Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-409

Foundations of a Metadata Repository for Databases of Registers and Trials

Jürgen STAUSBERG [a,1], Matthias LÖBE [b], Philippe VERPLANCKE [c], Johannes DREPPER [d], Heinrich HERRE [b], Markus LÖFFLER [b]
[a] *IBE, Medical Faculty, Ludwig-Maximilians-Universität München, Germany*
[b] *Institute for Medical Informatics (IMISE), Universität Leipzig, Germany*
[c] *XClinical GmbH, Munich, Germany*
[d] *Telematikplattform für Medizinische Forschungsnetze (TMF) e. V., Berlin, Germany*

Abstract. The planning of case report forms (CRFs) in clinical trials or databases in registers is mostly an informal process starting from scratch involving domain experts, biometricians, and documentation specialists. The Telematikplattform für Medizinische Forschungsnetze, an umbrella organization for medical research in Germany, aims at supporting and improving this process with a metadata repository, covering the variables and value lists used in databases of registers and trials. The use cases for the metadata repository range from a specification of case report forms to the harmonization of variable collections, variables, and value lists through a formal review. The warehouse used for the storage of the metadata should at least fulfill the definition of part 3 "Registry metamodel and basic attributes" of ISO/IEC 11179 Information technology – Metadata registries. An implementation of the metadata repository should offer an import and export of metadata in the Operational Data Model standard of the Clinical Data Interchange Standards Consortium. It will facilitate the creation of CRFs and data models, improve the quality of CRFs and data models, support the harmonization of variables and value lists, and support the mapping of metadata and data.

Keywords. clinical trials, documentation standards, metadata, registries

1. Introduction

Systematic and accurate planning and management of the variables collected in trials and registers are prerequisites to successfully achieve the underlying objectives. In their framework of procedures for the assurance of data quality in medical registries, Arts et al. mentioned unclear/ambiguous data definitions, unclear data collection guidelines, poor layout of case report forms (CRF), data overload, and insufficient data checks as causes of insufficient data quality [1]. For the maintenance of registers, a proposal was made how to adapt interventions to the current level of data quality [2]. Leiner and Haux presented an approach for a systematic planning of clinical documentation based on a "documentation protocol" [3]. This protocol assists projects in a reasonable development of the data structure, the design of the documentation system, and its implementation. Nevertheless, real practices do not reflect these thoughts. Typically, projects start from scratch. Data managers or statisticians meet with clinicians or

[1] Corresponding Author: Jürgen Stausberg, MD, PhD, Ludwig-Maximilians-Universität München, IBE, Marchioninistraße 15, D-81377 München, Germany; E-mail: juergen.stausberg@ibe.med.uni-muenchen.de.

scientists and define the variables in an informal process. On the one hand, the CRF is the main structure guiding this process in clinical or epidemiological trials. On the other hand, database models and entity-relationship (ER) diagrams are used to tailor the variables to the conditions of databases in registry projects. Preexisting variable collections are rarely reused, terminological standards sometimes neglected.

That is the current reality in many projects covered by the *Telematikplattform für Medizinische Forschungsnetze* (TMF) e. V. in Germany (see also [4]). The principal aim of the TMF is to improve the organization and infrastructure for networked medical research, i.e., clinical, epidemiological and translational research. Under the umbrella organization of the TMF, expert opinions, studies, concepts, requirements specifications, services, and tools are created. Consequently the TMF addressed the above mentioned challenges and developed solutions, for example guidelines for adaptive management of data quality in cohort studies and registers or generic data protection solutions for medical research networks (available via the TMF office, see http://www.tmf-ev.de/). In 2008, a number of research groups joined under the supervision of the TMF and developed the concept of a metadata repository for databases of registries and trials. Aims of this metadata repository are:

- to make the creation of CRFs and data models easier,
- to improve the quality of CRFs and data models,
- to harmonize variables and value lists,
- and to support the mapping of metadata and data.

As a first step, the group analyzed the foundations of a metadata repository and developed a concept for a national service. The results are outlined in this paper.

2. Use Cases of a Metadata Repository

A metadata repository should support a top-down as well as a bottom-up approach. It can be used top-down by funding organizations, national agencies or research networks to distribute standards for variable collections, variables and terminologies. If required, for example in case of regulatory drug affairs, a metadata repository can force standards like the Clinical Data Acquisition Standards Harmonization (CDASH) Standard. But a metadata repository can also be used bottom-up by single projects, maintaining their specific variables, combining local variables with national or international standards, linking the variables to controlled vocabularies, and maintaining data sets over time. Therefore, use cases can be identified on the level of a single trial or register, across several projects, but restricted to a center or research network, and on a national or international level. Similar to the work of the Lawrence Berkeley National Laboratory in the eXtended MetaData Registry (XMDR) Project (cf., http://www.xmdr.org/use-cases.html) we identified eight high-level use cases for a metadata repository of the TMF (assigned to **s**-ingle organization, **m**-ultiple organizations, **n**-ational):

- Maintain metadata – to specify a CRF or a catalogue of variables (s)
- Import metadata – to establish a pool of metadata (s)
- Mapping of metadata – to transform data from one model to another (s)

- Queries for multiple metadata – to support pooling of data (s,m,n)
- Select projects – to identify projects covering specific variables (m,n)
- Review of metadata – to harmonize variables, and value lists (m)
- Maintain standards – to harmonize CRFs and databases (n)
- Improve standards – to identify gaps in data standards or terminologies (n)

3. Metamodel of a Metadata Repository

Data dictionary: Linarsson and Wigertz proposed a data dictionary as means for the integration of electronic health records (EHR) and knowledge based systems (KBS) [5]. They defined a data dictionary as “a set of terms with their properties and relationships organized according to the structure of the “real world” as it may be represented in the clinical database and the medical knowledge base”. This data dictionary includes denominations and keys of variables, data types such as string or integer, administrative information such as author, and relationships to controlled vocabularies such as classifications and terminologies. The focus at building a bridge between an EHR and a KBS [6] was later expanded to the support of cooperative environments for the maintenance of medical documentation systems [7]. Consequently, Bürkle discusses dictionary servers as independent middleware in distributed environments in his review [8]. A transition to a metadata repository was outlined by Niland [9].

Clinical Data Interchange Standards Consortium (CDISC): If data dictionaries represent a foundation for a metadata repository from the point of view of medical informatics, the work within CDISC (cf., http://www.cdisc.org/) represents this foundation from the point of view of clinical research. CDISC describes its mission as “to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare”. Data collection is addressed by CDASH. Core of CDASH are tables organizing variables like “method of ECG” into domains like “ECG test results”. Variables are defined with six attributes: data collection field (“method of ECG”), variable name (“EGMETHOD”), definition, CRF completion instructions, additional information for sponsors, CDASH core (“optional”). Value lists are provided for a small number of variables. Clearly, the structure used by CDASH for the definition of variables is a throwback to the concept of a data dictionary. Another CDISC standard, the Case Report Tabulation Data Definition Specification (CRT-DDS, also called define.xml) describes a metamodel for the definition of variables. CRT-DDS allows the specification of variables (called items), value lists and coding references in relationship to item groups, CRFs and studies. XML is used as syntax, and XML schema for the validation of the structure.

ISO/IEC 11179 Information technologies – Metadata registries: An international approved standard is available with ISO/IEC 11179 (cf. http://metadata-standards.org/). A metadata registry is “an information system or database for registering metadata”. In the context of our project, “information system or database” denotes databases of registers and trials, “metadata” is defining the variables used in these databases, e.g., CRFs or ER-diagrams. The concepts corresponding to the structure of a data dictionary is evolved in part 3 of this standard: “Registry metamodel and basic attributes”. A metamodel is “a data model that specifies one or more other data models”, the latter in our case implicit as CRF or explicit as ER-diagram. Figure 1 shows the “high-level metamodel". Gender as example: Humans are split up into two categories according to gender, category A and B. A and B establish an Enumerated_Conceptual_Domain specifying the Data_Element_Concept gender. Values are defined as 1, 2, M, F. M and 1 are permissible values for A; F and 2 permissible values for B. M and F establish one Enumerated_Value_Domain, 1 and 2 another one. Then a Data_Element "sex" is related to the Data_Element_Concept gender and the Value_Domain "1 and 2".

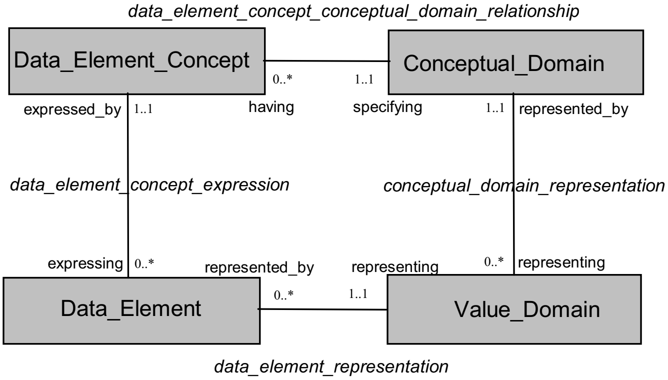

Figure 1. High-level metamodel of ISO/IEC 11179 part 3 [10]

4. Discussion

An expressive metamodel for a metadata repository is available with the ISO/IEC 11179. Successful implementations have been published [11]. But this metamodel does also not fully support the above mentioned high-level use cases. For example, the ISO reflects a hierarchical organization of a registry and is not able to manage multiple organizations explicitly. The representation of controlled vocabularies is limited to a flat structure of term lists in the second edition of ISO/IEC 11179. Thus, a usage of part 3 of ISO/IEC 11179 for a metadata repository for databases of registers and trials will require some additions. Some implementations are available, commercial, governmental and as open source solutions. Previously mentioned was the XMDR. The National Cancer Institute (cf., http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr/) offers an implementation of ISO/IEC 11179 with the Cancer Data Standards Repository (caDSR). Similar to the XMDR, the caDSR extends ISO in introducing the entities forms (for CRFs) and protocols. As presented by Nadkarni and Brandt [12], caDSR excludes Conceptual_Domain from the high-level metamodel of ISO (cf., Figure 1). The same simplification is implemented in a commercial implementation (OneData MDR with "Lite" model, see http://www.datafoundations.com/).

The coverage of all meaningful data structures and use cases will be one success factor for a metadata repository for databases of registers and trials. The integration into existing software applications used for the planning and maintenance of these databases will be of equal importance. Especially clinical trial and data management systems are widely used for CRF-design and study maintenance. It will be a knockout criterion for such a service if the users have to document their metadata twice, once for the metadata repository and once for a trial or data management system. We recognize the CDISC Operational Data Model as most important standard in this domain for the import and export of metadata. The availability of relevant content will be another factor of success. Variable collections proposed by CDISC should be available as well as controlled vocabularies like ICD-10, LOINC, and SNOMED. Keeping in mind these key issues, a metadata repository based on the metamodel of ISO/IEC 11179 and implemented as a Web service could successfully achieve the objectives mentioned in the beginning: to make the creation of CRFs and data models easier, to improve their quality, to harmonize variables and value lists, and to support the mapping of metadata and data.

Acknowledgements. The work was partially funded by the Telematikplattform für Medizinische Forschungsnetze (TMF) e. V., Berlin, Germany, under contract TMF-project V063-01. We would like to thank the other participants of this project namely Irene Droll, Silvia Groß, Johannes Hüsing, Magnus Knuth, Wolfgang Kuchinke and Roland Mücke. As external reviewer, Thomas Bürkle provided suggestions considered in this paper.

References

[1] Arts, D.G., De Keizer, N.F., Scheffer, G.J. (2002) Defining and improving data quality in medical registries: A literature review, case study, and generic framework. *Journal of the American Medical Informatics Association* 9:600–611.

[2] Stausberg, J., Nonnemacher, M., Weiland, D., Antony, G., Neuhäuser, M. (2006) Management of data quality – Development of a computer-mediated guideline. In Hasman, A., Haux, R., van der Lei, J., De Clercq, E., Roger-France, F.H. (Eds.) *Ubiquity: Technologies for Better Health in Aging Societies*, IOS Press, Amsterdam, 477–482.

[3] Leiner, F., Haux, R. (1996) Systematic planning of clinical documentation. *Methods of Information in Medicine* 35:25–34.

[4] Stausberg, J., Schütt, A. (2008) Einrichtungsübergreifende Datenbestände der medizinischen Forschung in Deutschland. In Jäckel, A. (Hrsg.) *Telemedizinführer Deutschland*, Ausgabe 2009, Medizin Forum AG, Bad Nauheim, 193–196.

[5] Linnarsson, R., Wigertz, O. (1989) The data dictionary – A controlled vocabulary for integrating clinical data-bases and medical knowledge bases. *Methods of Information in Medicine* 28:78–85.

[6] Stausberg, J., Wormek, A., Kraut, U. (1995) Terminological reference of a knowledge-based system: The data dictionary. In Greenes, R., Peterson, H., Protti, D. (Eds.) In *Proceedings of the Eight World Congress on Medical Informatics MEDINFO'95*, IMIA, Edmonton, 157–161.

[7] Knaup, P., Garde, S., Haux, R. (2007) Systematic planning of patient records for cooperative care and multicenter research. *International Journal of Medical Informatics* 76:109–117.

[8] Bürkle, T. (2000) Can we classify medical data dictionaries? In Hasmann, A., Blobel, B., Dudeck, J., Engelbrecht, R., Gell, G., Prokosch, H.U. (Eds.) *Medical Infobahn for Europe, Proceedings of MIE2000 and GMDS2000*, IOS Press, Amsterdam, 691–695.

[9] Niland, J.C. (2001) Creating a metadata repository in support of clinical research. In *Proceedings of Seoul 53rd Session*, International Statistical Institute, http://isi.cbs.nl/iamamember/CD2/pdf/1084.pdf.

[10] ISO/IEC 11179-3+COR1 (2003) *Information Technology – Metadata Registries (MDR) Part 3: Registry Metamodel and Basic Attributes*. Second edition 2003-02-15 Incorporating COR1. http://jtc1sc32.org/doc/N1151-1200/32N1168-ISO-IEC11179-3-2003COR1.zip.

[11] Park, Y.R., Kim, J.H. (2006) Metadata registry and management system based on ISO 11179 for cancer clinical trials information system. *AMIA Annual Symposium Proceedings 2006*, 1056.

[12] Nadkarni, P.M., Brandt, C.A. (2006) The common data elements for cancer research: Remarks on functions and structure. *Methods of Information in Medicine* 45:594–601.