Ontological Foundations of Medical Information Systems

H. HERRE[#], B. HELLER*

Research Group Onto-Med

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

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

University of Leipzig, Germany

Abstract. The modeling of knowledge and implementing it in software systems is not sufficiently supported by proper guidelines and well-established principles. An interdisciplinary approach seems to be necessary because the development of guidelines for ontologically sound knowledge modeling – in particular in medicine – requires knowledge and expertise from areas such as linguistics, informatics, medicine, cognitive science and philosophy. Moreover, the ever increasing amount of knowledge in the various fields of medicine and the increasing need for differing forms of presentation in very different electronic media including the Internet require a harmonization of systems of concepts which transcends national borders. These challenges should be tackled by a research program which, in equal measure, takes account of the theoretical foundations, the medical application areas and the use of software in medicine. In the present paper we discuss important aspects of such an approach and outline some results of the work of the Onto-Med research group. In particular, we expound a methodology which supports the ontological foundation of medical information systems. This method is based on ontological mappings using reference top-level ontologies, and is inspired by rigorous logico-philosophical principles.

1. Introduction

Many years of experience with conventional methods, with models from the areas of knowledge representation and processing and with the realization of large, computer supported knowledge bases in the area of medicine have shown that the systematic, in-depth modeling and representation of knowledge is only supported with qualifications, and even then only on the scale of small examples.

The lack of a standard for semantically sound mappings of knowledge in computer-based applications causes losses in the quality of this knowledge which are hard to accept. The correct implementation of domain-specific concepts in software systems is necessary for the quality assurance of software. This is especially important for applications with patient data in the field of medicine. There is no procedure for modeling knowledge and implementing it in software systems which is supported by proper guidelines. An interdisciplinary approach is necessary because the development of guidelines for ontologically sound knowledge modeling requires knowledge from areas such as linguistics, informatics,

medicine, cognitive science and philosophy. Moreover, the ever increasing amount of knowledge in the various fields of medicine and the increasing need for differing forms of presentation in very different electronic media including the Internet require a harmonization of systems of concepts which transcends national borders. These challenges should be tackled by a research program which, in equal measure, takes account of the theoretical foundations, the medical application areas and the use of software in medicine. In the present paper we discuss some of these points and outline some results in this direction.

Our paper is structured as follows. In the following section we introduce a general notion of an information system and give an overview about terminology systems and medical data dictionaries. Section 3 is devoted to the notion of ontology and its formalization, and furthermore the role of top-level ontologies in knowledge modeling. Section 4 outlines the formal framework of GOL which contains a top-level ontology as a part. In section 5 the theory of ontological mappings is presented in some detail and section 6 contains a comparison with other approaches to mapping. In section 7 the architecture of an ontologically founded data dictionary is outlined and in the remaining sections examples and conclusions as well as future work are discussed.

2. Information Systems

In modeling or specifying a concrete domain *D* we start with a body of source information about *D*, denoted by *SI(D)*, which is, usually, presented in different media (including natural language) and often in a non-structured form. From *SI(D)* there is constructed a specification *Spec(SI)* (which takes the form of a set of expressions) with the aim to capture the knowledge-content of *SI(D)*. Usually, *Spec(SI)* is expressed in a (formal) modeling/representation language, but also in natural or semi-formal languages. Examples of such languages are: KIF [1], Description Logics (cf. [2]), Conceptual Graphs [3], Semantic Networks, but also modeling languages like UML (Unified Modeling Language, [4]) or OPM (Object Process Methodology, [5]).

A specification *Spec(SI)* in the above-mentioned sense of any source information is called an *information system*. In general, the system *Spec(SI)* is not sufficiently founded ontologically, and the task remains to translate *Spec(SI)* into an ontologically founded, and strictly formal, knowledge base which is formulated in a target language *OL* (Ontology Language). An ontological mapping translates the expressions of *Spec(SI)* into the language *OL* resulting in the knowledge base *OKB(Spec(SI))*, which captures more precisely the ontological content of *Spec(SI)*. In this case we say that *OKB(Spec(SI))* is an ontological foundation of *Spec(SI)*.

A terminology system *TS* may be considered as a special information system *TS = (Tm, Rel, Def)* consisting of a set *Tm* of terms which denote concepts, a set *Rel* of relation symbols denoting relations between concepts or instances, and a function *Def* associating to every term *t* of *Tm* a definition *Def(t)* in a natural or semi-formal language which describes the meaning of the concept denoted by the term *t*.

The different terminology systems can be distinguished into *nomenclatures, classification systems* and *data dictionaries*. These systems are based on different architectures and methods for the representation of concepts. In the sequel we restrict the discussion to the medical domain, which is sufficiently rich to present all types of terminology systems. We give a short summary of our evaluation results concerning context representation and ontological foundation, concentrating on the most relevant terminology systems.

Within SNOMED CT [6] contexts are defined as “information that fundamentally changes the type of thing it is associated with”. An example for a context is <family history of> because it changes e.g., the type of the concept <myocardial infarction> which is a

heart disease to the new concept <family history of myocardial infarction> which is not a heart disease.

UMLS [7] integrates concepts and concept names (terms) from many controlled vocabularies and classification systems using a uniform representation structure with 134 different semantic types. In UMLS the context-dependency of concepts is not explicitly elaborated. UMLS uses contexts only to describe structural features of sources, e.g., the use of siblings and multiple hierarchical positions of concepts.

In GALEN [8] the entity-types modality and role can be interpreted as context-representing entities. An example for modality is <FamilyHistory>, by means of which, in combination with the concept <Diabetes> the new concept <FamilyHistory of Diabetes> can be derived. Examples for role are <Steroid which playsRole HormoneRole> or <plays-Role Drug-Role>. These examples describe the contexts <drug>, <hormone> which by implication are given by the denotations of the corresponding roles but can be derived explicitly.

In addition, the multi-axial classification of concepts can be considered as a representation form for contexts in which the root of a classification axis would correspond to a context; whereas a multiple assignment of concepts to superordinate concepts does not have influence on its attributes/relations.

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 Columbia-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).

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-center clinical trials; second step: international multi-center clinical trials
Tools	Apelon, Inc. Terminology Development Environment and Workflow Manager	Internet-based data dictionary tool *Onto-Builder*
Process	development process with eight steps	three interacting cycles (knowledge acquisition cycle in natural language, quality assurance cycle (see [9]), ontological foundation cycle (see [10]))
Output	caCORE distribution flat file / 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)

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 [11]. The Giessen MDD was constructed originally to store descriptive knowledge about drugs [12]. In the further evolution an independent data dictionary server (GDDS) was developed which supports the context-sensitive presentation of information sources in medical applications [13] [14]. A well-known approach in the USA is the NCI Metathesaurus [15]. Table 1 illustrates the characteristics of the NCI Metathesaurus in comparison to our data dictionary approach.

In summary, 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 that they have no serious ontological foundation.

3. Ontologies

Formal Ontology is the science which is concerned with the systematic development of axiomatic theories of forms, modes, and views of being of different levels of abstraction and granularity. On the most general level of abstraction formal ontology is concerned with the kinds, modes, views, and structures which apply to every area of the world. We call this level of description *General Ontology*, in contrast to the various *Domain Ontologies* which are applicable to more restricted fields of interest. We assume that every domain-specific ontology must use as a framework some general ontology, sometimes called top-level ontology, which describes the most general categories of the world.

Ontologies are becoming increasingly important for software and knowledge processing. In the meantime, their great significance has become clear in diverse areas such as e-commerce, qualitative modeling, database design and medical information sciences. Each of these areas requires its own domain-specific ontology. However, communication between different areas and the semantic foundation of domain knowledge require a uniform framework based on an interdisciplinary, general ontology, a so-called top-level or reference ontology. *Ontologies* are formal specifications of conceptualizations, described by axioms and explicit definitions, and are used for the precise semantic representation of concepts from different areas of knowledge. The representation of an ontology in a formal language is an important precondition for its implementation on computers as well as its availability in knowledge-based systems of different application domains. In this sense, an ontology can be described as “an explicit specification of a conceptualization” [16].

Recently, formal ontology has been applied in various areas where the notion of an ontology is used in a very broad sense. In general, a particular ontology is understood to be a description of a given domain which can be accepted and reused in all information systems referring to this domain. Sometimes even terminologies are considered as ontologies, whereas we take a position which is more narrow. At a minimum, the backbone of an application ontology is usually a taxonomy of concepts which is based on the subsumption link.

An ontology *Ont* – understood as a formal knowledge base – is given by an “explicit specification of a conceptualization” [16]; it consists of a structured vocabulary *V(Ont)*, called an ontological signature, and a set of axioms *Ax(Ont)* about *V(Ont)* which are formulated in a formal language *L(Ont)*. Hence, an ontology (understood as a formal object) is then a system $Ont = (L, V, Ax)$; the symbols of V denote categories, and relations between categories or between their instances. L can be understood as an operator which associates to a vocabulary V a set $L(V)$ of expressions which are usually declarative formulas. We assume the following condition: $V \subseteq V_1$ implies $L(V) \subseteq L(V_1)$, and $L(L(V)) = L(V)$. An ontology may be augmented by a derivability relation, denoted by $\vdash$, and by a semantic conse-

quence relation, denoted by $\models$. Then, such an ontology takes the form of a knowledge system *(L, V, Ax, Mod(V),* $\vdash$ *,* $\models$ *)* which includes a class *Mod(V)* of interpretations which serves as a semantics for the language *L(V).*

In this restricted sense, an ontology is a formal theory which describes the phenomena of a domain, and the question arises whether there is indeed a need to introduce a new term "ontology" instead of using the term "theory". In our opinion, the use of the notion of ontology is justified for general ontologies whose categories and relations can be applied to every domain. We believe that formalization can be achieved for ontologies of the most general level or for so-called *top-level ontologies*. Furthermore, such theories can be distinguished from arbitrary domain-specific theories because they are much more stable, they have a higher degree of evidence, and because they exhibit a universal applicability. The situation is different for large domain ontologies and for the integration problem. Nevertheless, we believe that an appropriate use of top-level ontologies in developing domain ontologies may contribute to improving the quality of the resulting knowledge bases.

4. The GOL Project

The GOL project was launched in 1999 as a collaborative research effort of the *Institute for Medical Informatics, Statistics and Epidemiology* (IMISE) and the *Institute for Informatics* (IfI). *General Ontological Language* (GOL) is intended to be a formal framework for building and representing ontologies. The main purpose of GOL is to provide a library of formalized and axiomatized top-level ontologies which can be used as a framework for building more specific ontologies. The GOL-Framework consists of three components representing different levels of abstraction. Meta-GOL contains basic principles of semantic choice, a general view on categories and classes, methods of semantic transformations, and principles for meta-logical analyses. GOL-Software tools contain a number of systems which support the development, the evaluation, the mapping and the integration of ontologies, but also application software (Onco-Workstation, Onto-Builder, SOP-Creator) based on ontological principles.

GOL on the object level consists of a basic logic and a representation language RGOL. RGOL has a built-in ontology which is called abstract core ontology, denoted by ACO. ACO contains the basic entities categories, classes, and concrete entities, and as relations identity, membership and instantiation; we believe that ACO is an indispensable part of every top-level ontology. The core of GOL is intended to be a library of top-level ontologies which extend ACO. The first of these ontologies, which is called General Formal Ontology (GFO), is under development and will be outlined in this section. There is a debate as to whether the top-level ontology should be a single, consistent structure or whether it should be considered as a partial ordering of theories which may be inconsistent with theories that are not situated on the same path of the partial ordering [17]. Concerning the partial ordering approach there are different kinds of distinctions between such ontologies. One kind of distinction is based on the fact that different ontologies may use different basic categories of entities. But even if two ontologies use the same basic categories they may differ with respect to the axioms formulated about the categories. Then the question arises as to which of the axioms should be included in the top-level axiomatization. Our general philosophy is to admit a restricted version of the partial ordering approach. We want to have only a restricted selection of ontologies with different basic systems but we are more liberal with respect to the admitted systems of axioms within a fixed system of ontological categories.

In the following sections we briefly discuss certain ontologically basic categories and relations of GOL which support the development of domain-specific ontologies. The ontological categories, basic relations and some axioms of GOL are expounded in greater detail in [18, 19].

4.1 Hierarchy of GOL Categories (Excerpt)

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

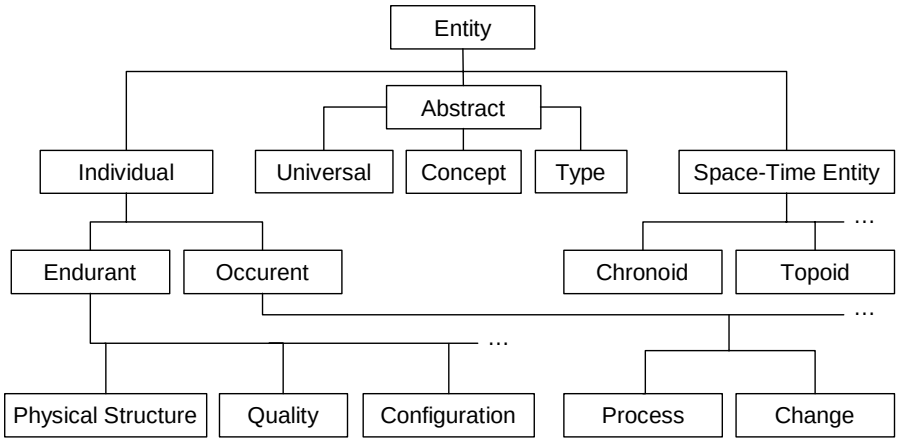

Figure 1: Hierarchy of the top-level categories in GOL (excerpt)

4.2 Classes, Categories, and Concrete Entities

In what follows we will discuss ontologically basic entities and some basic distinctions needed to structure the world. The main distinction we draw is between *classes, categories,* and *concrete entities.* Categories and concrete entities are *class-urelements*, i.e. they have no members and are different from the empty class. We classify concrete entities into *individuals* and *entities of space or time*. Individuals are further classified into *presentials* and *occurrents*. A category *c* is said to be primitive if its instances are concrete entities.

4.3 Individuals and Universals

An *individual* is a single thing thought of in contrast to universals, and hence we assume that universals are instantiated by individuals. A primitive category is an entity that can be instantiated by a number of different individuals, and hence a universal is always a primitive category. There are several kinds of universals: *immanent universals*, *conceptual structures* and *symbolic structures*. Usually, the individuals covered by a universal are similar in some respect. We assume that immanent universals exist in the individuals (*in re*) but not independently of them; thus, our view of immanent universals is Aristotelian in spirit (cf. [20]). Moreover, immanent universals are considered as a subcategory of the category of immanent categories. On the other hand, humans as cognitive subjects conceive of universals of any sort by means of concepts that are in their mind. Hence we hold that mental notions cannot be eliminated from ontology and, accordingly, we postulate a relation between immanent categories, concepts, and – concerning their representation – symbolic structures. For every category *U* there is a class *Ext(U)* containing all instances of *U* as elements.

We assume the axioms that classes of individuals, of space-time entities, and of categories are pairwise disjoint.

4.4 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 [21] and R. M. Chisholm [22]. Similar to Borgo [23] 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.

4.5 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 time-boundary at which they exist.

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.

4.6 Physical Structures, Physical Objects, Qualities and Properties

Physical structures 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. A physical structure is said to be a physical object if its parts are strongly connected. We assume that every physical object has a closed boundary.

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 physical structures. Inherence is a relation between individuals, which implies that inhering properties are themselves individuals. We call such individual properties *qualities* and assume that they are endurants. Qualities include *individual colors*, *forms*, *roles*, and the like. Examples of physical structures are *an individual patient, a microorganism, a heart* (each considered at a time-boundary).

We assume that the spatial location occupied by a physical object is a *topoid*, which is a 3-dimensional space region. Physical structures may have (physical) boundaries; these are dependent entities, which are divided into *surfaces*, *lines* and *points*.

Examples of qualities are this color, this weight, this temperature, this blood pressure, and this thought. According to our present ontology, all qualities have in common that they are dependent on physical structures, where the dependency relation is realized by inherence. Qualities are instances of properties which are considered as concepts. For example, "this (individual) red" of "this (individual) rose" is an instance of the property *red*.

4.7 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 [24] 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.

4.8 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 [18, 19].

Table 2: Basic relations in GOL

Basic Relation	**Denotation(s)**	**Brief Description**
Membership	$x \in y$	set *y* contains *x* as an element
Part-of	*part(x, y)* *mpart(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 material 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)*	quality *x* inheres in physical structure *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* $y =_{df} \forall u\,(u ::x \rightarrow (u ::y)$
Instantiation	$x :: u$ $x : y$ $x ::_i y$	individual *x* instantiates universal *u* list x instantiates relation y higher order instantiation, $i \geq 1$
Participation	*partic(x, y)*	*x* participates in process *y*, where *x* is a physical structure
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)*	*x* occupies region *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 [18].

5. Ontological Mappings

In this section we describe and discuss some basic formal principles which are important for the task of constructing a formal knowledge base out of an information system which is specified in a certain source language. These principles result in the notion of an *ontological mapping*. An ontological mapping translates the source information system into an ontologically founded formal knowledge base. The notion of *ontological foundedness* makes use of a top-level ontology.

We expound in the remaining part of the current section in more detail the construction of the formal knowledge bases *OKB(Spec(SI))* assisted and supported by a top-level reference ontology *Ont*; here, *Spec(SI)* is an information system as introduced in section 2. Generally, a formal ontology *Ont = (L, V, Ax(V))* consists of a structured vocabulary *V*, called an ontological signature, which contains symbols denoting categories, individuals, and relations between categories or between their instances, and a set of axioms *Ax(V)* which are expressions of the formal language *OL*. The set *Ax(V)* of axioms captures the meaning of the symbols of *V* implicitly. A definitional extension $Ont^d = (L, V \cup C(DF), Ax(V) \cup DF)$ of *Ont* is given by a set *DF* of explicit definitions over the signature *V* and a new set *C(DF)* symbols introduced by the definitions. Every explicit definition has the form *t := e(V)*, where *e(V)* is an expression of *L* using only symbols from *V* (hence the symbol *t* does not occur in *e(V)*).

As our starting point for *Spec(SI)* we consider terminology systems as introduced in section 2. A terminology system *TS* is an information system *TS = (Tm, Rel, Def)* consisting of a set *Tm* of terms which denote concepts, a set *Rel* of relation symbols denoting relations between concepts or instances, and a function *Def* associating to every term *t* of *Tm* a definition *Def(t)* in natural or a semi-formal language which describes the meaning of the concept which is denoted by the term *t.* An *ontological mapping M* of *TS* into *Ont* is given by a pair *M = (tr, DF)* consisting of a definitional extension *DF* of *Ont* and function *tr* which satisfies the following condition:

> For every term $t \in Tm$ the function *tr* determines an expression *tr(Def(t))* of the extended language $L(V \cup C(DF))$ such that *Def(t)* and *tr(Def(t))* are semantically equivalent with respect to the knowledge base $Ax(Ont) \cup DF$.

Then the set $OntMap(TS) = Ax(V) \cup DF \cup \{tr(Def(t)) : t \in Tm\}$ is a formal knowledge base which captures the meaning of *TS*, i.e. *OntMap(TS)* corresponds to *OKB(Spec(SI))*. The notion of semantic equivalence with respect to a knowledge base is used here informally because a strict formal semantics for natural language sentences does not yet exist; the notion has to be read "the meaning of the natural language (or semi-formal) sentence *Def(t)* is equivalent to the meaning of the expression *tr(Def(t))*.

An expression *e* is considered as ontologically founded on an ontology *Ont* if it is expressed in some definitional extension Ont^d of *Ont*. Hence, an ontological mapping of a terminology system *TS* associates to every term of *TS* an equivalent formal description which is based on a formal ontology *Ont*. Semantic translations can be used as a formal framework for schema matching, which is a basic problem in many database application domains, compare [25].

We now consider the fine-structure of an ontological mapping based on a top-level ontology *TO*. A definition *D* of a concept *C* of a terminology system is – usually – given as a natural language expression $e(C_1,..,C_n, R_1,\ldots, R_m)$ which includes concepts $C_1,...,C_n$ and relations $R_1,\ldots, R_m$. The concepts $C_1,..,C_n$ and relations $R_1,\ldots, R_m$ are in turn defined by other (natural language) expressions based on additional concepts and relations. In order to avoid this infinite regress we select a certain number of concepts $D_1,.., D_k$ and relations $S_1,\ldots,S_l$ – which arise from *e* – as primitive. An embedding of $\{D_1,...,D_k\}$ into *TO* is a function *emb* which associates to every concept D_i a category $emb(D_i) = F_i$ of *TO* which subsumes D_i, i.e. every instance of D_i is an instance of $emb(D_i)$. The problem, then, is to find a logical expression e_1 based on $\{F_1,...,F_k\}$ and the relations of *TO* which is equivalent to the initial expression *e*; such an expression is called a *local ontological mapping based on TO*. An ontological mapping based on TO, then, is a complete system of local ontological mappings covering all terms of the source system *TS*. It may be expected that – in general – the system *TO* is too weak to provide such ontological mappings. For this reasons *TO* has to be extended to a suitable system TO_1 by adding further categories and relations, and axioms about them. TO_1 should satisfy certain conditions of naturalness, minimality (the principle of Occam's razor), and modularity. The construction of ontological mappings includes three main tasks:

1. Construction of a set *PCR* of primitive concepts and relations from the set $\{Def(t) : t \in Tm\}$ (*problem of primitive base*)
2. Construction of an extension TO_1 of *TO* by adding the new categories *Cat* and relations *Rel* and a set of new axioms. $Ax(Cat \cup Rel)$ (*axiomatizability problem*)
3. Construction of equivalent expressions for $Def(t) \cup PCR$ on the base of TO_1 (*definability problem*).

We conclude this section with a summary of the modularization principle applied to a natural language text *T*. In analyzing the natural language text *T* one should satisfy the following basic modularity principle: Firstly, construct a primitive base *PCR* for the set *CR* of concepts and relations which are associated to *T;* usually, *PCR* is a proper subset of *CR*. Note that *PCR* is not uniquely determined. The explicit knowledge contained in *T* should then be represented as the union of two disjoint modules:

a) a set $Ax(PCR)$ of axioms about the concepts and relations of *PCR*
b) a set of explicit definitions $Def(CR - PCR)$ of the non-primitive concepts and relations which are contained in $CR - PCR$.

The knowledge associated to *T* and with respect to the selection *PCR* and *CR*, denoted by *KB(T,PCR)*, is defined by $KB(T) = Ax(PCR) \cup Def(CR - PCR)$. The difficult task is to find the set $Ax(PCR)$ and to select *PCR*. If we do not introduce axioms about *PCR*, i.e. if $Ax(PCR)$ is empty, then the knowledge system *KB(T)* becomes trivial. This phenomenon is sometimes overlooked in the field of knowledge engineering.

6. Comparison with other Approaches to Mapping

We suppose that a *semantic transformation* maps knowledge formulated in a source language to some equivalent expression in a target language. This very broad understanding comprises *knowledge capturing* on the basis of natural language texts as well as translations between formal languages, which we called *ontological mappings*. It is common in both cases that the semantics is to be preserved by the transformation.

In the present scientific landscape, these two types of tasks (knowledge capturing vs. ontological mapping) are strongly separated. Ontology-related communities in computer science usually deal with transformations of knowledge expressed in formal languages, e.g. transformations between ontologies based on description logics as is popular in the Semantic Web area.

The problem of how to integrate several formal ontologies in order to use them in combination has been recognized in a number of fields. As a result, a number of approaches ranging from theory-oriented works to implemented tools have been developed. Recently, some overviews of approaches and problems were published [26], [27]; cf. also section 3.6 of [28] and related works discussed in [29]. Schema matching in the database area is frequently considered a similar task, and it is reviewed in [25]. Therefore, we refrain from giving an extensive comparison of individual publications. Some of the major works as regards appearance in the literature are FCA-Merge [30], OntoMorph [31], Chimaera [32] and the tools of the PROMPT suite [29] which is developed at Stanford University.

Note that none of these works have alleviated the need for a terminological standardization. This is still one problem of the emergent area of ontological mapping. This can also be recognized by the collections of terms presented in [26], [27].

Apart from considering several ontologies in one language, one may want to combine ontologies which are stated in different languages. Another task which is closely related to this type of ontology integration problem is that of comparing formalisms themselves. [33] presents an attempt of a unifying approach. This is also important because each formalism contains itself certain basic ontological assumptions.

The second task from above, i.e. knowledge capturing, often refers to either knowledge acquisition or fields like natural language processing or computer linguistics. Knowledge acquisition pursues the development of methodologies for human users. In contrast, linguistic-related approaches employ a variety of methods for automated text understanding, from purely statistical approaches to machine learning, which is rooted more deeply in computer science.

One of the closer relationships to the field of ontology with respect to automation arises in WordNet [34]. WordNet is a linguistic resource with explicit semantic relationships connecting its synsets, which can roughly be understood as concepts. It has been used directly as an “ontology”, which is debatable, and it has been related to a formal ontology (cf. [35]). Together with sample text corpora tagged with WordNet synsets such an alignment may allow for an improved automated formalization of natural language texts.

7. Ontologically Founded Data Dictionaries

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 figure 2). In the following, we briefly describe their interaction within our ontological approach. Our data dictionary structure consists of two layers, which are depicted in the figure 2.

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.

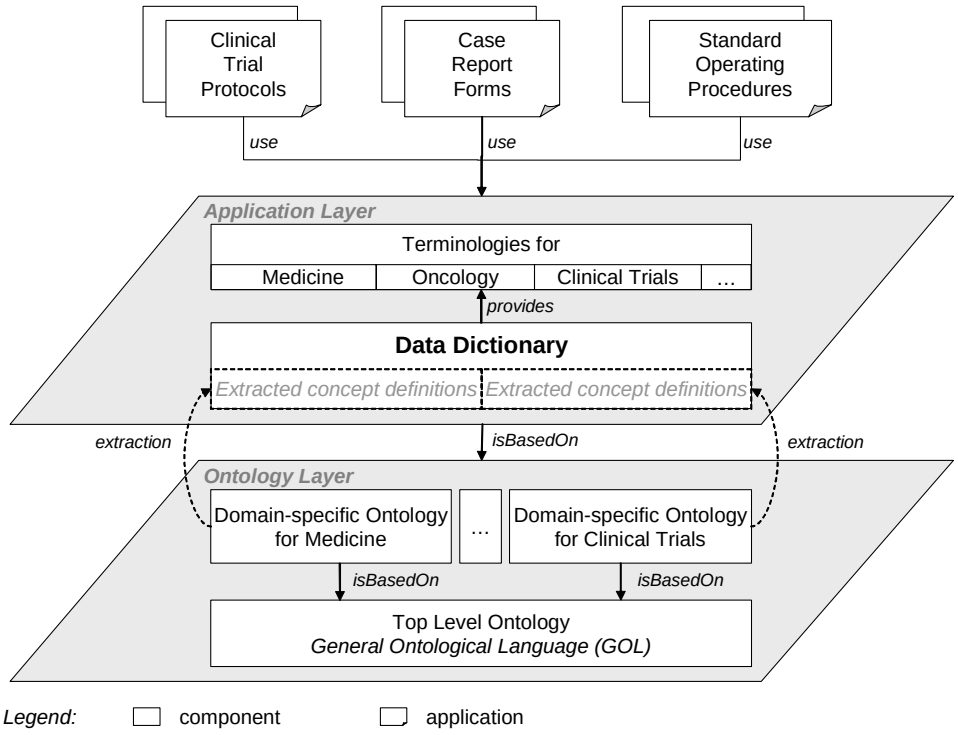

Figure 2: Two-layer model for an ontologically founded data dictionary

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, 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 the method of ontological mappings. 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.

8. Conclusions and Future Work

We have outlined an architecture of an ontologically founded data dictionary. The architecture 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. The evaluation of the applica-

tion 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.

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 mapping 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 at solving the ontological mapping problem.

Our future work includes – according to our research program – the following tasks:

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

9. Acknowledgement

We want to thank our medical experts of the *Competence Network Malignant Lymphomas* (Grant No.: 01GI9995/0) and the *Coordination Center for Clinical Trials, Leipzig* for their fruitful discussions concerning ontological foundation in clinical trials. 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 the top-level ontology of GOL. Last but not least we thank Evan Mellander and Frank Loebe for their effort in the editorial work of this paper.

10. References

[1] Genesereth MR, and Fikes R. Knowledge Interchange Format, Version 3.0, Reference Manual. Logic Group Report. Stanford: Computer Science Department, Stanford University; 1992. Report No. 92-1.

[2] Baader F, Calvanese D, McGuinness D, Nardi D, and Patel-Schneider P, eds. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge (UK): Cambridge University Press; 2003.

[3] Sowa JF. *Conceptual Structures: Information Processing in Mind and Machine.* Reading (Massachusetts): Addison-Wesley; 1984.

[4] Rumbaugh J, Jacobson I, and Booch G. *The Unified Modeling Language Reference Manual*. Reading (Massachusetts): Addison-Wesley; 1999.

[5] Dori D. *Object-Process Methodology: A Holistic Systems Paradigm.* Berlin: Springer; 2002.

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

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

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

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

[10] Heller B, Herre H, Lippoldt K, and Loeffler M. Standardized Terminology for Clinical Trial Protocols Based on Ontological Top-Level Categories. In: Kaiser K, Miksch S, Tu S, eds. Symposium on Computerized Guidelines and Protocols (CGP-2004); 2004; Prague: IOS Press; 2004. p. 113-117.

[11] Cimino JJ, and James J. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *Journal of American Medical Informatics Association* 2000; 7(3):288-97.

[12] 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.

[13] Ruan W, Bürkle T, and Dudeck J. A Dictionary Server for Supplying Context Sensitive Medical Knowledge. In: Overhage MJ, ed. Proceedings of the 2000 AMIA Annual Symposium; 2000 Nov 4-8; Los Angeles, USA: American Medical Informatics Association; 2000. p. 719-723.

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

[15] 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).

[16] Gruber TR. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 1993; 5(2):199-220.

[17] Niles I, and Pease A. Towards a Standard Upper Ontology. In: Welty C, Smith B, eds. Formal Ontology in Information Systems: Collected Papers from the Second International Conference; 2001 Oct; New York: ACM Press; 2001. p. 2-9.

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

[19] Heller B, and Herre H. Ontological Categories in GOL. *Axiomathes* 2004; 14(1):57-76.

[20] Aristoteles. *Philosophische Schriften*. Hamburg: Felix-Meiner Verlag; 1995.

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

[22] Chisholm RM. Boundaries as Dependent Particulars. *Grazer Philosophische Studien* 1983; 20:87-96.

[23] Guarino N, Masolo C, and Borgo S. 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.

[24] Barwise J, and Perry J. *Situations and Attitudes*. Cambridge (Massachusetts): MIT Press; 1983.

[25] Rahm E, and Bernstein PA. A survey of approaches to automatic schema matching. *The Very Large Databases Journal* 2001; 10(4):334-350.

[26] Kalfoglou Y, and Schorlemmer M. Ontology mapping: the state of the art. *The Knowledge Engineering Review* 2003; 18(1):1-31.

[27] Klein M. Combining and relating ontologies: an analysis of problems and solutions. In: Workshop on Ontologies and Information Sharing, IJCAI'01; 2001; Seattle, USA; 2001.

[28] Corcho O, Fernández-Lopéz M, and Gómez-Pérez A. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering* 2003; 46(1):41-64.

[29] Musen MA, and Noy NF. The PROMPT suite: interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies* 2003; 59(6):983-1024.

[30] Stumme G, and Maedche A. FCA-MERGE: Bottom-Up Merging of Ontologies. In: Nebel B, ed. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001); 2001 Aug, 4-10; Seattle, Washington, USA: Morgan Kaufmann; 2001. p. 225-234.

[31] Chalupsky H. OntoMorph: A Translation System for Symbolic Knowledge. In: Cohn AG, Giunchiglia F, Selman B, eds. Proceedings of the 7th International Conference on Knowledge Representation and Reasoning (KR2000); 2000; Breckenridge; 2000. p. 471-482.

[32] McGuinness DL, Fikes R, Rice J, and Wilder S. An Environment for Merging and Testing Large Ontologies. In: Cohn AG, Giunchiglia F, Selman B, eds. Proceedings of the 7th International Conference on Knowledge Representation and Reasoning (KR2000); 2000 April 11-15; Breckenridge, Colorado, USA: Morgan Kaufmann; 2000. p. 483-493.

[33] Flouris G, Plexousakis D, and Antoniou G. On a Unifying Framework for Comparing Knowledge Representation Schemes. In: Bry F, Lutz C, Sattler U, Schoop M, eds. Proceedings of the 10th International Workshop on Knowledge Representation meets Databases (KRDB 2003); 2003 September 15-16; Hamburg, Germany: Technical University of Aachen (RWTH); 2003.

[34] Fellbaum C, ed. WordNet: An Electronic Lexical Database. Language, Speech and Communication Series. Cambridge (Mass.): MIT Press; 1998.

[35] Gangemi A, Navigli R, and Velardi P. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In: Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003); 2003 Nov 3-7; Catania, Italy; 2003. p. 820-838.