A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki

Guoqian Jiang[1], Harold R. Solbrig[1], Dave Iberson-Hurst[2], Rebecca D. Kush[2], Christopher G. Chute[1]

1 Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN, 55905
2 Clinical Data Interchange Standards Consortium (CDISC), Austin, TX, 78746.

Abstract

Semantic interoperability among terminologies, data elements, and information models is fundamental and critical for sharing information from the scientific bench to the clinical bedside and back among systems. To meet this need, the vision for CDISC is to build a global, accessible electronic library, which enables precise and standardized data element definitions that can be used in applications and studies to improve biomedical research and its link with health care. As a pilot study, we propose a representation and harmonization framework for clinical study data elements and implement a prototype CDISC Shared Health and Research Electronic Library (CSHARE) using Semantic MediaWiki. We report the preliminary observations of how the components worked and the lessons learnt. In summary, the wiki provided a useful prototyping tool from a process standpoint.

Introduction

While tremendous progress has been made in biomedicine through the application of information technology, the information generated in biomedicine remains largely disconnected and disjoint [1]. In efforts to address this problem, a number of large projects in the biomedical research community have explored and built infrastructure and data systems utilizing an architecture that facilitates system interoperability [2-4]. The hope of such interoperable systems is that the speed and impact of research will be increased [5]. In this context, semantic interoperability among terminologies, data elements, and information models is fundamental and critical for sharing information from the scientific bench to the clinical bedside and back among systems.

In addition, there has been an increasing need to standardize the way certain data are collected and stored, transferred or reported across the institutions involved [6-7]. For instance, the National Cancer Institute (NCI) supports a broad initiative to standardize the common data elements (CDEs) used in cancer research data capture and reporting [8]. Notably, the Cancer Data Standards Repository (caDSR) was developed to address these needs and NCI caDSR chose the ISO/IEC 11179 standard for metadata registries to represent the common data elements (CDEs) in the database and implemented a set of APIs and tools used to create, edit, control, deploy and find the CDEs for metadata consumers and for UML model development. This infrastructure is being leveraged for cancer research by the National Cancer Institute's cancer Biomedical Informatics Grid (caBIG) [4, 9]. Since the representation of models within caBIG is complex and getting more complex, the community is facing the harmonization-scaling problem and the need for improved tooling to navigate the model space is urgent [5]. Additionally, to form better community adoption and governance, a more open, scalable and collaborative platform is desired.

Facing similar interoperability challenges, the stakeholders of Clinical Data Interchanges Standards Consortium (CDISC) [10] have made it clear that there is a pressing need to fill the gaps in the content of the existing standards, to bring those standards into semantic alignment while at the same time developing related therapeutic area standards. In addition, the ability to use EHR data in medical research is becoming increasingly attractive, which emphasizes the importance and value of harmonized common vocabularies/definitions across research and healthcare data. To meet these needs, the vision for CDISC [10] is to build a global, accessible electronic library, which enables precise and standardized data element definitions that can be used in applications and studies to improve biomedical research and its link with health care.

Wiki as a collaborative system provides tools for user participation into common tasks within a community, e.g., discussion pages. Combined with semantic web technology, *semantic wiki* provides the ability to capture (by humans), store and later identify (by machines) further meta-information or metadata about those articles and hyperlinks, as well as their relations [11] and has been demonstrated as an appropriate platform for knowledge engineering methods to work on the different levels of the continuum [12] (described in more detail in next section). For instance, a platform known as LexWiki [13] based on Semantic MediaWiki [14] enables the wider community to make both structured and unstructured proposals on the definitions of classes and property values, suggest new values, and

corrections to the current ones. LexWiki currently is at the core of community-based development of Biomedical Grid Terminology (BiomedGT) [15].

In this pilot study, we propose a collaborative framework for representation and harmonization of clinical study data elements (i.e. a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes). We implement a prototype of CDISC Shared Health and Research Electronic Library (CSHARE) using Semantic MediaWiki. We report the preliminary observations and evaluations of how the components worked and the lessons learnt.

Background

The mission of CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. Over the past decade, CDISC has fulfilled its mission by publishing and supporting a suite of standards that enable the electronic interchange of data throughout the lifecycle of a clinical research study. Specifically, CDISC has developed standards for use across the various points in the research study lifecycle: Protocol Development (Protocol Representation Model Version 1); data collection: Clinical Data Acquisition Standards Harmonization (CDASH); exchange of operational data: Operational Data Model (ODM); exchange of clinical laboratory data: (LAB); and data submission to regulatory agencies: Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM). As adopters have realized the benefits of these standards, it has become apparent that there is a need for a foundational standard to support *computable semantic interoperability* – the predictable exchange of meaning between two or more systems - across multiple standards including, but not limited to, those developed by CDISC.

Semantic MediaWiki

A number of semantic wikis exist in many different flavors. The most notable system currently is Semantic MediaWiki (SMW) [13] which is an extension of the popular wiki engine MediaWiki. SMW provides an extension that enables wiki users to semantically annotate wiki pages, enabling content browsing, searching, and reuse in novel ways. For example, the following text: Rochester is a [[is_a::City]] located in the southeastern part of [[is_located_in::Minnesota]] establishes the facts "Rochester is a city" and "Rochester is located in Minnesota". Here, "is_a" and "is_located_in" are called properties and defined in wiki pages in "Property" namespace. The formal semantics of annotations in SMW is given via a mapping to the OWL DL ontology language. Most annotations can easily be exported in terms of OWL DL, using the obvious mapping from wiki pages to OWL entities: normal pages correspond to abstract individuals, properties correspond to OWL properties, categories correspond to OWL classes, and property values can be abstract individuals or typed literals.

Methods

Representation and Harmonization Framework

We worked with the CSHARE community to determine the best format for loading the contributed community content. After some discussion and review, we settled on two spreadsheets – one for the data element descriptions and a second for loading the code lists (or value sets). Once the model was determined, we created a formal UML model that was used to map the spreadsheet content into the wiki (see Figure 1). The UML model also described how the loaded content was mapped to terminology, data types and the points at which the content would be aligned.

After a series of iterative explorations a prototype harmonization process was arrived at. This process involved into three steps: 1) Annotation - description and categorization of the individual data elements. This step involved adding names, definitions and semantic categorization to the individual data elements that were supplied by the evaluation community. This step was done by individual community members who were familiar with the use and purpose of the elements. 2) Selecting and sorting the annotated data elements to locate those that were closely related. This step has been referred to as "slicing and dicing" (i.e. analyzing the data elements in different views and perspectives). 3) Locating or, if necessary, creating one or more common data elements that represent the community semantics represented by the selected elements. This step also involved establishing the closeness of the match between the community data elements and common element.

Prototype Implementation

The CSHARE evaluation wiki was based on the Mediawiki software stack. For the purposes of the CSHARE evaluation, the baseline Mediawiki software was enhanced with several extensions, including SMW [14]. One of the extensions is SMW Halo [16] – an add-on to SMW that enables Ajax based query of wiki semantics and in-line text annotation. The Semantic Forms extension [17] enables forms based wiki data entry and the LexWiki extension [18] that provides a model and a set of access methods for thesauri, classification schemes and ontologies.

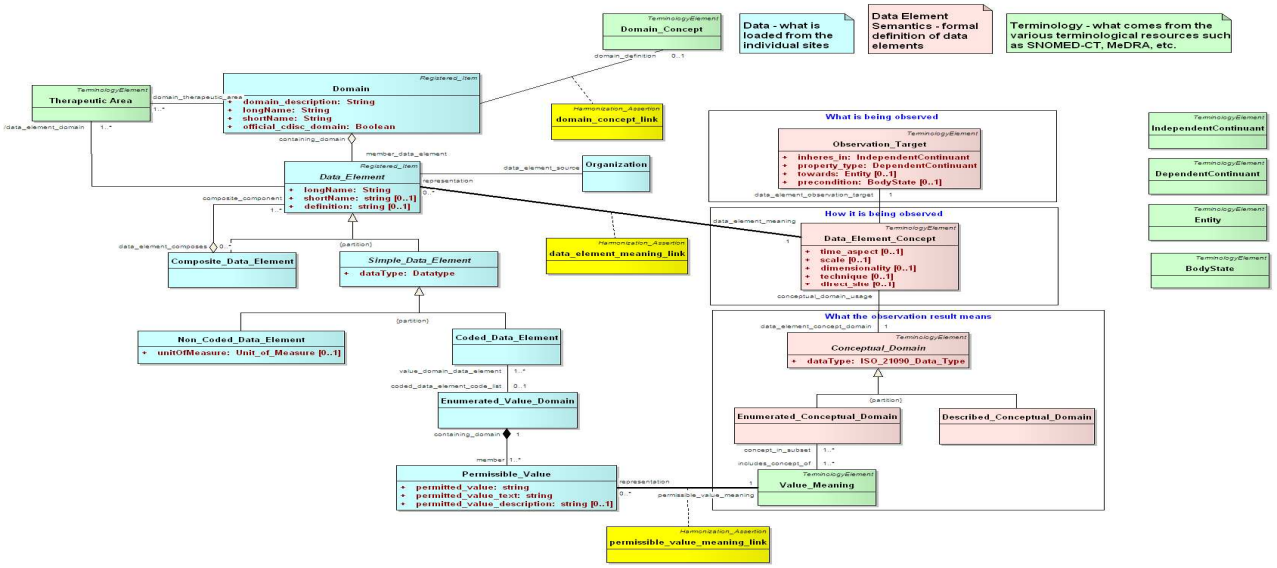

Figure 1. CSHARE wiki data element representation and harmonization model.

The evaluation wiki was customized where necessary to accommodate specific CSHARE requirements. Examples include enhancement of the SMW Halo search capabilities to provide more sophisticated terminology search, software to load the Excel spreadsheet content into the wiki, tools to map concept codes from and two different coding schemes and enhancements to the Semantic Wiki query tool to allow queries from object to source as well as the built in source to object.

The CSHARE wiki was loaded with a subset of NCI Metathesaurus [19] containing the following terminologies: SNOMED CT, NCI Thesaurus, HL7 Version 3, ICD-9-CM, MeDRA (subset), ICD-10, CDISC CDASH Terminology, CDISC SDTM Terminology, and CDISC SEND Terminology. In addition, the BRIDG model [20] and ISO 21090 [21] (datatypes) were loaded; the BRIDG model was represented as both vocabulary and data elements.

The wiki was also loaded with data elements and codelists from 11 domains (including Adverse Events, Lesion Measurement, Blood Products, etc.) in the oncology therapeutic area submitted from the community members in the standardized spreadsheet format. Considerably later in the evaluation process, it was determined that it would be very useful to have both the CDASH and SDTM content available in the wiki as well. These were loaded as data elements from the CDISC namespace.

The wiki environment was tailored to attempt to meet the needs of the above harmonization process. In particular, it became heavily dependent on a tool, Exhibit [22], produced by the MIT Simile environment. This tool formed the framework of the "slicer and dicer", and was one of the more successful elements of the prototype, although it would certainly need to be enhanced and streamlined to function usefully in a production environment.

Preliminary Findings

Terminology Components

The terminology served four roles in the harmonization process:

1) Classification: The slicing and dicing process depended on "semantic keywords" to determine whether two or more components were related. Formal terminology such as SNOMED-CT, the NCI Thesaurus, etc. provided controlled terminologies from which these keywords could be drawn.

2) Definition: Terminological resources provided the potential for formally defining the intended meaning of both the community supplied data elements and the harmonized data elements. Note that this is not the same as classification, as the purpose is to provide a formal and precise definition of the particular resource, where a classification is to provide a list of terms that might be used in conjunction with similar related elements.

3) Value Meanings: Each of the individual values for enumerated data elements needed to be linked to a terminological element that indicates their intended meaning. As an example, a "1" in a Mayo patient gender data element might mean "male", and needs to be mapped to a corresponding concept code in a standard terminology.

4) Value Sets: Value sets represent collections of value meanings. As an example, a value set might represent possible anatomical locations, either in a

particular or general context. The ability to determine the nearest value set that contained all of the value meanings for a particular data element turned out to be quite valuable when it came to determining when two or more data elements might be related.

Process Components

The wiki environment served well as a vehicle for discussing the prototype. The availability of all of the terminological components in a single form, the ability to locate specific and sets of data elements, etc. and the ability to rapidly change the layout and content of forms proved to be very useful.

The wiki environment seemed less than ideally suited for much of the harmonization process. Semantic MediaWiki is a relatively free-form, customizable medium for publication and discussion. It is less than ideal for processing large lists of values, batch mapping, sorting and selecting, etc. It did, however, present considerable potential for the purposes of discussion, evaluation and dissemination. We believe that a hybrid model, based in part on enhanced spreadsheets, customized applications and Semantic Mediawiki may provide a workable platform for the harmonization process. It should also be noted that while the Semantic Mediawiki appears to be a useful mechanism for publishing harmonized content, it is probably not the ideal vehicle for communicating formal mappings and/or providing repository services. We would recommend creating an Operational Data Model (ODM) [23] import/export mechanism and a set of enhanced ODM based services for that. Our experience also suggested replacing Excel spreadsheets as the primary import format with loading the individual organization forms directly into the wiki and doing the extraction and annotation process directly within the wiki.

Discussion

While this evaluation is obviously very limited in nature, we observed that:

1) It was difficult to find the set of terminological components that were needed for classification. A search on almost any term name ("lesion size", "disease stage", etc.) yielded tens or even hundreds of possible terminological matches. We believe that there are at least two tasks that must be completed before this sort of terminological annotation becomes viable: a) Terminology must be pre-vetted for classification. A community subject matter expert needs to create a list of classification "value sets" from which classification elements for a particular domain should be drawn. This needs to be done in such a way that missing elements can be added as needed. It also should be noted that it isn't obvious that it is necessary for these value sets to be drawn from existing terminology, although there will be benefits if it could be b) Terminology tools need to be considerably more sophisticated than what is available from SMW Halo or even the Mayo extensions. Users need to be able to search by name, definition, code system, parent code, related code, and need to be able to easily display the details of a particular concept – both its textual and its associations with other concepts within selection dialog box.

2) Definitions require a model. A "pile of concepts" are not sufficient to define the intended meaning of a data element or common data element. A model, such as that observable model being developed by the IHTSDO [24] community identifies the various components that are needed to completely define data elements while simultaneously limiting the possible selections for the various aspects of the model. The model also normalizes the granularity of various definitions.

3) The ability to map value meanings to common terminology increases the ability to discover overlap. If, for instance, one community maps information to NCI Thesaurus codes and a second to SNOMED CT codes, the mapping work done by the NLM and NCI in the NCI Metathesaurus makes it possible to discover overlap and potential shared content.

4) None of the terminologies carried good value set definitions, though it was often possible to map individual elements. As an example, Eli Lilly provided a rich value set called "Lesion Method of Measure". While most of the individual values in this set had matching meanings in the terminology space (e.g. "103" maps to SNOMED CT Code 289935006), there didn't seem to be any useful upper level container that represented all of the possible methods. This may be significant, as set members do not necessarily correspond to ontological ordering.

5) Data types played a key role in classification. This said, the ISO 21090 data types appeared to be overkill, as we seemed to be interested in a very limited set (text, date/time, coded, numeric, …), and the nuances such as flavors of null, SET vs. BAG, CD vs CS, PQ vs PQR vs. INT, etc. went beyond what was needed for classification. Note, however, that the mapping from data elements to common data elements, a step that was discussed but not implemented in this prototype may draw heavily on the details of the ISO 21090 types.

6) The BRIDG model, by and large, was too coarse to add much significant information to what was already known. As with the ISO data types, it appeared that the BRIDG model could play a key role in subsequent model alignment steps, but was of little value from the harmonization perspective.

7) Units, as represented by the HL7 V3.0 UCUM (Unified Codes for Units of Measures) system,

played an insignificant role in the harmonization process. It appeared, however, that the notion of dimensionality (e.g. length, area, pressure, concentration, etc.) might play a useful role in the harmonization of quantitative data elements. Doing this, however, would require the selection of a baseline set of dimensions along with mapping to and from UCUM.

In general, the terminological component added significant value. It is particularly interesting to compare some of the annotations that have been done in the context of the caDSR with those done within this prototype – they are quite similar in coverage and quality. Not unexpectedly, however, the terminology is no "silver bullet". It is both too much and too little, and tools would need to be provided that aided in the selection of the right concept(s) from the terminology when they existed, and in the construction of post-coordinated concepts and sets when they didn't. In addition, tooling which did reasoning across the terminologies would be invaluable – both in discovering similar broader/narrower elements and in comparing pre- and post coordinated terms.

It should be noted, however, that SNOMED CT, the NCI Thesaurus, HL7V3.0 and UCUM each potentially play a different role. SNOMED CT provided broad coverage, for categorization and has the potential to be a primary candidate source for definitions, due to alignment with the IHTSDO model and the strong formal semantics. The NCI Thesaurus was the primary source of value sets, which is not unexpected as the NCI Thesaurus is where the CDISC data elements have been recorded to date. The HL7V3.0 terminology provides alignment with HL7 V3 specific messages.

Summary

The wiki was loaded with approximately 380,000 terms drawn from 9+ terminologies. While the terminology proved extremely useful in locating potentially similar data elements, the process was not nearly as efficient as it could be were more refined tooling and domain-appropriate subsets available. Many of the data elements required more than one code to categorize and/or define, meaning that reasoning capability will be needed to be able to match "pre-coordinated" with "post-coordinated" terms. A formal observables model such as that being developed by IHTSDO would potentially be useful from both completeness and appropriate level of granularity aspect. The NCI Thesaurus provided most of the value sets that were found, and would make the best candidate for registering future value sets. The NLM and NCI mappings between code systems appear to provide considerable value. From a process standpoint, the Wiki provided a useful prototyping tool, but was less than ideally suited for many of the batch sorts of tasks.

References

[1] Buetow KH. Cyberinfrastructure: empowering a "third way" in biomedical research. Science. 2005 May 6;308(5723):821-4.
[2] The Biomedical Informatics Research Network (BIRN): http://www.nbirn.net.
[3] The myGrid project: http://www.mygrid.org.uk.
[4] The cancer Biomedical Informatics Grid (caBIG): http://caBIG.nci.nih.gov.
[5] Kunz I, Lin MC, Frey L. Metadata mapping and reuse in caBIG. BMC Bioinformatics. 2009 Feb 5;10 Suppl 2:S4.
[6] Winget MD, Baron JA, Spitz MR, Brenner DE, Warzel D, Kincaid H, Thornquist M, Feng Z. Development of common data elements: the experience of and recommendations from the early detection research network. Int J Med Inform. 2003 Apr;70(1):41-8.
[7] Patel AA, Kajdacsy-Balla A, Berman JJ, Bosland M, Datta MW, Dhir R, Gilbertson J, Melamed J, Orenstein J, Tai KF, Becich MJ. The development of common data elements for a multi-institute prostate cancer tissue bank: the Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. BMC Cancer. 2005 Aug 21;5:108.
[8] Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003:1048.
[9] Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado S, Reeves DM, Hadfield JB, Ludet C, Covitz PA. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform. 2008 Feb;41(1):106-23. Epub 2007 Apr 2.
[10] The CDISC URL: http://www.cdisc.org/.
[11] Kamel Boulos MN. Semantic Wikis: A comprehensible introduction with examples from the health sciences. Journal of Emerging Technologies in Web Intelligence. 2009; (1): 94-96.
[12] Baumeister J, Reutelshoefer J, and Puppe F. Engineering on the Knowledge Formalization Continuum. Proceedings of the Forth Semantic Wiki Workshop (SemWiki 2009), co-located with 6th European Semantic Web Conference (ESWC 2009). Hersonissos, Heraklion, Crete, Greece, June 1st, 2009
[13] Jiang G, Solbrig H. LexWiki framework and use cases. The first meeting of Semantic MediaWiki users. November 22-23, 2008. Boston, Massachusetts, USA. The sildes are available at https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexWiki#Presentations.
[14] Krötzsch M, Vrandecic D, Völkel M, Haller H, Studer R. Semantic Wikipedia. Journal of Web Semantics 5: 251--261. September 2007. ISSN: 1570-8268
[15] BiomedGT URL: http://biomedgt.nci.nih.gov/index.php/Main_Page .
[16]. Halo extension URL: http://www.mediawiki.org/wiki/Extension:Halo_Extension.
[17] Semantic Forms extension URL: http://www.mediawiki.org/wiki/Extension:Semantic_Forms.
[18] LexWiki extension URL: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexWiki.
[19] NCI Metathesaurus URL: http://ncimeta.nci.nih.gov/.
[20] BRIDG model URL: http://bridgmodel.org/.
[21] ISO 21090 URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=35646.
[22] Semantic Result Formats – Exhibit extension URL: http://www.mediawiki.org/wiki/Extension:Semantic_Result_Formats/exhibit_format.
[23] ODM URL: http://www.cdisc.org/models/odm/v1.3/index.html.
[24] IHTSDO URL: http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-hierarchies/observable-entity/.