping
Valletta, Malta 2012

DM-CMAP: A CONCEPT MAP-BASED TOOL TO DESCRIBE, ENHANCE, AND SHARE HIGH LEVEL DATA MODELS WITH INFORMATION CONSUMERS

Rodrigo Carvajal, Scott Morgan, Carlos Pérez & David Fenstermacher
Moffitt Cancer Center & Research Institute, USA
www.moffitt.org

Abstract. When presenting a data model to information consumers (non-data processing personnel such as statisticians, molecular biologists, geneticists, oncologists, and epidemiologists), often times we find ourselves describing information systems and table relationships verbally, while standing in front of and pointing at a large printed or projected copy of the model on a conference room wall. The software development life cycle (SDLC) slows down because the engineers don't have the "right tool" to communicate and explain the system architecture and data structures to the researchers. An analogy between concept maps (Novak & Gowin, 1984) and database schemas has been used to describe complex data models as concept map-based knowledge domains using IHMC CmapTools (Cañas et al., 2004). The use of concept maps facilitates the communication between system developers and information consumers. This article introduces DM-Cmap, an algorithm that represents Data Models as Concept Map.

1 Introduction

The implementation of Comparative Effectiveness Research (CER) as a decision-making process for physicians and researchers to determine the best diagnostics and therapeutics for subpopulations of patients in a medical facility requires the design of complex database infrastructures to incorporate large observational patient-based data from multiple sources including electronic medical records, molecular data, radiometric images, insurance claims, registry information, and molecular characterizations of disease (Fenstermacher, Morgan & Carvajal, 2011).

Development of CER systems involves continual interaction between information technology engineers and researchers from different disciplines. Observations recorded while gathering system requirements indicated that some users had trouble identifying the relationships between the different tables, the data types or attributes, and the availability of queries and reports in the database. Data modeling tools (i.e. Erwin, SLQ Server ER Diagrams) and database diagramming tools do not provide the desired level of data navigation and comprehension needed by the users.

By using IHMC CmapTools (Cañas et al., 2004) an initial set of concept maps (Novak & Gowin, 1984), were created to describe at a summary high level, and subsequent levels of detail, the data model abstractions and relationships used for the CER database. By using non-technical linking phrases, researchers and information consumers are able to understand how the data has been organized and navigate to deeper levels of detail in the areas they are most interested in. Taking advantage of the IHMC CmapTools feature that allows adding links to other pertinent content, a set of detailed data models, data dictionary reports, and sample data listings (reports) were made easily accessible as resource links from the concept maps (Fenstermacher, Morgan & Carvajal, 2011).

2 Data Models and Entity Relationships Diagrams

For the purpose of this article, we understand an Entity – Relationship (ER) Diagram as the abstraction of a database. ER Diagrams can represent data models. Data models typically fall into one of three categories:

- Conceptual Data Models group together subject areas of information and show the relationships between the subject areas in high-level business terms. It includes all the components to be included within the system.
- Logical Data Models define how collected data will be organized and stored. Tables (entities) collect information in columns (attributes). Columns are identified by types (i.e. string, numeric, date, time, boolean) and sizes. Unique row identifier keys are identified and relationships between tables are built by linking those keys. Data quality rules are defined and specified using data types, default values and other constraints ("referential integrity") to improve accuracy and reliability.
- Physical Data Models represent what the physical database looks like. The information is stored in the schema associated to the database. The physical data models facilitate the implementation of the Logical Data Model within a specific computing environment, adhering to the environment standards and syntax.

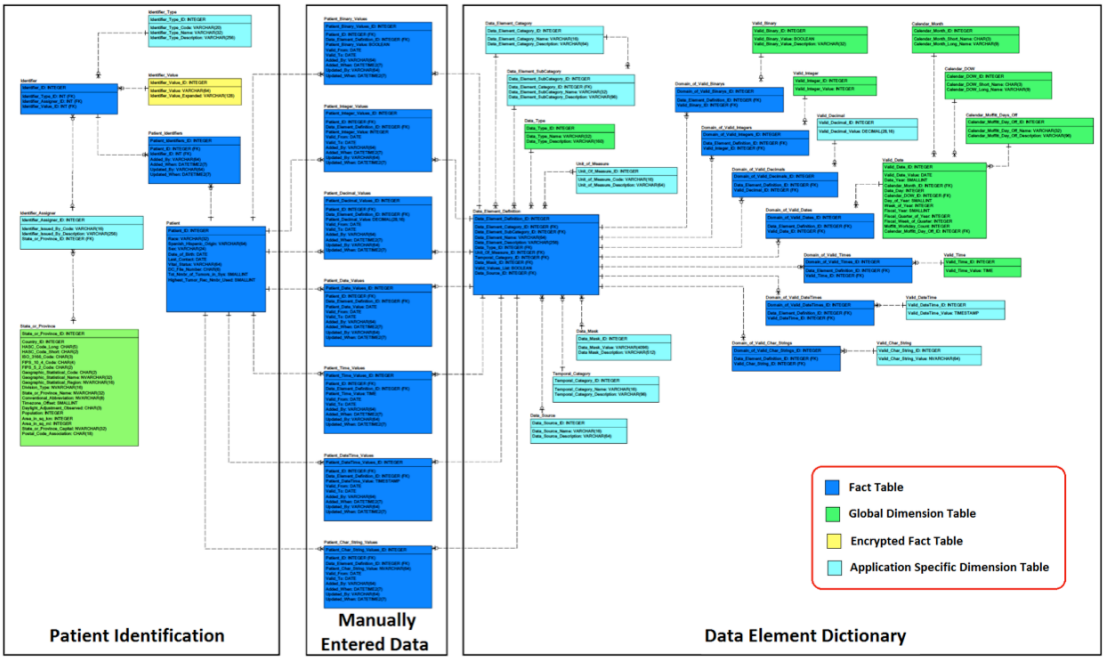

Figure 1. Entity Relationship Diagram: Abstraction of the data model including relationship between tables.

An example of an ER Diagram created with CA ERwin can be found on "**Figure 1**. Entity Relationship Diagram: Abstraction of the data model including relationship between tables." In the diagram, three subject areas are defined in the Conceptual Data Model: Patient Identification, Manually Entered Data, and Data Element Dictionary.

3 Concept Maps and Data Models Analogy

Concept maps are graphical tools for organizing and representing knowledge (Novak & Gowin, 1984). In our case concept maps are used for organizing and representing databases. Two or more concepts (nodes) connected with linking phrases or verbs establish a meaningful statement termed a semantic unit or proposition. A concept map can be defined as a collection of propositions. Each concept can contain links to additional concept maps creating a data model and metadata hierarchy along with links to digital resources such as informative documents, images, web pages, business intelligence reports, and ER diagrams.

Based on Cañas et al. (2005): "*Concepts tend to be nouns and linking phrases are usually verbs, and it is recommended that both consist of as few words as possible. Linking phrases can express any type of relationship, and are not limited to a defined set (e.g., is-a, part-of, etc.) as in other diagramming techniques such as semantic networks*" an analogy was developed in order to map data models, including the tables definitions and their relationships. See "Figure 2. Concept map and table definition generated using DM-Cmap". Four main elements from the database schema are represented as concepts:

- Subject Areas: They are defined in the Conceptual Data Model. A concept map is created per Subject Area and this concept map contains links to all its table members and to other Subject Areas sharing at least one table. In a table concept map, a Subject Area is represented as a concept with linking phrases to its associated tables and a resource link to a concept map describing the Subject Area. A different color is assigned to concepts representing subject areas.
- Tables: A table is a set of data elements (fields or columns). Each table is represented as a concept. Table nodes have linking phrases to and from Subject Areas and to columns. Concepts representing tables have resource links to web pages describing the metadata associated with the table, to business intelligence reports and other supportive documentation.
- Data Elements or columns. Each column is represented as a concept. Depending on the data type and meta-information stored in the schema, a different color is assigned. Data Elements have incoming

links from table concepts. When the column refers to a Foreign Key (i.e. Patient_ID), a resource link to the table (schema and concept map) is generated.

- Data Types. A concept per data type specified in a table can be added to the concept map. Only column names have links to data type concepts.

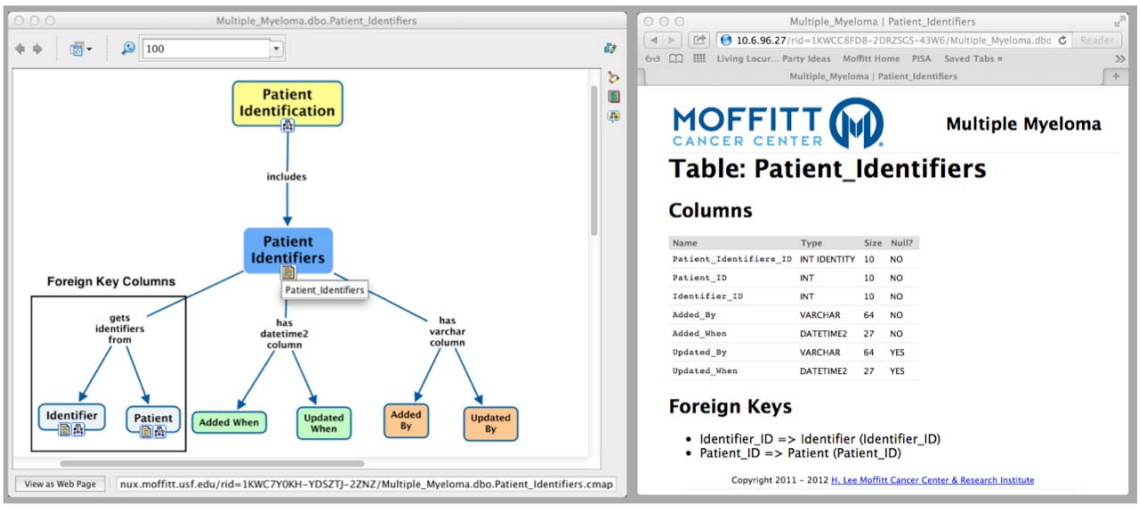

Figure 2. Concept map and table definition generated using DM-Cmap

Using Concept Mapping Extensible Language (CXL) (Cañas et al., 2006) and XML-based configuration files, the CmapTools Styles (color, font, line, background, text alignment, etc.) are set for each group of elements as described previously. As part of the configuration file, linking phrases between the elements can be configured as well. Linking phrase types are described as follows:

- From Subject Area to Subject Area. Tables can be member of more than one Subject Area. The linking phrase denotes that two Subject Areas share at least one table.
- From Subject Area to Table. It identifies the belonging of a table to a Subject Area.
- From Table to Table. Identifies foreign keys or identifiers used by a table.
- From Table to Column. The data type of the column can be part of the linking phrase, otherwise a concept labeled with the data type name is added to the concept map.

In "Figure 2. Concept map and table definition generated using DM-Cmap" the details of the data dictionary are revealed and enhanced with a link to a web page generated by DM-Cmap based on the schema stored in the database. Concept maps generated by DM-Cmap provide an intuitive navigation across data domains and one can drill down to access detailed information as desired.

4 DM-Cmap Architecture

Given the constant changes in the data model, especially during the initial tasks (planning, requirement analysis, and implementation), an automatic way to generate the concept map-based version of the data model was needed. A program called "DM-Cmap" (from Data Model to Concept Map) was created to generate concept map-based knowledge domains of databases.

DM-Cmap uses Java as the programming language, SQL Server as the RDBMS, CA ERwin as the data modeling tool, IHMC CmapServer (Cañas et al., 2004) are used to generate and store the concept maps. Concept maps generated using DM-Cmap can be customized based on two configuration files. The CXL Style Sheets file defines the web pages of the table schemas and the style palette components available in CmapTools. Those most used in our implementation are: connection types, colors, fonts, arrowheads, and alignment. The application configuration file is Java properties file. It provides the customization of the linking phrases grouped by type of data element (i.e. subject area, table, column data type), the connection string to the database (server name, database, username and password), and the CmapServer information (server name or IP address, username, password, and folder to host the knowledge model).
The process to generate the concept maps is as follows:

- The database schema is loaded
- Table definitions are mapped to the configuration files
- Text and webpages files containing table definitions are generated
- Concept maps are generated and stored in the CmapServer
- Links between concept maps and resources are created

5 Conclusions and Future Work

DM-Cmap generates a highly interactive concept map-based interface that reveals a comprehensive data dictionary detailing subject areas, tables, and columns of a database, while also exposing the physical and contextual metadata data elements through links to digital resources. Using DM-Cmap around 400 tables have been represented as concept maps and made available from a private CmapServer.

Data Models represented as concept maps allow the information consumers or researchers to "browse" the data model at their leisure from their own work area, using either the web version of the concept maps or IHMC CmapTools, whenever they have a question about what data is stored where and what does it look like.

At the current implementation, DM-Cmap is only used by demand by the developers when a significant change is made to the data model. The next step in the development of DM-Cmap will be the implementation of an automatic mechanism that triggers the generation of the concept maps when the data model changes. Also, the DM-Cmap will be integrated with the CA-ERwin Web Portal in order to add custom ERwin reports and views as resources associated with the table's nodes.

6 Acknowledgements

This Research Project was supported by the National Cancer Institute – Developing Information Infrastructure Focused on Cancer Comparative Effectiveness Research 5-UC2-CA148332-02.

7 References

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Eskridge, T., Arroyo, M., Carvajal, R. (2004). CmapTools: A Knowledge Modeling and Sharing Environment. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping (Vol. I, pp. 125-133). Pamplona, Spain: Universidad Pública de Navarra.

Cañas, A. J., Carff, R., Hill, G., Carvalho, M., Arguedas, M., Eskridge, T., Lott, J., Carvajal, R. (2005) Concept Maps: Integrating Knowledge and Information Visualization. In Knowledge and Information Visualization: Searching for Synergies, S.-O. Tergan, and T. Keller, Editors. Heidelberg / New York: Springer Lecture Notes in Computer Science.

Cañas, A. J., G. Hill, L. Bunch, R. Carff, T. Eskridge, C. Pérez, KEA: A Knowledge Exchange Architecture Based on Web Services, Concept Maps, and CmapTools, In A. J. Cañas, J. D. Novak (Eds.), Concept Maps: Theory, Methodology, Technology, Proceedings of the Second International Conference on Concept Mapping, San José, Costa Rica (September 5-8, 2006), Editorial Universidad de Costa Rica, pp. 304-310.

Fenstermacher, D., Morgan, S., Carvajal, R. (2011) Creating a Concept Map Interface To Visualize and Interpret a Comparative Effectiveness Data Model. 19th Annual International Conference on Intelligent Systems for Molecular Biology. Vienna, Austria.

Novak, J. D., & Gowin, D. B. (1984). Learning How to Learn. New York: Cambridge University Press.