Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-564

Interoperability in Clinical Research: From Metadata Registries to Semantically Annotated CDISC ODM

Philipp BRULAND[a,1], Bernhard BREIL[a], Fleur FRITZ[a], Martin DUGAS[a]
[a]*Institute of Medical Informatics, University of Münster, Germany*

Abstract. Planning case report forms for data capture in clinical trials is a labor-insensitive and not formalized process. These CRFs are often neither standardized nor using defined data elements. Metadata registries as the NCI caDSR provide the capability to create forms based on common data elements. However, an exchange of these forms into clinical trial management systems through a standardized format like CDISC ODM is currently not offered. Thus, our objectives were to develop a mapping model between NCI forms and ODM. We analyzed 3012 NCI forms and included common data elements regarding their frequency and uniqueness. In this paper, we have created a mapping model between both formats and identified limitations in the conversion process: Semantic codes requested from the caDSR registry did not allow a proper mapping to ODM items and information like the number of module repetitions got lost. Summarized, it can be stated that our mapping model is feasible. However, mapping of semantic concepts in ODM needs to be specified more precisely.

Keywords. clinical research, CRF, data elements, mapping, metadata registry

1. Introduction

An essential part of conducting randomized clinical trials is the collection of subject data documented on case report forms (CRFs). Planning those CRFs is a cumbersome, labor-intensive and mostly informal process. Several forms and data elements have to be specified for these questionnaires. The workload increases when studies are conducted at multiple centers and several individuals are involved in designing CRFs for data collection and capture. To address the challenge of data standardization and to support the process of CRF planning, metadata registries as the National Cancer Institute's (NCI's) Cancer Data Standards Registry and Repository (caDSR) [1] provide access to common data elements (CDEs) [2] that can be used to create CRFs. CDEs address the problem of inconsistent data representation of similar or identical concepts that have been used for different purposes. CDEs from caDSR are encoded with semantic concepts of the Enterprise Vocabulary Services [3] that provide two major terminology resources: NCI Thesaurus and NCI Metathesaurus. The caDSR – based on ISO standards [4] for metadata registries – summarizes a compilation of tools as the CDEBrowser, FormBuilder and NCI Term Browser. The aim of this registry is

[1] Corresponding Author: Philipp Bruland, Institute of Medical Informatics, University of Münster, Germany; e-mail: philipp.bruland@uni-muenster.de

to define a comprehensive set of standardized metadata descriptors for cancer research data applied for data collection and analysis, which can be accessed via web interfaces. With a total amount of >3,000 forms and >45,000 CDEs, the registry provides a substantial and promising contents base for cancer research. All data elements in forms of NCI-sponsored clinical trials are specified in the caDSR. Forms within the FormBuilder correspond to CRFs used in clinical trials and are accessible via the web front end or can be downloaded as Excel-spreadsheets. To build new forms within the FormBuilder portal, CDEs are placed in a cart and can be used to create questions on a form.

Promoting the standardization and structured representation of CRFs, the Clinical Data Interchange Standards Consortium (CDISC) has published the Operational Data Model (ODM), a vendor-neutral transport and presentation layer for trial metadata and clinical data [5]. It also offers archiving of study databases [6] and most EDC (Electronic Data Capture) systems make use of ODM in processing and communicating clinical research data. In the context of data collection, larger trials tend to be more likely to adopt EDC. In 2006/2007, 40% of Canadian clinical trials were using EDC systems [7]. The interoperability of forms and data elements between metadata registries and clinical trial management systems is desirable to close the gap between trial planning and execution. The exchange of study metadata between different systems is shown by Kuchinke et al. [8] and Brandt et al. have discussed the mapping of trial database tables into CDISC ODM version 1.1.0 [9]. Unfortunately, the rich pool of standardized forms and data elements within caDSR currently does not support exchange with common clinical trial management systems. This leads to our following objectives:

The purpose of this paper is to develop and implement a mapping model between NCI forms and semantically annotated CDISC ODM files. To assess the feasibility of this approach, the number of converted forms, containing elements and semantic concept codes are analyzed.

2. Methods

We have analyzed the technical specifications of the NCI caDSR registry [10] and CDISC ODM [5]. For our study, we have downloaded all NCI forms with “released” as workflow status (in total 3,012 Excel-files). Based on these analyses, we have derived the mapping between both formats. Our analysis comprises frequency and uniqueness of CDEs and semantic concept codes within the forms. Therefore, we have developed a Java™ 7 application and generated classes using Java Architecture for XML Binding (JAXB 2.2.3), which allows a proper generation of ODM XML files. The Java classes are based upon the ODM-XML schema description version 1.3.1.

Additional properties like data element length and semantic concept codes are not listed within the Excel-files. These attributes have been requested from the caDSR interface [1]. After receiving metadata from the NCI forms, additional attributes and semantic codes, all forms have been converted into semantically annotated CDISC ODM format.

3. Results

In the following, we describe the mapping model between both form definitions and the results of the analyses to quantify the feasibility of the conversion, respectively.

3.1. Mapping model between NCI forms and CDISC ODM

Due to our analyses we created a mapping model as shown in figure 1. NCI protocols are divided into four levels: Forms, Modules, CDEs and Permissible Values.

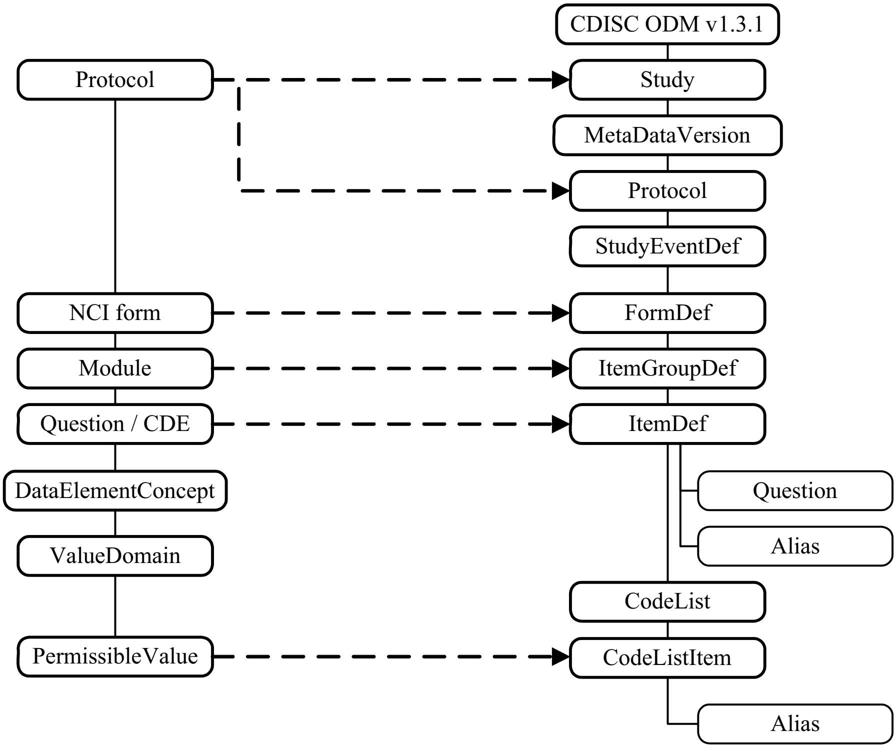

Figure 1. Mapping model between NCI forms on the left side and CDISC ODM on the right. The dashed line shows the transformation from NCI form elements into ODM elements.

ODM offers multiple hierarchical sections to describe the structure of clinical studies. One NCI study protocol can be represented by an individual ODM file. The ODM section *Protocol* contains the definition of NCI protocols. *MetaDataVersion* and *StudyEventDef* definitions are schema conform and inserted per default. All levels in the Excel-files contain a name and a description, which are mapped into the respective ODM elements. In addition, for each level a public ID and a version number are given that we use to create a unique identifier. Each element in ODM can be identified and referenced by an OID. This ID is generated from the caDSR registration authority identifier and the respective public ID and version. *ItemGroupDefs* are derived from modules and every module contains the number of repetitions, which is expressed through a Boolean field in ODM. The display order is defined by a link between

ItemGroup and *FormDef*. *DataTypes* are mapped between the CDE's *Value Domain Data Type* and XML data types in ODM. The *name* attribute is populated with the "CDE" column and the *Question* element in ODM with the identically named column in the Excel-file. If the following row(s) in the item list contain one or more *permissible values*, new *CodeList* elements will be referenced for parent *ItemDefs* and each *permissible value* represents a *CodeListItem*.

According to the ISO/IEC 11179 metadata registry standard, all CDEs in the caDSR registry consist of three classes, which are semantically annotated with concept codes: object class, property and representation. The mapping occurs between these three classes and the *Alias* element in ODM. Due to the missing protocol definition, form category, CDEs length and semantic codes in the Excel-files, these attributes are requested via the NCI's caDSR interface.

3.2. Frequency and categories of forms, common data elements and codes

We have determined the frequency of data elements and codes utilized for the NCI forms. CDEs provide the basis to create questions within a respective form. As of December 16, 2011, in total 93,170 questions are placed in all NCI forms with a "released" workflow status. The whole amount of included CDEs totals 91,685, which implies that 1,485 questions are free-text and not defined as common data elements. Our analysis showed that the total amount of unique CDEs inside the caDSR registry was not covered in all NCI forms. The determined number of distinct data elements amounts to 13,859. Table 1 shows the distribution of semantic concept codes, which are used in the CDEs.

Table 1. Number of codes, which are used in 91,685 CDEs, divided into three data element classes.

Classes	Total number of codes	Number of uniquely used codes
Object Class	106,409	1,790
Property	127,916	2,701
Representation	107,471	520
∑	341,796	5,011

All data elements and their semantic concepts from 3,012 NCI forms could be successfully mapped to ODM files.

4. Discussion

In this paper we address the transformation of clinical trial metadata based on the caDSR registry into semantically annotated CDISC ODM. We have shown that it is feasible to map forms and their respective data elements into a standardized transport format like ODM that is mentioned by Stausberg et.al. [11]. ODM elements like *FormDef*, *ItemGroupDef*, *ItemDef* and *CodeLists* were mapped to the NCI form structure. Semantic concepts were inserted within ODM *Alias* elements.

The mapping model shows two limitations: First, the number of module repetitions gets lost through the Boolean expression in ODM. Second, attributes in the *Alias* element could comprise any content, so this is not suitable for direct automatic processing. The ODM schema definition is extensible and for future use it is possible to integrate additional specifications [12] for external concept codes and further definitions on *CodeLists* and its items. The suitability of this extension should be assessed regarding semantic concepts for *CodeListItems*.

The results from our CDE analyses are also suitable for data re-use and comparison in different domains, for instance in projects regarding the secondary use of routine medical data for clinical research and modeling of eligibility criteria [13-15]. For this purpose, a mapping of frequently used data elements in clinical trials with the electronic health record environment is necessary.

5. Conclusion

Metadata registries are essential for standardizing and harmonizing data elements used on CRFs for data collection in clinical research. A mapping between NCI forms and CDISC ODM is possible with only minor limitations; therefore exchange of clinical research forms between registries and clinical trial management systems is feasible.

References

[1] Cancer Data Standards Registry and Repository (caDSR) https://cabig.nci.nih.gov/concepts/caDSR/ (last access: January, 10. 2012)

[2] Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. Methods Inf Med. 2006;45(6):594-601.

[3] NCI Enterprise Vocabulary Services http://evs.nci.nih.gov/ (last access: December, 28 2011)

[4] ISO/IEC 11179 Specification and standardization of data elements Parts 1–6.

[5] CDISC Operational Data Model http://www.cdisc.org/odm (last access: December, 21. 2011)

[6] Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med. 2009;48(5):408-13. Epub 2009 Jul 20.

[7] El Emam K, Jonker E, Sampson M, Krleza-Jerić K, Neisa A. The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. J Med Internet Res. 2009 Mar 9;11(1):e8.

[8] Kuchinke W, Wiegelmann S, Verplancke P, Ohmann C. Extended cooperation in clinical studies through exchange of CDISC metadata between different study software solutions. Methods Inf Med. 2006;45(4):441-6.

[9] Brandt CA, Morse R, Matthews K, Sun K, Deshpande AM, Gadagkar R, Cohen DB, Miller PL, Nadkarni PM. Metadata-driven creation of data marts from an EAV-modeled clinical research database. Int J Med Inform. 2002 Nov 12;65(3):225-41.

[10] NCI FormBuilder https://wiki.nci.nih.gov/display/caDSR/Form+Builder (last access: January, 05. 2012)

[11] Stausberg J, Löbe M, Verplancke P, Drepper J, Herre H, Löffler M. Foundations of a metadata repository for databases of registers and trials. Stud Health Technol Inform. 2009;150:409-13.

[12] CDISC Terminology http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc (last Access: January, 08. 2012)

[13] Dziuballe P, Forster C, Breil B, Thiemann V, Fritz F, Lechtenbörger J, Vossen G, Dugas M. The single source architecture x4T to connect medical documentation and clinical research. Stud Health Technol Inform. 2011;169:902-6.

[14] El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform. 2011 Dec;44 Suppl 1:S94-S102. Epub 2011 Aug 25.

[15] Electronic Health Records for Clinical Research: http://www.ehr4cr.eu/ (last access: January, 05. 2012)