Journal of Biomedical Informatics 53 (2015) 49–57

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A multi-technique approach to bridge electronic case report form design and data standard adoption

Ching-Heng Lin, Nai-Yuan Wu, Der-Ming Liou *

Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 19 December 2013
Accepted 30 August 2014
Available online 6 September 2014

Keywords:
Data standard
Common data elements
Case report form
Natural language processing
Ontology-based knowledgebase

ABSTRACT

Background and objective: The importance of data standards when integrating clinical research data has been recognized. The common data element (CDE) is a consensus-based data element for data harmonization and sharing between clinical researchers, it can support data standards adoption and mapping. However, the lack of a suitable methodology has become a barrier to data standard adoption. Our aim was to demonstrate an approach that allowed clinical researchers to design electronic case report forms (eCRFs) that complied with the data standard.

Methods: We used a multi-technique approach, including information retrieval, natural language processing and an ontology-based knowledgebase to facilitate data standard adoption using the eCRF design. The approach took research questions as query texts with the aim of retrieving and associating relevant CDEs with the research questions.

Results: The approach was implemented using a CDE-based eCRF builder, which was evaluated using CDE- related questions from CRFs used in the Parkinson Disease Biomarker Program, as well as CDE-unrelated questions from a technique support website. Our approach had a precision of 0.84, a recall of 0.80, a *F*-measure of 0.82 and an error of 0.31. Using the 303 testing CDE-related questions, our approach responded and provided suggested CDEs for 88.8% (269/303) of the study questions with a 90.3% accuracy (243/269). The reason for any missed and failed responses was also analyzed.

Conclusion: This study demonstrates an approach that helps to cross the barrier that inhibits data standard adoption in eCRF building and our evaluation reveals the approach has satisfactory performance. Our CDE-based form builder provides an alternative perspective regarding data standard compliant eCRF design.

1. Introduction

The rapidly development of new research area and the wider adoption of informatics systems have resulted in the exponential growth of biological and clinical data. Although “big data” creates new research opportunities [1], researchers also face the difficulty of obtaining data as well as the high cost of data collection. Therefore, it has been inevitable that an urgent need for data harmonization, which would facilitate the subsequent data aggregation and sharing, has arisen.

The use of a data standard is a critical requirement for such harmonization, and is also the first step towards data integration. A data standard is an agreed upon set of rules that allow information to be shared and processed [2]. It could be classified as semantic standard (i.e., terminology artifacts), syntax standard for data representation and format (i.e., markup language), and content standard, such as minimum information checklist or common data elements (CDEs) [3–5].

As the National Institute of Health (NIH) encourages the use of CDEs [6], some researchers have designed their CRFs based on CDEs [7]. Several CDEs have been developed, for example, the Cancer Bioinformatics Grid (caBIG) [8], the National Institute of Neurological Disorders and Stroke (NINDS) common data element project [9], the Parkinson Disease Biomarker Program (PDBP) [7,10], as well as a number of other clinical CDE for a variety of different purposes [11–13]. The CDE is a logical unit of data that provides for the definition of data, including an identifier, an element type to indicate the value type, and detailed information, which is the meta-data that fully defines the semantics of the CDE [14]. To define the CDE in formal representation, the ISO/IEC 11179, which is a metadata repository standard, provides the standard syntax and grammar need to describe data element metadata. Many

* Corresponding author. Address: No. 155, Sec. 2, Li-Nong St., Beitou District, Taipei City 112, Taiwan. Fax: +886 2 2820 2508.

E-mail address: dmliou@ym.edu.tw (D.-M. Liou).

http://dx.doi.org/10.1016/j.jbi.2014.08.013
1532-0464/© 2014 Elsevier Inc. All rights reserved.

efforts have been made to adopt this standard, for example, the National Cancer Institute (NCI) Cancer Data Standards Repository (caDSR) implements the ISO/IEC 11179 standard for metadata registries when presenting CDEs in their repository [15]. The cancer Common Ontologic Representation Environment (caCORE) created by National Cancer Institute Center for Bioinformatics (NCICB) is an interoperability infrastructure that is based on model driven architecture and contains a metadata repository based on the ISO/IEC 11179 standard to allow semantic interoperability [16]. Another effort is the CDISC Shared Health and Research Electronic Library (CSHARE) and this utilizes the ISO/IEC 11179 standard as the semantic basis for its metadata and has adopt the ISO 21090 for the formal presentation of CDE data type [17].

The CDE should be able to not only standardize data collection, but also should facilitate the follow-up comparison of results across multiple studies [18]. Nevertheless, CDEs are center-specific and are not a global standard; therefore such an approach, which is called computable semantic interoperability, may exhibits scalability problems when applied beyond a well-defined domain [19]. As a result, using CDEs is still a compromise solution in terms of current research domains. To address the issue of computable semantic interoperability, Payne et al. developed the Translational Research Informatics and Data Management Grid (TRIAD), which leverages the caGrid [20] middleware and extended this to support working interoperability. Such working interoperability means that stakeholders are able to negotiate and use context-relevant semantic models that enable better semantic exchange [19]. In the TRIAD, a CDE metadata registry repository called the MDR (metadata repository) Core is one of the system's four major components.

In clinical studies, the case report form (CRF) is an important tool for collecting data. The CRF is usually designed by researchers based on their study objective, for example, demographic information, medical history, and/or the results of clinical examination. Many clinical data capturing systems support electronic CFR (eCRF) design [21,22]. Through use of eCRFs, clinical research data is able to be captured and stored in clinical data repositories. For data integration and sharing purposes, Brandt et al indicated that there is a requirement for an information tool that will aid researchers in creating comprehensive and valid CRFs that can be mapped to a data standard [23]. Such an approach would enable the adoption of a data standard that can be used for clinical research applications, particularly if there is a tool supporting the retrieval and reuse of existing standard items [24].

How to efficiently and precisely select data elements from a CDE repository in order to build an eCRF that is able to accurately reflect the study question is the challenge that needs to be met in this context, especially when some researchers might not be familiar with the application of CDEs. Most commercial available clinical data capturing systems do not allow users to associate their research questions with CDEs, but merely provide a list of hundreds of CDE for selection or allow simple searching of the CDEs. The lack of an informatics tool that is able to substantially increase efficiency has become a barrier that inhibits data standard adoption.

To cross this barrier, we developed a multi-technique approach that included the creation of an ontology-based knowledgebase, the development of natural language processing and the creation of an information retrieval technique. In this study, we demonstrated this approach by implementing an eCRF builder that supports researchers and helps them design CDE compatible eCRFs.

2. Materials and methods

There were mainly three parts to the implementation of the multi-technique approach (shown in Fig. 1): (1) the creation of an ontology-based knowledgebase of the CDEs, (2) the development of an information retrieval strategy for suggesting the CDEs and (3) the linking of the CDEs to the clinical questions.

2.1. Creating an ontology-based knowledgebase of CDEs

This study took PDBP CDEs [25] as the example for demonstrating the process of creating an ontology-based knowledgebase. Originally, the PDBP CDEs were hosted in a straightforward relational database format. Our approach is compatible with the relational database format; however, such a format does not support formal semantic definitions. The ontology technique has been widely adopted in the clinical studies to allow semantic interoperability. Some studies have utilized ontology to harmonize their data standards [26] or to model the entities and relationships within study designs [27], while others have presented a clinical data element model using Web Ontology Language (OWL) [28,29]. In this research we would like to develop the CDE ontology to allow further semantic interoperability and to demonstrate the compatibility of our approach with semantic web technology.

Even through PDBP CDEs are not ISO/IEC 11179 compliant; they still have a well defined structure. In this study, we developed a program using the Protégé API [30] that build this ontology using the PDBP CDE relational database. The CDE information contains general details, such as identifier, title, variable name and description, data definitions, which includes element type, the text of the suggested question, guidelines and pre-descriptions, categorization and classification. The categorization and classification predicate the restricted hierarchical structure of the CDEs. The hierarchy is composed of disease, domain and subdomain. Each disease contains specific domains and each domain contains specific subdomains; furthermore, each CDE element belongs to a specific subdomain. To represent these restrictions, we adopted the OWL sequence extension [31] to express the restricted hierarchical structure of the CDEs. The OWL sequence extension uses the *hasNext* property to point to the next member in the sequence and to identify that the content of the member is associated through the *hasContents* property. In this study, we created four OWL classes: *Disease*, *Domain*, *Subdomain* and *CDE*. Those classes are linked with each other in sequence using the OWL object properties (*hasDomain*, *hasSubDoaim* and *hasCDE*) and the owl:individual of the owl:class is associated through the *hasIndividual* property. By setting the rdfs:domain and rdfs:range of the object property, each owl:individual belonging to a specific owl:class will inherit the restriction. The general details and data definition information is then stored in each CDE entity via the OWL annotation property. There are 426 CDE entities under CDE OWL class. An example of CDE ontology structure is shown in Fig. 2.

2.2. Information retrieval strategy for suggesting CDEs

In order to allow researchers to adopt the data standard when carrying out eCRF design, this study developed an information retrieval strategy that provides question relevant CDEs to its researchers. The study question, which is input by the user, is treated as the query for CDE information retrieval. Since the question is able to be in a variety of formats, the use of a pattern matching search approach might not be appropriate. Our information retrieval strategy included three major steps (Fig. 1). Firstly, we need to index the CDEs from the knowledgebase to allow information retrieval. The second step was to generate the query from study question, which is in free text, and then to perform searching. Thirdly, we evaluate the quantity of searching results obtained and refined the query if necessary. An open source and full-featured text search engine, Apache Lucene, was adopted for implementing the information retrieval strategy [32].