A Knowledge-Based System for the Specification of Variables in Clinical Trials

Matthias Löbe, Barbara Strotmann, Kai-Uwe Hoop, Roland Mücke

Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
University Leipzig
Härtelstraße 16-18
04107 Leipzig
matthias.loebe@imise.uni-leipzig.de

Abstract: Study variables in clinical trial specifications are often defined manually, depending on the knowledge and experience of the author rather than on commonly agreed methods and standards. Therefore, we argue that a knowledge-based system can support this task with regard to data quality, consistency and completeness. We present a model based on Semantic Web technology that is flexible enough to represent different granularities and views as well as mapping medical terminologies. That model has been implemented in a software application called Trial Item Manager (TIM).

1 Introduction

Controlled clinical trials are the preferred instrument to prove the effectiveness of new drugs or therapies. A core part of the clinical trial specification is to create Case Report Forms (CRFs). They usually consist of hundreds of clinical variables, so called *items*, which are representations of the medical concepts to be observed. Clinical trials as long-lasting, strongly regulated and costly experiments obviously require a clear definition which covers all relevant aspects of an individual item. Inaccurate definitions result in queries during the trial and may have a negative impact on result analysis and meta-analysis. Trial protocol and CRF creation is part of the work of physicians and biometricians. Today, this process is in many cases affected by personal preferences and experiences in previous trials. Although there are a lot of medical terminologies, only a few standard item datasets exist. CDISC's *Clinical Data Acquisition Standards Harmonization* (CDASH) [CD08] is the most current effort to build a list of standardized CRF data collection fields. CDASH however does specify only few parts of an item, basically the general clinical concept, a variable name and a recommendation category for collection. Other aspects including valid code lists, measurement units and observation methods are not continuously defined. The main reason is that an appropriate detailed level of clinical consensus is hard to achieve [Re99]. The system presented in this paper tries to overcome some of these problems. It utilizes a generic semantic model that enforces core rules but allows local tailoring. It supports different levels of abstraction, therefore providing granular views depending on the user's context, for instance its role in the trial specification process (biometricians, data manager, database programmer) [Ci98]. The model is implemented in the software application *TIM* using Semantic Web technology. TIM can improve data quality with regard to its support for reuse of well-proven and reviewed items out of an item repository. It assures consistency by using an ontology and logical reasoning. Furthermore, an adequate level of item detailing can be obtained by self-defining rules for data completeness and referencing external terminologies.

2 Meta-model

The meta-model contains all concepts necessary for describing trial components including parts and relations between each other as well as concepts for the declaration of rules. It uses the semantics of OWL [MH04] which is utilized by automatic reasoning in TIM.

The meta-model consists of five basal entities:

(1) trial components and

(2) trial components types represent and classify entities like items;

(3) characteristics,

(4) characteristic types and

(5) characteristic values analogous for attributes and their values.

Figure 1 shows the meta-model as an UML diagram. Associations are painted in different colors to symbolize certain aspects. Black lines represent *types*, red lines *range restrictions*, blue lines *values* and green lines *UML class restrictions*.

A *trial component* is a container for characteristics. The set of all characteristics describes the trial component. Every trial component has one or more trial component types which determine its processing and the interpretation of the characteristics. Every trial component is uniquely identified by an URI. Trial components correspond roughly to the data elements defined in ISO 11179 [IS03]. Every trial component forms a part of the specification of the items of a clinical trial. By the means of structural characteristics multiple trial components can be composed and items or other parts of a trial specification such as modules or CRF sheets can be constructed. Trial components correspond to resources in RDF [MM04].

The *trial component type* defines the mode of use of a trial component, i.e. how to interpret its characteristics, which characteristics are necessary, which are mandatory, and which trial components of what type can be connected by structural characteristics. The trial component types are used for the classification of trial components and can form a poly-hierarchy. Every trial component has at least one trial component type, but usually has several ones. On one hand this is a result of the hierarchical structure of trial component types and on the other hand it results from the necessity to classify a trial component with respect to different interpretational facets. Trial component types can be abstract which means that they can not be declared explicitly as a type of a trial component but will be assigned implicitly by the characteristics of the trial component. Trial component types correspond to classes in OWL.

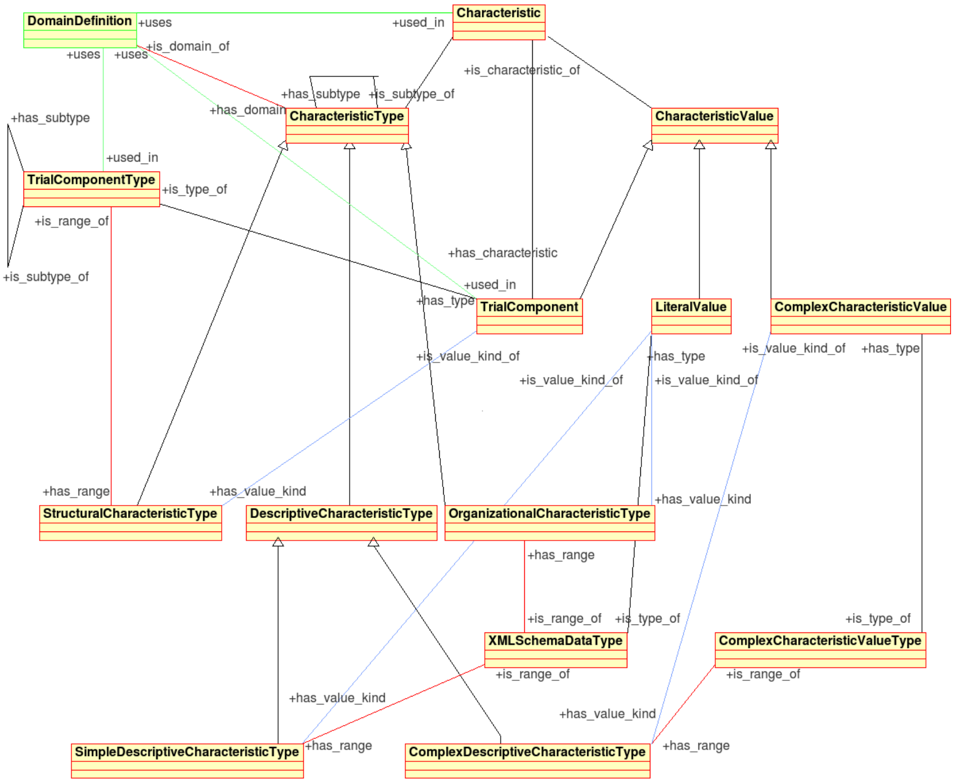


figure 1: Meta model

A *characteristic* assigns a characteristic value to a trial item and indicates – by the means of a characteristic type – how a characteristic value should be interpreted in respect to the trial component and its trial component type(s). A characteristic can be understood as a proposition of the form subject-predicate-object, whereas the trial component corresponds to the subject, the characteristic type to the predicate and the characteristic value to the object. A trial component can have any number of characteristics. The set of all characteristics forms a complete description of a trial component. The assignments of trial component types to trial components are not characteristics because these assignments are partly deduced implicitly from the set of explicitly declared characteristics. There is a distinction between structural, descriptive and organizational (administrative) characteristics: structural and descriptive characteristics are created explicitly by the user, organizational characteristics are produced automatically by the system. Thus, a trial component can have multiple characteristics with the same characteristic type but different characteristic values. Moreover, it is also possible to limit the number of characteristics with the same characteristic type (e.g. only one characteristic with the characteristic type “label”). Characteristics correspond to statements (triples) in RDF.

The *characteristic value* is the entity which is assigned to a trial component by a characteristic. Depending on the characteristic type, the characteristic value can be a simple literal, a complex characteristic value or another trial component. Characteristic values themselves have types, for instance literals are characterised by XML schema data types, complex characteristic values by complex characteristic value types and trial components by trial component types. These characteristic value types determine the interpretation and processing of characteristic values. Characteristic values correspond to objects in RDF statements. They can be either literals or resources.

A *characteristic type* defines how to interpret and process a characteristic of a trial item. The characteristic type is used to manage which trial components could be assigned to characteristics of that particular characteristic type. Furthermore, the range of the characteristic type defines which characteristic values are allowed. Vice versa, the declaration of domain and range can be used to automatically classify the characteristic values of existing characteristics of that characteristic type (that is the intrinsic intention of the RDF/S semantics). There exist three subclasses of characteristic types: structural characteristic types for the structural and hierarchical relations between trial components, descriptive characteristic types for the non-structural and associative description of trial components and organizational characteristic types used for internal administration of trial components. Characteristic types can form a poly-hierarchy, hence, multiple characteristic types can be implicitly assigned to a characteristic. Characteristic types correspond to the predicates of statements and to properties in RDF/S and OWL respectively.

3 Data Model (Ontology) and Rules

A data model is an instance of a meta-model. Therefore, the data model consists of instances of *trial component type* and *characteristic type*. The assembly of these is regulated by a system of rules.

There are three categories of trial component types (see figure 2). *Structural component* is the super class for any trial component that is used for hierarchy building. *Trial item* is the most important structural component; further ones are container components like *module* or *study*. Trial items are versatile; they can have characteristics of all categories of characteristic types (structural, descriptive, organizational), they can consist of sub-items to represent complex items like “blood pressure systolic/diastolic” and they can contain atomic form components and code list components. *Atomic form component* is the super class for representing form fields like *input field* for string values or *checkable field* for binary values. Atomic form components store no layout preferences; they only have syntactical/grammatical characteristics like the set of allowed characters, the minimal and maximal length or simple data types. One example is a component representing a date using the format “##.##.####”. Any information related to the clinical concept is stored in trial components, not in form components, so they can easily be reused. Atomic form components are called *atomic* because they must not have sub-components. *Code list component* is the super class for single choice and multiple choice code lists. Code lists themselves aren’t trial items but lists of items.

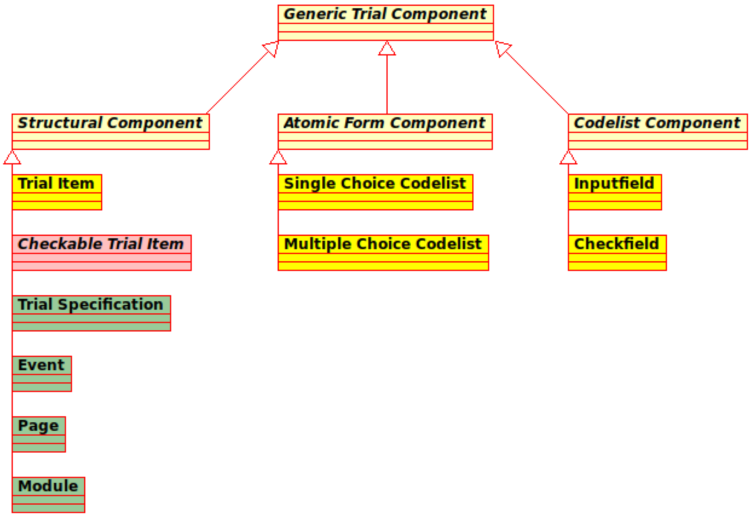


figure 2: Data model for trial component types

Characteristic types can be simple or complex. They are named *simple* if their value is an XML-Schema data type, for instance “label” (string) or “length” (integer). *Complex* descriptive characteristic types have complex characteristic values. An example is “#kg”, a unit of measure that has multiple characteristics of its own, e.g. labels in different languages, and types of “measuring unit” and of “SI base unit” (International System of Units).

Trial component types and characteristic types can be *core types* or *domain-specific types*. Core types are mandatory for every rule system. They provide elementary features. Domain-specific types are assertions that are true for a certain user, an institution or application domain and can be modified as needed. The hierarchical order of items in a trial can serve as an example. Items are often grouped in containers like modules or item groups. These containers are themselves grouped by higher level containers. The meta-model allows to specify any intended hierarchy, e.g. “Study-Event-Page-Module” or just “Study-Form-ItemGroup”, by defining domain-specific structural components and aligning them with consistent hierarchical relations.

In figure 2, core trial component types are colored in yellow, and domain-specific types are colored in green. The red colored trial component type symbolizes a type classified by the reasoner (see section 2). It is, as well as the other types in italics font style, an abstract type that is assigned implicitly.

4 Item Example

Data are instances of a data model. In our case, these data are individual trial components and characteristics. In figure 3, an example of the expressiveness of the model with regards to terminologies is shown (for the sake of clarity, form components are omitted).

A trial component is typed as the trial component type “trial item” (1), has a simple descriptive characteristic “creatinine” of type “label” (2) and a complex descriptive characteristic type “measurement unit” with its value “#micromole-per-liter” (3). Furthermore, it is annotated with several characteristics from the LOINC terminology (4); each complex descriptive characteristic value corresponds to a LOINC name field value. One of the LOINC characteristics, the *component* (analyte) (5), is more precisely elaborated and its value represents a concept of the NCI thesaurus, which itself can refer to external properties and links to other terminologies, e.g. it possesses UMLS.

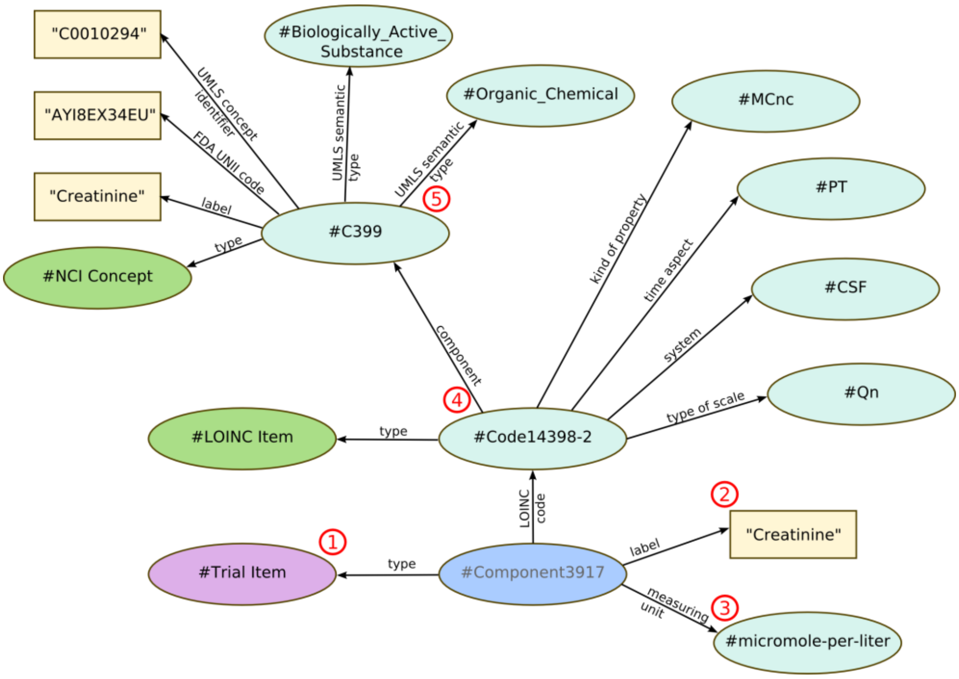


figure 3: Trial item referencing LOINC and NCI Thesaurus

Another example is the item shown in figure 4. It is an item that can likely be found on a real-world CRF. It’s a similar concept compared to figure 3, but this time it illustrates the expressiveness of the model with regards to form components (type input field), formatting instructions (a string of length 3 containing only digits) and normal ranges (two complex characteristics, one for males and one for females).

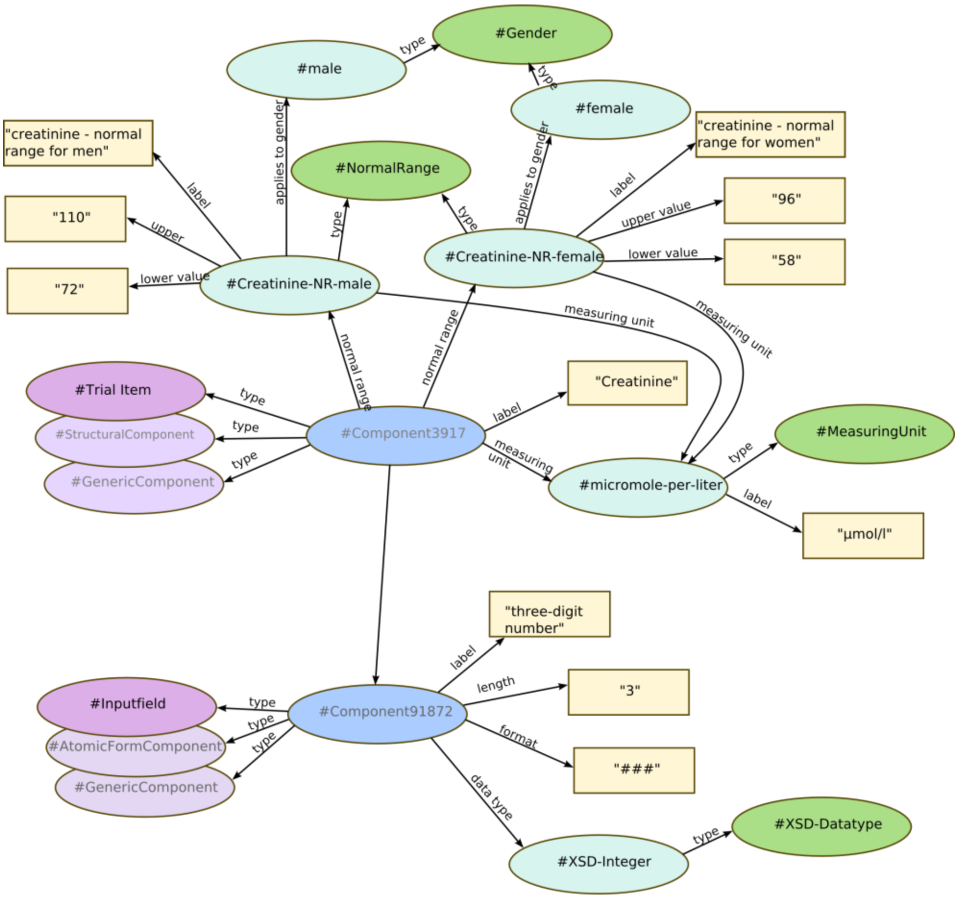
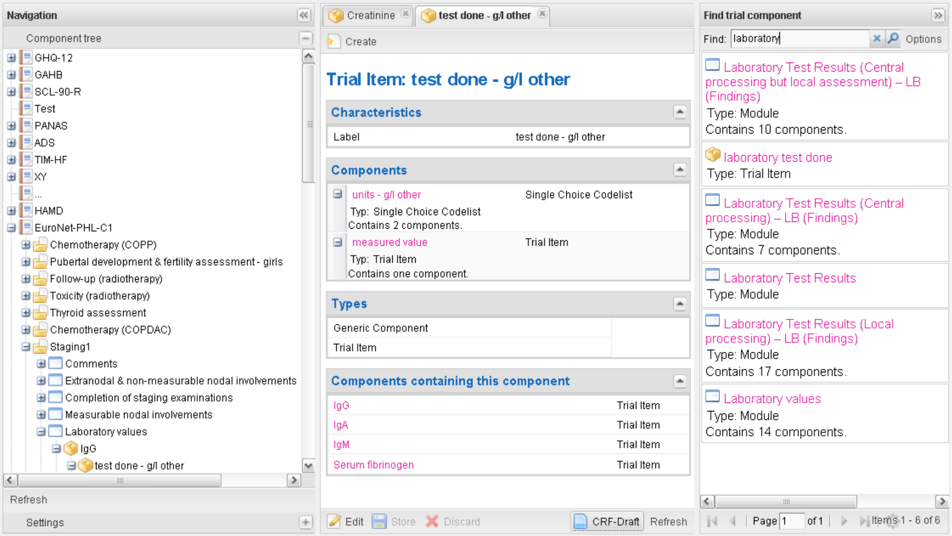


figure 4: Trial item with form components, formatting instructions and normal ranges

5 Conclusion

The model described in this paper is part of an Ajax-based web application called Trial Item Manager (TIM) [Mu07]. Currently, the item repository contains 5 clinical trials with about 2,000 trial items. Evaluation has shown that the primary advantage compared to traditional EDC software is the flexibility to extend the ontological model at runtime to provide support for user-specific needs and the ability to reuse existing item or modules.



Trial Item: test done - g/l other

Characteristics

Label	test done - g/l other

Components

units - g/l other — Single Choice Codelist
Typ: Single Choice Codelist
Contains 2 components.

measured value — Trial Item
Typ: Trial Item
Contains one component.

Types

Generic Component
Trial Item

Components containing this component

IgG	Trial Item
IgA	Trial Item
IgM	Trial Item
Serum fibrinogen	Trial Item

Find: laboratory

Laboratory Test Results (Central processing but local assessment) – LB (Findings)
Type: Module
Contains 10 components.

laboratory test done
Type: Trial Item

Laboratory Test Results (Central processing) – LB (Findings)
Type: Module
Contains 7 components.

Laboratory Test Results
Type: Module

Laboratory Test Results (Local processing) – LB (Findings)
Type: Module
Contains 17 components.

Laboratory values
Type: Module
Contains 14 components.

figure 5: Immunoglobulin G item in TIM’s graphical user interface

Acknowledgements: This work has been financed by the German Federal Ministry of Education and Research (BMBF), grants no.: 01KN0702, 01GI0495, 01GI0205.

References

[CD08] 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

[Ci98] Cimino, J. J.: Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 37, S. 394-403 (1998).

[IS03] ISO/IEC JTC 1 SC 32. Information technology — Meta-data 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, 2003–2005.

[Mu07] Mücke, R.: Trial Item Manager: Towards an Ontology based Specification of Items for Clinical Trials. 12th World Congress on the Internet in Medicine (MedNet2007), Leipzig.

[MH04] McGuinness, D. L., van Harmelen, F.: OWL Web Ontology Language overview. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge, Massachusetts, 2004.

[MM04] Manola, F.; Miller, E. (editors): RDF Primer. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge, Massachusetts, 2004.

[Re99] Rector, A. L. Clinical terminology: why is it so hard? Methods Inf Med 38, 239-252 (1999).