DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications2010

The Question about Questions: Is DC a Good Choice to Address the Challenges of Representation of Clinical Research Questions and Value Sets?

James E. Andrews
School of Library and Information Science
University of South Florida, U.S.A.
jimandrews@usf.edu

Denise Shereff
Division of Bioinformatics and Biostatistics, University of South Florida (USF), U.S.A.
Denise.Shereff@epi.usf.edu

Timothy B. Patrick
Health Care Informatics and Administration,
University of Wisconsin, Milwaukee, U.S.A.
tp5@uwm.edu

Rachel L. Richesson
Division of Bioinformatics and Biostatistics, University of South Florida (USF), U.S.A.
Rachel.Richesson@epi.usf.edu

Abstract

Question and answer sets are the core of clinical research. The [RD] PRISM (Patient Registry Item Specifications and Metadata for Rare Disease) project will provide a library of standardized questions across a broad spectrum of rare diseases that can be used for a variety of clinical information and data collection purposes, such as registries. Questions will be encoded using well-established clinical terminologies to enable cross-indication and cross-disease analyses, facilitate collaboration, and generate meaningful results for rare disease patients, physicians, and researchers. Encoded question and answer sets will also be indexed to facilitate information retrieval by subject matter, data type, and time interval. This project will outline issues and challenges related to indexing questions for future use and for data sharing; to explore possible metadata and terminological standards for indexing them; and, determine if Dublin Core (DC) is a viable alternative to other schemes for a library of standardized rare disease research questions.

Keywords: metadata; question and answer sets; standardized question library; MeSH, UMLS, SNOMED CT; indexing questions; health informatics; data standards; interoperability; registries; case report forms

1. Overview / Introduction

A central feature of clinical research is that it revolves around the asking and answering of questions. For clinical research studies, a principal investigator or team of investigators identifies research questions and hypotheses, which are tested by the analysis of data collected systematically throughout the study. These data are defined at the start of each study by the investigators as data items or questions on data collection forms (paper or electronic). Within a distributed research network, the opportunities for investigators to share questions administered from data collection forms or standardized instruments are limited by their ability to understand and access the content of questions previously used by themselves or other investigators. Addressing this much-needed ability to understand and access the content of standardized questionnaires could also increase the use of standards, and relieve new investigators in generating their own question content (which can risk creating duplicate or poorly structured questions). Brandt et al. (2004) stress the importance of standards for representing the content of questions and questionnaires for the maintenance and curation of data libraries that support the clinical research process. They also speculate that such standards could allow intelligent aggregation and analysis of multiple question forms that attempt to measure the same construct in different settings. In clinical research, tools have been developed to support the re-use of

https://doi.org/10.23106/dcmi.952109935

questions and their permissible answers as part of the development of new data collection forms, but rarely use controlled terminologies to improve their efficiency or promote data sharing and concept reuse (Duftschmid, Gall, Eigenbauer, & Dorda, 2002). Still, we are not aware of any comparison of controlled terminologies for this use. The consistent and accurate application of data standards, including terminologies, to represent and organize clinical research questions is a critical challenge facing clinical research informatics.

Because question items represent the vehicle of data collection/entry in clinical research studies, any successful implementation of data standards in the clinical research domain must include the indexing of questions and their permissible answers that facilitates retrieval of desired items by multiple attributes, such as subject matter, data type, and time interval. A model for indexing and retrieving questions is critical to identifying and reducing variation in settings where multiple parties control content, and is pivotal to permeation of standards in the clinical research domain.

The purpose of this paper is to outline the issues and challenges related to indexing questions and their permissible answers for future use and for data sharing; to explore possible metadata and terminological standards for indexing them; and, determine if Dublin Core (DC) is a feasible alternative to be explored for a library of standardized questions across a broad spectrum of rare diseases.

2. Issues and Challenges

Representing and organizing clinical research questions is essentially an indexing challenge. This may seem fairly straightforward. The question and answer set is a short information object with a well-defined context, as opposed to a lengthy scholarly article (within a vast collection of articles) with multiple concepts addressed and of an often highly technical nature. However, there are syntactic and semantic ambiguities and challenges inherent in research questions and their corresponding answer sets such as:

- Context (type of study, disease or treatment of interest, etc.)
- Format of questions, and location of semantics
- Who is asking the question (patient, relative, doctor)
- Audience or person being asked the question
- Relevant data standards for specific answer sets

There are a host of other issues and challenges stemming from questions and answers. For example, similar questions may be presented in tabular format versus a linear/vertical format of questions. In a patient registry (a tool used to collect patient-reported data for various studies or related purposes) discreet answer choices may be presented for a user's selection in one registry, whereas a write-in section may be presented in a similar form designed for another registry. The set of answer choices presented to the end-user may be subtly different between data-collection instruments, meaning the results may not be directly comparable. Other data collection instruments may link a series of unrelated answer choices under a single question-like heading, whereas another instrument might break up the answer set into two or more different questions, perhaps with a different set of semantic codes. Some data collection instruments may present mutually-exclusive answer choices, or allow the selection of inconsistent answers, whereas others may have created mutually exclusive, non-overlapping answers sets. Some forms may use constructs, such as "check all that apply" to list all positive findings, whereas others may require that the end-user explicitly state that a given condition is not present, or cannot be answered. Even simple differences such as variations in the use of font size, color, bold-face text, and question placement or order on a form can subtly affect how a form is completed by an end-user. The grouping of questions into form sections may also influence the response to certain questions, due to variation in context. Lastly, the data produced by related non-standardized

https://doi.org/10.23106/dcmi.952109935

instruments can be fundamentally incomparable without a standard approach to modeling question/answer sets and registry form construction.

A critical and largely unaddressed problem for registries, in particular, and clinical research data collection in general, is the need for tools that allow data collection forms and their component questions and answers – which are the currency of clinical research – to be indexed in such a way that they can be retrieved for re-use (e.g., to support the rapid development of another related rare disease registry or to create a related case report form). Useful standardization of elements making up registry forms should enable unambiguous, consistent, and reliable re-use of questions, answers, and groups of question/answer sets among different registries. The implementation of common sets of questions and answers derived from approaches (e.g., CDISC, caDSR/caBIG) between studies is still not common, and the encoding of questions/answers with standard terminologies is not done consistently (Nadkarni & Brandt, 2006; Richesson & Krischer, 2007). Moreover, semantic encoding of data elements (i.e., question + answer + definition) is prone to inter-coder variability (Richesson, Andrews, & Krischer, 2006; Andrews, Richesson, & Krischer, 2007), and makes consistent querying based on these "standard" codes difficult and unreliable. Subsequently, mechanisms for storing the collected data are also subject to considerable variability, often using fundamentally different data model architectures. In these approaches, data sharing often requires complicated conversions between data models, or the complex and error-prone use of terminological inference (i.e., using computer-based inference method, such as a transitive closure) approaches in determining the similarity or equivalence of data stored according to different data architectures.

3. Questions and Answers

One context where the challenges related to the representation of questions and answers are evident is that of patient registries. Patient registries can be sponsored and developed by governments, academic scientists, clinical investigators, or pharmaceutical companies, or Patient Advocacy and Support groups (PAGs). For rare diseases in particular, patient groups are often the first to sponsor and develop a registry – usually as a means to get an early look at the numbers of people affected and their characteristics. The content of these registries (i.e., the questions) may change over time as more becomes known about a disease and its clinical variations. Often the registries (and the supporting questions) are developed ad hoc by the PAGs themselves, yet there is currently no clear specification for standards or the organization of banks of existing questions for patients to access. In clinical research, registries and case report forms (CRFs) may be created impromptu, but they are ostensibly based on supporting evidence in the literature (i.e., they are derived from the protocol, which must be based on evidence). Usually, neither is represented with a surrogate, such as an indexing record.

3.1. Use case for registry

The first use case to illustrate our project involves a new registry: a Vasculitis Reproductive Health Survey, a hypothesis-generating registry that collects two types of information directly from patients regarding reproductive health and vasculitis-specific information. For this registry, new questions were developed based on pregnancy research experience and fertility outcomes. The registry features a subset of disease-specific (multiple vasculitides) questions from Vasculitis Clinical Research Consortium forms used in various natural history studies for the previous five years. Relevant stakeholders for this registry include: the patient or the public, domain experts from the clinical research consortium, the Data Management and Coordinating Center (DMCC) at the University of South Florida (USF) and the College of American Pathologists (CAP).

Questions include demographic information; specific vasculitis information for men and women, including detailed questions about medications; fertility; questions covering the spectrum of reproductive health; reproductive intent; and pregnancy history.

https://doi.org/10.23106/dcmi.952109935

3.2. Use case for Case Report Form

The second use case involves the creation of CRFs for an existing registry: the Urea Cycle Disorders Consortium (UCDC). Each CRF is a data collection form. Collectively, CRFs comprise the data collection for a given research study. A team of physicians and clinicians with extensive clinical research experience designed the studies over a one year period. In this use case, question sources would involve a subset of questions from these UCDC forms used in a multi-site natural history study for the previous five years. Although the questions have been used in research, they are new to the PRISM library.

Questions include baseline assessments, diagnostics for medical record review, medication, eligibility, demographics, family history, interim events, laboratory results, medical and developmental history, neuropsychological testing, nutritional information, and physical and neurological examinations.

For both of the aforementioned use cases, there is significant value in limiting value set options/permissible value sets for the question and answer sets, as both examples involve very specific disease research with few common elements with other studies.

4. Controlled vocabularies

Terminology control, when implemented correctly and consistently, can dramatically improve the quality of search results in most contexts. Since many controlled vocabularies in healthcare cover very specific domains, it is important to select one that will best meet the needs of the task.

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) is the CHI-recommended (Consolidated Health Informatics; http://www.hhs.gov/healthit/chi.html) standard for findings, and previous research has indicated that SNOMED CT is well suited for clinical concepts, though possibly less suited for representing the full amount of information collected on CRFs (Richesson, Andrews, & Krischer, 2006). Given the potential linkages that are often required between clinical research and electronic medical records, SNOMED CT is a likely candidate terminology for representing the content of research questions and permissible value sets. However, other terminologies are better for representing special information, such as RxNorm for pharmaceuticals, or LOINC (Logical Observation Identifiers Names and Codes) for lab results. An alternative for representing clinical content of CRFs or registry questions is MeSH (Medical Subject Headings). While primarily used for indexing the medical literature, it has been used in a variety of health contexts. The detailed documentation and training materials that exist for MeSH make it relatively easy to use for information professionals, though it is problematic (as most vocabularies are) for non-experts. Still, in comparison to SNOMED CT, MeSH has clear limits in terms of the clinical descriptiveness possible, and so is generally not used for data that are clinically rich. The level of granularity required for the use cases suggested here is not certain; however, it is likely that if the goal is for greater recall, MeSH should be sufficient.

Lastly, if multiple terminologies are employed to represent the concepts embodied in CRFs and registries, further exploration of the benefits of the UMLS Metathesaurus will also be necessary. The Metathesuaurs is a powerful tool that enables, among other functions, interoperability among medical terminologies.

5. Dublin Core

A key question in this project is to determine what metadata elements are necessary to effectively index clinical research questions and permissible value sets given the previously described context. In particular, since a number of organizations across healthcare might engage in representing questions and answers for reuse and sharing, we plan to explore Dublin Core (DC) as a candidate scheme to better enable non-experts to effectively index their information (an underlying theme driving the development of DC).

https://doi.org/10.23106/dcmi.952109935

5.1 Advantages

Dublin Core is a small, generic set of metadata elements that is potentially useful in any context, and using Dublin Core is inexpensive and relatively easy to maintain. If one is considering making a particular resource open access, it is important to consider that OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting) providers must deliver DC XML metadata records as a minimum (Open Archives Initiative, 2008). Dublin Core appears to meet requirements for interoperability through mapping and crosswalks (Day, 2002). Additionally, Dublin Core encourages use of a number of controlled terminologies, and the usage guide recommends using controlled vocabularies for improving search results (Dublin Core, 2005), something central to our effort. There are precedents in using Dublin Core with MeSH, for instance, including the CISMeF project (http://www.chu-rouen.fr/cismef/cismefeng.html).

The ability to extend Dublin Core to accommodate special contexts is an attractive feature. As noted, clinical research questions and answers do not exist in isolation, but in the context of a particular disease, the study's protocol and requirements, and often as part of a scientifically tested instrument (such as some psychometric measures). Initial review of potential extending elements covered in Dublin Core reveal a number of opportunities to fully represent our information objects.

5.2 Disadvantages

Despite the advantages of using Dublin Core, there are several concerns that will need to be explored before making a final determination as to its usefulness to our project. Metadata schemes in healthcare abound, though are often in early development and acceptance stages. In healthcare, it is critical to understand the interrelationships of where information is generated and with what other systems it is likely to be shared. In other words, Dublin Core in and of itself may not be robust enough to enable this level of interoperability. It may need to be augmented by other types of metadata to address specific needs and minimize the loss of information. Future tests and comparisons will help elucidate these challenges.

6. Conclusion

The consistent and accurate application of data standards to represent and organize clinical research questions and permissible answer sets is a critical challenge facing clinical research informatics. The benefit is that standards help facilitate semantic interoperability, questions reuse, information and data management, accurate external reporting, and translation of research results into practice. Representing and organizing clinical research questions and their permissible answers is basically an indexing problem. However, context plays a major role in indexing rare disease research questions. Dublin Core may have the requisite flexibility to index the encoded question and answer pairs, but may only be effective once interoperability with other metadata efforts in healthcare is established.

References

Andrews, J. E., Richesson, R. L., & Krischer, J. (2007). Variation of SNOMED CT coding of clinical research concepts among coding experts. *Journal of the American Informatics Association*, 14(4), 497-506.

Brandt, C. A., Cohen, D. B., Shifman, M. A., Miller, P. L., Nadkarni, P. M., & Frawley, S. J. (2004). Approaches and informatics tools to assist in the integration of similar clinical research questionnaires. *Methods of Information in Medicine*, 43(2), 156-162.

Day, M. (2002). Metadata. Mapping between metadata formats. Retrieved April 9, 2010, from http://www.ukoln.ac.uk/metadata/interoperability/.

Dublin Core® Metadata Initiative (2005). Using Dublin Core. Retrieved April 6, 2010 from http://dublincore.org/documents/usageguide/.

Duftschmid, G., Gall, W., Eigenbauer, E., & Dorda, W. (2002). Management of data from clinical trials using the ArchiMed system. *Medical Informatics and the Internet in Medicine*, 27(2), 85-98.

https://doi.org/10.23106/dcmi.952109935

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(6), 594-601.

Open Archives Initiative. (2008) The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved April 6, 2010 from http://www.openarchives.org/OAI/openarchivesprotocol.html.

Painter, Jeffrey and Natalie Flowers. (2009) CodeSlinger: An Interactive Biomedical Ontology Browser. *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine*, AIME 2009, Verona, Italy, July 18-22, 2009; proceedings. Berlin: Springer. http://dx.doi.org/10.1007/978-3-642-02976-9.

Richesson, R. L., Andrews, J. E., & Krischer, J. P. (2006). Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. *Journal of the American Informatics Association*, 13(5), 536-546.

Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: gaps, overlaps, challenges and future directions. *Journal of the American Informatics Association*, 14(6), 687-696.

https://doi.org/10.23106/dcmi.952109935