The Theory of Top-Level Ontological Mappings and its Application to Clinical Trial Protocols

Barbara Heller[*], Heinrich Herre[#], Kristin Lippoldt[*]

Onto-Med Research Group
[*]Institute for Medical Informatics, Statistics and Epidemiology (IMISE),
[#]Department of Formal Concepts, Institute for Informatics (IfI),
University of Leipzig, Germany
Liebigstrasse 27, 04103 Leipzig, Germany
Phone: +49 (0)341 9716104, Fax: +49 (0)341 9716130
herre@informatik.uni-leipzig.de
{barbara.heller, kristin.lippoldt}@onto-med.uni-leipzig.de

Abstract. In the present paper we expound a methodology for the development of terminology systems and the construction of semantically founded knowledge systems. This method is based on ontological mappings using reference top-level ontologies, and is inspired by rigorous logico-philosophical principles. We outline a framework consisting of a system of formal tools designed to support the development of data dictionaries, taxonomies and knowledge systems. The core-module of this system named *Onto-Builder* is an internet-based software application for building context-sensitive data dictionaries. To ensure the broad acceptance of context-dependent descriptions within a diverse group of domain experts, a multistage quality assurance cycle has been established. Ontological mappings based on top-level ontologies are the foundation of a further module of the system, which assists the construction of knowledge systems out of terminology systems. The framework is intended to be applied to the medical domain, in particular to the field of clinical trials.

1. Introduction

In achieving good medical care, quality assurance has become increasingly important in the last few years. This is particularly the case in the area of clinical research, where national and international projects have been undertaken to develop strict guidelines for carrying out clinical trials. The high level of documentation, which such guidelines require, however, is time-consuming and can result in enormous human resource expenditures. In part, this is due to the absence of standardized clinical trial protocols and corresponding CRFs[1]. Another factor is the unavailability of explicit definitions of the medical concepts used in these documents. This is particularly troublesome in multidisciplinary areas of medicine such as oncology,

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

where cooperating experts often interpret medical data differently, each according to his or her individual area of expertise. These multiple interpretations or views often result in the ambiguous, misleading, or incorrect use of concepts in clinical trial protocols, CRFs, and other trial documentation, especially during the preparation of consecutive clinical trial protocols. This in turn can lead to misinterpretations of medical facts and incorrect diagnoses, diminishing the overall quality of health care in general and clinical research in particular.

To address these problems, software applications have been developed in recent years to support diagnostic and therapeutic guidelines [2] as well as the documentation and management process of clinical trials, e.g., eClinical Trial®[2]. Such systems do not currently support context-dependent definition variants for concepts, however, which are particularly useful in large, multidisciplinary projects such as clinical trials, which involve multiple users with diverse backgrounds, specializations and levels of expertise.

As discussed above, it is precisely under these conditions where a single concept can have different meanings for different users as a function of the specific context in which a concept is viewed. Against this background, our approach is focused on the development and implementation of a computer-based framework for a standardized medical terminology with the following aims:

- Reusability of precise definitions of medical concepts to optimize the development of clinical trial protocols and CRFs as well as to achieve better comparability of clinical trial results.
- Availability of a consistent concept base, which supports the harmonization of clinical trial databases as well as the interchange between clinical trial management software and local clinical information systems.
- Availability of a domain-specific ontology for clinical trials, which is based on a top-level ontology and executable on the computer.

The remainder of our paper is structured as follows. In the following section we review three medical terminology systems – SNOMED, UMLS and GALEN – and situate our proposal in the context of current research in terminology management. Section 3 describes the data dictionary model, introduces our software system *Onto-Builder* [3, 4], and discusses the quality assurance cycle. In section 4 the theory of ontological mappings, which are based on top-level ontologies, is expounded; the underlying formal principles are presented in some detail. The discussion on the chosen method and the outlook for further work in the area of ontological research are provided in the last two sections.

2. Terminology Systems

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 to the medical domain, which is sufficiently rich to present all types of terminology

[2] http://www.ert.com/products/eresearch_network.htm

systems. The following authors [5, 6] [7, 8] [9] give a summary of different medical terminology systems and discuss the features of these systems with regard to requirements for concept taxonomies. For our objective to construct 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. In the following we give a short summary of our evaluation results concerning the context representation and concentrating on the most relevant terminology systems.

Within SNOMED CT [10] 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 [11] 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 [12] the entity-types modality and role can be interpreted as context-representing entities. An example for modality is `<FamilyHistory>` whereas in combination with the concept `<Diabetes>` the new concept `<FamilyHistory of Diabetes>` can be derived. Examples for role are `<Steroid which playsRole HormoneRole> or <playsRole 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 super ordinate concepts does not have influence on its attributes/relations.

Our terminology systems analysis has shown that the underlying models of SNOMED, UMLS, GALEN do not fit our requirements with regard to a context-dependent description of concepts. To achieve our goal, namely the definition of a semantically founded and context-dependent data dictionary, we have conceived a terminology model of our own.

3. Terminology Building and Knowledge Acquisition

Our approach aims, in the first step, at the construction of context-sensitive data dictionaries. The innovation of this approach lies in the ontological foundation of the underlying terminology model for basic and domain-specific concepts and relations. The terminology framework is partly based on a generic, domain-independent top-level ontology, described in [13] [14].

The *Onto-Builder* [3] is the core-module of our general framework; it is an internet-based software application, which we have developed as a first prototype for the construction of terminology systems. The *Onto-Builder* offers the possibility to represent natural-language, as well as semi-formal concept descriptions. An ontologically founded frame-work [13] [14] is made available by basic and domain-specific entities for the representation of semi-formal concept descriptions. These concept descriptions are created according to our terminological guidelines [4] which contain lexical and semantic rules for defining medical concepts and relations.

Another module of the system assists the extraction of formal knowledge from several sources; it is intended, in particular, to support the translation of terminology systems and taxonomies into ontologically founded formal *knowledge systems*. Here we use the newly developed theory of ontological mappings, which is based on top-level ontologies. The resulting formal knowledge base is equipped with deductive machinery, which allows for intelligent queries and automatic problem solving abilities.

3.1 Model of the Data Dictionary

The model of the data dictionary is based on the following main entities: concept, denotation, description, context and relation, which are described below.

Concept, Denotation, and Term: A concept is an abstract unit of meaning which is constructed over a set of common qualities [15] 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. In our approach two pairs of opposing concepts are distinguished: generic/domain-specific (e.g., `<disorder>/<disease>`) and primitive/derived concepts (e.g., `<therapy>/<supportive therapy>`). A concept is called generic if it has the same general meaning in different domains (e.g., the concept `<disorder>` has the general meaning that something is deficient or has a defect, independently of the domain in which the concept `<disorder>` is used). The general meaning of a concept is derived of its domain-independent qualities/properties (e.g., in case of `<disorder>` the property `<cause of disorder>` is a general property).

Contrary to this, a domain-specific concept has a concrete meaning only in the domain affiliated to it, (e.g., the concept `<disease>` only has a meaning in the domain of `<living beings>` and not in the domain `<computer science>`). Primitive concepts are concepts which do not reference other concepts and therefore cannot be expressed on the basis of other concepts. In contrast to this, derived concepts reference other concepts. Further ontological categories are discussed in [13].

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. The representation method can be natural-language, semi-formal (e.g., attributes, relations, and rules) or formal (axioms).

Context: With regard to the various discussions on the notion of context, e.g., in [16] 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 generic if concepts which have general properties/qualities are available in it (e.g., a generic context is `<process>` which contains 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 context (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>`.

Relation: according to [14] 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. Our method handles at the present stage 11 basic relations (e.g., `<instantiation>, <membership>, <part-of>, <inherence>, <association, <denotation>, <ontical connectedness>`). These relations are defined and available in our representation language GOL[3] [14]. Examples for domain-specific relations are: `<treatedBy>, <SideEffectOf>` as well as for terminological relations: `<synonymy>, <homonymy>, <polysemy>.`

The basic entities and relations of the data dictionary model are represented in figure 1. The syntax of the model in figure 1 follows the UML[4] syntax, whereas rectangles represent classes (here: entities), rhombus n-ary associations (here: relations) and lines represent relations between the entities.

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 of the corresponding `Context`. If a `Concept` is not yet assigned to a `Context`, a default `Denotation` is given.

[3] 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].

[4] Unified Modeling Language [17]

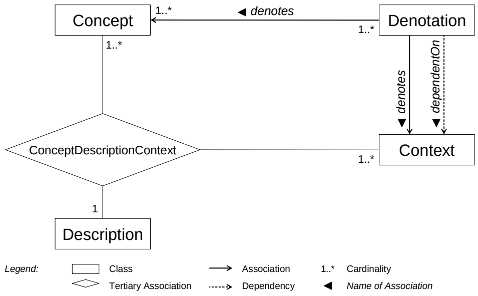

Fig. 1. Excerpt of the data dictionary model

The next two examples show context-dependent descriptions with regard to different granularities on the one hand, and to status and process-oriented aspects on the other.

Example 1: Remission of a tumor

```
<concept>: remission
  <context>: hematology
    <denotation>: hematological remission
      <description>: There are no signs of diseases using examination
      methods which identify variances on the cellular level.
  <context>: cytology
    <denotation>: cytological remission
      <description>: There are no signs of diseases [...] variances on
      the chromosomal level.
```

The difference between the two concept descriptions in example 1 is seen in the different granularity levels (here: the cellular and chromosomal level).

Example 2: Staging

```
<concept>: staging
  <context>: (status(documentation-results))
    <denotation>: staging
      <description>:  Examination  results  of  obligate  examinations:
      anamnesis, clinical and laboratory examinations, gastroscopy, etc.
      <source>: CRF of RICOVER-60 Protocol [18]
  <context>: process
    <denotation>: staging
      <description>: Detection of the anatomic extent of the tumor, both
      in  its  primary  location  and  in  metastatic  sites  through
      exploratory  surgery  or  biopsy  and  assignment  to  the  TNM
      classification stages ...
      <source>: definition derived from [19]
```

Example 2 shows the interpretation of a concept description according to the contexts <status> (here: <documentation-results>) and <process>. It shows also the difference between generic and domain-specific descriptions of the very same concept (here: <staging>). In our example, the process-oriented description is generic for the oncological area, the status-oriented description is specific only for one disease (here: Aggressive Non-Hodgkin's Lymphoma) [18]. Different relations are valid in the various contexts (e.g., in the context (<status>(<documentation-results>)) the relation <has ExaminationResult> is valid and in the context <process> the relation <hasMethod> is valid).

3.2 Quality assurance cycle

The quality assurance cycle guarantees the broad acceptance of context-dependent descriptions within a group of domain experts. This cycle is based on five user roles, which are dependent on the following aspects: function, organization, experience, qualification, and language. A personal profile is derived from the information about the respective aspects for every user (e.g., person A has the following profile `<function>:` editor, `<organization>:` EORTC, `<experience>:` expert in medicine, `<qualification>:` principle investigator, `<language>:` English). According to this user profile, person A is authorized to work on difficult descriptions of medical concepts in the context of clinical trials within the consensus process of the EORTC (European Organization for Research and Treatment of Cancer). To reach a consensus, a workflow with integrated iterative steps has been established. According to the complexity of the concept descriptions, these steps can be modified dynamically with regard to multiple checks of concept descriptions and different user roles. The result of the whole quality assurance cycle is expected to be a consistent and accepted terminological basis. Real consistency cannot be guaranteed; so methods must be developed for consistency checks. In case no consensus can be found, the terminological basis will be tested against our ontological framework.

4. Ontological Mappings Based on Top-Level Ontologies

In this section we describe and discuss some formal basic principles, which are important for the task of constructing a knowledge base out from a terminology system. The ontological mappings, which are introduced and considered in the sequel, are centered on a top-level ontology *TO*. Hence, the implementation of ontological mappings according to our approach presupposes some fixed top-level ontology. The research group Onto-Med is developing a top-level ontology which is called *GFO* (General Formal Ontology) and which is part of the GOL-project of the University of Leipzig [13], [14]. The module of the *Onto-Builder*, which supports knowledge extraction, is based on the top-level ontology *GFO*.

4.1 Formal Principles

We expound in more detail the construction of a formal knowledge bases assisted and supported by top-level reference ontologies. Generally, a formal ontology $Ont = (L, V, Ax(V))$ consists of a structured vocabulary V, called 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 L. 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)).

A terminology system *TS* may be considered as a 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 Ont^d of *Ont* by (the set of definitions) *DF* 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.*

The notion of *semantical 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*. Ontological mappings can be used as a formal framework for schema matching, which is a basic problem in many database application domains, compare [20]. An advanced elaboration of this theory, which is being investigated by the Onto-Med group, will be presented in [21].

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 $e(C_1,..,C_n, R_1,\dots, R_m)$ which includes concepts $C_1,...,C_n$ and relations $R_1,\dots, R_m$ .The concepts $C_1,..,C_n$ and relations $R_1,\dots, R_m$ 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 $S_1,\dots,S_l$ – 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 reason *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 out from the set $\{Def(t) : t \in Tm\}$ (*problem of primitive base*)

2. Construction of an extension TO_1 of TO by adding 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*).

A developed theory of ontological mappings based on top-level ontologies is in preparation and will be expounded in [21].

4.2 The Basic Modularization

In analysing a natural language text *T* one should satisfy the following basic modularity principle: Firstly, we 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 be then represented as the union of two disjoint modules:

1. a set *Ax(PCR)* of axioms about the concepts and relations of *PCR*
2. 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)* und 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.

4.3 Example

To illustrate some aspects of ontological mappings we consider the following short example. We focus on the first reduction step of selecting a set of primitive domain-specific concepts. Therefore we will give a preliminary definition of primitive domain concepts.

Definition: A set of concepts *C* is called primitive concept base for a class *DOM* of domains (of the same granularity) iff every concept $d \in C$ is generic with respect to all domains from *DOM* and if there does not exist a concept $d \in C$ which is derivable from the set of concepts $C – \{d\}$ on the same granularity level.

Application

Tissue in the medical sense is to be seen as contained in a primitive domain-specific concept base because its meaning and interpretation is the same in different medical domains (e.g. pathology, endocrinology). The domain-specific concept tissue can be interpreted as a "part of an organism consisting of an aggregate of cells having a similar structure and function"[5]. Normally the concept tissue can be partly derived from the more granular concept cell. In our approach the derivation of concepts is limited to concepts of the same level of granularity and therefore the concept tissue is

[5] [http://www.hyperdictionary.com/]

not derivable from the concept cell. In contrast to tissue the concept fatty tissue should not be considered as a primitive concept. It has the same meaning in different contexts but can be derived directly from the concept tissue and the concept fatty on the same granularity level. Further examples for primitive domain-specific concepts are body, cell, organ, tumour, disease, therapy. To give an example for the main ideas of a local ontological mapping ontological sketched above we consider the

concept *C* organ system
and its
definition *D* A group of organs, vessels, glands, other tissues, and/or pathways which work together to perform a body function within a multicellular organism.

In the first step we analyze the natural language definition *D* with regard to the concepts and relations it includes. These concepts and relations must be classified in primitive and derived concepts and relations. In the given definition the following concepts should be included, among others, in a primitive domain-specific concept base: organ, vessel, gland, tissue, organism. For further analysis let us consider the primitive concept tissue and focus on its structural aspects. The concept tissue has to be classified within the top-level ontology *GFO* as physical endurant. This assignment is part of the ontological embedding of the base of primitive concepts into the hierarchy of categories of *GFO* i.e. `(tissue is-a substance)`.

Further steps of the construction of an ontological mapping have to take into consideration suitable extensions of *GFO* to finally achieve formal expressions (in the framework of *GOL*) which are semantically equivalent to the concepts included in the primitive concept base C.

5. Comparison with other Approaches

We suppose that a *semantic translation* maps knowledge formulated in a source language to some equivalent expression in a target language. This very broad understanding comprises *knowledge extraction* on the basis of natural language texts as well as translations between formal languages. It is common in both cases that the semantics is to be preserved by the transformation. *Ontological mappings* in the sense of the current papers are semantic translations, which are based on top-level ontologies. The meaning of the term *ontology mapping* differs from the meaning of our ontological mappings; ontology mappings are semantic translations between formal knowledge bases (which in many cases are called ontologies).

In the present scientific landscape, two types of tasks (knowledge capturing/extraction vs. ontology mapping) are rather separated. Ontology-related communities in computer science usually deal with translations of knowledge expressed in formal languages, e.g. translations 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 [22] [23]; cf. also section 3.6 of [24] and related works discussed in [25]. Schema matching in the database area is frequently considered a similar task, and it is reviewed in [20]. Therefore, we refrain from giving an extensive comparison of single publications. Some of the major works as regards appearance in the literature are FCA-Merge [26], OntoMorph [27], Chimaera [28] and the tools of the PROMPT suite [25], which is developed at Stanford University.

Note that all of these works have not solved the need for a terminological standardization. This is still one problem of the emergent area of ontology mapping. This can also be recognized by the collections of terms presented in [22] [23].

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. [29] 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/extraction, 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 by WordNet [30]. 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. [31]). Together with sample text corpora tagged with WordNet synsets such an alignment may allow for an improved automated formalization of natural language texts.

We may summarize that ontological mappings as introduced in the current paper can be understood as semantic translations which are centered around top-level ontologies. The target language is always a formal language in which the top-level ontology and its extensions are formalized. Hence, almost all of the mentioned approaches can be interpreted as special cases of ontological mappings.

6. Results and Conclusions

The software tool *Onto-Builder* is the core module of our general framework and has been developed as a first prototype to construct terminology systems. In 2002, the *Onto-Builder* was introduced in the *Competence Network Malignant Lymphoma* and in the *Coordination Centers for Clinical Trials, Cologne* and *Leipzig,* Germany.

Initially a multilingual data dictionary was constructed for the area of clinical trials. In the present version it includes approximately 13 contexts, 1000 domain-specific concepts and 2500 concept descriptions. The evaluation of the data dictionary in the medical research networks has shown that it can be efficiently adapted to different medical domains (here: malignant lymphoma, cardiovascular diseases). The experience gained has shown that explicit concept descriptions are of great use for applications in the domain of clinical trials, e.g., by saving time and improving quality assurance. By integrating medical experts into the development process, a high degree of acceptance of the concept definitions in the data dictionary was reached using the quality assurance cycle.

The explicit separation between the entity types concept, context and relation within our terminology framework permits a high degree of flexibility with regard to extendibility and adaptability. The concept descriptions existing in the first version of the data dictionary still allow for slightly different interpretations despite the assignment of concepts to contexts. This is due to the absence of a completed domain-specific ontology on the basis of which clear descriptions (statements) can be made about the concepts, as well as the absence of an ontological mapping method. In our opinion a clear context-dependent concept definition can be reached if the definition is available in a semi-formal representation language and if this language is based on the basic- and domain-specific entities of the ontological framework. On the way to a representation of concept descriptions which is semi-formal and based on ontologies, we have to be concerned following problems:

- Finding an adequate degree of ontological mapping to make applicability possible.
- Finding clear criteria for the distinction between primitive and derived concepts as well as between general and domain-specific concepts.
- Finding solutions for linguistic problems (e.g., handling of synonyms, homonyms, polysems).
- Finding an intermediate representation level of semantics, which is able to close the gap between natural-language representations and formal ontological propositions while remaining consistent with the top-level ontology of our ontological framework.

If we overcome these problems we will achieve a deeper semantic foundation of concept descriptions in contexts. Our data dictionary is merely a concept base for clinical trials at the present stage and not yet a fully developed and formalized domain ontology. The reason for this lies in 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. Ontological mapping is a current research topic in the science of *Formal and Applied Ontology*. Against this background in the present paper we discussed the following fundamental issues of ontological mapping:

- Formal principles and ontological mapping tasks
- Basic modularization of knowledge systems.

Future work consists in the further development of the theory of ontological mappings, in the explicit representation of semi-formal descriptions for domain-specific concepts as well as in the expansion of the theoretical framework by further basic categories (e.g., situations, views, qualities).

Acknowledgements

We want to thank our medical experts of the *Competence Network Malignant Lymphoma* and the *Coordination Centers for Clinical Trials*, *Cologne* and *Leipzig* for their support in creating and analyzing medical concepts. Many thanks to the members and the Ph.D. students in the Onto-Med research group for fruitful discussions and implementing software modules for the *Onto-Builder*. In particular we thank Frank Loebe for providing his analysis results concerning the comparison to other approaches (chapter 5).

References

[1] ICH. 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, Löffler M, Musen M, and Stefanelli M, eds. Computer-Based Support for Clinical Guidelines and Protocols. Amsterdam/Berlin/Oxford: IOS Press; 2001.

[3] Heller B, Lippoldt K, and Kuehn K. Onto-Builder - A Tool for Building Data Dictionaries. Onto-Med Report. Leipzig: Forschungsgruppe Ontologies in Medicine, Universität Leipzig; 2003. Report No. 3.

[4] Heller B, Lippoldt K, and Kuehn K. Guideline for Creating Medical Terms. Onto-Med Report. Leipzig: Research Group Ontologies in Medicine, University of Leipzig; 2003. Report No. 4.

[5] 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. *Journal of American Medical Association* 1997; 4:238-251.

[6] Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods of Information in Medicine* 1998; 37(4-5):394-403.

[7] Rector AL. Clinical Terminology: Why Is it so Hard? *Methods of Information in Medicine* 1999; 38:239-252.

[8] de Keizer NF, Abu-Hanna A, and Zwetsloot-Schonk JHM. Understanding terminological systems. I: Terminology and typology. *Methods of Information in Medicine* 2000; 39(1):16-21.

[9] de Keizer NF, and Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. *Methods of Information in Medicine* 2000; 39(1):22-29.

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

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

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

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

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

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

[16] Bouquet P, Ghidini C, Giunchiglia F, and Blanzieri E. Theories and uses of context in knowledge representation and reasoning. In: Journal of Pragmatics: Elsevier Science; 2003: p. 455-484.

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

[18] 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.

[19] Braunwald E, Isselbacher KJ, Petersdorf RG, Wilson JD, Martin JB, and Fauci AS, eds. Harrison's Principles of Internal Medicine. 11 ed. New York: McGraw-Hill Book Company; 1987.

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

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

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

[23] 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.

[24] Gómez-Pérez A, Fernández-López M, and Corcho O. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Berlin: Springer; 2004.

[25] 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.

[26] 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.

[27] Chalupsky H. OntoMorph: A Translation System for Symbolic Knowledge. In: Proceedings of 7th International Conference on Knowledge Representation and Reasoning (KR2000); 2000; Breckenridge; 2000. p. 471-482.

[28] 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.

[29] 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.

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

[31] Gangemi A, Navigli R, and Velardi P. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In; 2003 Nov 3-7; Catania, Italy; 2003. p. 820-838.