e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-368

A Reference Data Model of a Metadata Registry Preserving Semantics and Representations of Data Elements

Martin LÖPPRICH[a,b,1], Jennifer JONES[a], Marie-Claire MEINECKE[b], Hartmut GOLDSCHMIDT[b], and Petra KNAUP[a]
[a] *Institute of Medical Biometry and Informatics, Heidelberg University, Germany*
[b] *Department of Internal Medicine, Division of Hematology/Oncology/Rheumatology, Heidelberg University Hospital, Germany*

Abstract. Integration and analysis of clinical data collected in multiple data sources over a long period of time is a major challenge even when data warehouses and metadata registries are used. Since most metadata registries focus on describing data elements to establish domain consistent data definition and providing item libraries, hierarchical and temporal dependencies cannot be mapped. Therefore we developed and validated a reference data model, based on ISO/IEC 11179, which allows revision and branching control of conceptually similar data elements with heterogeneous definitions and representations.

Keywords. metadata registry, reference data model, heterogeneous data collection systems, data integration

Introduction

The integration and analysis of clinical data collected in various systems and over a long period of time has become a major challenge in medical care and research [1, 2]. The increased use of data warehouses simplifies this process by aggregating current and historical data into a central data repository [3]. However, data recorded in different systems, relational databases and spreadsheets is likely to be heterogeneous in its structure, inconsistent in its definition and often described insufficiently, which makes data integration a complex and time consuming task.

Metadata registries (MDR) can help to define the characteristics of data elements like their meaning and system-specific representation. Within *ISO/IEC 11179, Information Technology -- Metadata registries (MDR)* the structure of an MDR is described in detail, but it does not address the characteristics of metadata [4]. Thus, several solutions emerged, implementing ISO/IEC 11179 according to their specific purpose [5, 6]. Since all of these solutions focus on describing data elements to establish domain consistent data definitions, they are limited in representing variations of heterogeneous data elements across multiple data sources and over time. This makes it impossible to track a single data element, recorded e.g., for decades in various definitions and systems, with its modified semantics and representation. As an example,

[1] Corresponding Author: Martin.Loepprich@med.uni-heidelberg.de

the documentation of the data element "diagnosis" changes when the classification renews, diagnostic procedures improve or simply the number of included diagnoses extends. Managing and mapping complex and heterogeneous data elements like "diagnosis", without a tool that supports the harmonization process, is even with modest-sized datasets almost impossible [5].

The main objective of this paper is to find a reference data model, based on ISO/IEC 11179, which allows generalizing data elements, preserving their semantics and representations, and identifying hierarchical and temporal dependencies. According to this specification we developed an MDR and validated it by entering all data elements recorded over two decades in multiple data collection systems in the Department of Multiple Myeloma, Heidelberg University Hospital. The data elements result from clinical routine care, observational and interventional studies of diagnosed and treated patients with multiple myeloma.

1. Methods

The data model of our MDR is based on open-source tools and on *ISO/IEC 11179, Information Technology -- Metadata registries (MDR)*, published by the International Organization for Standardization (ISO).

1.1. ISO/IEC 11179, Information Technology -- Metadata registries (MDR)

ISO/IEC 11179 defines a conceptual model to describe metadata, but does not specify an implementation. Metadata is regarded as data that defines and describes other data. A *Data Element*, which is the atomic unit of data that exists in a data collection system, is associated with a so-called *Data Element Concept* and a *Value Domain* according to the ISO/IEC standard. While the *Data Element* corresponds to a particular object described by its identifier, name, definition and other attributes, the *Data Element Concept* represents its semantic component and the *Value Domain* its representational component. A *Data Element Concept* is an abstract unit of knowledge, described independently of any representation that can be associated with multiple data elements. For example, certain conditions may require linking the data element "blood sugar" and "blood glucose level" to the same concept. The *Value Domain*, includes the data type (e.g., number, character, or date), unit of measure, permissible values, and permissible length (e.g., number of characters). In case of an enumerated list, the *Value Domain* is extended by a *Conceptual Domain*, which provides the permissible values of the enumeration.

1.2. Development Tools

For implementation of our MDR we used Valentina Studio as a visual database design tool, and PostgreSQL as a relational database management system because both are specific suited in their functionality to develop and edit a reference data model in a flexible manner and to convert it to a relational database for validation purposes. Both tools can be replaced by products with similar functions.

1.3. Verification datasets

The Department of Multiple Myeloma, Heidelberg University Hospital, has established a documentation system for recording myeloma-specific clinical data since 1992. Over these years, one of the most comprehensive and detailed clinical dataset for diagnosis and therapy of multiple myeloma has emerged. The dataset is used for scientific evaluations and enhanced by cytogenetic, imaging, or molecular data. Datasets exist in various file formats, were recorded with different purposes and guidelines, and are heterogeneous in data structure, definition, and quality. Therefore data have to be generalized and harmonized carefully before integration into a data warehouse.

2. Results

2.1. Reference data model of a metadata registry

The reference data model of our MDR is shown in Figure 1. It consists of five tables, with *DataElement*, *DataElementConcept* and *ValueDomain* as the most important ones.

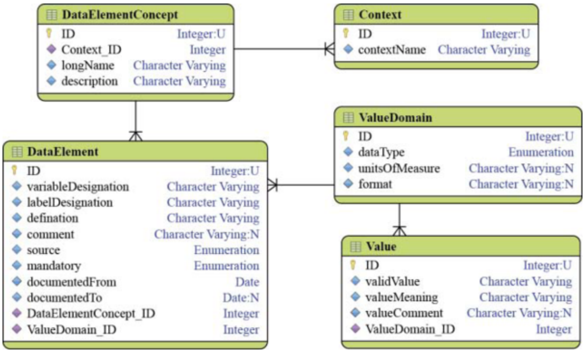
DataElementConcept
ID Integer:U
Context_ID Integer
longName Character Varying
description Character Varying
Context
ID Integer:U
contextName Character Varying
ValueDomain
ID Integer:U
dataType Enumeration
unitsOfMeasure Character Varying:N
format Character Varying:N
DataElement
ID Integer:U
variableDesignation Character Varying
labelDesignation Character Varying
defination Character Varying
comment Character Varying:N
source Enumeration
mandatory Enumeration
documentedFrom Date
documentedTo Date:N
DataElementConcept_ID Integer
ValueDomain_ID Integer
Value
ID Integer:U
validValue Character Varying
valueMeaning Character Varying
valueComment Character Varying:N
ValueDomain_ID Integer

Figure 1. Reference data model of a metadata registry preserving semantics and representations.

All tables are identified with a numeric and unique **ID**. The remaining attributes describing a data element in table *DataElement* are:

- **variableDesignation:** symbolic name or designation of a data element used in the source code; normally only known by the administrator
- **labelDesignation:** refers to the data element´s name as it is visible on the graphical user interface
- **definition:** text field to describe meaning and semantics of a data element
- **comment:** text field in case of additional information, that is not part of the definition but critical to know
- **source:** a selection list, created by the administrator of the MDR, consisting of all possible data collection systems from which data elements are registered
- **mandatory:** three-valued logic (yes, no, unknown) to indicate that the data element is mandatory
- **documentedFrom:** date of recording the data element
- **documentedTo:** date up to which the data element has been recorded; if documentation is still ongoing, the unknown end date can be left blank

The characteristics of a data element concept, which can be associated with multiple data elements, in table *DataElementConcept* are the following:

- **Context_ID:** to assign multiple concepts to a single domain, we use a table named *Context*, which consists of a unique identifier and a name, to serve as a selection list, created and managed by the user of the MDR; the context is thereby on a more abstract level than the concept
- **longName:** abstract designation related to the corresponding data elements; in case a concept is only associated with a single data element, the concept's name corresponds to the data element's label
- **description:** abstract description related to the corresponding data elements; in case a concept is only associated with a single data element, the concept's description corresponds to the data element's definition

The representational component of a data element, in terms of its system-specific view, is described by table *ValueDomain*, which consists of these fields:

- **dataType:** selection list, created by the administrator of the MDR, consisting of all possible data types to which a data element can be assigned
- **unitsOfMeasure:** selection list, created by the administrator of the MDR, or a text field of the units of measurement. Whether a selection list or a free text is more appropriate should be decided depending on the expected number of different units and the level of standardization.
- **format:** to specify the permissible length of characters and of the number of decimal places; wildcards such as ?, * and # and decimal marks are used

If the value domain comprises a list of enumerated or ordinal items, the permissible values are specified in the table *Value*. The fields are described as follows:

- **validValue:** designation of the permissible value or the value itself; since values are often coded by an integer in data collection systems, and only the integer is stored instead of the value itself, both are accepted
- **valueMeaning:** a text field to describe meaning and semantics of a valid value
- **comment:** a text field in case of additional information, that is not part of the meaning but critical to know

2.2. Verification of the reference data model

For verification of the reference data model we entered all data elements from seven data collection systems used over the last two decades in the Department of Multiple Myeloma, Heidelberg University Hospital. Upon completion we obtained 660 data elements, assigned to 450 data element concepts with 500 value domains.

As an example, the data element "diagnosis" was recorded continuously in all data sources. Over the time it was expanded from Boolean to an enumeration list with three permitted values, later with five values, and was modified by adding the cancer stage to the diagnosis and lastly changed in 2013 from ICD-10 to ICD-11.

The different versions of the similar data elements were mapped in our reference data model by adding all seven data elements individually with their specific attributes and representations, and then linked to the data element concept "diagnosis" and described as "relevant diagnoses to be documented". Evaluation of the "diagnosis" data can easily be achieved by comparing the different representations, and deciding if mapping and integration is appropriate.

3. Discussion

We successfully developed a reference data model that preserves semantics and representation and still performs a generalization by mapping to a conceptual level, to identify hierarchical and temporal dependencies. The advantage is that changes in metadata can be tracked across data source and over time. This provides a revision control of data elements, in terms of replacing a data element by its new version in one source, as well as branching, meaning that two or more conceptually similar data elements exist parallel in multiple sources. Analyzing revisions and branches of a data element can support harmonization, mapping and finally integration.

We used the conceptual model of ISO/IEC 11179 that describes the structure of metadata, but does not specify its characteristics and attributes. Our focus was preserving semantics and representations we deviated from the standard by not considering the relationship between Conceptual Domain and Date Element Concept. According to ISO/IEC this connection is used for Data Elements that share the same Data Element Concept, so that the conceptually related representations are shared as well. This was not necessary in our specification.

In many other groups implementing ISO/IEC 11179 an MDR is used as description tool to establish consistent data definitions and to provide an item library for setting up data collection forms for clinical trials [5, 6]. Since we did not want to cover an entire domain, but to depict the existing elements in an extensive and comprehensible way, our implementation differs. Our data scheme can be used in other domains as introduced above.

The distinctive feature of our implementation is that multiple revisions and branches of conceptually similar data elements with heterogeneous definitions and representations, can be managed, stored and overviewed.

Next, we plan to extend the data reference model to include validation rules to ensure logical integrity between different data elements and their values. This will allow us not only to represent semantic and representational dependencies within a data element concept, but also dependencies within the content.

References

[1] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, P. Tarczy-Hornoch, Data integration and genomic medicine. *Journal of Biomedical Informatics* **40** (2007), 5–16.

[2] W. Sujansky, Heterogeneous Database Integration in Biomedicine. *Journal of Biomedical Informatics* **34** (2001), 285–298.

[3] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc* (2006), 1040.

[4] International Standards OrganizationISO/IEC 111179, Information Technology -- Metadata Registries. Available from URL:http://metadata-stds.org/11179/

[5] P.M. Nadkarni, C.A. Brandt, The Common Data Elements for cancer research: remarks on functions and structure. *Methods Inf Med* **45** (2006), 594–601.

[6] J. Stausberg, M. Löbe, P. Verplancke, J. Drepper, H. Herre, M. Löffler, Foundations of a metadata repository for databases of registers and trials. *Stud Health Technol Inform* **150** (2009), 409–413.