TIM: A Semantic Web Application for the Specification of Metadata Items in Clinical Research

Matthias Löbe[1], Magnus Knuth[2], Roland Mücke[3]

[1] Clinical Trial Center Leipzig, University of Leipzig, Germany
matthias.loebe@zks.uni-leipzig.de
[2] Institute for Medical Informatics, University of Leipzig, Germany
magnus.knuth@imise.uni-leipzig.de
[3] itemis AG, Leipzig, Germany
roland.muecke@itemis.de

Abstract. This paper presents a Semantic-Web-based application for specifying metadata items in clinical research and life sciences. Items are complex data structures with a multitude of characteristics, most of them annotations linking to other concepts. They are utilized as templates to instantiate variables for biomedical experiments and therefore should be detailed, consistent, and – in order to ensure the comparability of results from different experiments – commonly used. We developed an ontology for item modeling and an application for managing a repository of items based on W3C standards.

Keywords: metadata repository, item management, clinical trials, CRF creation

1 Introduction

Randomized controlled clinical trials are called the gold standard in clinical research. They are studies on human subjects that test a hypothesis by statistical means. The data base for the necessary calculations is a number of answers to medical questions collected by use of case report forms (CRFs). However, it is not easy to specify data elements – question-value-pairs often called *trial items* – precisely and consistently, since it requires not only knowledge about the medical domain but also about additional constraints, like conventions of value coding, units of measurement, normal ranges, or conceptual relations. Furthermore, specifying a clinical trial is a concerted effort involving many people with different backgrounds. Physicians initiate a new trial after identifying a new scientific problem, biometricians help to write the study protocol and ensure that all relevant parameters are recorded, database programmers set up a database for storing trial items and consistency rules, and data managers design CRFs based on the specification. In order to coordinate that effort a lot of communication is required, especially in academic institutions focusing on non-commercial trials where people involved in specifying a trial often work in different departments. Another issue is the interoperability of software used. As people involved in the process may use different software stacks lacking interoperability there is a risk that changes cannot be propagated and have to be made manually in every single document or program.

Yet another point is that with the clinical problem varying from trial to trial even experienced biometricians will have difficulties in keeping abreast of all items and item variants used. This may lead to incomplete or inconsistent specifications delaying the overall specification process or, even worse, affecting the subsequent analysis of trial results. There are initiatives to define global item libraries [1] or disease-specific local core-data sets, still they cover only a fraction of the required items and lack some semantics, too.

2 Application

We present a tool called Trial Item Manager (TIM). It is based on Semantic Web technology and can be used to create CRFs and to manage clinical trial items. TIM is a Web 2.0 application written in Java featuring a rich Ajax client that allows collaborative editing. Thus it offers the usual advantages of web applications (accessible from every computer connected to the internet, central repository, software-as-a-service, no client-side modifications necessary). It is based on the Jena Framework [2] and utilizes Pellet [3] as a reasoner.

2.1 Model

TIM's data model arises from a conceptual approach; we state that everything applicable to structural nesting is subsumed by the notion of a *Component*, e.g. items, trials, CRFs, etc. Moreover, components can exhibit arbitrary *characteristics* or be related via these characteristics. A special characteristic is that which models the nesting of components, called *contains*.

The application utilizes a set of ontologies: a vocabulary and a component store. TIM's data model is based on the Resource Description Framework (RDF) [4] and is supplemented by a vocabulary ontology expressed in the Web Ontology Language (OWL) [5]. A corresponding RDF model contains the core entities that classify the classes (*rdfs:Class ComponentType*, *rdfs:Class CharacteristicValueType*) and properties (*rdfs:Class CharacteristicType*) of the data model; it serves as a linkage between the hard-coded part of the data model and the ontology entities of the vocabulary store.

The vocabulary ontology contains all statements about items and other components of a trial, e.g. containers for the hierarchical composition of items, like *CRF*s consisting of *Module*s, containers for data values like *Checkfield*s or *Codelist*s. It formally defines how components can be composed and by which characteristics they can be described. Furthermore, the ontology may be modified or extended to adapt to external requirements and the application will recognize this change and readapt accordingly. Finally, the actual component instance data is stored separately from the vocabulary in an RDF Model – the component store. Every trial component is a resource in terms of RDF, and arbitrary statements may be associated with it. This approach offers a maximum of flexibility.

2.2 System Architecture

The architecture of the developed software tool corresponds to ideas described in the previous section, where ontologies control the behavior of the application.

The vocabulary ontology is used for a reasonable processing of the statements. It serves two purposes: First, it defines rules to restrict the assembling of individual statements and how they can be combined; second, it is interpreted by the software to adapt the applications behavior under certain conditions. For example, a trial item which contains a check field is classified as *Checkable Item*. Furthermore, all "Single Choice Items" may solely contain such "Checkable Items". TIM utilizes this kind of information from the ontology to guide users in the process of item specification. Classifying classes and properties, applying rules, and checking the consistency of the knowledge base is done automatically by the reasoner; which also means that these tasks do not need to be coded in software.

Only a few fundamental concepts for processing RDF models and OWL ontologies (classes, instances, properties, domains, ranges) and the core entities from the TIM vocabulary ontology (components, the containment relation and basic characteristics) are hard-coded in the source code. All other entities used to describe trial items and to adapt functionalities, e.g. conventions for hierarchical structuring or mapping to medical terminologies, are loaded from the ontologies at the runtime of the application. Thus, the behavior of TIM can be adjusted to local needs simply through modification of the underlying ontology. In order to implement these dynamics, we apply SPARQL [6] queries for filtering (search) and reverse containment retrieval of components, as well as for accessing the model's component and characteristic types. Simple modifications like the creation of new characteristics, such as a mandatory database field name, can be made via the configuration frontend within the application.

In some aspects, TIM follows a frame-oriented approach. Since RDF is not frame-based, there may be a discrepancy between facts about a component a user explicitly stated and statements the reasoner deduced. For example, when we apply a characteristic of the type *eudraCTnumber*, which has the domain *ClinicalTrial*, to a component of the type *TrialItem*, the reasoner will deduce that the component is "Trial Item" (explicitly stated) as well as "Clinical Trial" (inferred by the domain). Therefore, it is necessary to differentiate between explicit and inferred facts. To achieve this we employ asserted and inferred RDF models.

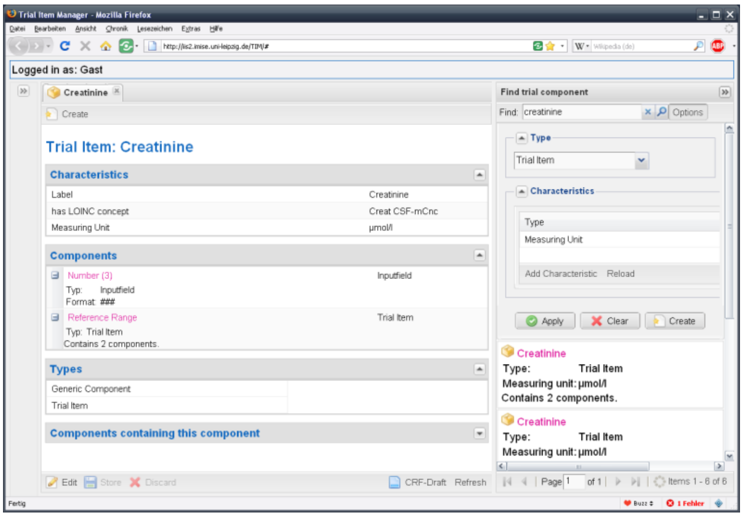
Internally, we handle five RDF stores, i.e. explicit statements about the components and about the vocabulary exist in separated raw models and are also available in a combined model. Additionally, inferred models, one for components and one for the vocabulary, contain up-to-date generated entailments. Queries can thus be directed to either the original RDF model which comprises only assertions or to several derived models based on the available ontologies.

3 Discussion

TIM benefits from Semantic Web technologies in numerous ways. The graph-based model of RDF is much more flexible and "conceptually closer" to the problem domain than a traditional relational approach (which requires a lot of mapping tables and has few possibilities of explicitly specifying relations or inheritance). OWL has expressive constructs that allow checking for valid domain and range values (not only primitive data types) without additional programming. It is easily extensible, at least for a knowledge engineer, so that modifications can be made without necessarily recompiling the source code.

The availability of reasoners allows checking for inconsistencies and classifying objects based on their properties rather than on explicit statements. Furthermore, the application can support the user in the process of item specification through reasonable assumptions, e.g. that a German-speaking user is likely to prefer a label in German or that blood pressure should have a measurement unit that is a *Pressure Unit*.

A very important issue is searching and navigating in the item repository. Handling libraries of thousands of items is surely a usability issue. Of course, it should be possible for a user to find a desired item in far less time than it would take to specify the item anew. Textual search is not enough. On the one hand, many labels, descriptions and definitions can contain the search string without being conceptually related to the search item. On the other hand, the set of search results can nevertheless be very large. LOINC [7], for instance, has 1,300 hits for "creatinine" in the COMPONENT axis alone. TIM has an advanced search interface that combines textual search with semantic search. Figure 1 shows an example search for "creatinine" that restricts the result set to resources of the type "Trial Item" with the characteristic "Measuring Unit".


Trial Item Manager - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://lis2.imise.uni-leipzig.de/TIM/#
Wikipedia (de)
Creatinine
Create
Edit Store Discard
CRF-Draft Refresh
Find trial component
Options
Add Characteristic Reload
Apply Clear Create
Fertig
Buzz
1 Fehler

Logged in as: Gast

Trial Item: Creatinine

Characteristics

Label	Creatinine
has LOINC concept	Creat CSF-mCnc
Measuring Unit	µmol/l

Components

Number (3) Typ: Inputfield Format: ###	Inputfield
Reference Range Typ: Trial Item Contains 2 components.	Trial Item

Types

Generic Component
Trial Item

Components containing this component

Find: creatinine

Type: Trial Item

Characteristics

Type
Measuring Unit

Creatinine
Type: Trial Item
Measuring unit: µmol/l
Contains 2 components.

Creatinine
Type: Trial Item
Measuring unit: µmol/l

Page 1 of 1
Items 1 - 6 of 6

Fig. 1. Right side: example search for items labeled "creatinine" using a type (trial item) and a characteristic (measuring unit) restriction; left side: detailed view of the first item of the result set including characteristics and subcomponents

However, we also observed some limitations of a Semantic Web approach. While the number of items in the repository increased steadily, the performance of the application did the contrary, especially when changes were stored in the knowledge base. In addition, it is currently hard to find mature RDF data on the Web, which is the case with both existing trial specifications and medical terminologies.

Currently, the item repository contains 5 pre-existing clinical trials with about 2,500 genuine items from domains like oncology, cardiology, and infectious diseases. While no formal evaluation of usability has been done so far, biometricians experienced in using web applications have been successful in applying TIM in order to model own CRFs, given a short introduction. Frequent reuse rates were found for individual code lists (216 times "yes-no"), as well as some modules within the same trial (16 times "patient characteristics"). We expect that planning new trials will also elevate the item reuse rates.

4 Related Work

Several software systems exist that are designed (or can be used) as repositories for data element collection. The most well-known effort is National Cancer Institute's (NCI) Cancer Data Standards Registry and Repository (caDSR) [8], a part of the Cancer Common Ontologic Representation Environment (caCORE) framework [9], which has been developed in the context of the Cancer Biomedical Informatics Grid (caBIG).

caCORE's main goal is to provide an infrastructure for syntactic and semantic interoperability. To achieve this, it utilizes "semantic metadata" elements stored and maintained in the caDSR. The basic information artifact that corresponds to an item in TIM is the Common Data Element (CDE), an instantiation of *Data Element* from the ISO 11179 "Metadata Registry" meta-model [10]. A CDE consists of two basic components, a *Data Element Concept* (DEC) – the specification of a concept – and a *Value Domain* (it's representation). caDSR uses the Enterprise Vocabulary Server (EVS) for mapping DECs to ontologies and terminologies, e.g. the NCI Thesaurus. A tool called Form Builder allows for creating CRFs as a collection of data elements.

Another implementation of ISO 11179 is the eXtended MetaData Registry (XMDR). Contrary to caDSR, it relies on Semantic Web languages and technologies. It uses an OWL implementation of ISO 11179, the Jena API, and the Pellet reasoner. Therefore, it can support formal ontologies and reasoning and is able to formally specify statements and relations between elements.

Over the last years, wikis gained a strong influence on terminology development and annotation of resources in Health Care and Life Sciences [11, 12]. Wikis are easy to use, in most cases open to a wide community; they support collaborative work and track changes in a version history. WikiHIT (www.wikihit.org) is an example for a platform for the development and refinement of clinical data definition. Semantic wikis can even more enhance the usability and offer valuable features: better information retrieval over structured, semantically tagged data, inference of "new" facts and maintaining consistency based on a formal knowledge representation. WikiProteins (proteins.wikiprofessional.org), BOWiki (bowiki.net), and WikiNeurons (neuroweb3.med.yale.edu/mediawiki/) are wikis for annotating biomedical data. The

LexWiki Distributed Terminology Development Platform (cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexWiki) is used in CDISC's Shared Health and Clinical Research Electronic Library (SHARE) pilot for developing a precise and standardized core data set for various disease domains to improve data quality and interoperability for biomedical research.

The openEHR methodology [13] separates the information model (which is expected to be mature and non-volatile) and the knowledge model (which may evolve over time). The knowledge is represented in the Archetype Definition Language (ADL). Archetypes are the basic semantic units. They can be composed into more complex data structures with constraints and terms referencing an external ontology. Archetypes could be shared in a repository similar to CDEs.

Finally, most clinical trial data management applications support the reuse of items by providing a "data dictionary" of specified items. The majority of these dictionaries have a vendor-specific architecture and proprietary data formats, but some support the CDISC Operational Data Model (ODM), an international standard for exchanging data and metadata about clinical trials that allows at least a syntactical conversation of data elements.

5 Outlook

In its current state, the application is no more than a proof of concept. Many features demanded by users do not relate to the Semantic Web part but to functionalities typical of commercial software: different import/export formats, better CRF drafts, copy-and-paste and so on. Furthermore, TIM lacks the "Web" of "Semantic Web" – currently it provides no SPARQL endpoint nor offers resolvable URIs or Linked Open Data.

For the future, we plan to enhance the application with respect to personalization. Different user groups focus on different tasks, e.g. a physician may prefer not to see database codes and formatting instructions. It should also prove advantageous to exploit and integrate existing ontologies from the OBO Foundry [14], because they contain concepts from the domain of clinical research as well as information artifacts. The same is true for maintained medical terminologies like LOINC [15], ICD [16], and SNOMED [17]. Another open issue is the reuse of items or substructures of items. An analysis has shown that users tend to reuse simple structures like date/time fields or "very good-good-neutral-bad-very bad" choices but do not do so in complex cases. Therefore, the application should support the rating and harmonization of item variants in order to assist in the development of sets of core-items for various clinical domains. That would not only improve the quality of the trial specification but would also lead to a better support of meta-analysis of results from different clinical trials.

References

1. Clinical Data Interchange Standards Consortium Inc. Clinical Data Acquisition Standards Harmonization (CDASH). Version 1.0, October 2008, http://www.cdisc.org/standards/cdash/index.html

2. Carroll J.J., Dickinson I., Dollin C., Reynolds D., Seaborne A.,Wilkinson K.: Jena: implementing the semantic web recommendations. In Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters. New York, NY, USA, May 19 - 21 (2004)
3. Sirin E. et al. Pellet: A practical OWL-DL reasoner, Web Semantics: Science, Services and Agents on the World Wide Web, 2007, Volume 5, Issue 2, 51–53
4. Manola F., Miller, E. (editors): RDF Primer. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge, Massachusetts (2004)
5. McGuinness D. L., van Harmelen, F.: OWL Web Ontology Language overview. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge, Massachusetts (2004)
6. Prud'hommeaux E., Seaborne A.: SPARQL Query Language for RDF. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge, Massachusetts (2008)
7. Regenstrief Institute, Inc. and LOINC Committee. Logical observation identifiers names and codes (LOINC ®). http://loinc.org (1994–2009)
8. NCI. The NCICB User Applications Manual. ftp://ftp1.nci.nih.gov/pub/cacore/NCICBapplications/NCICBAppManual.pdf
9. Komatsoulis G., Warzel D., Hartel F., Shanbhag K., Chilukuri R., Fragoso G., et al.: caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability, J Biomed Inform, 2007 April.
10. ISO/IEC JTC 1 SC 32: Information technology — Metadata registries (MDR). ISO Standard 11179, International Organization for Standardization (ISO), ISO/IEC JTC 1: Information technology, Subcommittee SC 32: Data management and interchange, Geneva, Switzerland, 2002–2005.
11. Cheung K.-H., Yip K., Townsend J., Scotch M.: HCLS 2.0/3.0: Health care and life sciences data mashup using Web 2.0/3.0, Journal of Biomedical Informatics, (41) 5, Semantic Mashup of Biomedical Data, October 2008, Pages 694-705
12. Boulos M.: Semantic Wikis: A Comprehensible Introduction with Examples from the Health Sciences. Journal or Emerging Technologies in Web Intelligence, 2009 August; 1(1)
13. Beale T., Goodchild A., and Heard S.: EHR Design Principles. openEHR Foundation; 2001
14. Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007 November;25(11):1251-1255 (2007)
15. Huff S., Rocha R., Mcdonald C., De Moor G., Fiers T., Bidgood W., et al.: Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5(3):276-292.
16. WHO. International statistical classification of diseases and related health problems: Tenth revision (ICD10). Version 2007, World Health Organization (WHO), Geneva, Switzerland (2007)
17. IHTSDO. Systematized nomenclature of medicine – Clinical terms (SNOMED-CT). International Health Terminology Standards Development Organisation (IHTSDO), Copenhagen, Denmark (1999–2009)