Standardized Terminology for Clinical Trial Protocols Based on Ontological Top-Level Categories

B. Heller*, H. Herre[#], K. Lippoldt*, M. Loeffler*

**Institute for Medical Informatics, Statistics and Epidemiology (IMISE)*

[#]Institute for Computer Science, Department of Formal Concepts

University of Leipzig, Germany

Abstract. This paper describes a new method for the ontologically based standardization of concepts with regard to the quality assurance of clinical trial protocols. We developed a data dictionary for medical and trial-specific terms in which concepts and relations are defined context-dependently. The data dictionary is provided to different medical research networks by means of the software tool *Onto-Builder* via the internet. The data dictionary is based on domain-specific ontologies and the top-level ontology of *GOL*[1]. The concepts and relations described in the data dictionary are represented in natural language, semi formally or formally according to their use.

1. Introduction

Medical care is increasingly based on clinical guidelines and clinical trial protocols. Clinical trials are carried out to gain insights into the etiology and progression of diseases as well as to analyze new diagnostic and treatment procedures and in particular to test new drugs. They are basic instruments of knowledge attainment and quality assurance in medicine. For this reason, the number of clinical trials is increasing and more and more international multi-center clinical trials are being carried out. Furthermore, general analyses of clinical trials are being made with regard to the international comparability of clinical trial results. Both the design and definition of trial protocols and the management of multi-center clinical trials are laborious processes in which different experts are involved. There already is an international guideline "Guideline for Good Clinical Practice" [1] for the execution of clinical trials which also was provided as an EU guideline and is being realized in national laws of the European countries currently. No standards are available, however, for the structuring of trial protocols or for reusable concepts in the clinical trial context. There exists no uniform terminology for trial-relevant concepts, for example.

[1] General Ontological Language is a formal framework for building ontologies. GOL is being developed by the Onto-Med research group at the University of Leipzig [http://www.onto-med.de].

The missing standards are one reason for additional labor expenditures of work arises in the design and the definition of new clinical trials, since one discusses the structure of trial protocols and the definition of relevant concepts again and again. In connection with this, it is our aim to provide templates for trial protocols and CRFs[2], on the one hand. Among other things, the therapy management tool *Onco-Workstation* [2] which makes short protocols of clinical trials available in a standardized form has been developed for this task. On the other hand, we are developing methods and software tools to advance the harmonization of concepts which are used in clinical trial documents and standard operating procedures. We have developed and implemented the software tool *Onto-Builder* which provides a data dictionary for clinical trials. This data dictionary is a terminological framework for clinical trial concepts which is partly based on the top-level ontology of GOL [3] [4]. The project GOL (General Ontological Language) was launched in 1999 as a collaborative research project of the Institute for Medical Informatics, Statistics and Epidemiology (IMISE) and the Institute for Computer Science (IfI) at the University of Leipzig. The project is aimed, on the one hand, at the construction of a formal framework for building and representing complex ontological structures, and, on the other hand, at the development and implementation of domain-specific ontologies in several fields, especially medical science [5].

Our paper is structured as follows. In the following section we situate our proposal in the context of ongoing research in terminology management and current approaches in the development of medical data dictionaries. In section 3 we show how the data dictionary can be integrated into the development process of clinical trials. Following this, we introduce our methodology in section 3 and define the relevant components. Sections 4-7 give a deeper insight into our approach by describing the model of the data dictionary, introducing the relevant ontological categories and relations of GOL and discussing our idea of ontological reduction. In the last two sections we discuss the chosen method and outlook the further work in this area of ontological research.

2. State of the Art

In the medical domain there are many medical terminology systems (nomenclatures, classification systems and data dictionaries) with different structures and representation of concepts. Many authors have given an overview of medical terminology systems and discussed their properties, e.g. [6, 7] [8, 9] [10]. For our goal – the construction of an ontologically founded context-sensitive data dictionary - in the first step it was necessary to analyze medical terminology systems with regard to reusability for the construction of a context-sensitive data dictionary model. Therefore we analyzed medical terminology systems among other things concerning their context representation methods and their relation to top-level ontologies. The evaluation included among others the following terminology systems: Systematized Nomenclature of Medicine (SNOMED) [11], Unified Medical Language System (UMLS) [12] and Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine (GALEN) [13]. Because of limited space in this paper only a short resume of our analysis of the considered terminology systems [14] can be given. The underlying models of SNOMED, UMLS, GALEN do not fit our requirements with regard to ontological foundation because they are limited with respect to the precise representation of relations, to the inclusion and adequate treatment of different views, and to the representation of context-dependent concepts.

A further analysis focused on medical data dictionaries which are developed by and used in medical institutions. Examples are the Medical Entities Dictionary (MED) of Co-

[2] A Case Report Form (CRF) is a printed, optical or electronic document designed to record all of the protocol required information to be reported to the sponsor on each trial subjects [1]

lumbia-Presbyterian Medical Center (New York), the Medical Data Dictionary (MDD) developed at the Giessen University, and the Metathesaurus of the National Cancer Institute NCI (Bethesda, USA). MED is constructed to serve the primary purpose of a repository for codes and terms used by clinical applications to represent data in the clinical data repository [15]. The Giessen MDD was constructed originally to store descriptive knowledge about drugs [16]. In the further evolution an independent data dictionary server (GDDS) was developed which supports context-sensitive presentation of information sources in medical applications [17] [18]. A well known approach in the USA is the NCI Metathesaurus [19]. The following table illustrates the characteristics of the NCI Metathesaurus in comparison to our data dictionary approach.

Table 1: Comparison of NCI Metathesaurus and our data dictionary

	National Cancer Institute (NCI)	data dictionary (Onto-Builder)
aim	increase the interoperability of information systems, development of a Thesaurus for NCI	increase of quality assurance based on standardized terminology, development of an ontologically founded generic data dictionary
target group	specific with respect to NCI, extended to bioinformatics	first step: national multi-centre clinical trials, second step: international multi-centre clinical trials
tools	Apelon, Inc. Terminology Development Environment and Workflow Manager	internet-based data dictionary tool *Onto-Builder*
process	development process with 8 steps	Three interacting cycles (knowledge acquisition cycle in natural language, quality assurance cycle (see [20]), ontological foundation cycle (see [21])
output	caCORE distribution flat flile / XML / Ontology in OWL light	xml-based prototype of GOL (GOL Markup Language *GOML)*
method	based on the UMLS Metathesaurus	based on the top-level ontology of GOL
structure	entities: kind, role, property, concept	top-level entities: basic categories (inclusive concept, denotation, term, description, context, and basic relations (see section 6)

Comprising it can be stated that these three medical data dictionaries are institution-specific, applied to specific applications (e.g. hospital information systems), limited in context-representation and they have no serious ontological foundation. To achieve our goal, namely the definition of a semantically founded context-dependent generic data dictionary, we elaborated a terminology model which is based on the ontological top-level categories of *GOL*. In the present paper the data dictionary is focused on the domain of clinical trials.

3. Application Environment

The design and definition of new clinical trials requires the preparation of different paper-based documents (clinical trial protocols, CRFs) and computer based tools for the administration of the clinical trial data (clinical trial databases, entry masks). To support the use of a uniform concept base for these trial documents and software tools we have developed a data dictionary which makes context-dependent definitions of concepts available. The basic configuration of this data dictionary contains general concepts for medicine (e.g. `therapy`, `laboratory parameter`) and for clinical trials (e.g. `inclusion/exclusion criteria`, `randomization`). For the design and definition of a new clinical trial, relevant concept

definitions can be extracted and queried from the data dictionary. If no adequate definitions are available from the data dictionary, the basic concepts can be expanded by corresponding alternative definitions. On the one hand, the trial-specific concept definitions are used for the conception of the corresponding trial database and on the other hand for the construction of the necessary CRFs.

The use of an uniform concept base for protocols, trial databases and CRFs within a clinical trial, reduces or even minimizes inconsistencies occurring in the documentation and analysis of patient-related trial data. Furthermore, the explicitly described basic concepts permit a harmonization between different clinical trials, this means the unification of the meaning and interpretation of relevant medical concepts and clinical trial data. Additionally, meta-analyses about different clinical trials are possible on the basis of uniform concept use in clinical trials. These meta-analyses allow prospective statements about the success of clinical trials planned and are an important quality assurance instrument in the field of clinical trials [22].

The following figure gives an overview of the use of the data dictionary in the definition of a clinical trial x_1.

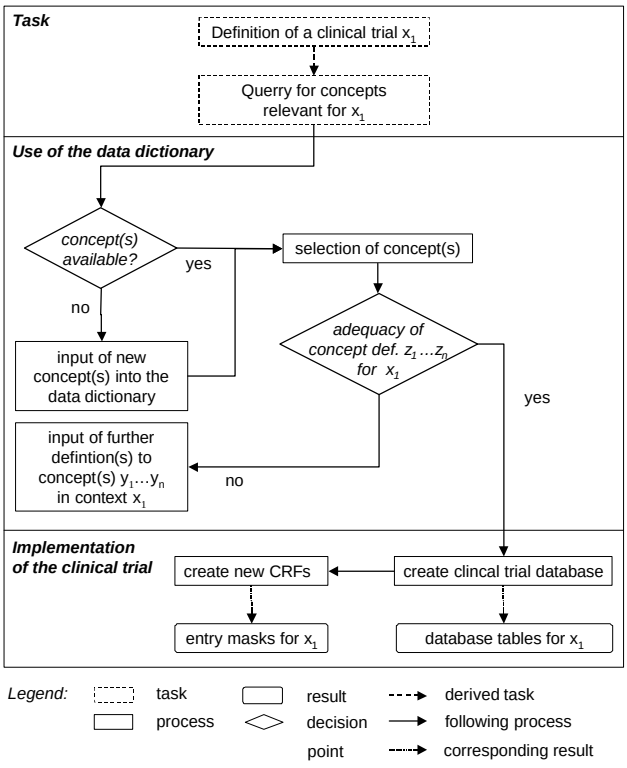

Figure 1: Use of the data dictionary in the clinical trial definition process

4. Definitions and Methodology

Our approach of an ontologically founded terminology is based on different interacting computer-based components, namely terminology, data dictionary, domain ontology, and top-level ontology (see also fig. 2). In the following, we briefly define these components and describe their interaction within our ontological approach:

Terminology: According to, [23] a terminology is the complete stock of the concepts, their definitions and names in a concrete domain. As an example of a very early medical terminology in the area of anatomy is the *Nomina Anatomica*. [24]

Data Dictionary: A data dictionary is to be understood as a collection of data which are described and interpreted as concepts with context. We claim that our notion of data dictionary is applicable on the one hand to different domains such as medicine, biology or technology and on the other hand to different application scenarios such as paper-based documents or software applications.

Domain Ontology: We use the notion of a domain ontology in accordance to Gruber [25]. A domain ontology provides formal specifications and computationally tractable standardized definitions of the terms used to represent knowledge of specific domains in ways designed to enhance communicability with other domains.

Top-Level Ontology:. A top-level ontology is concerned with the most general categories of the world, with their analysis, interrelations, and axiomatic foundation. On this level of abstraction ontology investigates kinds, modes, views, and structures which apply to every area of the world.

We assume as a basic principle of our approach that every domain specific ontology (here in the field of clinical trials and medicine) must use as a framework some top-level ontology which describes the most general, domain-independent categories of the world. Therefore our data dictionary structure consists of two layers which is depicted in the following figure.

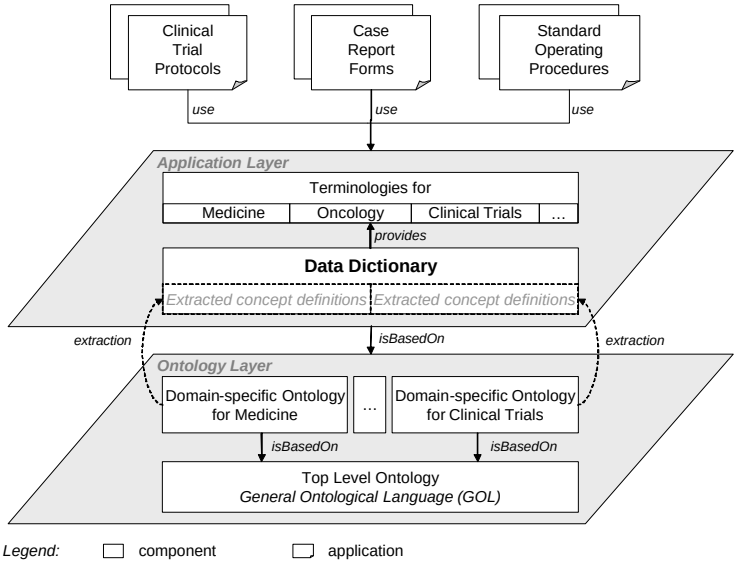

Figure 2: Two-layer model for an ontologically founded data dictionary

The *first layer* – called the application layer – contains two components: the data dictionary and the generic domain-specific terminologies in the field of medicine, oncology, clinical trials etc. (briefly: generic domain terminologies). The concept definitions of the generic domain terminologies are extracted from the identified and selected concept definitions of the data dictionary which are generic for the relevant domain. This domain generic information will be taken as a basis for the definitions which are included in the component of generic domain terminologies. This means, that these concept definitions are generic with respect to a confined area. The concepts of diagnosis, therapy and examination

for example, are defined generally in a terminology for medicine. In a special terminology e.g. for examination types concrete specializations of general definitions are, however, indicated with regard to single differentiable examination types.

The second component of the application layer consists of the data dictionary which contains context-dependent concept definitions as well as references to corresponding information (e.g. the relevant CRFs, radiographs, samples for a patient declaration of consent) and provides the main definitions of concepts for domain-specific terminologies. The applications (here: clinical trial protocols, case report forms, standard operating procedures) have access to the application layer from which they query relevant concept definitions and integrate them correspondingly.

The *second layer* consists of two types of ontologies, namely the domain-specific ontologies (here for clinical trials, oncology and medicine) and the top-level ontology of GOL. The domain-specific ontologies describe formal specifications of concepts which are associated to a specific application. According to our approach top-level concepts are used to build definitions of domain-specific concepts on a firm ground, and for this purpose we are developing a method of ontological reduction. An outline describing the steps of an ontological reduction is discussed briefly in section 8. The top-level ontology of GOL provides a frame-work with basic categories (e.g. universal/class, individual, quality, time, space, process and basic relations) which are described more precisely in section 6.

The two layers interact in the sense that the domain-specific concepts of the ontology layer are extracted from the data dictionary and are made available for the application-oriented concept descriptions which are provided for the application layer.

5. Application Layer

5.1 Main Entities of the Data Dictionary

In this section we describe the model of the data dictionary and focus in particular on the following main entities: concept, denotation or term, description, context and relation. Definitions, relevant typings/classifications as well as references to the other components (Terminology, Domain Ontology, Top-Level Ontology) are included in the descriptions of these entities.

Concept, Denotation, and Term*:* A concept is an abstract unit of meaning which is constructed over a set of common qualities [23] and which can also describe a cognitive entity (e.g., feeling, compliance, idea, thought). A denotation or term consists of one or several words and is the linguistic representation of a concept [20].

In the data dictionary model we distinguish between generic (e.g., `<disorder>`, `<process>`, `<treatment>`) and domain-specific (e.g., `<disease>`, `<symptom>`, `<medical treatment>`) concepts. A generic concept has a general meaning in different domains due to its domain-independent qualities. The concept `<treatment>`, for example, generally expresses that something or somebody is handled in a certain way. A concept is generic with respect to a class D of domains if it applies to every domain which is included in D. A domain-specific concept, however, has a meaning only in a certain domain. The concept `<medical treatment>` which is only relevant in the domain of medicine is an example of this kind of concept. A domain-specific concept of the data dictionary refers to at least one ontological category which is specific for this domain and which is included in the ontology related to this domain. The examples chosen also show that it is possible to change a generic concept into a domain-specific concept by adding an attribute. Rules for changing a concept type, composition and decomposition of concepts are the topics of a forthcoming paper [20].

Description: The description of a concept contains information about its meaning with respect to its qualities, its relations to other concepts, statements about its use, etc. [20]

Our model offers the possibility of handling alternative descriptions. There are different reasons for the occurrence of alternative descriptions, e.g. different granularity levels, static/dynamic aspects, subject area-related specifications, organization-dependent or institution-dependent differences as well as different expert opinions due to medical facts which have not yet been completely investigated. These different alternative definitions are represented with the help of contexts.

Context: With regard to the various discussions on the notion of context, e.g., in [26] we give here the following preliminary definition: A context is a coherent frame of circumstances and situations on the basis of which concepts must be understood.

As in the case of *concepts*, we similarly distinguish between generic and domain-specific contexts. A context is – roughly speaking - generic if concepts are associated to it whose description includes general properties/qualities (e.g., a generic context is `<process>` which includes the concept `<process course>` with among others the generic property `<process duration>`). Contrary to this, a domain-specific context includes concepts whose qualities/properties and their corresponding values specifically apply to this domain (e.g., a domain-specific context is `<disease>` which contains the concept `<course of a disease>` with among others the domain-specific property `<course expression>` and the values `<chronic>` or `<acute>` [20].

Relation: According to [3], relations are defined as entities which glue together the things of the world. We distinguish between three classes of relations: basic, domain-specific and terminological relations [20]. Our method handles at the present stage 12 basic relations which are briefly outlined in section 6. Examples for domain-specific relations are: `<treatedBy>`, `<SideEffectOf>`, and for terminological relations: `<synonymy>`, `<homonymy>`, `<polysemy>`.

5.2 Model of the Data Dictionary

A brief overview of the basic entities and relations of the data dictionary model is given in figure 3. The syntax of the model in figure 3 follows the UML[3] syntax, whereas rectangles represent classes (here: entities), rhombus n-ary associations (here: relations) and lines represent relations between the entities.

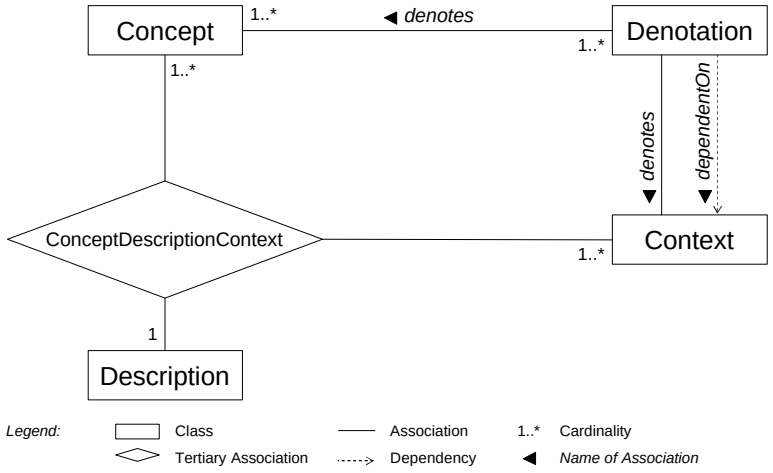

Figure 3: Data dictionary model (except)

[3] Unified Modeling Language [27]

In our model, one Concept can be assigned to many Description/Context pairs [1..n] and one Context can be assigned to many Concept/Description pairs [1..n]. A Concept can be defined only by one Description in one Context. Different descriptions for a concept apply in different contexts. The relation between Description, Concept and Context is expressed by the ternary association ConceptDescriptionContext which satisfies the above mentioned constraints. The entity Denotation describes Concepts and Contexts via the association *denotes*. The dependency (here: *dependentOn*) between Denotation and Context means that Denotation of a Concept can be dependent on the corresponding Context. If a Concept is not yet assigned to a Context, a default Denotation is given.

6. Ontology Layer

6.1 Domain-specific Ontology

A domain-specific ontology describes a specification of basic categories as these are instantiated through the concrete concepts and relations arising within a specific domain. For this reason, ways must be found to take into consideration different experts´ views on the domain concepts and relations, as well as different goals and contextually determined foci.

Domain-specific ontologies have a low portability. They can be transferred to other applications only to a very limited degree. Methods also have to be found to raise the degree of portability of domain-specific concepts, for example by using strictly modular description methods.

6.2 Top-Level Ontology GOL

The General Ontological Language GOL is intended to be a formal framework for building and representing ontologies. These ontologies are based on a system of formalized and axiomatized top-level ontologies which are provided by GOL.

In the following sections we discuss briefly certain ontologically basic categories and relations of GOL which support the development of domain-specific ontologies. A more detailed description of the ontological categories, the basic relations and some axioms of GOL are expounded in [3] [4].

6.2.1 Hierarchy of GOL Categories (Excerpt)

The following figure shows an excerpt of the categories in GOL.

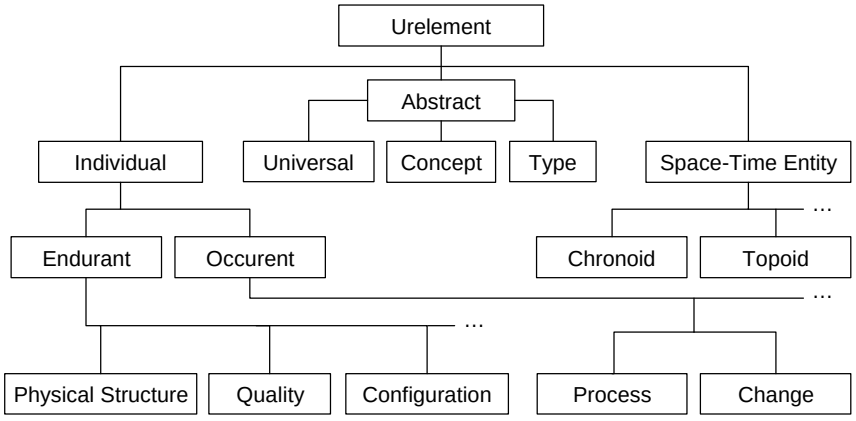

Figure 4: Hierarchy of top-level categories in GOL (excerpt)

6.2.2 Sets, Classes, and Urelements

The main distinction we draw is between *urelements* and *classes*. C*lasses (which include sets)* constitute a metamathematical superstructure above the other entities of our ontology.

6.2.3 Urelements

Urelements are entities of type 0 which are not classes. Urelements form an ultimate layer of entities lacking set-theoretical structure in their composition. Neither the membership relation nor the subclass relation can reveal the internal structure of urelements. We shall assume the existence of three main categories of urelements, namely *individuals, universals*, and *entities of space and time.* An *individual* is a single thing which is in space and time. A *universal* is an entity that can be instantiated by a number of different individuals. We distinguish several classes of universals: immanent universals, concepts and textual types. We assume that the universals exist in the individuals (*in re*) but not independently from them. On the other hand, humans as cognitive subjects conceive of universals by means of concepts that are in their heads. For this reason we include the class of concepts. The symbolic-linguistic representation of concepts is based on textual types which exhibit another kind of universal. Alongside urelements there is the class of *formal relations.* We assume that formal relations are classes of certain types.

6.2.4 Space and Time

In the top-level ontology of GOL, *chronoids* and *topoids* represent kinds of urelements. *Chronoids* can be understood as temporal intervals, and *topoids* as spatial regions with a certain mereotopological structure. Chronoids are not defined as sets of points, but as entities *sui generis*. Every chronoid has boundaries, which are called *time-boundaries* and which depend on chronoids, i.e. time-boundaries have no independent existence. We assume that temporal entities are related by certain formal relations, in particular the *part-of relation between chronoids,* the relation of *being a time-boundary of a chronoid*, and the relation of *coincidence between two time-boundaries.*

Our theory of topoids is based on the ideas of F. Brentano [28] and R. M. Chisholm [29]. Similar to Borgo [30] we distinguish three levels for the description of spatial entities: the *mereological level* (mereology), the *topological level* (topology), and the *morphological level* (morphology).

Topology is concerned with such space-relevant properties and relations as connection, coincidence, contiguity, and continuity. Morphology (also called qualitative geometry) analyses the shape, and the relative size of spatial entities.

6.2.5 Endurants and Processes

Individuals are entities which are in space and time, and they can be classified with respect to their relation to space and time.

An *endurant* or a continuant is an individual which is in time, but of which it makes no sense to say that it has temporal parts or phases. Thus, endurants can be considered as being wholly present at every a time-boundary at which they exists.

Processes, on the other hand, have temporal parts and thus cannot be present at a time-boundary. For processes, time *belongs to them* because they *happen in time* and the time of a process is built into it. A process *p* is not the aggregate of its boundaries; hence, the boundaries of a process are different from the entities which are sometimes called *stages* of a process.

Substances, Substantials and Objects

Substances are individuals which satisfy the following conditions: they are endurants, they are bearers of properties, they cannot be *carried by* other individuals, and they have a spatial extension.

The expressions *x carries y* and *x is carried by y* are technical terms which we define by means of an ontologically basic relation, the *inherence relation* which connects properties to substances. Inherence is a relation between individuals, which implies that inhering properties are themselves individuals. We call such individual properties *moments* and assume that they are endurants. Moments include *qualities*, *forms*, *roles*, and the like. Examples of substances are an individual patient, a microorganism, the heart (each considered at a time-boundary).

We assume that the spatial location occupied by a substance is a *topoid* which is a 3-dimensional space region. A *physical object* is a substance with unity, and a *closed substance* is a substance whose unity is defined by the strong connectedness of its parts. Substances may have (substantial) boundaries; these are dependent entities which are divided into *surfaces*, *lines* and *points*.

Individual properties, Qualities and Properties

Individual properties are endurants; in contrast to substances, they are entities which can exist only in another entity (in the same way in which, for example, hormone production exist only in the corresponding organ). Examples of individual properties are this color, this weight, this temperature, this blood pressure, this thought. According to our present ontology, all individual properties have in common that they are dependent on substances, where the dependency relation is realized by inherence.

6.2.6 Situoids, Situations, and Configurations

Situations present the most complex comprehensible endurants of the world and they have the highest degree of independence among endurants. Our notion of situation is based on situation theory of Barwise and Perry [31] and advances their theory by analyzing and describing the ontological structure of them.

There is a category of processes whose boundaries are situations and which satisfy certain principles of coherence and continuity. We call these entities *situoids*; they are the most complex integrated wholes of the world, and they have the highest degree of independence. Situoids may be considered as the ontological foundation of contexts.

6.2.7 Relations

We can distinguish the following basic ontological relations of GOL in table 2, which are needed to glue together the entities introduced above. A more detailed description of the relations is given in [3] [4].

Table 2: Basic relations in GOL

Basic Relation	**Denotation(s)**	**Brief Description**
Membership	*x Î y*	set *y* contains *x* as an element
Part-of	*part(x, y)* *tpart(x, y)* *spart(x, y)* *cpart(x, y)* *part-eq(x, y)* *tpart-eq(x, y)* *spart-eq(x, y)* *cpart-eq(x, y)*	*x* is part of *y* *x* is temporal part of *y* *x* is spatial part of *y* *x* is constituent-part of *y (y contains x)* the reflexive version of *part* the reflexive version of *tpart* the reflexive version of *spart* the reflexive version of *cpart*
Inherence	*i(x, y)*	moment *x* inheres in substance *y*
Relativized Part-of	*part(x, y, u)*	*u* is a universal and *x* is a part of *y* relative to *u*
Is-a	*is-a(x,y)*	*x is-a y* $=_{df}$ *" u (u ::x ® (u ::y)*
Instantiation	*x :: u* *x : y*	individual *x* instantiates universal *u* list x instantiates relation y

	$x ::_i y$	higher order instantiation, $i \geq 1$
Participation	*partic(x, y)*	*x* participates in process *y*, where *x* is a substance, an abstract substance or a substance process
Framing	*chr(x, y)* *chr(x)* *top(x, y)* *top(x)*	situoid *x* is framed by chronoid *y* denotes the chronoid framing *x* situoid *x* is framed by topoid *y* denotes the topoid framing *x*
Location and Extension Space	*occ(x, y)* *exsp(x, y)*	substance *x* occupies topoid *y* substance *x* has extension space *y*
Association	*ass(x, y)*	situoid *x* is associated with universal *y*
Ontical Connectedness	*ontic(x, y)*	*x* and *y* are ontically connected
Denotation	*den(x, y)*	symbol *x* denotes entity *y*

In table 2 the symbols x and y are entities. The concretization of the entities x and y depends on the type of the basic relation, e.g. tpart(x, y) means that x and y are *processes*. An exact specification of the admissible types of arguments of the basic relations in table 2 is presented in [4].

7. Example

The incremental ontological foundation of concepts is illustrated briefly below regarding the concept `remission` on the two layers of our model. The concept remission is defined in the domain medicine, sub-domain oncology under consideration of different contexts (here: course of a disease) in our data dictionary. In this case the data dictionary contains the following two definitions of remission which correspond to different stages of the course of a disease, whereas these definitions are parts of the terminology in the domain of oncology:

(a) "*Partial Remission (PR):* decrease by more than 50 percent of the sum of the products of the two largest perpendicular diameters of all measurable lesions, in the absence of growth of any lesion or appearance of a new lesion." [32]

"*Complete Remission (CR):* disappearance of all signs and symptoms, or recalcification of all osteolytic metastases during at least 1 month." [32]

With regard to an ontological reduction, the natural language definitions (a) are translated in the first step into a semi-formal representation. At this stage, subtleties in the definition are lost in favour of reduced interpretation possibilities. An example for the partial and complete remission as part of the domain ontology is shown in the following:

(b)
```
<concept>: remission
  <context>: course of an oncological disease
     <stage>: partial remission
        <criteria>: (decrease by more than 50 percent of the sum of
        the products of the two largest perpendicular diameters of
        all measurable lesions)
        AND NOT (growth of any lesion)
        AND NOT (appearance of a new lesion)
     <stage>: complete remission
        <criteria>: (disappearance of all signs and symptoms)
        OR (recalcification of all osteolytic metastases during at
        least 1 month)
```

Against the background of the examples (a) and (b) the data dictionary for clinical trials in the field of malignant lymphoma would include a more detailed context-dependent definition, as follows:

(c) *Partial Remission (PR):* The following criteria must be met in partial remission:

1. Lymphoma tissue still present (histological confirmation in all doubtful cases), but a clear reduction at all involved sites and reduction of the total lymphoma volume by at least 50%
2. No new lymphoma manifestations
3. Normalisation of blood counts

Context: disease: Aggressive Non-Hodgkin's Lymphoma; clinical trial: RICOVER-60 [33]

On the basis of the semi-formal definitions of (b) the next steps can be taken toward ontological foundation, namely the definition of relations between relevant concepts of the corresponding domain ontology followed by the reduction of definition contents to categories of the top-level ontology and its extensions.

8. Ontological Reductions and Semantic Transformations

An *ontological reduction* of an expression *E* is a definition of *E* by another expression *F* which is considered as *ontologically founded on a top-level ontology.* An expression is considered as ontologically founded on the top-level ontology of *GOL* if it is built up from atomic formulas whose meaning is inherited from the categories included in *GOL*. Ontological reductions exhibit a special case of semantic transformations. A semantic translation of a knowledge base *K* into a knowledge base *M* is a semantics-preserving function *tr* from the specification language *SL(K)* underlying *K* into the specification language *SL(M)* underlying *M*. Semantic translations can be used to compare the expressive power of ontologies and is an approach to the integration problem for ontologies. An outline of this theory which is being elaborated by the Onto-Med group is presented in [34].

We sketch the main ideas concerning the notion of an ontological reduction based on a top-level ontology of *GOL*. A definition *D* of a concept *C* for example is – usually – given as a natural language expression $E(C_1,..,C_n)$ which includes concepts $C_1,...,C_n$. The concepts $C_1,..,C_n$ are – in turn – defined by other expressions based on further concepts. In order to avoid this infinite regress we select a certain number of concepts $D_1,..,D_k$ – which arise from *E* – as primitive. An embedding of $\{D_1,...,D_k\}$ into *GOL* is a function *tr* which associates to every concept D_i a category $tr(D_i) = F_i$ of *GOL* which subsumes D_i, i.e. every instance of D_i is an instance of $tr(D_i)$. The problem, then, is to find a logical expression E_1 based on $\{F_1,...,F_k\}$ which is equivalent to the initial expression *E*; such an expression is called an ontological reduction based on *GOL*. It may be expected that – in general – the system *GOL* is too weak to provide such equivalent expressions. For this reasons *GOL* has to be extended to a system GOL_1 by adding further categories. GOL_1 should satisfy certain conditions of naturalness, minimality (the principle of Occam's razor), and modularity. The problem of ontological reduction includes four tasks:

a. construction of a set of primitive concepts (initialisation problem)
b. construction of an ontological embedding into *GOL* (embedding problem)
c. construction of an extension GOL_1 of *GOL* (extension problem)
d. finding an equivalent expression (definability problem).

A developed theory of ontological reductions based on top-level ontologies is in preparation and will be expounded in the paper [35].

9. Results and Discussion

With regard to the construction of a standardized terminology for clinical trial protocols and CRFs we have developed a methodology of an ontologically founded data dictionary. The methodology is based on two layers – the application layer and the ontology layer. The application components and theories at the two layers have been developed in parallel since 1999. One result of our work on the ontology layer is the development of the top-level ontology of GOL with approx. 50 basic categories and 12 basic relations. In the area of the domain-ontology we have started with the definition of domain-specific concepts which are partly based on top-level categories.

Concerning the application layer we constructed a data dictionary for clinical trials which contains context-dependent concept descriptions. This data-dictionary has been implemented as the web-based software tool *Onto-Builder* [36, 37]. This tool is provided to several research networks with approximately 500 medical experts via the internet. Against this background, the handling of different expert views is indispensable within the *Onto-Builder*. This requirement is fulfilled with the availability of contexts into the data dictionary model which handle different expert views, granularity aspects as well as special aspects of clinical trials. The present version of the data dictionary includes approximately 13 contexts, 1000 domain-specific concepts and 2500 concept descriptions.

Our evaluation of the data dictionary in the medical network for Malignant Lymphoma with about 300 different medical experts has shown a higher level of harmonization of concepts and concept descriptions in different clinical trial protocols. This has been possible due to the availability of a terminological concept base which has led in turn to an improved quality assurance in the clinical trial context.

10. Conclusion and Future Work

The evaluation of the application and theory components has shown that the underlying models of the data dictionary and the top-level ontology of GOL can be adapted to other domains and to other ontologies (e.g. DOLCE [38]).

Our data dictionary is merely a concept base for clinical trials at the present stage and not yet fully based on domain ontologies. The reason for this lies on the one hand in the extraction of domain-specific concept descriptions from the ontological layer which has not yet been realized completely. On the other hand this is connected to the problem of the ontological reduction of natural-language concept definitions via a semi-formal definition to formal propositions based on the built-in top-level ontology and its extensions. In our methodology we have already developed and partly integrated the first attempts of solving the ontological reduction problem.

Our future work consists in:

- the expansion of the theoretical framework by further basic categories, e.g. situations, views and qualities
- the elaboration of a theory of contexts and its evaluation in the area of clinical trials
- the incremental refinement of domain-specific concept descriptions with top-level categories
- the development of criteria for the specification of domain-specific concept types
- the explicit representation of semi-formal descriptions of domain-specific concepts
- the adaptation of the data dictionary to accommodate clinical trials in further medical research networks.

11. Acknowledgement

We want to thank our medical and biometrics experts of the *Competence Network Malignant Lymphoma* (Grant No.: 01GI9995/0) and the *Coordination Centers for Clinical Trials, Cologne* and *Leipzig* for their fruitful discussions in the field of clinical trials and medicine. Many thanks to the members and the Ph.D. students in the Onto-Med research group for implementing software modules for the *Onto-Builder* terminology management system and for the numerous discussions in building a top-level ontology. Last but not least we thank Evan Mellander for his large effort in the editorial work of this paper.

12. References

[1] ICH Harmonised Tripartite Guideline: Guideline for Good Clinical Practice (GCP) E6: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; May 1996.

[2] Heller B, Krüger M, Löffler M, Mantovani L, Meineke F, and Mishchenko R. OncoWorkstation - Ein adaptives Agentensystem für das Therapiemanagement klinischer Studien. In: 47. Jahrestagung der GMDS; 2002 Sept. 8-12; Berlin: München: Urban & Fischer; 2002. p. 380.

[3] Heller B, and Herre H. Ontological Categories in GOL, in press. *Axiomathes* 2003.

[4] Heller B, and Herre H. Formal Ontology and Principles of GOL. Leipzig: Research Group Onto-Med, University of Leipzig; 2003. Report No. 1.

[5] Heller B, and Herre H. Research Proposal. Leipzig: Research Group Onto-Med, University of Leipzig; 2003. Report No. 2.

[6] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, and Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. *JAMA* 1997; 4:238-251.

[7] Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Meth Inform Med* 1998; 37(4-5):394-403.

[8] Rector AL. Clinical Terminology: Why Is it so Hard? *Meth Inform Med* 1999; 38(4-5):239-252.

[9] de Keizer NF, Abu-Hanna A, and Zwetsloot-Schonk JHM. Understanding terminological systems. I: Terminology and typology. *Meth Inform Med* 2000; 39(1):16-21.

[10] de Keizer NF, and Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. *Meth Inform Med* 2000; 39(1):22-29.

[11] SNOMED. SNOMED® Clinical Terms Content Specification.: College of American Pathologists; 2001. Report No. DRAFT version 004.

[12] NLM. *UMLS Knowledge Sources*. 14 ed: National Library of Medicine (NLM); 2003.

[13] Rogers JE, and Rector AL. Extended Core model for representation of the Common Reference Model for procedures. Manchester, UK: OpenGALEN; 1999.

[14] Heller B, and Lippoldt K. Ontological Foundations of Medical Terminologies - Possibilities and Limitations. forthcoming.

[15] Cimino JJ, and James J. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *J Am Med Inform Assoc* 2000; May-June, 7(3):288-97.

[16] Prokosch HU, Bürkle T, Storch J, Strunz A, Müller M, Dudeck J, Dirks B, and Keller F. MDD-GIPHARM: Design and Realization of a Medical Data Dictionary for Decision Support Systems in Drug Therapy. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 1995:250-261.

[17] Ruan W, Bürkle T, and Dudeck J. A Dictionary Server for Supplying Context Sensitive Medical Knowledge. In: Overhage MJ, ed. AMIA Annual Symposium; 2000; Los Angeles, USA; 2000. p. 719-23.

[18] Bürkle T. Klassifikation, Konzeption und Anwendung medizinischer Data Dictionaries [Habilitationsschrift]. Giessen: Klinikum der Justus-Liebig-Universität Gießen; 2000.

[19] Golbeck J, Fragoso G, Hartel F, Hendler J, Parsia B, and Oberthaler J. The National Cancer Institute's Thesaurus and Ontology. *Journal of Web Semantics* 2003; 1(1).

[20] Heller B, Herre H, Lippoldt K, and Loeffler M. Terminology Management for Clinical Trials (submitted).

[21] Heller B, Herre H, and Lippoldt K. Domain-Specific Concepts and Ontological Reduction within a Data Dictionary Framework (submitted).

[22] Loeffler M, Brosteanu O, Hasenclever D, Sextro M, Assouline D, Bartolucci AA, Cassileth PA, Crowther D, Diehl V, Fisher RI, Hoppe RT, Jacobs P, Pater JL, Pavlovsky S, Thompson E, and Wiernik P. Meta-Analysis of chemotherapy versus combined modality treatment trials in Hodgkin's disease. *Journal of Clinical Oncology* 1998; 16(3):818-829.

[23] Deutsches Institut für Normung e.V. DIN 2342 Teil 1: Begriffe der Terminologielehre. Berlin: Deutsches Institut für Normung e.V.; 10/1992.

[24] International Anatomical Nomenclature Committee. *Nomina Anatomica*. São Paulo; 1997.

[25] Gruber TR. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies* 1995; 43(5/6):907-928.

[26] Bouquet P, Ghidini C, Giunchiglia F, and Blanzieri E. Theories and uses of context in knowledge representation and reasoning. *Journal of Pragmatics* 2003; 35:455-484.

[27] Booch G, Jacobson I, and Rumbaugh J. *The Unified Modeling Language User Guide.* Amsterdam: Addison-Wesley; 1999.

[28] Brentano F, ed. Philosophische Untersuchungen zu Raum, Zeit und Kontinuum. Hamburg: Felix-Meiner Verlag; 1976.

[29] Chisholm RM. *Boundaries as Dependent Particulars. Grazer Philosophische Studien 20*; 1983.

[30] Borgo S, Guarino N, and Masolo C. A Pointless Theory of Space Based on Strong Connection and Congruence. In: Aiello C, Doyle J, Shapiro SC, eds. Principles of Knowledge Representation and Reasoning (KR96); 1996; Cambridge, Massachusetts: San Francisco: Morgan Kaufmann; 1996. p. 220-229.

[31] Barwise J, and Perry J. *Situations and Attitudes*. Cambridge, MA, USA: Bradvord Books, MIT Press; 1983.

[32] Peckham M. *Oxford Textbook of Oncology*: Oxford Univ. Press; 1998.

[33] Pfreundschuh M. Randomised Study Comparing 6 and 8 Cycles of Chemotherapy with CHOP at 14-day Intervals, both with or without the Monoclonal anti-CD20 Antibody Rituximab in Patients aged 61 to 80 Years with Aggressive Non-Hodgkin's Lymphoma. RICOVER-60: German High-grade Non-Hodgkin's Lymphoma Study Group; 1999.

[34] Heller B, Herre H, and Loebe F. Semantic Transformation of Ontologies. forthcoming.

[35] Heller B, Herre H, and Loebe F. Ontological Reductions Based on Top-Level Ontologies. forthcoming.

[36] Heller B, Kuehn K, and Lippoldt K. Onto-Builder - A Tool for Building Data Dictionaries. Leipzig: Research Group Onto-Med, University of Leipzig; 2003. Report No. 3.

[37] Heller B, Lippoldt K, and Kuehn K. Handbook Onto-Builder: Part I: Construction of Medical Terms. Technical Report. Leipzig: Research Group Onto-Med, University of Leipzig; 2003. Report No. 5.

[38] Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A, and Schneider L. Wonderweb Deliverable D17. Preliminary Report, Version 2.0. Padova [Italy]: ISTC-CNR; 2002.