Modeling the International Classification of Diseases (ICD-10) in OWL

Manuel Möller, Daniel Sonntag, and Patrick Ernst

German Research Center for AI (DFKI),
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
manuelm@manuelm.org
sonntag@dfki.de
patrick.ernst@dfki.de

Abstract. Current efforts in healthcare focus on establishing interoperability and data integration of medical resources for better collaboration between medical personal and doctors, especially in the patient treatment process. In covering human diseases, one of the major international standards in clinical practice is the International Classification for Diseases (ICD), maintained by the World Health Organization (WHO). Several country- and language-specific adaptations exist which share the general structure of the WHO version but differ in certain details. This complicates the exchange of patient records and hampers data integration across language borders. We present our approach for modeling the hierarchy of the ICD-10 using the Web Ontology Language (OWL). OWL, which we will introduce shortly, should provide a formal ontological basis for ICD-10 with enough expressivity to model interoperability and data integration of several medical resources such as ICD. Our resulting model captures the hierarchical information of the ICD-10 as well as comprehensive class labels for English and German. Specialities such as "Exclusion" statements, which make statements about the disjointness of certain ICD-10 categories, are modeled in a formal way. For properties which exceed the expressivity of OWL-DL, we provide a separate OWL-Full component which allows us to use the hierarchical knowledge and class labels with existing OWL-DL reasoners and capture the additional information in a Semantic Web format.

1 Introduction

Over the last decades healthcare has changed from isolated treatments towards a distributed treatment process. This process depends greatly on the cooperation of specialized medical disciplines. Moreover, medicine questions require to take an enormous amount of expert knowledge into account before decisions are made. To facilitate information exchange and knowledge sharing in medical domains, standardization is playing an important role. The goal is to increase the interoperability within all domains of the healthcare industry so that the interchange of documents can be simplified and medical workflows can be improved.

The difficulty in the area of medical knowledge management is the high diversity of knowledge about single entities. Let us consider a patient in a clinical environment. Even if he does not have a complex disease, he would have to undergo a high number of examinations in different clinical departments. In each step of this treatment process, huge amounts of metadata are created and stored, based on single specific models every time. The challenge is to integrate these *islands of information* [10], so that an overall knowledge base can emerge.

A second problem of this integration process is *semantic heterogeneity*, which means, that there are disagreements about the semantics and names of concepts between the terminologies. Additionally, medical knowledge is very complex and evolves continuously over time. Therefore, new architectures and standards are needed that deal with these problems [18], [17].

Standardized terminologies have a long history in medicine. For human diseases, the first approaches date back to the 18th century. The roots of the modern International Classification of Diseases (ICD) can be traced back to the Bertillon Classification of Causes of Death. The ICD was introduced in 1893 at the International Statistical Institute in Chicago. Five years later, the American Public Health Association (APHA) recommended that Canada, Mexico, and the United States should also adopt it. Many other countries joined subsequently. It was revised several times over the last 100 years. The sixth revision included morbidity and mortality conditions and was renamed the "Manual of International Statistical Classification of Diseases, Injuries and Causes of Death (ICD)." Since 1948 the World Health Organization assumed the responsibility for maintaining and publishing revised versions of the ICD. The current version is ICD-10 from 2006.

Although the overall structure of the ICD-10 was accepted by numerous countries, different versions exist which are maintained by national institutions. For instance, the German version of the ICD-10 is maintained by the DIMDI[1] which is under the authority of the German Federal Ministry of Health. While major parts of the ICD-10 hierarchy are equal both in the DIMDI version and the WHO version, we found out that the structure and content of certain parts of the ICD-10 varies. Section 3.8 provides details of these differences.

The aim of the work presented here is to generate an ontology covering the domain of human diseases based on the classifications of the two country specific ICD-10 versions described above. The ultimate goal is to leverage technologies from the Semantic Web to ease the work of medical experts (1) by supporting them in making medical image data as well as patient records available for semantic search, and (2) by providing intelligent annotation suggestions based on rich formal models for medical domain knowledge. This research was triggered by the broader effort within the research projects MEDICO and RadSpeech (http://www.dfki.de/RadSpeech/). From our discussions with clinicians we learned that a representation of the ICD-10 is an absolute necessity for the efficient semantic radiological image annotation in the everyday practice of the university hospital participating in MEDICO [12].

[1] Deutsches Institut für Medizinische Dokumentation und Information

2 OWL for Medical Ontologies and Related Work

OWL is based on First Order Logic and currently is the most commonly used language to represent formal ontologies in the Semantic Web. It is similar to XML Schema [7] in so far as it allows for a specification of constraints for the structure of XML documents. Additional to that, OWL provides information about the interpretation of RDF statements and constructs for representing classes, subclasses as well as typed properties and sub-properties. Domains and ranges (restrictions regarding the subjects and objects of properties) can be specified and cardinality constraints can be formulated. Concrete objects are instances belonging to a certain class. In a simple example ontology, the actual liver of the author and the liver of the reader are both *instances* of the *class* liver. These instances might differ in size, weight, etc., and are composed of different physical matter, of course. But they share certain features like function, being part of the respective human body, and so on to qualify them as belonging to the same class liver. A class in an ontology can have any number of instances.

Typical questions that an OWL reasoner can answer can be roughly divided into two groups: (1) questions that are limited to the class level model and (2) questions also involving instances. Typical reasoners for OWL-DL are Pellet [16] and KAON2 [9]. A typical task for reasoners is to check for the global consistency of a (medical) ontology. Consistency means that there are no logical contradictions in the ontology. This is usually used to detect modeling mistakes. An examples for the application of such tests for finding incorrect modeling in the medical terminology SNOMED[5] (Systematized Nomenclature of Human and Veterinary Medicine) can be found in [4].

The initial idea for generating an OWL version of ICD-10 from data available on the web dates back to a similar approach for generating an OWL model for ICD-9 as presented in [11]. Biomedical ontologies and terminologies received high attention in the last decade and provide promising technologies for data integration. Bodenreider et al. evaluated popular large scale ontologies such as SNOMED, FMA, and Gene Ontology and stated that "ontologies play an important role in biomedical research through a variety of applications" [2]. In this context, a number of semi-structured medical terminologies and classification systems have been converted to formally structured formats recently. For SNOMED, an OWL ontology was created and used to detect weaknesses in the original modeling [14,15]. Noy and Rubin have presented an approach for translating the Foundational Model of Anatomy ontology (FMA) to OWL [13]. From their approach we adopted the idea to split the generated ontology into an OWL-DL and an OWL-Full component. Cardillo et al. presented an approach for a formal representation of mappings between ICD-10 and the International Classification of Primary Care version 2 (ICPC-2) [3]. However, their focus was on the formal representation of mappings between ICD-10 and ICPC-2. The work presented in this paper tries to complement their efforts by providing a formal model of additional relations within the ICD-10.

3 Modeling Approach

This section describes our general approach for the generation of the ICD-10 in OWL. Figure 1 shows the data flow during the ontology generation process. Following the elements in this diagram, the subsequent sections will discuss the different processing steps and give details about the applied techniques and algorithms.

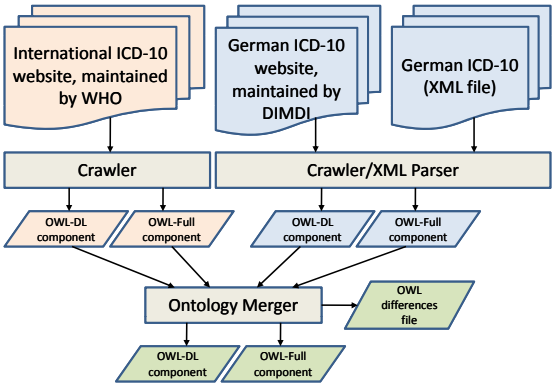

Fig. 1. Data flow of the ontology generation process

3.1 Data Sources

The OWL ontology which we generated is based on data available via the websites of the organizations responsible for maintaining the respective ICD versions. As we will show, the data which is publicly available on the Internet is well suited to generate a rich formal model of the classification of human diseases. The websites are highly structured and contain enough information to fit the use case in the MEDICO project. Another advantage is that we can reflect updates of the ICD published on the websites by re-running our crawlers.

For the English version of the ICD we used the official WHO website.[2] The website only partly reflects the original hierarchical structure of the ICD-10. As an additional source we used the ICD-10 manual [21]. Figure 2 (a) shows a screenshot covering the first of "Nutritional anaemias (D50-D53)."

From the different German sources available we chose to use the current ICD-10-GM, "GM" being the "German Modification" (see Section 3.8). Figure 2 (b) shows the same fragment of the ICD-10 as the previous screenshot, but this time in German. Our starting points for the German ICD-10 is the respective website and a publicly available XML file which is structured using the Classification

[2] `http://apps.who.int/classifications/apps/icd/icd10online/`

Nutritional anaemias (D50-D53)

D50 **Iron deficiency anaemia**
Includes: anaemia:
· asiderotic
· hypochromic

D50.0 **Iron deficiency anaemia secondary to blood loss (chronic)**
Posthaemorrhagic anaemia (chronic)
Excludes: acute posthaemorrhagic anaemia (D62)
congenital anaemia from fetal blood loss (P61.3)

D50.1 **Sideropenic dysphagia**
Kelly-Paterson syndrome
Plummer-Vinson syndrome

D50.8 **Other iron deficiency anaemias**

D50.9 **Iron deficiency anaemia, unspecified**

(a) Example for an English entry from the WHO ICD-10 website

Alimentäre Anämien (D50-D53)

D50 **Eisenmangelanämie**
Inkl.: Anämie:
· hypochrom
· sideropenisch

D50.0 **Eisenmangelanämie nach Blutverlust (chronisch)**
Posthämorrhagische Anämie (chronisch)
Exkl.: Akute Blutungsanämie (D62)
Angeborene Anämie durch fetalen Blutverlust (P61.3)

D50.1 **Sideropenische Dysphagie**
Kelly-Paterson-Syndrom
Plummer-Vinson-Syndrom

D50.8 **Sonstige Eisenmangelanämien**

D50.9 **Eisenmangelanämie, nicht näher bezeichnet**

(b) Respective entry from the German DIMDI ICD-10 website

Fig. 2. Language-specific ICD-10 online versions: *Nutritional anaemias* in English and *Alimentäre Anämien* in German

Markup Language (ClaML). As the name suggests, ClaML is special language designed to represent classification hierarchies [8]. It provides special notations to state super- and subclass relations, declare attributes, and to specify metadata elements, among other things. To interpret the notation correctly, we use the WHO manual [21] and a supplementary documentation for attributes exclusively stated in the DIMDI version [6] using the the DIMDI website.[3]

3.2 OWL Model Generation

The general structure of the ICD-10 is as follows. It consists of "Chapters" using Roman numerals from I to XXI. The chapters again contain "Blocks of categories" (e. g., Chapter III: "Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism") which specify a range of categories of a particular aspect (e. g., D50-D89). These blocks then contain "Categories," denoted by an ICD-10 code, a capital letter, and Arabic numbers (e. g., D50-D53: Nutritional anaemias; D55-D59: Haemolytic anaemias; etc.). They are further subdivided into "Subcategories." The subcategories are coded by attaching an additional digit after the decimal point (e. g., D50.1: Sideropenic dysphagia) . These codes differentiate from the specific language or writing system of the different ICD-10 versions.

Contrary to the WHO website, the DIMDI XML file constrains the subcategories by using an additional decimal number, which is appended to the codes of their parent categories. These subcategories are also defined using "Modifiers" which specify a set of "ModifierClasses" as subclasses. Each of these classes possesses a number, a label, and additional information such as "Exclusions" or "Inclusions." If a category contains a "Modifier," it will be specialized by generating new categories with each particular "ModifierClass." These relations are represented in OWL by creating a new OWL class for each "Modifier," which is a subclass of `icd10:Modifier`, and defining the appropriate ModifierClasses as subclasses. The combination is denoted using an `owl:unionOf` of the specific category and each ModifierClass. This form of specification is used extensively. Our analysis has shown that 4488 of the 16214 classes are specified in this way by DIMDI.

ICD-10 is not only a classification of diseases, but the terminology also includes links to other related aspects, such as symptoms, signs and consequences of other external causes. Therefore, the manual describes an additional level of order which groups certain chapters according to their particular aspects. For example, "Chapters I to XVII relate to diseases and other morbid conditions." It is worth mentioning that this systematic level is not available from the website but only from the manual.

Using this information about chapters and groups of chapters we modeled the first two hierarchy levels by hand. The OWL class `icd10:Entry` is the super class of the bilingual ICD-10 hierarchy. As mentioned before, differences between

[3] `http://www.dimdi.de/static/de/klassi/diagnosen/icd10/htmlgm2009/index.htm`

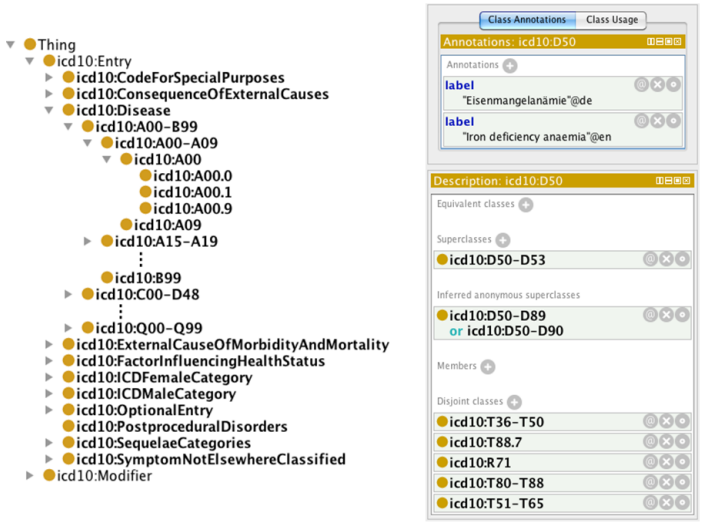

Fig. 3. (Left) General class hierarchy of the OWL model; (Right) Screenshot of the OWL-DL version of the ICD-10 in the Ontology Editor Protégé

the German and English ICD-10 exist. Our analysis shows that there are ICD-10 categories which are present in the German ICD-10 but not in the English ICD-10 and vice versa. Section 3.8 gives details about these differences.

In addition, the origins of the concepts are also encoded in our OWL model. The first approach was to add a super class for each class to denote its origin. But this proved to be incorrect. Let us consider the block R00-R99, which is present in both terminologies and thus gets both super classes. The symptom R65, which has the super class R00-R99, is only stated by the DIMDI. Consequently, a reasoner would infer that R65 is in the DIMDI and WHO version, because it would build the transitive closure and R65 would get both super classes. For that reason, we decided to denote the provenance using the two boolean OWL-Full properties `icd10:isDIMDIEntry` and `icd10:isWHOEntry`.

Figure 3 shows an (abbreviated) example of the generated class hierarchy. We use two HTTP crawlers, implemented in Java, to generate OWL models for each of the two input sources. OWL classes and axioms are generated using the Jena Ontology API.[4] Other libraries—such as the OWL API [1]—were not able to handle the OWL-Full expressivity of our modeling.

The generated OWL model consists of two components. The OWL-DL part contains the hierarchy of the ICD-10 according to the hierarchical structure described above. All ICD-10 categories and subcategories are reflected by OWL classes. The hierarchical information is reflected by `owl:subClassOf` axioms. For a discussion of the contents of the OWL-Full part see further below.

[4] `http://jena.sourceforge.net/ontology/`

We will explain the next steps by giving an example. Figure 2 shows the first part of the WHO ICD-10 website about "Nutritional anaemias (D50-D53)." We will focus on the entry "D50.0 Iron deficiency anaemia secondary to blood loss (chronic)" as a guiding example throughout this paper.

Each class is identified by an URL, which consists of a specific ICD-10 name space and the special term as the URL anchor. The terms for the categories are simply the particular ICD-10 codes. For blocks and chapters, a range pattern is used which covers their content, e. g., D50-D53. From this we create an OWL class with the local name "D50.0." The ".0" indicates that this is a sub-category of "D50 Iron deficiency anaemia." Thus, we add an `owl:subClassOf` axiom which represents this relationship. The bold-faced name of the sub-category becomes the English `rdfs:label` of this class. Later, by merging with the OWL model of the German ICD-10, we can also add the German labels "Eisenmangelanämie nach Blutverlust (chronisch)." For some concepts, the DIMDI specifies up to three labels, which differ in their length and detail. The smaller labels are, thereby, necessary because some print formats require them. In our case we can neglect this limitation and use always the most detailed label available. We use the standard XML language tags to differentiate between these languages.

3.3 ICD-10 Characteristics

Despite a specialization hierarchy, multiple characteristics are stated in the WHO manual [21] and the DIMDI supplement [6], and they can be shared by different classes. These are:

Dagger and Asterisk categories: Statements containing information about an underlying disease with a particular additional manifestation can be expressed thanks to asterisk and dagger codes. Underlying diseases are marked with a dagger and are the primary criterion. Therefore, they have to appear in the diagnostic statement, whereas the manifestation marked with an asterisk is only additional. These circumstances are represented in OWL using two properties, namely `icd10:hasAdditionalManifestation` with its inverse `icd10:hasUnderlyingDisease`. The first one's domain is all classes which represent a dagger category and have a range of all asterisk categories. The restriction that an additional manifestation needs at least one underlying disease is expressed by the property's cardinality, which is at least one.

Optional concepts: The DIMDI defined a supplemental characteristic and this is only applied in their version of the terminology. Optional concepts are similar to dagger and asterisk categories. If marked as optional, a concept will be mandatory for some diagnoses but only supplemental for other ones.

Categories limited to one gender: The ICD-10 contains several categories which are only applicable to either males or females. Consider, for instance, diseases of the genitals, like "D40 Neoplasm of uncertain or unknown behavior of male genital organs" or conditions which occur during the pregnancy of women, e. g., "O00 Ectopic pregnancy." The facts are represented by one super class for each gender.

Sequelae categories: Sequelae categories are used for mortality cause encoding. They indicate that the death is not caused by the main effect of a given disease. Instead it is caused by residual effects.

Postprocedural disorders: Categories which fall under this characteristic point out conditions and complications which occur after treatment, e. g., surgical wound infections or shock.

Contrary to dagger and asterisk categories, each characteristic is represented using a specific super class. All classes which share the characteristic are subclasses of this class.

3.4 Handling ICD-10 Exclusions

For some ICD-10 categories so-called "Exclusions" also exist. According to the ICD-10 manual [21], they exclude certain conditions that, "although the rubric title might suggest that they were to be classified there, are in fact classified elsewhere." The example in Figure 2 (a) lists two such excludes for D50.0: "acute posthaemorrhagic anaemia" with a link to ICD-10 category D62 and "congenital anaemia from fetal blood loss" with a link to category P61.3. We capture this information by adding `owl:disjointWith` axioms between D50.0 and D62 as well as between D50.0 and P61.3. This can be expressed using the expressivity of OWL-DL (see Figure 4).

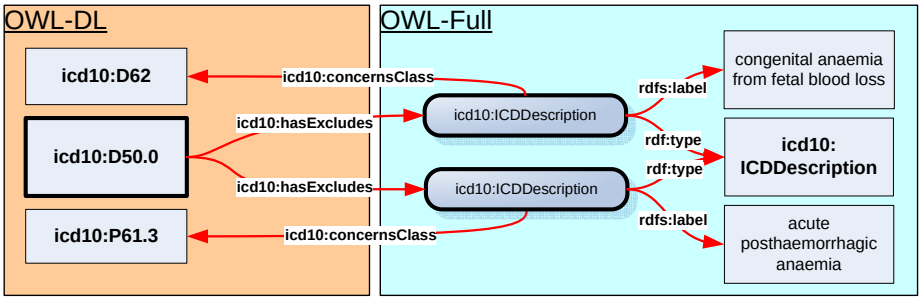

Fig. 4. Structure and relationship of OWL-DL and OWL-Full component by example

However, by relying exclusively on `owl:disjointWith` axioms, we would lose important information. As the ICD-10 manual states, exclusions can be extended using additional strings and constructions of braces. They indicate that neither the words that precede them nor the words after them are proper terms. Thus, a more precise qualification has to be applied [21]. If we compare the brace constructs with the encoding in the XML file, it becomes clear that they are used to provide a more comprehensive structuring of the data. The XML file reflects them by splitting up the "Exclusions" and adding a new fragment for each brace element to the "Exclusions." Figure 5 depicts "Inclusions" for concept

"O71.6: Obstetric damage to pelvic joints and ligaments" using braces and figure 6 shows them using the German XML encoding. As we see, without proper post-processing we are not able to relate that the "Exclusion" of concept M54.1 is shared by multiple Exclusions.

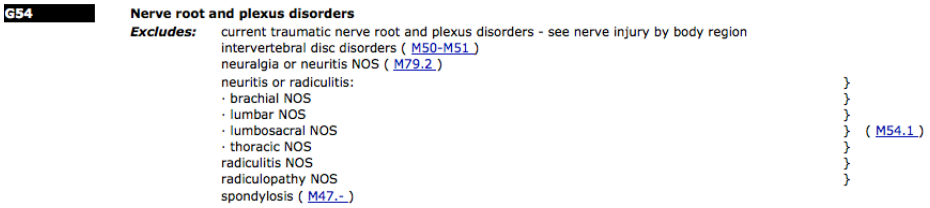

Fig. 5. Example for ICD-10 category with "Exclusions" that do not point to other ICD-10 categories

These qualifications are covered in additional OWL individuals of the upper level class `icd10:Description`. The individuals can have several properties of type `icd10:concernsClasses`. Because this property concerns other ICD-10 categories, it needs to have a class-valued range. Thus, the individuals require OWL-Full expressivity. To encode the string information of an "Exclusion," we use `rdfs:label`. For each excluded statement, we generate one individual. We also encode the information contained in the brace constructs, which appear in the WHO version. Therefore, we create an individual for the information which occurs after the braces. They are related to each Exclusion using an OWL-Full property `icd10:qualifiedBy`. Extracting this information for the DIMDI data is among our next steps.

Additionally, a closer look at the ICD-10 revealed that for numerous categories, the "Exclusions" do not point to other ICD-10 categories but to arbitrary descriptions of certain medical symptoms. Figure 5 gives an example showing ICD-10 subcategory "O71.6: Obstetric damage to pelvic joints and ligaments." As we cannot generate proper disjointness axioms for these exclude expressions we decided to store them using the exclude individuals described above without pointing to another ICD-10 category.

```
<Rubric kind="exclusion">
    <Label xml:lang="de" xml:space="default">
        <Fragment type="list">Neuritis oder Radikulitis:</Fragment>
        <Fragment type="list">brachial o.n.A.<Reference class="in brackets" code="M54.1">M54.1-</Reference></Fragment>
    </Label>
</Rubric>
<Rubric kind="exclusion">
    <Label xml:lang="de" xml:space="default">
        <Fragment type="list">Neuritis oder Radikulitis:</Fragment>
        <Fragment type="list">lumbal o.n.A.<Reference class="in brackets" code="M54.1">M54.1-</Reference></Fragment>
    </Label>
</Rubric>
```

Fig. 6. Example for ICD-10-GM braces constructs within "Exclusions"

3.5 Handling ICD-10 Inclusions and Notes

Similar to the "Exclusions" two other properties for categories are part of the ICD-10. "Inclusions" are additions to the rubric in which they occur. The ICD-10 manual describes them as a guide and provides examples to formulate diagnostic statements. "Inclusions" are represented using the individuals in OWL-Full in the same way as "Exclusions."

In addition, it is possible that a note is provided for an ICD-10 element. These notes give hints how to use the particular category, block, or chapter. For example, a physician who is writing a medical report sees from these hints that the category "G09 Sequelae of inflammatory diseases of central nervous system" is to be used to indicate conditions whose primary classification is G00-G08 (i. e., excluding those marked with an asterisk) as the cause of sequelae, themselves classifiable elsewhere. The only purpose of the notes is to support human beings in interpreting the ICD-10. Also, they are not interpretable for reasoners because they only contain continuous text. For that reason, we do not relate the notes individuals to classes with OWL-Full properties, instead we use `owl:AnnotationProperties`.

3.6 Merging the English and German OWL Models

To merge the two ICD-10 variants, we have to distinguish between the OWL-DL and Full parts. The two OWL-Full parts are merged by just importing them into a new ontology. This is possible because they only contain properties of the classes defined in the DL versions and the class definitions were not altered during the merging process.

The merging of the DL ontologies can be divided into two phases. First, an automatic integration is performed, which is then refined by a manual step. The automatic merging process starts with the WHO ontology. As stated in the last paragraph of section 3.2, all classes are identified using their particular ICD-10 code. In most of the cases, these codes are the same in the DIMDI and WHO version. Therefore, we begin the integration by checking if each DIMDI class is present in the WHO version. If so, we add the label and `owl:subClassOf` properties of the class. Adding `owl:subClassOf` axioms is necessary because only some of them exist in the DIMDI version. For example, if a block only appears in the DIMDI version, all `owl:subClassOf` relations concerning this block only occur in the DIMDI ontology. If the class is not contained, it will be created and all properties will be copied.

In addition, a few classes exist which are semantically very similar, but differ by their ICD code. For example, the DIMDI chapter D50-D90 only varies in the range and the existence of the class D90 from the WHO chapter D50-D89, but both concern the same diseases. To merge these classes, we first manually determine all possible pairs $(classA, classB)$ which differ in this sense. After that, we define a new super class $unionAB$ for each pair. This is the `owl:unionOf` of the pair's classes and gets the a concatenation of the local names of both. To determine the location of the new class, we search the first super class which

classA and *classB* have in common. *unionAB* is then added as a subclass of this class. This traversing is necessary, because there can be super classes not covering the entire range of *unionAB*. For example, we merge the blocks D80-D89 and D80-D90 and the direct super classes are D50-89 and D50-D90 respectively, which are only present either in the WHO or DIMDI version. Therefore, we have to traverse the hierarchy one step further and find the appropriate super class, which is `owl:unionOf` of the classes D50-D89 and D50-D90.

Besides the new integrated ontology, a difference ontology is generated during the merging process, which distinguishes between WHO and DIMDI classes only occurring in one version. We know that this knowledge is already contained in the ontology by the label and `icd10:isDIMDIEntry` or `icd10:isWHOEntry`. However, we decided to produce an explicit representation of the differences, because it makes the merging process more transparent and the differences are easier to examine. For these reasons, we denote the differences in both ontologies using OWL-Full properties in the difference ontology. We are using one property for every ICD source to denote the exclusiveness and one property to denote the classes later manually added by the merging process. Figure 3 shows a screenshot of the generated OWL class for ICD-10 category D50.0 in Protégé[5].

3.7 Results and Discussion

By the definition given in the ICD-10 manual, the ICD is a classification system with "a hierarchical structure with subdivisions." And further: "A statistical classification of diseases should retain the ability both to identify specific disease entities and to allow statistical presentation of data for broader groups, to enable useful and understandable information to be obtained." From this we concluded that the ICD is based on a hierarchical system of classes. The relations between these classes are proper subset relations in the sense of set theory. Thus, we decided to represent the relations of the ICD using OWL and its `subClassOf` relations. Table 1 lists some general metrics for the generated ontology. It also lists differences between the German and the English ICD-10 versions in terms of number of classes. The majority of all ICD-10 categories, i. e., about 60%, share the same ICD-10 code (for details see Section 3.2) and thus could be mapped using this as an identifier. However, there were some discrepancies between the WHO and German versions. One reason were different modeling granularities producing more categories in some branches. These differences are discussed subsequently.

We decided to split our OWL model into two components similar to the approach of Noy and Rubin in [13] for translating the Foundational Model of Anatomy ontology to OWL. The OWL-DL component allows to perform DL-reasoning using standard OWL-DL reasoners like Pellet [16]. Information from the ICD which requires modeling in OWL-Full is still available in the OWL-Full component. To use the complete model, the OWL-Full component can be loaded. This variant imports the OWL-DL model.

[5] `http://protege.stanford.edu`

	WHO ICD-10	*German ICD-10*
OWL classes	11,308	16,214
disjointness axioms (see Section 3.4)	13,094	27,899
excludes pointing to another category	5,150	4,417
excludes without a proper link to other categories	35	73

Table 1. Metrics for the generated ICD-10 ontology

3.8 Differences between WHO and DIMDI versions of ICD-10

In both terminologies we located classes that either only occur in the DIMDI or WHO version. These differences can be classified into two categories: (1) classes which only appear in one ICD-10 variant and have no particular counterpart in the other; and (2) classes which have a slightly different identification, but can be merged manually.

The first category is the most extensive. We identified 1,145 classes which are exclusively part of the WHO version and 5,707 classes exclusively part of the DIMDI version. It is important to note that we are only regarding the classes of the actual ICD-10 entries and not the classes for constructs like modifiers; these will be discussed later. Furthermore, there are blocks which differ in both versions. This means that one version specifies some parts of its terminology with more granularity than the other or that some concepts were simply left out. For example, the WHO subdivides the block "V01-X59 Accidents" into 27 sub-blocks using two hierarchy levels. In contrast, the DIMDI version does not make any further subdivisions here. Moreover, the DIMDI describes the block "U60-U61 Stadieneinteilung der HIV-Infektion" ("Staging of HIV-Infection"), which is not present in the WHO version at all.

We derived the second category during a manual examination of all differences. Hereby, we identified five blocks, which vary in their ICD-10 code. Table 2 lists them and opposes the WHO blocks with the details which the DIMDI contains. In all cases, there are differences in the range of the blocks. This is interesting, as even though a block can specify a broader range, it can be semantically more restricted. We will illustrate this by an example. The WHO version has the block "U80-U89 Bacterial agents resistant to antibiotics." The German version "U80-U85 Infektionserreger mit Resistenzen gegen bestimmte Antibiotika oder Chemotherapeutika" has almost the same range ("U80-U89" vs. "U80-85"). It could be assumed that the block with the smaller range is also more specific. But in this example the opposite is true since the German term "Infektionserreger" ("infectious agent") includes diseases caused by both bacteria and viruses. In contrast, the WHO block only covers diseases caused by bacteria. The generated OWL ontology is used in the research project MEDICO to allow for semantic annotation and retrieval across medical documents and images annotated with ICD-10 terms both in English and German.

WHO ICD-10	DIMDI ICD-10
D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D90 Krankheiten des Blutes und der blutbildenden Organe sowie bestimmte Störungen mit Beteiligung des Immunsystems
D80-D89 Certain disorders involving the immune mechanism	D80-D90 Bestimmte Störungen mit Beteiligung des Immunsystems
V01-Y98 External causes of morbidity and mortality	V01-Y84 Äußere Ursachen von Morbidität und Mortalität
O80-O84 Delivery	O80-O82 Entbindung
U80-U89 Bacterial agents resistant to antibiotics	U80-U85 Infektionserreger mit Resistenzen gegen bestimmte Antibiotika oder Chemotherapeutika

Table 2. ICD-10 blocks, which can be merged together, although they exclusively appear in the DIMDI or WHO version

4 Conclusion and Future Work

We presented an OWL model of ICD-10. Our model captures hierarchical information of ICD-10 as well as comprehensive class labels for both English and German. Peculiarities such as "Exclusions" statements, which make statements about the disjointness of certain ICD-10 categories, are provided in a separate OWL-Full component. This component allows us to use hierarchical knowledge and class labels with existing OWL-DL reasoners. Our automatic generation and merging method also revealed systematic differences between the German DIMDI and the English WHO version. The goal of this approach was to reduce *Semantic Heterogeneity* in healthcare by integrating two semi-formal terminologies. The current ontology represents the main and most important parts of both ICD-10 variants.

We plan to combine the results with additional conceptualizations, so that ontology networks can be created and interconnected. For example, Cardillo et al. describe an approach to map the ICD with the International Classification of Primary Care Version 2 (ICPC-2) [3]. This would foster the creation of expressive medical knowledge bases which would improve the retrieval and reuse of knowledge by reducing ambiguity. The XML file states additional information to be integrated in the future. For example, the dagger and asterisks are distinguished depending on the treatment. Different asterisk and dagger terms have to be used to formulate diagnoses for clinical treatments if diagnosis reports are created for accounting documents. During our examination of the generated OWL files and the respective ICD-10 websites, we recognized that certain verbal structures occur very often, e. g., "Injury of X," where X is an anatomical designation, like arm or leg. Moreover, the manuals describe predefined terms which are used very often, for example, the two acronyms NOS, meaning "not otherwise specified,"

and NEC, standing for "not elsewhere classified." Linguistic analysis can exploit this information to extract relations, e. g., to concepts represented in anatomical ontologies. This approach has already been applied successfully to other corpora within the MEDICO project [20,19] and we plan to extend it to the ICD-10 as well. Another possibility to enhance our OWL models is to include more languages. This would generate an international representation of the ICD-10. For example, it would be possible to include the French version[6] of the ICD-10. The website is structured like the German DIMDI version. Consequently, either our HTML crawler could be used to extract the necessary information or, if a XML encoding is available, this could be parsed directly. However, one can see at first sight that the French version also differs from the other two version, e. g., it covers only 21 chapters and omits the chapter about "Codes for special purposes." Therefore, an examination of the differences would be necessary.

Acknowledgements This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016. The responsibility for this publication lies with the authors.

References

1. Bechhofer, S., Volz, R., Lord, P.W.: Cooking the semantic web with the OWL API. In: International Semantic Web Conference. pp. 659–675 (2003)
2. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research 32 (Database Issue), D267–D270 (2004), http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D267
3. Cardillo, E., Eccher, C., Serafini, L., Tamilin, A.: Logical analysis of mappings between medical classification systems. In: Dochev, D., Pistore, M., Traverso, P. (eds.) AIMSA. Lecture Notes in Computer Science, vol. 5253, pp. 311–321. Springer (2008), http://dblp.uni-trier.de/db/conf/aimsa/aimsa2008.html#CardilloEST08
4. Ceusters, W., Smith, B., Kumar, A., Dhaen, C.: Mistakes in medical ontologies: where do they come from and how can they be detected? Stud Health Technol Inform 102, 145–163 (2004), http://view.ncbi.nlm.nih.gov/pubmed/15853269
5. Cote, R., Rothwell, D., Palotay, J., Beckett, R., Brochu, L.: The systematized nomenclature of human and veterinary medicine. Tech. rep., SNOMED International, Northfield, IL: College of American Pathologists (1993)
6. Deutsche Krankenhausgesellschaft: Deutsche kodierrichtlinien - allgemeine und spezielle kodierrichtlinien für die verschlüsselung von krankheiten und prozeduren. Tech. rep., Institut für das Entgeltsystem im Krankenhaus (InEK GmbH) (2009)
7. Fallside, D.C., Walmsley, P.: XML Schema Part 0: Primer Second Edition. W3C Recommendation (28 October 2004), http://www.w3.org/TR/xmlschema-0/
8. Hoelzer, S., Schweiger, R.K., Liu, R., Rudolf, D., Rieger, J., Dudeck, J.: Xml representation of hierarchical classification systems: from conceptual models to real applications. Proc AMIA Symp pp. 330–4 (2002)

[6] http://www.dimdi.de/dynamic/en/klassi/diagnosen/icd10/htmlfren/fr-icd.htm

9. Hustadt, U., Motik, B., Sattler, U.: Reducing SHIQ- Description Logic to Disjunctive Datalog Programs. In: Proc. of the 9th International Conference on Knowledge Representation and Reasoning (KR2004). pp. 152–162. Whistler, Canada (June 2004)
10. Lenz, R.: Information management in distributed healthcare networks. Data Management In a Connected World 3551, 315–334 (2005)
11. Möller, M., Mukherjee, S.: Context-Driven Ontological Annotations in DICOM Images – Towards Semantic PACS. In: Azevedo, L., Londral, A.R. (eds.) Proceedings of the Second International Conference on Health Informatics, HEALTHINF. pp. 294–299. INSTICC Press (2009)
12. Möller, M., Sintek, M., Buitelaar, P., Mukherjee, S., Zhou, X.S., Freund, J.: Medical image understanding through the integration of cross-modal object recognition with formal domain knowledge. In: Proc. of HEALTHINF 2008. vol. 1, pp. 134–141. Funchal, Madeira, Portugal (2008)
13. Noy, N.F., Rubin, D.L.: Translating the Foundational Model of Anatomy into OWL. Web Semantics: Science, Services and Agents on the World Wide Web 6(2), 133–136 (2008)
14. Schulz, S., Suntisrivaraporn, B., Baader, F.: SNOMED CT's problem list: Ontologists' and logicians' therapy suggestions. In: Proc. of The Medinfo 2007 Congress. Studies in Health Technology and Informatics (SHTI-series), IOS Press (2007), http://lat.inf.tu-dresden.de/research/papers/2007/SchSunBaa-Medinfo-07.pdf
15. Schulz, S., Suntisrivaraporn, B., Baader, F., Boeker, M.: SNOMED reaching its adolescence: Ontologists' and logicians' health check. International Journal of Medical Informatics 78(Supplement 1), S86–S94 (2009)
16. Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 51–53 (June 2007), http://dx.doi.org/10.1016/j.websem.2007.03.004
17. Sonntag, D.: Ontologies and Adaptivity in Dialogue for Question Answering. AKA and IOS Press, Heidelberg (2010)
18. Sonntag, D., Wennerberg, P., Buitelaar, P., Zillner, S.: Pillars of ontology treatment in the medical domain. Journal of Cases on Information Technology (JCIT) 11(4), 47–73 (2009)
19. Wennerberg, P.: Aligning medical domain ontologies for clinical query extraction. In: Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL '09). pp. 79–87. Association for Computational Linguistics, Morristown, NJ, USA (2009)
20. Wennerberg, P., Möller, M., Zillner, S.: A linguistic approach to aligning representations of human anatomy and radiology. In: Proc. of the International Conference on Biomedical Ontologies (ICBO 2009) (July 2009), http://precedings.nature.com/documents/3521/version/2
21. WHO: International statistical classification of diseases and related health problems. Tech. rep., World Health Organization (2004), http://www.who.int/classifications