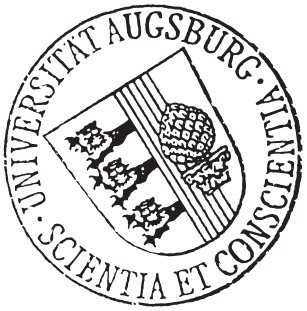
Integrated Representation of Clinical Data and Medical Knowledge
An Ontology-based Approach for the Radiology Domain

Heiner Oberkampf

Dissertation
for the degree of
Doctor of Natural Sciences (Dr. rer. nat.)



University of Augsburg
Department of Computer Science
Software Methodologies for Distributed Systems

Software Methodologies
for distributed systems

October 2015

Integrated Representation of Clinical Data and Medical Knowledge
An Ontology-based Approach

Supervisor: **Prof. Dr. Bernhard L. Bauer**, Department of Computer Science, University of Augsburg, Germany

Prof. Dr. Elisabeth André, Department of Computer Science, University of Augsburg, Germany

Advisor: **Prof. Dr. Sonja Zillner**, Siemens AG Corporate Technology, Munich, Germany

Date of defence of dissertation (oral examination): March 17th, 2016

Contents

Part I.

INTRODUCTION, BASICS, RELATED WORK

1

Introduction

Within the healthcare sector there is a tremendous increase of available data. Firstly, all data about the patient which is documented in some electronic health record. This encompasses all data about the health status of a particular patient such as finding descriptions and reports or images, demographic information, medications etc. This data is referred to as *clinical data*. Secondly, there is an increase of scientifically proved *medical knowledge*, i.e. knowledge about the structure and functioning of the human organism in general. For example knowledge about diseases and their manifestations in symptoms and clinical findings. The combination of both data resources, i.e. the clinical data with medical knowledge, bears the potential better and more effective decision making by clinicians. The assumption here is simply that the more data one has about some patient, the better is the understanding of the patient's health status and thus the better this patient can be treated. In order to understand why this potential is not already unlocked, a closer look at clinical data and medical knowledge is needed.

Clinical data is from various clinical domains such as radiology, microbiology, genetics, pathology or care documentation and of different formats (e.g. free text reports or images). Already for long time the most fundamental data, i.e. diagnosis has been stored in structured form. For instance, the classification of diseases reaches back to the 18th century and resulted in the International Classification of Diseases (ICD). The initial motivation for this classification system was analysis of mortality in populations through a common classification and thus understanding of diseases. Today, classification and coding systems are used to consistently refer to certain clinical data and thus provide the basis for data interoperability. Using a common coding system to store lab-values, it becomes possible to easily compare results from consecutive examinations even when they were not performed by the same healthcare provider. This fact is underlined by a recent Code of Federal Regulations published by the US government in the context of the meaningful use initiative specifying "standard vocabularies and coding systems to represent electronic health information" such as lab-values, problems, procedures or medications [14b]. In current clinical information systems, classification and coding systems are to be used for this kind of data. The overwhelming amount of clinical data however is still unstructured. That is, the challenge here is to get structured representations for more granular clinical data in order to unlock its full potential

for higher quality and more efficiency in healthcare.

Additionally to the increase in clinical data, *medical knowledge* is growing similarly. Most of it is contained in literature such as scientific publications about clinical studies or reference books. However there is also basic medical knowledge formalized in so called ontologies which define entities and relations between them. More generally, an ontology is *"an explicit specification of a conceptualization"* [Gru93]. Most ontologies within the biomedical domain cover exactly one particular domain. For instance, the Foundational Model of Anatomy (FMA) covers the human anatomy and defines more than 80,000 entities which are connected by 185 different relations. Similar to classification systems mentioned above, ontologies provide a controlled terminology that can be used as a standard reference for coding clinical data. But additionally the relations between entities can be specified and thus different kinds of medical knowledge can be represented in ontologies. At the time of writing, the BioPortal [14h] contains more than 400 different biomedical ontologies.

The increase in available clinical data and availability of medical knowledge provides the basis for various applications which aim to increase the quality and safety of healthcare. Personalized treatment, e.g. the optimal medication with fewest side effects, less unnecessary examinations and mistreatment can be realized based on the available data. Today however enhanced quality of treatment on the one hand and lower cost on the other hand are in conflict: in the current situation more data means more efforts for the clinicians and thus higher costs. This is mostly because of three problems: Firstly, clinical data is stored case-centric and distributed across different systems. It is not longitudinal integrated. Secondly, only small amount of the data is structured while high percentage of clinically relevant data is unstructured. Thirdly, existing data is not sufficiently aligned to medical knowledge and thus not on the appropriate level of detail for decision support systems. As a result of these problems most of the available data is simply not used in their full strength and no personalized treatment is applied. Today healthcare providers do not achieve improvements in quality *and* efficiency. A detailed description of the mentioned problems and their underlying challenges is given in section 1.1.

The thesis has three objectives targeting parts of the aforementioned problems. The first objective is the creation of a semantic model for clinical information which is based on established upper ontologies[1]. In particular, it is focused on the representation of clinical findings from radiology examinations. The second objective is to extract structured representations from radiology reports using formalized medical knowledge. The extracted information is stored in the semantic model. The third objective is to enrich the data using inference techniques and formalized medical knowledge to allow realization of different views on the data,

[1] Upper ontologies define the basic entities and their relations without being specific to one domain.

needed by clinicians for more efficient decision making. In particular, radiology findings extracted from unstructured reports are classified as normal/abnormal and finding descriptions are linked to disease information. A longitudinal view of radiology finding data is realized through a prototype implementation of a report viewer which relies on medical knowledge about the human anatomy, the semantic representation of finding descriptions and the meta-data of the original radiology reports.

The remaining part of this chapter is organized as follows: In section 1.1 the main problems and challenges are described, which provides the motivation for this work. Section 1.2 describes the main objectives of this thesis in detail, the approaches taken as well as corresponding contributions. Finally, the structure of the thesis is outlined in section 1.4 before a brief overview of publications is given in section 1.3.

1.1. Problems and Challenges

Clinical data is diverse, i.e. from different medical domains and different formats. Further, large parts are unstructured, e.g., free text reports and images. Only a longitudinally integrated view allows effective usage of this wealth of available data in clinical decisions. This integration relies mainly on the use of a common vocabulary and data structure. Initially computer systems were used in the healthcare domain to codify clinical data for statistics such as mortality analysis and administrative purpose such as billing. In this context coding systems for diseases and clinical procedures and examinations are widely used. As described in the introduction, today coding systems are also used to store basic patient information such as lab-values and medications. Standard coding systems such as those defined in the Code of Federal Regulations the US government [14b] provide the basis for interoperability in healthcare information systems. In current systems however they are only used for very high level information and large parts of the data remain unstructured.

Coding systems and ontologies often cover one particular domain. For instance, the International Classification of Diseases (ICD) covers diseases, Logical Observation Identifiers Names and Codes (LOINC) covers medical laboratory observations, the Foundational Model of Anatomy (FMA) covers the human anatomy and Gene Ontology (GO) genetics. The domains of different ontologies might also overlap: For instance Radiology Lexicon (RadLex) covers the radiology domain and contains anatomical entities as well as diseases and other entities relevant within the radiology domain. An overview of biomedical ontologies is given in the next chapter in Section2.2. The most comprehensive clinical healthcare terminology is Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) containing more than 311,000 entities [Int14]. But even SNOMED CT cannot contain codes for any clinical information artefact: As explained by [Cor09], in SNOMED CT for each disease *"one of five episodicities, one of six severities and one of seven courses can be chosen, which would result in over 300 combinations for a single disease"* - and there are several thousand diseases. Similarly, descriptions of clinical finding can require to use even more combinations. No terminology or ontology can define entities for all finding descriptions, diseases etc. Instead, the role of the ontology is to provide the basic entities allow their post-coordination, i.e. the multiple entities to express information. Or, as argued by [DSM02] *"modern terminologies such as SNOMED CT are compositional, allowing concept expressions to be pre-coordinated within the terminology or post-coordinated within the medical record"*. Here, the role of the medical record is to define the data structure in which entities from the coding system are combined. That is, terminologies, coding systems and ontologies are not enough. Additionally, models, which define the data structure are needed. This data structure is commonly referred to as the *information model*. To capture the content of clinical information in general a combination of an information model with terminologies is needed [Ben10]. Sometimes these information models are

also referred to as *structural* information models, since they define the schema according to which the terminology is used. While the ontology or vocabulary is the conceptualization of an understanding of the world, *"the function of the information models is to make it possible to specify and test the validity of data structures so that they can be exchanged and re-used in different information systems"* [RQM06]. As motivated in [Ben10] (chapter 7) the information model can be seen as the grammar that defines the usage of the vocabulary (such as entities from SNOMED CT). There are however situations, where one can express information more than one way even when using only one common terminology. Consider the finding "enlarged lymph node". It could be represented by using the either the entity `lymphadenopathy` or the combination of two entities `lymph node` and `enlarged`. Thus, the data structure has to ensure a consistent representation through logical bindings to the terminologies. Further, classes and relations need to be well defined in order to ensure correct usage.

Problem 1: Current Information Models are Case-centric and Lack Clear Semantic Definitions

As explained above the role of the information model is to provide a clear structure for the usage of entities from existing reference terminologies. Thus, a semantically well founded information model is needed. There are two information models which are well known: the HL7 Reference Information Model v3 (HL7 RIM) and the OpenEHR. Both models however lack clear semantic definitions and bindings to terminologies. As described in [SC06] definitions in HL7 Reference Information Model are even incoherent, which leads to ambiguous and wrong data representations. OpenEHR, on the other hand, is more focused on the structured entry of clinical data.

- **Challenge 1 - Integrated Semantic Model for Clinical Information:** For integrated clinical data it is essential to have a semantic information model with precise definitions of its entities and relations. The model has to cover meta data of the clinical data to enable efficient retrieval of information. The information model has to facilitate the usage of existing ontologies and vocabularies. In particular structured representations for detailed descriptions for clinical findings are needed which facilitate their reuse in different application scenarios.

Problem 2: Most of the Clinically Relevant Data is Unstructured

Having a semantic model for clinical information and a methodology to represent information in it by using terminologies, data still needs to be mapped to the

model. The problem however is that large amounts (about 80%) of clinical data is unstructured [Ter13]. Even though some data can be entered by clinicians in structured form such as medication, diagnosis, procedures etc., structured reporting in general is not widely applied in practice. Especially finding descriptions and observations are commonly documented in free text clinical reports or in images. In subsequent decision processes however, clinicians do not have the time to review all potentially valuable images and reports. To efficiently include information from unstructured data in clinical decisions one needs to extract structured representations. In the context of Natural Language Processing (NLP) information extraction from text involves the following five steps [Cun06]: Named Entity Recognition (NER), co-reference resolution, template element construction, template relation construction and template scenario production. The first two steps extract entities from text. Biomedical ontologys are commonly used within NER, which is also referred to as semantic annotation. The ontology then allows the semantic understanding of the extracted entities. For instance from a sentence "lymph nodes in the mediastinum with a size up to 1.6 cm" the entities lymph node, mediastinum, mediastinal lymph node as well as the measurement 1.6 cm is extracted. The problem however is that a *set* of extracted entities does not represent the asserted content precisely enough to be useful in further applications. Thus, the remaining steps resolve the *relations* between the extracted entities. For instance that the measurement describes the size of the mediastinal lymph node (see figure 1.1).

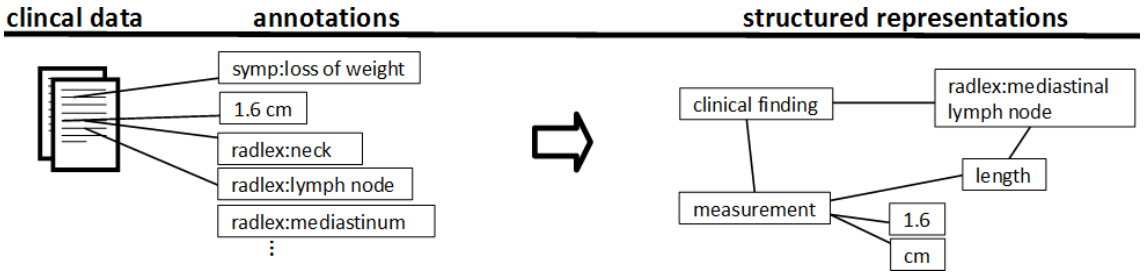


Figure 1.1.: To obtain structured representations from clinical data, relations between extracted entities need to be resolved.

The resolution of relations between entities from clinical reports is challenging since the sentences are often very long, they lack proper grammar and much information is contained only implicitly. This is because the intended audience of these reports are clinicians who have a good understanding of the domain. Consequently formalized *"domain knowledge is an essential resource in Information Extraction (IE) from free text"* [Ang10].

- **Challenge 2 - Integration of Semantic Annotations:** Since there are different annotators available, annotations need to be integrated into a common structure. This includes filtering and normalization of annotations.

- **Challenge 3 - Extract Structured Representations:** Based on annotations,

structured representations of the content of the annotated data need to be extracted. The challenge here is that annotations might be incomplete or false. Extraction of correct representations relies on reasoning and the application of medical knowledge.

Problem 3: Data is not linked to Clinical Knowledge, not on the Appropriate Level of Detail for Usage in Clinical Decisions and not longitudinal integrated

If the content from reports is represented in structured form, one has a description of the patients health status. The problem however is that only a small part of the data is relevant in clinical decision making. A clinician is interested only in certain parts of the data to answer specific questions. That is findings need to be classified and linked to other resources to allow different views necessary in clinical decision making. In diagnosis, a clinician is interested in what is currently abnormal or pathological. In the case of lab-values this information is available: for instance it is automatically asserted whether results from blood counts are within or outside of the normal range. For radiology findings such an interpretation is missing. In a radiology report, the size of the spleen might be simply specified as "spleen 7 x 13 x 12 cm" without any interpretation. To automatically classify normality status of anatomical entities, data needs to be linked and combined with medical knowledge. For instance to infer the spleen measurement above describes an enlarged and thus abnormal spleen. In treatment evaluation the clinician might be interested only in the *change* of findings in comparison to a previous examination. That is, finding types such as new, changed/unchanged, increase, decrease needs to be inferred. Since finding data is currently not longitudinally integrated, automatic inference of the change of the health status over time is not possible.

- **Challenge 4 - Integration of Medical Knowledge with Clinical Data:** In the current situation medical knowledge and clinical data are not integrated. The main challenge here is to formalize and represent medical knowledge in a similar way as the clinical data itself, i.e. using entities from existing medical ontologies in a similar semantic data structure.

- **Challenge 5 - Reasoning about Clinical Data Using Medical Knowledge:** Since a clinician is interested only in certain parts of the data to answer specific questions, clinical findings need to be classified using medical knowledge to allow corresponding selections. These classifications have to be patient-specific. Further, medical knowledge has to be used to provide different views on the data: for instance access finding information over time for comparison or interpretation of findings in the context of diseases.

- **Challenge 6 - Longitudinal Integration of Findings from Consecutive Examinations:** Since direct references to the related finding from previous examinations are not always made, longitudinal integration of findings relies mainly on the usage of knowledge from medical ontologies to identify similar findings. The challenge here is that one needs to link findings which are described in different levels of detail. Further, it is essential that the representation of the findings allow direct comparison so that change (what is new, getting better/worse) can be automatically inferred.

1.2. Objectives and Approach

Data representation in the healthcare domain is an enormous problem with several deep challenges. The overall objective of this thesis is to demonstrate how semantic technologies can be used to resolve specific parts of the problems described in the previous section. That is, along clinical use-cases, semantic technologies are used for data representation, integration of unstructured data in order to enable more efficient clinical decision making. Thus, the main objective of the thesis is the creation of a semantic information model which can be used in combination with terminologies to express clinical data, focussing on radiology findings. Further the model is used to express medical knowledge that is used later to enrich corresponding finding descriptions. Since most of the clinically relevant information is currently contained in unstructured data such as images or free text reports, structured representations need to be extracted from these sources. The structured representation in a semantic model is the first step towards enhanced data access. To include data in clinical decisions, it is also necessary to provide the clinician only the data needed in certain contexts. This requires classification of finding data using medical knowledge.

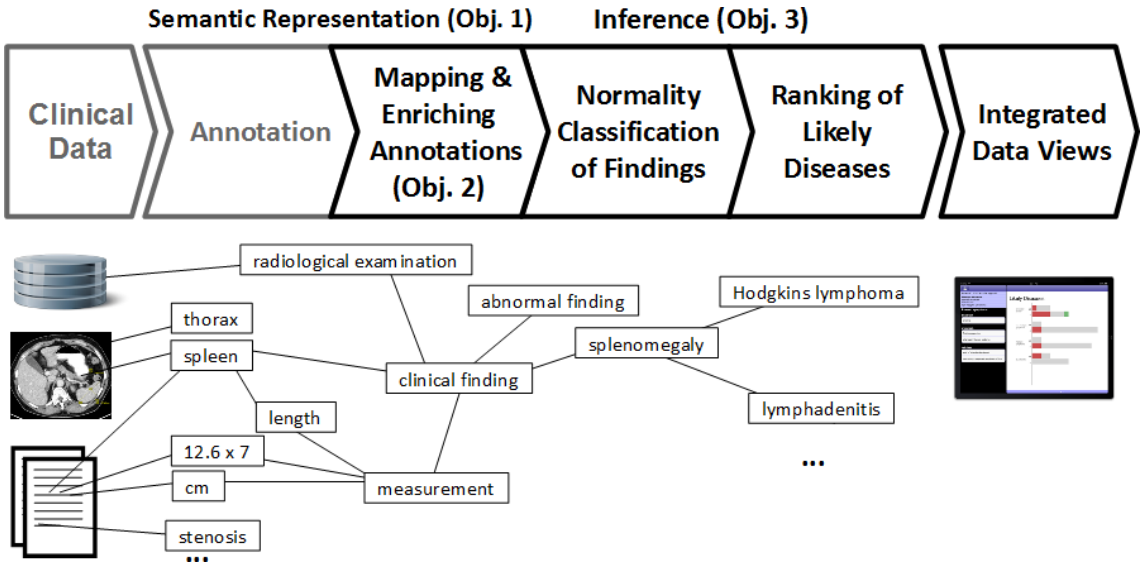


Figure 1.2.: Overview of the three objectives of the thesis.

Objective 1: Creation of a Semantic Model for Clinical Information

The main objective of this thesis is the creation of a semantic Model for Clinical Information which is used in combination with reference terminologies. The model has to cover basic patient data, meta data about provided examinations and

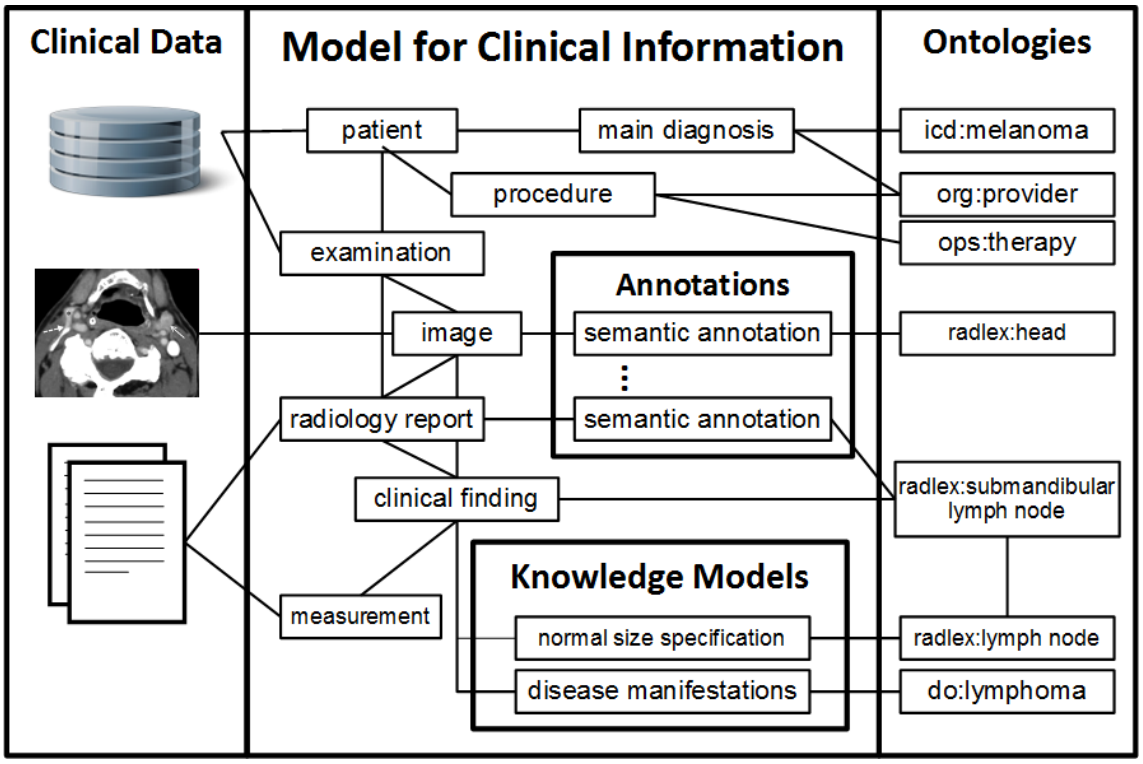


Figure 1.3.: The Model for Clinical Information represents clinical data by using entities from reference ontologies. Semantic annotations of images and reports as well as medical knowledge on normal size specifications and disease manifestations are represented in MCI.

diagnosis as well as a general structure for representation of clinical findings and their integration over consecutive examinations. Since clinical findings are central information objects, it is concentrated on them, focusing on the representation of findings from radiology examinations. Further, the model has to facilitate the representation of medical knowledge that is necessary to enrich finding descriptions automatically.

Approach: The model has to be based on a well founded upper level ontology and other existing models should be reused whenever possible. Further, established semantic modelling approaches are followed and existing reference terminologies are reused. The coverage of the model is defined by the data of the use cases and discussions with clinicians in which important user needs are analysed. The model has to facilitate the linkage to formalized medical knowledge, so that additional information about findings can be automatically inferred. Technically, Resource Description Framework (RDF) and the Web Ontology Language (OWL) are used to express the model and corresponding standard reasoning techniques are employed. Further, rules are implemented by SPARQL Protocol and RDF Query Language (SPARQL) update queries.

Contribution: The semantic Model for Clinical Information (MCI) consists of the following components:

- Foundation of the model based on upper- and mid-level ontologies of the Open Biological and Biomedical Ontologies (OBO) library. Selected ontologies are integrated and customised. The usage of upper ontologies provides a solid, semantically well defined basis for the model.

- Additionally to the reused upper ontologies, classes and properties required for granular representation of the clinical data of the use cases are defined. In particular, a structure for representation of radiology findings and corresponding image measurements is presented.

- The model defines classes and properties to express medical knowledge: Firstly, a structure for normal size specifications of anatomical entities and secondly a model for the interrelations of diseases, their manifestation in symptoms and clinical findings as well as corresponding examinations. The created knowledge models are linked to existing reference ontologies and were discussed with clinical experts to ensure their quality.

- The representation of findings allows application of standard reasoning techniques to automatically classify findings. This allows more efficient retrieval of findings in clinical decision making.

- Definition of bindings to terminologies: Bindings to reference ontologies are created, so that validation of their correct usage is possible.

Objective 2: Mapping Data to MCI and Enriching Semantic Annotations

Structured and extract facts from unstructured data are mapped to the semantic Model for Clinical Information. As explained in section 1.1 this requires two main steps: extraction of entities and extraction of relations between these entities. The first step is not an objective of this thesis. It is accomplished using available NLP technologies. That is, starting point for this work is a set of extracted entities detected in text – so called *annotations*. These annotations are mainly from an ontology (i.e. *semantic* annotations), but information artefacts such as sentence boundaries, date values, measurements etc. are also provided. These annotations help to understand what the data is about. However, they do not represent the *content* of the data sufficiently since the relations between these annotations are missing. The extraction of these relations is the main focus of this objective. In particular the extraction of measurement information (i.e. relations between measurements and anatomical entities) from radiology reports.

Approach: Mapping of structure and coded data to MCI which involves reading or querying the original data and transformation to patterns of MCI. Additionally unstructured data is annotated by using existing tools and then mapped to MCI. For resolution of relations between annotations of unstructured texts formalized medical background knowledge is used. Firstly, medical knowledge from existing biomedical ontologies is used. Secondly, a knowledge model about normal specifications of anatomical entities (see first objective) is used. This knowledge-based approach is taken since clinical reports lack grammatical structure and tend to be very long. Further, report meta-data is mapped to MCI and define their representation within the triple store.

Contribution:

- Existing clinical data from the relational database i2b2 and report data is mapped to the Model for Clinical Information. Further, annotations of unstructured data are mapped to a common schema. In particular, image annotations and text annotations of different schema are aligned in MCI.

- Based on annotations of unstructured radiology reports relations between measurements and associated anatomical entities are resolved. The knowledge model about normal size specifications and knowledge of the RadLex ontology are used to accomplish the resolution for radiology image findings.

Objective 3: Enrich Clinical Data using Medical Knowledge

To enhance quality and efficiency of healthcare, data on clinical findings needs to be better integrated into the process of decision making by clinicians. Besides structured representations of findings, this requires that findings are linked to medical knowledge and automatically classified, so that clinicians can selectively retrieve those which relevant for their task. In discussions with clinicians corresponding three main user needs were identified that guided the focus of this objective: Firstly, distinction between normal and abnormal findings. Secondly, transparency of the change of clinical findings between consecutive examinations. Thirdly, interpretation of finding information with respect to diseases.

Approach: Different knowledge resources are used to enrich clinical data that is stored in MCI: In particular, medical knowledge from existing biomedical ontologies and the knowledge models created in the context of the first objective. For the classification of findings as normal/abnormal the knowledge model about normal specification of anatomical entities is used. Additionally, meta-data of the reports is used for the longitudinal integration of finding data. Retrieval of

longitudinal data from different clinical perspectives such as anatomical site, type of findings or diseases is also realized using knowledge from existing ontologies.

Contribution:

- Measurement findings are enriched using formalized medical background knowledge in order to provide different views on the data, needed by clinicians for more efficient decision making. In particular, measurement findings from radiology reports are classified as normal or abnormal.

- Using the knowledge model of disease and their manifestations in clinical findings and symptoms, likely diseases of a particular patients based on finding information are inferred. The developed ranking algorithms takes as input the status of findings (present, absent, unknown), their type of relation to particular diseases (leading symptom or not) as well as the incidence of the disease. Thus, finding data is automatically interpreted with respect to diseases and help the clinician to make a diagnosis and plan further examinations.

- It is demonstrated that the structured representations of findings from the first objective are suitable to provide an integrated view on finding data from consecutive radiology examinations.

All contributions mentioned in this section are implemented in several demo applications and evaluated using different data sets.

1.3. Publications

In this section, the publications related to the thesis are briefly described. Instead of a chronological order, publications are described along the objectives presented in section 1.2, before further publications are described. An estimation of the personal contribution is given in brackets.

Main Publications

In "An OGMS-based Model for Clinical Information (MCI)" [Obe+13] (90%) the initial version of the integrated semantic model is presented. The reuse of upper level ontologies from the OBO library is motivated. Further, the functional scope of the model, how it references entities from existing terminologies and the most important classes and properties used for representation of basic patient data and important types of clinical findings are described.

To include information from unstructured clinical data one commonly used semantic annotations with entities from biomedical ontologies. However, successful semantic annotation highly depends on the vocabulary provided by the respective ontologies. Working with German texts brings the following challenge: Most of the existing ontologies have either no or only few German labels which leads to a very low annotation coverage of German reports. Since translations need expert review, a full translation of large ontologies is often not feasible. In "Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation" [Bre+14] (45%) it is proposed to use the corpus to be annotated to identify high occurrence terms and their translations to extend respective ontology concepts. Using this approach, the translation of a subset of ontology concepts is sufficient to significantly enhance annotation coverage. For evaluation, RadLex ontology concepts were automatically translated from English into German. In RadLex only about 25% of the entities have German labels. Even very basic entities such as `lesion` lack German translations! It is shown, that by adding translations for a rather small set of entities (here 433 from more than 40,000), which were identified by corpus analysis, it is possible to enhance the amount of annotated words from 27.36% to 42.65%. The result of this work is an extension to the official RadLex version which is used for better annotation coverage in later work.

A large percentage of relevant radiologic patient information is currently only available in unstructured formats such as free text reports. In particular, measurements are important since they are comparable and thus provide insight into the change of the health status over time, for example in response to some treatment. In radiology most of the measurements in reports describe the size of anatomical entities. Even though it is possible to extract measurements and

anatomical entities from text using standard information extraction techniques, it is difficult to extract the relation between the measurement and the corresponding anatomical entity. In "Knowledge-based Extraction of Measurement-Entity Relations from German Radiology Reports" [Obe+14] (90%, acceptance rate: 30%) a knowledge-based approach to extract information about size measurements using a model about typical size descriptions of anatomical entities in combination with hierarchical knowledge of RadLex is presented. The approach was evaluated on two data sets of German radiology reports reaching an F1-measure of 0.85 and 0.79 respectively.

In "Semantic Representation of Reported Measurements in Radiology" [Obe+15b] (80%), the usage of MCI for integrated representation of image findings and medical knowledge about the normal size of anatomical entities is demonstrated. An integrated view of radiology findings is provided by a prototype implementation of a ReportViewer. Further, RECIST (response evaluation criteria in solid tumours) guidelines are implemented by SPARQL queries on MCI.

Often clinicians are interested only in specific information. For instance in diagnosis a clinician would like to retrieve symptom information for certain disease. In "Interpreting Patient Data using Medical Background Knowledge" [Obe+12a] (90%) an the initial ontology containing lymphoma-related diseases and symptoms as well as their relations is presented. The manually created ontology is used to infer likely diseases of patients. In this way, symptom information can be understood in the context of likely diseases and help the clinician to make a diagnosis.

The algorithm underlying the ranking of likely diseases is described in "Towards a Ranking of Likely Diseases in Terms of Precision and Recall" [Obe+12b] (90%). Basically, the patient's symptom information is matched to typical disease symptomatology. The matching is based on the knowledge model of diseases and their manifestation in symptoms and clinical findings. The disease symptom model used in the disease ranking was created manually. In *"From Symptoms to Diseases – Creating the Missing Link"* a disease symptom graph is created by automatically integrating available knowledge from different ontologies [Obe+15a] (40%).

Further Publications

In the paper "Knowledge Engineering Requirements for Generic Diagnostic Systems" [Mue+12] (20%) a general view on diagnosis is presented by investigating typical diagnostic examples from the industrial and medical domain and describe the basic requirements for a generic diagnostic knowledge representation language (DKRL). DKRL is intended to facilitate the generic representation, handling, and interchange of diagnostic knowledge required for performing diagnostics without

regard to specific diagnostic approaches.

In healthcare big data faces several problems. In "Towards a Technology Roadmap for Big Data Applications in the Healthcare Domain" [Zil+14] (15%) it is argued that seamless access to the various healthcare data pools is only possible in a very constrained and limited manner. For enabling the seamless access several technical requirements, such as data digitalization, semantic annotation, data sharing, data privacy and security as well as data quality need to be addressed. A detailed analysis of these technical requirements for Big Data applications in healthcare is introduced and it is shown how the results of the analysis lead towards a technical roadmap.

The Linked Open Data (LOD) paradigm is becoming the de-facto standard for sharing and connecting pieces of data, information and knowledge on the Web. In "UIMA2LOD: Integrating UIMA Text Annotations into the Linked Open Data Cloud" [BOZ15] (15%) an approach for conceptual representation of textual annotations which distinguishes linguistic from semantic annotations is presented. It is shown how standard UIMA text annotations need to be adapted in order to obtain RDF graphs with proper conceptual representations and links to existing LOD datasets.

In contrast to software development, in the context of ontologies, modular reuse of existing components is not well formalized. In "Management of Variability in Modular Ontology Development" [Lan+13] (20%) and "Change and Version Management in Variability Models for Modular Ontologies" [Lan+14] (20%) it is evaluated how feature models could be applied to variability and version management of modular ontologies.

Patents

The following six US patents were created in the context of this thesis: two accepted patents titled *"Method and system for supporting a clinical diagnosis"* US13540952 and US13473134 and patents with pending status: *"A Method of Composing an Integrated Ontology"* US14029270, *"A method for extending concept labels of an ontology"* US14038927, *"System and method for extracting measurement-entity relations"* US14250326 as well as *"Method and apparatus for generating a knowledge data model"* 201501544.

1.4. Outline

The thesis is structured in parts I-VI and the chapters are organized as follows:

Part I: Introduction, Basics, Related Work

Chapter 1 – Introduction is this chapter which presented an overview of the challenges and objectives of this thesis.

Chapter 2 – Basics provides a background of knowledge representation and current formats. Important biomedical ontologies and their role for annotation of unstructured data are described.

Chapter 3 – Related Work describes data representation in the healthcare domain and in particular the combination of information models with terminologies. An overview of existing (semantic) data models is given and existing work related to the application scenarios is discussed.

Part II: The Model for Clinical Information

Chapter 4 – Model Foundation gives an overview of the upper- and mid-level ontologies from the Open Biological and Biomedical Ontologies library which are reused by the Model for Clinical Information. For each ontology the general scope and intended role and reused entities and adaptations are described.

Chapter 5 – The Model for Clinical Information presents all classes and properties that were defined for the representation of clinical data are described. The motivation for them and their intended usage are explained along numerous examples.

Chapter 6 – Knowledge Models describes how medical knowledge that is necessary to enrich finding descriptions can be formally represented with MCI. Firstly, the model on normal size specifications is presented and secondly, the model about diseases and their manifestations in clinical findings.

Part III: Usage of MCI for Storage, Inference and Integrated Data Views

Chapter 7 – Architecture provides an overview of the different technical components and their interactions.

Chapter 8 – Inference describes different inference techniques and their role for enriched data representation. It is focused on inference techniques that employ the knowledge models described in chapter 6. The knowledge model on normal size specifications is used to classify findings and the diseases model is used for ranking likely diseases.

Chapter 9 – Integrated Data Views shows how integrated views on the data can be provided by using the semantic data structure of MCI in combination with reference ontologies.

Part IV: Case Studies

Chapter 10 – Data Sets describes the data resources employed for evaluation and prototype implementations.

Chapter 11 – Evaluation provides the evaluation of presented algorithms. In particular, the relation resolution for measurement entity relations, the normality classification and the ranking of likely diseases is evaluated.

Part V: Conclusion

Chapter 12 – Conclusion summarizes the contribution of the thesis and provides directions for future research challenges.

Part VI: Annex

Provides the detailed big picture of MCI and some background information on human anatomy and notations used in this thesis. Further, the bibliography, a glossary as well as lists for acronyms, figures and tables are given.

1.5. Acknowledgements

I thank all the people who supported me in writing this thesis. First of all I thank my supervisor Prof. Dr. Bernhard Bauer for giving me the opportunity of doing a Ph.D. His friendly guidance over the whole period of my Ph.D. time and giving me freedom to follow my own ideas and topics provided the basis for successful completion of this thesis. I would also like to thank him for the financial support for travels to workshops and conferences.

I thank my advisor at Siemens, Prof. Dr. Sonja Zillner for motivating me to work on a semantic model for patient data. In numerous meetings, I learned from her to write papers and present my ideas in a well structured form.

I also thank Prof. Dr. Elisabeth André, who also accepted to be an supervisor of my thesis.

Special thanks go to Matthias Hammon (M.D.), the clinical expert who did all the clinical evaluations of the developed algorithms and models. In discussions with him I got the medical understanding necessary for accomplishing this work.

I thank all OBI-developers, especially James A. Overton (Ph.D.) for explaining me the principles underlying the OBO ontologies and in particular the representation of measurements.

I thank Claudia Bretschneider who provided me with an annotator for German clinical reports.

I also kindly thank Bernard Gibaud (Ph.D.) for inviting me to the CrEDIBLE workshop to Nice which gave me the opportunity to discuss my work with other workshop participants.

I also thank Prof. Dr. Ulrike Sattler for giving me the opportunity to visit the University of Manchester and for the discussions with her on OWL representations.

I also thank my colleagues at Siemens and at the university of Augsburg with whom I always enjoyed to work.

Special thanks go to my parents Hanne and Jörg for their continuous support and advice. They encourage me by their positive attitude to life.

Last but not least I thank especially my wife Nadja for her understanding for my work and supporting me all the time.

Notes:

Example sentences from clinical reports are translated from German into English and partly shortened.

2

Basics

In this chapter the basics underlying the presented work are described. It is started with a brief overview of knowledge representation by defining different representation systems from simple vocabularies to very expressive ontologies. Then the underlying logic of ontologies (Description Logic (DL)) and an overview of the semantic web technology stack are presented. Further, the most important knowledge modeling initiatives within the biomedical domain namely the Unified Medical Language System (UMLS) and Open Biological and Biomedical Ontologies (OBO) are described and we also give insights in how ontologies are used for semantic annotation of heterogeneous clinical data such as images and free text reports.

2.1. Knowledge Representation

Knowledge representation is a main research field within the artificial intelligence research domain. Its core research challenge is *"to figure out how to represent knowledge in computers and to use it algorithmically to solve problems"* [Har+07]. The field of knowledge representation is very broad and includes satisfiability (SAT) solvers, DL, constraint programming, conceptual graphs, non-monotonic logics, answer set logic, techniques for believe revision, qualitative models of continuous systems, problem solvers as well as representation of uncertainty, temporal and spacial aspects, agent's knowledge and beliefs and multi-agent systems, situation and event calculus, query answering, the semantic web, automated planning and cognitive robotics [Har+07]. The following description is focused on knowledge representation within the context of DL and semantic web technologies.

2.1.1. Terminologies, Classification Systems and Ontologies

There exist many different formal systems to represent knowledge. They are distinguished by expressiveness, i.e. the type of information that can be represented using the corresponding systems. Figure 2.1 gives an overview of different knowledge representation systems with increasing expressiveness. The figure is adapted from the figure of [Wel07] regarding the usage of knowledge representation systems within the biomedical domain. The authors of [SS10] define the following four types of representation systems (mapping to the systems of figure 2.1 in brackets): lexico-semantic representations (thesauri), representation of types of entities (taxonomies), representation of background knowledge (ontology) and representation of individuals (not in the figure 2.1).

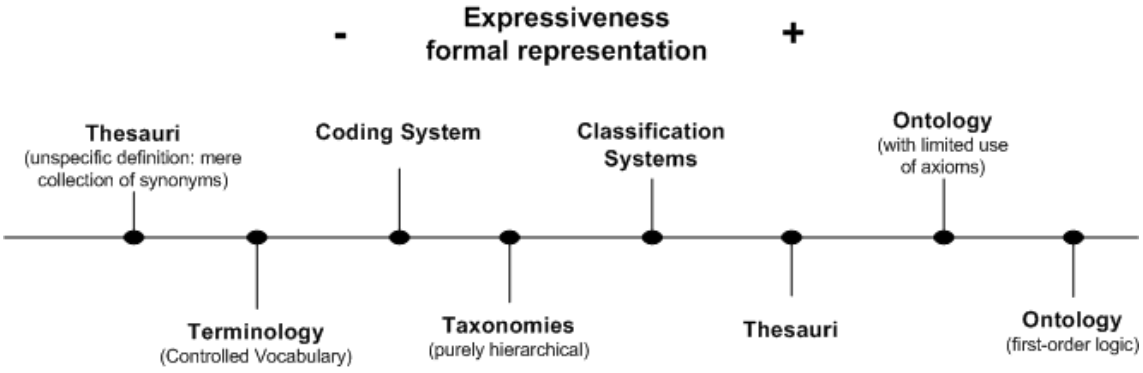


Figure 2.1.: Knowledge representation systems ordered by increasing expressiveness and formality. Modified from a figure by [Wel07] regarding their usage within the biomedical domain.

In the following the term *entity* is used to refer to a concrete class or instance defined in one ontology, terminology or data set. This abstraction is necessary

since knowledge representation differs across repositories and domains. The term *concept* is used to describe the semantic of an entity (i.e. the abstracted meaning) on a conceptual level without reference to any concrete implementation, such as some particular ontology. For example, the two entities `radlex:Hodgkin lymphoma` from RadLex and `do:Hodgkin's lymphoma` from Human Disease Ontology (DOID) represent the same concept *Hodgkin lymphoma*.

Terminologies and Classification Systems

The basic entity of terminologies is a *term*, which consists of a word or a compound of words. A *vocabulary* is a set of terms used in a specific context or domain. A *terminology* is a controlled vocabulary, i.e. a fixed set of terms which is issued by some institution of group. Terminologies often provide textual definitions or descriptions for the contained terms to explain their semantics. However, since their definitions are in free text from, the semantic of terms is understandable only for humans. Terminologies are used to establish a common understanding (semantic) of a set terms. A *thesaurus* is a terminology where the terms are additionally grouped by their semantic similarity. For instance, synonym information is contained in a thesaurus and bound to the corresponding entity. An example of a biomedical thesaurus is the National Cancer Institute Thesaurus (NCIT). A *taxonomy* is a hierarchically structured terminology. The hierarchical structure is represented by *is-a* relations between entities. A *classification system* is a hierarchically structured set of classes where sibling classes are disjoint. A very prominent example of a classification system is the ICD. Regarding a more formal representation the separation of lexical information (i.e. the term) and the identifier of an entity are often separated. Almost all terminologies, thesauri within the biomedical domain provide this separation. Systems, where the usage of the identifier as a standard reference to codify data is predominant, are sometimes also called *coding systems*. An example of a coding system is Logical Observation Identifiers Names and Codes (LOINC), which defines codes for measurement results of laboratory examinations such as blood counts etc.

Ontologies

Ontologies allow to express more knowledge about a certain domain, than pure classification or coding systems. A very popular definition of ontology comes from [Gru93]: *"An ontology is an explicit specification of a conceptualization."* Basically, an ontology defines a set of entities and their relations. Furthermore, logical axioms are used to formalize the semantic of entities. Thus, in contrast to terminologies the semantics of entities are *machine*-understandable. Furthermore, axioms can be used to check the consistency of the represented knowledge. As described in [Wel07] ontologies can be distinguished by their expressiveness. While some

ontologies make only limited use of axioms, others fully employ first-order logic constructs. Additionally, different types of ontologies are distinguished by their role for the representation of knowledge. A so called *upper ontology* is an ontology that provides the basic concepts for knowledge representation independent of any particular domain. Typical entities of an upper level ontology are *entity* or *process*. Furthermore, there are so called *domain* ontologies, *application* ontologies and *reference* ontologies. Even though their distinction is not sharp, they can be described roughly as follows: A *domain ontology* is an ontology which covers the concepts of a certain domain. The scope of the domain however can be very different: For example, the Infectious Disease Ontology (IDO) covers the domain of infectious diseases, while RadLex covers the domain of radiology including diseases, anatomical instances and other concepts with relevance to the *radiology* domain. An *application ontology* is an ontology with a limited term coverage and expressiveness in order to be efficiently used in certain applications. Its reuse for completely different applications is mostly not possible without adaptations. A *reference ontology* is often a large ontology which has the purpose to be reused in many different scenarios for the representation of different types of data. Their role is more to provide a structured reference terminology. The FMA or the DOID can also be seen as reference ontologies.

Obviously, not all existing knowledge representation systems can be unambiguously mapped to exactly one of these systems. Thus, there are several fuzzy terms further describing the usage or type of corresponding systems: e.g. reference terminology/ontology, application ontology, foundational ontology, domain ontology, upper level ontology etc. Many biomedical "ontologies", which can be found at the BioPortal [14h], are in fact rather terminologies, taxonomies, thesauri or classification systems than ontologies. However, as described in [SS10] *"what were formerly referred to as* terminology systems *or* vocabularies *are today often vaguely referred to the name* ontology*"*. Thus, for simplicity the distinctions might be omitted and it is referred to different representation systems as *reference ontologies* or simply *ontologies*.

Ontology Design Patterns

As in software engineering, there are also basic design patterns for the creation of ontologies. Some important of them are described here – for a comprehensive overview it is referred to [14k]. A common problem of classification systems is that they often contain classes representing the *unspecified* types in one hierarchy level. For instance the ICD-10 contains a class `Other disorder of facial node` as a subclass of `Facial nerve disorders`. In general this is considered as bad practice, since the semantic scope of that class is defined by the siblings, not by the class itself. Thus, change of the siblings changes the scope of the class. This is especially problematic with respect to the interpretation of legacy data. Another

important pattern (called *"Don't Repeat Yourself (DRY)"*) is to avoid repetition of class hierarchies. This is especially relevant when entities have many possible modifications. If for each modifier the class hierarchy is duplicated, this would lead to an extremely large ontology containing redundant information. For instance, the *intensity* of a disease (severe, moderate etc.) is such a modifier. According to the Don't Repeat Yourself (DRY) design pattern, an ontology should not define classes such as *severe lymphoma*, *moderate lymphoma* since this would repeat the information of the ontology. Instead, one should define that a disease *can have* a associated severity which is one of a defined set of entities denoting severities. The specific entity is then created by the user through *post-coordination* of entities. A combination of several entities within the original ontology is referred to as *pre-coordination*. Even though pre-coordination allows a more direct usage of the concepts, it leads to an tremendous increase of entities and thus does not scale. A good description of that problem is given by [Cor09], where it is shown that in SNOMED for *"many disease one of five episodicities, one of six severities and one of seven courses can be chosen"* which leads to 300 possible combinations.

2.1.2. Description Logic and Reasoning

According to [Baa+10], Description Logic (DL) originates from *semantic networks* (nodes and links) and *frame-based systems*. While semantic networks mainly express IS-A relationships which defines a generalization hierarchy, the semantics of other relations is not formally defined. In contrast to this, a *"characteristic feature of Description Logics is their ability to represent other kinds of relationships that can hold between concepts, beyond IS-A relationships"* [Baa+10]. For instance, DL classes can be defined by their relation to other classes by value restrictions, number restrictions or by equivalence to the intersection of other classes. The role of a DL *language* is to define a structure and the type of relations that can be used to describe classes logically by these kind of restrictions. As for other logic languages, DL *terms* or *expressions* are constructed based on atomic concepts and atomic roles to form more complex expressions. A basic application of DL reasoning is to infer subclass relationships (subsumption) and membership of individuals (instances) in classes, i.e. the finding the different *types* for instances. Further, reasoners are used to check the consistency of defined models, i.e. the absence of contradicting expressions and satisfiability of class expressions. There are different DL languages (such as DL lite and DL full) distinguished by their expressiveness and corresponding trade-off regarding reasoning complexity.

With respect to modeling languages of classical databases, a "distinguishing feature of DL is the open-world assumption" [Baa+10], which basically means that absence of a fact does not imply that the fact is *false*, but only that it is *unknown*. Thus, inferred conclusions are never falsified when new facts are added. This interpretation is suitable for representation of clinical findings. Simply because

a symptom such as fever is not mentioned in a report, does not mean that the patient does not have fever, but only that the status of fever is unknown.

2.1.3. The Semantic Web Technology Stack

This section describes the basic technologies and W3C standards of the semantic web. For a more detailed description of the mentioned technologies it is referred to the respective official W3C websites. The semantic web is an *extension* not a replacement of the current web. While the classical web is focused on layout and the possibility to link web pages, the semantic web is focused on *meaning* [BHL01] and data. By [BHL01] the semantic web has the following three components:

1. **Expressing Meaning:** usage of common vocabulary to represent data
2. **Knowledge Representation:** reasoning through rules
3. **Ontologies:** formal definition of relations among terms

Here ontologies are considered to be mainly taxonomies. In the following, entities are denoted in typewriter font. The IRI of an entity has two parts: the namespace and the local identifier. Within one document the namespace might be associated by a shorter prefix. For instance the namespace IRI `http://www.w3.org/2002/07/owl#` is commonly associated with the prefix `owl:` and one can write `owl:Class` instead of the full Internationalized Resource Identifier (IRI). According to common practice entities are denoted by using a prefixed notation, i.e. one writes `owl:Class` instead of the full IRI `http://www.w3.org/2002/07/owl#Class`. A detailed description on the notation is given in appendix A.3. Logical axioms are given in Manchester syntax [Hor+06; HP12].

As shown in figure 2.2 the semantic web is built on the usage of unicode and the usage of Uniform Resource Identifier (URI) as identifiers. A well known subset of URI is Uniform Resource Locator (URL), which are used to identify web pages. The semantic web however uses identifiers not only for web pages but more general for any resource of interest. Extensible Markup Language (XML) is a markup language to describe structured information. Furthermore, XML Schema is a language, which can be used to define validity of XML documents, and XML namespaces are used to qualify element names. Namespaces are identified by URIs. In the context of the semantic web, *entities* can be of the following types [MPP12]: class, object property, data property, annotation property, datatype and named individual. Each entity is specified by exactly one IRI. Additionally to entities there are literals (strings, integers etc). The nodes can be IRIs, literals or blank nodes, however literals must not be used as a subject in a triple. *Typed*

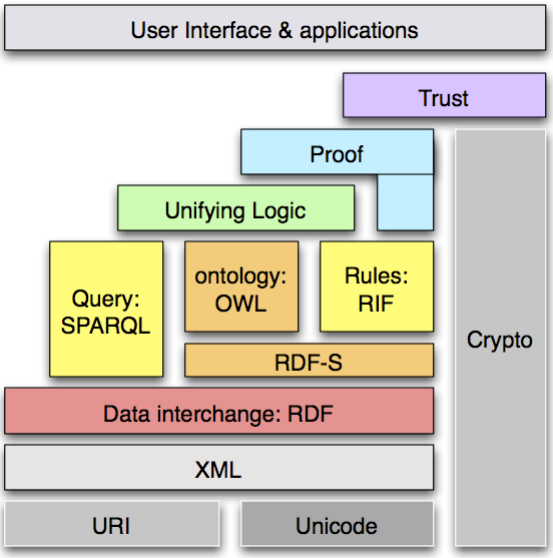


Figure 2.2.: The semantic web technology stack as presented in [06a].

literals have an assigned data type, while plain literals do not have an assigned data type.

Resource Description Framework

The Resource Description Framework (RDF) is a framework for representing information in the Web [CWL14]. It provides the basic vocabulary to describe graph-like data: nodes and relations between nodes are represented as triples (subject, predicate, object). A triple is also called *RDF statement* and a set of triples is called an *RDF graph*. RDF is the exchange format of the semantic web. For instance, the relation specifying that an entity is an instance of a class is defined in RDF by `rdf:type`. RDF built in vocabulary defines for example `Property` and `Statement` as well as properties to relate an instance of statement to respective subject, predicate and object resources. One serialization of RDF documents is RDF/XML. An RDF document serialized in XML is a XML document satisfying a certain syntax defined by the RDF XML Syntax [GS14]. There are serializations in other formats such as turtle [PC14], JSON [Spo14], N-Triples [CS14] etc. Since turtle is most human readable serialization, it is used throughout this document for example code.

Resource Description Framework Schema

Resource Description Framework Schema (RDFS) is used to define entities which are commonly used to describe taxonomies. It is based on RDF and defines

concepts such as `Class`, `Resource`, `DataType`, `Literal` etc. as well as properties that are used to specify the domain and range of properties. Further, properties to hierarchically order classes `rdfs:subClassOf` and properties `rdfs:subPropertyOf` are provided by RDFS.

Web Ontology Language

Then come more expressive languages – namely Web Ontology Language (OWL) and OWL2 [MPP12]. OWL builds upon RDFS and contains more concepts to logically define the semantics of classes and properties by logical definition (called *axioms*). Axioms are *"statements that say what is true in the domain"* [MPP12] and they are used to express logical relations between entities. For instance, the OWL vocabulary contains a class `AllDisjointClasses` to allow the specification of a set of disjoint classes. This is especially useful when one defines a classification system where an entity can be an instance of only one class out of several sibling classes. Properties in OWL are distinguished as owl:ObjectProperty (relation between individuals), owl:DatatypeProperty (relation between individual and literal) and owl:AnnotationProperty (relation between a class or individual and class, individual or literal). There are different expressiveness levels specified: OWL Lite, OWL DL and OWL full. The OWL 2 specification [MPP12] defines the following types of axioms [MPP12]:

- **class axioms:** SubClassOf, EquivalentClasses, DisjointClasses and DisjointUnion
- **property axioms:** SubObjectPropertyOf, ObjectPropertyDomain, ObjectProperyRange, SymmetricObjectProperty, InverseObjectProperty etc.
- data property axioms: SubDataPropertyOf, DataPropertyDomain, DataProperyRange, DisjointDataProperties, FunctionalDataProperty etc.
- **data type definitions:** additionally to the built in data types such as xsd:integer one can define new data types based on existing ones
- **keys:** axioms that are used to identify individuals by object and/or data type property assertions
- **facts:** assertions, which are axioms about individuals, e.g., the ClassAssertion is used to express the membership of an individual within some class

Many different types of properties are used by reasoners to verify consistency and inference. For instance, a functional property define that one entity can be related by that property to maximal one object. OWL also provides the relation

`owl:imports` that can be used to reuse other ontologies by including all its entities and relations.

There are many different *profiles* of OWL 2 (also called fragments) which define "restrictions on the structure of OWL ontologies" [Mot+12]. Within the biomedical domain, EL++ is an important and widely used lightweight profile, which was designed for usage of large scale ontologies. The EL++ profile is well suited for application contexts, because *"ontology consistency, class expression subsumption, and instance checking can be decided in polynomial time"*[Mot+12] .

Semantic Web Rule Language

The Semantic Web Rule Language (SWRL) is an extension to OWL abstract syntax which allows the definition of so called *rule axioms*. *"A rule axiom consists of a* antecedent *(body) and a* consequent *(head) each of which consists of a (possibly empty) set of atoms"* [Hor+04]. That is based on the existence of a set of triples of a graph pattern specified in the rule body new triples (atoms) are inferred. The atoms of rules can be class membership assertions, property assertion, an identity assertion (owl:sameAs) or a distinction assertion (owl:differentFrom) [Hor+04]. In general, SWRL extension makes OWL DL undecidable [Hor+04].

SPARQL

RDF data is represented in form of triples that can be stored in a file or in so called *triple stores*. Some triple stores offer the usage of *named graphs* [CWL14] for further grouping of triples, which is essentially a representation as *quadruples*. RDF specifies data in graph format and SPARQL [HS] is the query language for these graphs, i.e. one defines a subset of the data based on graph patterns which are specified in a SPARQL WHERE clause. In total, SPARQL defines four types of queries: SELECT queries are used to retrieve specific data based on the variables defined in the WHERE clause. DESCRIBE queries are used to retrieve triples related to a certain resource. The scope of the returned triples is determined by the query processor. The advantage of a DESCRIBE query is that the user does not have to know the schema of the RDF data set. An ASK query determines the existence of at least one data set matching the graph pattern defined in the WHERE clause and accordingly returns *true* or *false*. CONSTRUCT queries are used to create new triples based on the variable binding defined in the WHERE clause, i.e. an RDF graph is obtained. A good introduction to SPARQL is given by [DuC11].

2.2. Biomedical Ontologies

In the biomedical domain ontologies have a long tradition and many well designed, large and semantically rich ontologies exist. At the time of writing, the BioPortal [14h], an ontology repository for the biomedical domain, contains 421 ontologies, where 49 have more than 10,000 entities. Table 2.1 gives an overview of selected biomedical ontologies, terminologies, coding systems and classification systems related to this work. For a comprehensive list of ontologies it is referred to the websites of the NCBO BioPortal [14h].

Table 2.1.: Overview of biomedical ontologies and classification systems ordered by the number of contained classes.

Name	Classes	Domain	Family	Type
SNOMED CT	401,200	health	UMLS	coding system
MESH	245,871	health	UMLS	coding system
LOINC	162,368	lab results	UMLS	coding system
NCIT	107,882	cancer research	UMLS	thesaurus
FMA	83,283	anatomy	OBO	ref. ontology
OMIM	76,717	genetics	UMLS	ref. ontology
RadLex	45,754	radiology	-	ref. ontology
ICD	12,451	diseases	UMLS	classific. system
OPS	15,411	procedures	-	classific. system
HP	10,593	phenotypes	OBO	ref. ontology
DOID	8,744	diseases	OBO	ref. ontology
ATC	7,666	medication	-	classific. system
OBI	2,797	experiments	OBO	ontology
CMO	2,627	measurements	OBO	ontology
PATO	1,536	phenotype	OBO	ontology
BT	390	general	-	upper ontology
UO	331	units	OBO	ontology
TMO	301	clinical studies	-	ontology
IAO	173	information	OBO	ontology
BSPO	136	spatial	OBO	ontology
OMRSE	127	social entities	OBO	ontology
OGMS	125	medicine	OBO	upper ontology
CARO	50	anatomy	OBO	upper ontology
BFO-1.1	39	general	OBO	upper ontology
BFO-2	36	general	OBO	upper ontology
RO	0	general	OBO	upper ontology

In the biomedical domain there exist two big families of ontologies: The Unified Medical Language System (UMLS) and the Open Biological and Biomedical Ontologies (OBO). The Unified Medical Language System (UMLS) is a set of different multilingual classification and coding systems, where concepts are aligned by

the UMLS Metathesaurus. The OBO Library consists of orthogonal ontologies (without domain overlaps) which are all based on a common set of upper-level ontologies. Thus, the fundamental distinction between the two approaches is that while the UMLS *aligns* existing terminologies and coding systems based on *language* (i.e. on concept labels), the OBO Library promotes the "coordinated evolution" based on sound ontological upper ontologies. Both families are described in the following subsections.

2.2.1. Unified Medical Language System

The UMLS is a collection of terminologies, coding systems and classification systems. Its purpose is to represent the language used in the clinical to domain in a structured and integrated form to *"promote creation of more effective and interoperable biomedical information systems and services, including electronic health records"* [14u]. The terminologies of the UMLS (also call source vocabularies) cover almost all clinically relevant domains from patient care, over health statistics and health service billing to clinical research. One of the most comprehensive ontologies is the SNOMED CT which contains more than 300,000 entities. The general idea and conceptualization of the UMLS is described in [Cam+98]. As described in [NLM09] there are several components forming the UMLS: the UMLS Metathesaurus, the UMLS Semantic Network as well as the SPECIALIST Lexicon and the Lexical Tools. The two most important are the following:

- The *UMLS Metathesaurus* integrates many different terminologies and coding systems through Concept Unique Identifier (CUI) (often simply called *CUIs*). It contains more than 1 million concepts from over 100 source vocabularies [14u]. Thus, the Metathesaurus serves as a central connection or mapping point between the different source terminologies.

- The *UMLS Semantic Network* is a model containing 133 different semantic types which are interlinked by semantic relationships. The semantic types provide high-level semantic categories for the concepts used in the Metathesaurus (all concepts of the UMLS Metathesaurus have at least one semantic type). Examples of the semantic types are `Organism` (with subtypes `Virus`, `Animal`, `Human` etc.), `Biologic Function` (with subtypes `Physiologic Function` and `Pathologic Function`), `Finding` (with subtypes `Laboratory or Test Result` and `Sign or Symptom`) or `Anatomical Structure` (with subtypes `Anatomical Abnormality`, `Cell`, `Tissue` etc.) [NLM09].

The UMLS is not a coding system itself. The meaning of a concept (CUIs) emerges from the linked concepts of the source terminologies. As described, these links are established based on *language* - without an ontologically sound framework. Thus, "in some cases the CUI's emergent meaning can differ significantly from

the original sources' intended meanings of terms linked by that CUI" [Cam+98]. The main difference of that modeling approach is that hierarchical relationships between entities of different semantic type are possible (not avoided). For example, a disease might be a subclass of a symptom which is obviously bad from a pure ontological perspective. In the following, some of the UMLS terminologies are described in more detail.

- **International Classification of Diseases:** The ICD hierarchy contains 12,541 classes representing diseases and other health related problems in a hierarchy of 5 levels. It is used to record diagnostic information (e.g. main and secondary diagnosis). The ICD is issued by the WHO, it exists (with some modifications) in 42 languages and is internationally accepted. The coded diagnostic information provides the basis statistics about mortality and morbidity by WHO Member States[14g].

- **Systematized Nomenclature of Medicine – Clinical Terms:** SNOMED CT is one of the most comprehensive clinical terminologies. It contains more than 400,000 classes covering anatomy, clinical findings, pharmaceutical products, procedures, specimens, social context of patients etc. It is the *"designated standard for use in U.S. Federal Government systems for the electronic exchange of clinical health information"*[1].

- **Logical Observation Identifiers Names and Codes:** LOINC is a standard coding system for describing laboratory tests measurements and observations. It covers 162,368 codes organized along the following six dimensions: `Component` (Analyte), `Property`, `Time`, `System` (Specimen), `Scale` and `Method` [Reg14].

- **Online Mendelian Inheritance in Man:** Online Mendelian Inheritance in Man (OMIM) is a *"comprehensive, authoritative compendium of human genes and genetic phenotypes"* [15l].

- **National Cancer Institute Thesaurus (NCIT):** The National Cancer Institute Thesaurus (NCIT) is a large reference terminology (biomedical ontology) for the domain of cancer research which is being actively developed since 2003 [GPS11].

2.2.2. Open Biological and Biomedical Ontologies

In contrast to the UMLS, the Open Biological and Biomedical Ontologies (OBO) Foundry [14t] follows the idea of a *"coordinated evolution"* of *orthogonal* ontologies,

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

which are based on a common architecture to support biomedical data integration [Smi+07]. That is, the OBO ontologies do not need to be mapped, because they are aligned right from the start. The common architecture that is necessary for this alignment is given by the usage of the same upper ontology – namely the Basic Formal Ontology (BFO) which is described below. Additionally, the OBO Foundry developed naming conventions described in [Sch+07] (and reviewed in [Sch+09]) that support the coordinated creation of ontologies in a common framework. Further, a methodology for reusing (parts of) existing ontologies – called *mireoting* – was developed and is implemented by the OntoFox web service [Xia+10]. Search for OBO can be performed by using Ontobee [Xia+11]. In the following the BFO and some other important OBO ontologies are briefly described. All ontologies, which were reused for the Model for Clinical Information (MCI), are described in detail in chapter 4.

Basic Formal Ontology

The Basic Formal Ontology (BFO) is a typical upper-level ontology providing the basic classes for other ontologies, without containing any domain-specific ones. It is used by almost all ontologies of the OBO library and thus defines the common class architecture of them. As shown in figure 2.3, the BFO distinguishes between *continuants* (e.g. anatomical entities or qualities) and *occurents* (e.g. processes). Both classes are subclass under the root class entity. Continuants are further distinguished as *dependent* continuants or *independent* continuants. For instance, a material entity is an independent continuant while a role, a function or a quality is a dependent continuant. The Basic Formal Ontology version 1.1 (BFO-1.1) contains no properties and is meant to be used with the Relation Ontology described next.

	Continuant				Occurrent
	Independent		Dependent		
Granularity					
Organ and organism	Organism (NCBI taxonomy or similar)	Anatomical entity (FMA, CARO)	Organ function (Physiology ontology, to be determined)	Phenotypic quality (PATO)	Organism-level process (GO)
Cell and cellular component	Cell (CL, FMA)	Cellular component (FMA,GO)	Cellular function (GO)		Cellular process (GO)
Molecule	Molecule (ChEBI, SO, RnaO, PRO)		Molecular function (GO)		Molecular process (GO)

Figure 2.3.: The OBO library as shown in [Smi+07].

Relation Ontology

The Relation Ontology (RO) contains only object properties. The defined properties are very general and reused by most of the OBO ontologies. For instance, RO contains the properties `has part`, `has quality`, `located in` and `participates in`. Most of the properties defined by RO have corresponding inverse properties. [Smi+07].

Basic Formal Ontology version 2

In 2012 a new version of the Basic Formal Ontology, the Basic Formal Ontology version 2 (BFO-2) was presented. The basic idea of BFO-2 [12c] is to integrate BFO classes with the properties of the Relation Ontology (RO), so that one can define e.g. the domain and range of the properties. The fundamental difference however is that in BFO-2 relations are temporalized, i.e.part_of is replaced by `part of continuant at some time` with subproperty `part of continuant at all times`. BFO-2 was also adapted to the OBO Foundry naming conventions. That is, while in BFO-1.1 the class entity had the local name `Entity` in BFO-2 it is `BFO_0000001`. Thus, BFO-2 it is not backward compatible with the BFO-1.1.

Shortcommings Since the two versions of the Basic Formal Ontology, BFO-1.1 and BFO-2, are *incompatible*, ontologies that are build on different versions are incompatible too. This is currently one major problem of the OBO library. While the class hierarchy of both versions is conceptually almost the same, there are two fundamental distinctions between BFO-1.1 and BFO-2:

- **Different URIs:** While BFO-2 now uses the OBO namespace, BFO-1.1 did not. For instance, the class `entity` has URI `http://purl.obolibrary.org/obo/BFO_0000001` in BFO-2, while `http://www.ifomis.org/bfo/1.1#Entity` in BFO-1.1. Thus, BFO-2 is not backward compatible with BFO-1.1.

- **Temporalized relations:** While BFO-1.1 did not contain any relations at all, BFO-2 integrates relations formerly defined in a separate Relation Ontology (RO). Furthermore, in BFO-2 many relations are either defined to hold *at all times* or *at some time*. For example, instead of `ro:part of` BFO-2 defines two relations `part of continuant at some time` and a subproperty `part of continuant at all times`. As described in [12a], the main problem of BFO-1.1 was the inability to *"annotate process measurement data (e.g. pulse rates"*.

In the future BFO-2 should be used by all ontologies of the OBO library, however it is noted that the introduction of temporalized relations by BFO-2 bears several problems as described in [Mun14] and that only the class hierarchy of BFO-2 can be considered to be stable and accepted. For those, who want to reuse only the class hierarchy of BFO-2, an official "class only" file is available [12d] which can be used in combination with RO. At the time of writing most OBO ontologies have already switched to BFO-2 (e.g. Ontology for General Medical Science (OGMS), Ontology for Biomedical Investigations (OBI) and Information Artefact Ontology (IAO)) using however only the class hierarchy of BFO-2.

Ontologies of the OBO-Library

The OBO-library consists of many ontologies. Here we briefly mention some of the well known and those which are important for this thesis:

- **Units Ontology:** There are different Units Ontology (UO) versions available, distinguished by the implemented reasoning level and the use of references to Phenotypic Quality Ontology (PATO) entities. The most comprehensive used OWL reasoning and references to PATO entities to define what qualities the units are for. As explained in [GSH12], units such as `meter` could be modelled as classes or instances. Both variants are provided by different UO versions.

- **Ontology for Biomedical Investigations:** The Ontology for Biomedical Investigations (OBI) is an integrated ontology for the description of life-science and clinical investigations [14r]. It is actively developed within a cross-community effort and provides an integrative framework [Bri+10]. It defines `planned process` with subclasses such as `investigation` or `freezing`. Furthermore, OBI defines various classes under `biological process`, some of them taken from the Gene Ontology. It is based on BFO-2.0 classes, however not using temporalized relations.

- **Foundational Model of Anatomy:** The Foundational Model of Anatomy (FMA) is a large ontology (83,281 classes) with the most detailed description of the human anatomy. Many relations like different part-of relations enrich the ontology and make the FMA useful for spatial reasoning. Only few classes have German labels.

- **Gene Ontology:** The Gene Ontology (GO) [14f] consists of three main parts: cellular component, biological process and molecular function [Ash+00]. It is one of the most famous ontologies of the OBO library and used for indexing of scientific publications.

- **Human Disease Ontology:** The Human Disease Ontology (DOID) contains *"descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts"* [Sch+12]. The ontology contains many links to ICD, NCIT, SNOMED CT and thus provides a bridge between the OBO ontologies and those of the UMLS.

- **Human Phenotype Ontology:** The Human Phenotype Ontology (HP) is a large ontology that describes human phenotypes by combining concepts from anatomy with qualities [Rob+08]. For instance, the class `splenomegaly` is defined as `pato:increase size` (quality) of `fma:spleen` (anatomical entity). The Human Phenotype Ontology (HP) is used for example in clinical

diagnosis in human genetics [Köh+09].

- **Symptom Ontology:** Symptom Ontology (SYMP) contains 840 classes defining various symptoms distinguished mainly by the affected anatomical region (e.g. abdominal symptom, digestive system symptom, head and neck symptom etc.) or body system functioning (e.g. nervous system symptom, neurological and physiological symptom, nutrition, metabolism, and development symptom).

2.2.3. Other Ontologies, Classification and Coding Systems

The following three ontologies are neither part of the UMLS nor of the OBO library. They are listed here, because they were used as reference terminologies in this work.

Radiology Lexicon

The Radiology Lexicon (RadLex) was initially a terminology for the radiology domain. RadLex contains classes for clinical findings, imaging modalities, imaging observations, procedures, processes, descriptors and report components. Many classes however describe anatomical entities which were mostly imported from the FMA.

Anatomical Therapeutic Chemical classification system

The Anatomical Therapeutic Chemical classification system (ATC) is a WHO international standard developed to describe drug information. It consist of five levels describing the (1) anatomical main group, (2) therapeutic main group, (3) therapeutic or pharmacological subgroup, (4) chemical, therapeutic or pharmacological subgroup and (5) chemical substance. The ATC is available in several languages including German. Since there is no RDF version of ATC available, it was created by transforming the Excel content to RDF.

German Procedure Classification

The German procedure classification (Operationen- und Prozedurenschlüssel) (OPS) is *"the official classification for the encoding of operations, procedures and general*

medical measures in the inpatient sector and for surgical procedures in the outpatient sector" [14m].

2.3. Semantic Annotations

In most general form an annotation is an association between distinct pieces of information [SCV13]. Thus, annotations can be seen as meta data serving very different purposes (such as comments or notes) as described in [AF07]. In the context of NLP, IE is a technique to find important information pieces in unstructured texts and to extract them as structured information [Mey+08]. For instance, IE is used to detect semantic entities such as date values, names, measurements, etc., in texts. Approaches for annotation range from automatic image parsing [Sei+09] and information extraction from DICOM headers and structured reports [MRS09] to (aided) manual approaches [Wen+08], [Cha+10] and [Rub+08]. We refer to *semantic annotation*, the annotation of unstructured data with concepts from ontologies. Here the ontological entities are mapped to the corresponding words in the text, based on the controlled vocabulary of an ontology. This task is also referred to as NER. Semantic annotation has the advantage, that the meaning of the annotation is defined through the relations within the ontology. Having an annotation with the concept `fma:Liver` one can refer to the FMA and that a liver is a subclass of `fma:Lobular organ` or further of `fma:Anatomical Structure` as shown in figure 2.4.

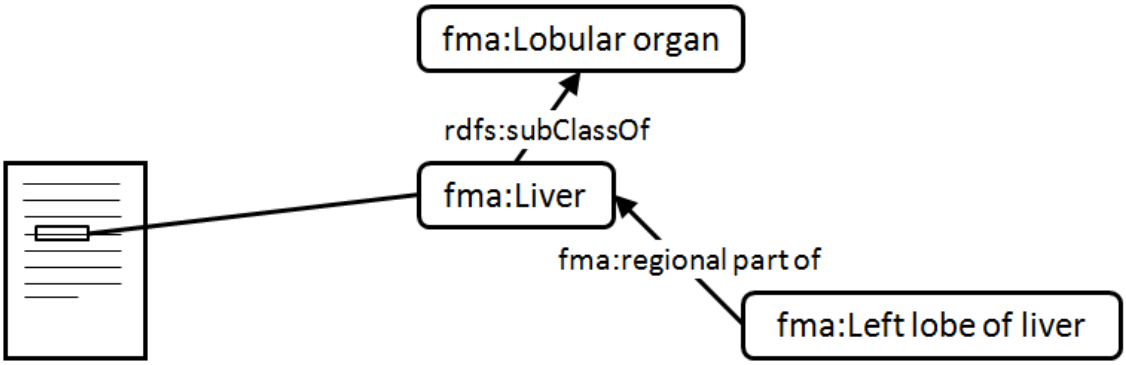


Figure 2.4.: A semantic annotation of an unstructured text with the FMA class for *liver*.

NCBO Annotator [JSM09] is a publicly available web service [2] which can be used to annotate text with ontology terms from UMLS and the NCBO BioPortal. In summary, semantic annotation is a technique to make heterogeneous clinical data better accessible - e.g. for search or other processes that aim to capture the *content* of the unstructured data. Since annotations have different types and context (provenance information etc.), one commonly needs a data model to store and structure annotations. Examples are the Open Annotation Data Model [SCV13] or the MEDICO Annotation Ontology [Sei+10]. The usage of an annotation model improves precision and recall in information retrieval in comparison to simple key-word tagging as pointed out in [OPS10]

[2] http://bioportal.bioontology.org/annotator

3

Related Work

In the previous chapter, the basics of knowledge representation, semantic web technologies and biomedical ontologies were described. Here, work related to the objectives that are listed in the introduction is presented. Mainly, this is the integrated representation of clinical data and medical knowledge. Since the field of knowledge and data representation in biomedicine is very broad, the following description is concentrated on related work that utilizes semantic technologies and existing reference ontologies for representation of clinical data. Data models (or information models) with a similar scope as the Model for Clinical Information (MCI), which is presented in this thesis, are described and differences to MCI are highlighted. The following models are analyzed with respect to their ability to express clinical findings: The HL7 Reference Information Model (RIM) version 3, OpenEHR, DICOM Structured Reporting, the OBO-based data models and ontologies Translational Medicine Ontology (TMO) and Computer-Based Patient Record ontology (CPR) and further the Quantitative Imaging Biomarker Ontology (QIBO). After this review of related data models, related work regarding the inference mechanisms (chapter 8) and corresponding case studies are described: Firstly, the knowledge-based extraction of measurement entity relations from free text radiology reports and the subsequent classification of normal findings and, secondly, the ranking of likely diseases.

3.1. The Relation between Data Models and Reference Ontologies

As described in the introduction (chapter 1), the representation of clinical data in future healthcare systems is realized through a combination of some standardized information model (that defines the data structure) with standardized controlled reference terminologies/ontologies (providing the vocabulary). A good description of this approach is given by [Ben10]. The most well known information models are the HL7 Reference Information Model (RIM) and the OpenEHR Entry Model (both are seen as standards [MSC08]), while the most well known and comprehensive reference ontology within the clinical domain is SNOMED CT which contains more than 400,000 entities. A combination of an information model with a reference ontology is needed since no terminology (or ontology) can provide precoordinated concept for every possible information artefact: The author of [Cor09] shows that in SNOMED for *"many disease one of five episodicities, one of six severities and one of seven courses can be chosen"* leading to 300 possible combinations for a single disease. Thus, the role of the information model is to define relationships and thus provide the structure for post-coordination of entities from the reference ontology. A *binding* of the information model to reference terminology is needed to ensure consistent and unambiguous usage of terms.

There are two main challenges here: Firstly, a clear definition and separation of the scope of an information model and reference ontologies. Secondly, a good approach to define the relation between (or binding of) entities of the information model with those of the reference ontology to avoid ambiguous usage and obtain normalized representations. In the following, important research contributions regarding these two problems are summarized.

In [KBS11] the relationship between information models and reference ontologies and their characteristic differences are analyzed. It is argued that the main difference between information models (EHR models) and biomedical (reference) ontologies is that the first has to express or include *"observations, opinions, instructions, proposals, requests etc."*, while ontologies define biomedical entities and their relations in a general and "mind-independent" way. Similar to the data model of OpenEHR, "achetypes" are used to specify constraints on the correct usage of the reference ontology.

The scope of existing reference terminologies is often too broad and thus hinders normalized representations: As argued in [Sch+10] existing reference ontologies such as SNOMED "contain all kind of linguistic expressions and entities which are not terms in a strict sense", e.g., the entity "Poor condition at birth without known asphyxia" can hardly be called a *term*. These so called "non-terms" of SNOMED CT are analyzed and remodelled as information artefacts using the structure of HL7 RIM and alternatively with IAO. This view of placing all meta-data such as

plans and observations in the realm of the information model is similar to [KBS11] and helps to clarify the scope and the role of reference ontologies.

The authors of [QKR07] highlight the fact that clear semantics of the data model (information model) is essential for quality and accuracy of data mapping to terminology codes. The authors of [QR07] further developed a *"semi-automated mapping system called the Model Standardisation using Terminology (MoST)"* that maps archetypes of the data model to a reference terminology. They demonstrate this by remodeling the "especially ambiguous" OpenEHR data model and achieve an increase of accuracy from 64.7 % to 80.55 % for their mapping system. The major issues with OpenEHR data model were *"ambiguous categorisation of top level data terms"* such as process, observation, action etc., *"ambiguously named terms"*, *"similar or duplicate labeling of terms with inadequate definitions"*, *"insufficient separation of meta data information from core clinical recording information"* and *"using post-coordination of terms adequately and in the right place"*. The results of the analysis of [QKR07], especially the clear separation of concepts of the data model (for meta data) from the vocabulary of the reference ontology and clear patterns for post coordination guided the development of the information model that is presented in this thesis.

The authors of [BH07] emphasize that "information models must satisfy many requirements when actually implemented including: computational efficiency and performance; economically viable implementation; maintainability; system scalability and extensibility; and the considerable privacy and security requirements of the health domain." Addressing these requirements, the authors present a four step approach for creation of a successful data model for healthcare information. A focus is set on the requirement of *recording* of information during the care process reflected by the created OpenEHR *Entry model* (details below in section 3.1.2) for which the following five types of information created during the care process were identified: observation, opinion, instruction, action and administrative event.

In [MSC08], the combination of HL7 and OpenEHR with SNOMED CT is compared. In particular the overlaps and gaps of the (structural) information model with the terminology are analyzed. Different facets of clinical information are listed and mapped to five categories for intended usage: *"terminology model only"* (e.g. general semantic relationships between concepts), *"terminology model preferred"*, *"gray area"* (e.g. for representation of context information), *"structural model preferred"* and *"structural model only"* (e.g. identifiable instances, data types etc.). As [RQM06], the authors of [MSC08] emphasize the importance of *bindings*. Ten Practical principles for terminology bindings such as *"understandability"*, *"reproducibility"* and *"transformability and normal forms"* are presented.

In [RQM06] it is argued that "a key problem for usage of medical ontologies for electronic health records and messages is to specify validation rules." The data model for the EHR or messaging systems are referred to as *"information models"*. Addressing this problem, authors present a methodology for the *binding*

of information models with ontologies, where the role of the binding is to define "which codes of the terminology (ontology) are to be used where". The presented "Code Binding Interface (CBI)" is implemented with OWL to express type and cardinality constraints on the allowed classes from the reference ontology. Further, "place holder classes" are introduced to express more complex bindings. The usage is demonstrated by binding of HL7 Reference Information Model with SNOMED CT. The CBI may be stored separately from information model and reference ontology and thus can be considered as the third major component for representation of EHR data. In the work of this thesis the definition of "place holder classes" for the binding to reference ontologies was adapted from the work of [RQM06].

In the following subsections different data models are analyzed in detail. For each model the size, its intended role and scope and the difference to MCI are described.

3.1.1. HL7 Reference Information Model Version 3

The HL7 Reference Information Model was developed to create the structure for coded clinical data and to provide the basis for all other information models of the HL7 V3 family. It should form the foundation of all information modeling within HL7 (in combination with data types and vocabularies) [15f]. HL7 Reference Information Model is specified in UML, however since 2010 an RDF version has been available at BioPortal [14h] containing 7501 classes. The RDF version does not contain any object or data properties and axioms are only used for definition of subclass relationships. In contrast to the HL7 v2.x standards, which standardize the exchange of health information (e.g. in defining the document structure) while offering flexibility regarding the transferred data, the HL7 Reference Information Model focusses on standardizing the representation of the data itself. The motivation for version 3 is to *reduce* flexibility and thus enforce standardization "focusing on semantic interoperability" [15f]. The RIM is complemented by a *data type specification* and a *vocabulary specification*.

The classes of the HL7 Reference Information Model are organized under 4 main classes `Entity` (person, organization, material, device etc.), `Role` (patient, employee, access etc.), `Participation` and `Act` (observation, procedure, supply, document, account etc.). For instance the `ActClass` has subclasses `act`, `disciplinary action`, `Specimen Collection` and 19 times `Retired Code`! The class hierarchy, as well as textual definitions, are often irritating. For example, the class `ActPriority` has 14 direct subclasses including

- `stat`: "With highest priority (e.g., emergency)."

- `ASAP`: "As soon as possible, next highest priority after stat."

- `emergency`: "An unforeseen combination of circumstances or the resulting state that calls for immediate action."

- `callback for scheduling`: "Filler should contact the placer (or target) to schedule the service. (Was "C" in HL7 version 2.3's TQ-priority component.)"

Firstly, the difference between `emergency` and `stat` is not clear: From the textual definitions, emergency is a specialization of `stat` – it is however a sibling of `stat`. Further, the order of priorities is unclear: `ASAP` is second after `stat` (highest priority), however `emergency` is mentioned in the definition of `stat`. As shown by the entity `callback for scheduling`, the textual definitions in HL7 Reference Information Model mix several aspects: the semantic definition of the class, intended action of healthcare staff, and legacy information. To give another example, consider the definition of `image` (a subclass of `ActCode`): *"Description: A characteristic representing a single file reference that contains two or more views of the same dosage form of the product; in most cases this should represent front and back views of the dosage form, but occasionally additional views might be needed in order to capture all of the important physical characteristics of the dosage form. Any imprint and/or symbol should be clearly identifiable, and the viewer should not normally need to rotate the image in order to read it. Images that are submitted with SPL should be included in the same directory as the SPL file."* [15g] Probably this class should represent images of medication, but even then: why is an image an act code? A better separation of meta data from process data (acts) are needed here. Further, HL7 RIM contains four classes for *patient*: one is a subclass of RoleClassRelationshipFormal and defined as *"Scoped by a provider"*, a second is a subclass of RoleActCode and defined as *"The recipient for the service(s) and/or product(s) when they are not the covered party."* [15g].

The irritating textual definitions and lack of a clear understandable class hierarchy make many classes of HL7 RIM very difficult to use and thus semantic interoperability will hardly be achievable in this form. A detailed review of the ontological issues and lack of consistency of HL7 Reference Information Model is given in [SC06]. The work of this thesis is an attempt to create an information model on the basis of well defined established upper-level ontologies with clear semantic interpretation of the high level concepts to overcome the above mentioned ontological issues. Note that HL7 RIM is 15 times as large as MCI: On the one hand, HL7 RIM has a much broader coverage than MCI, but on the other hand, many classes such as the priority codes are better placed in a reference terminology (coding system) than in an information model and thus were *intentionally not defined in MCI*.

Fast Healthcare Interoperability Resources

Recently, a new HL7 initiative called Fast Healthcare Interoperability Resources (FHIR) [15e] was started, that attempts to create the next generation standards framework. It should combine the *"best features of HL7's version 2, version 3 and CDA products"*. The work on RDF version of the data model is promising, however an official HL7 version was still under development at the time of writing.

3.1.2. OpenEHR

OpenEHR consists of a *"set of specifications defining a health information reference model, a language for building 'clinical models', or archetypes, which are separate from the software, and a query language"* [15m]. The OpenEHR Foundation has utilized the "European standard for Electronic Health Records (EN13606) for developing a range of highly-constrained clinical statement and record composition models (called archetypes and templates)" [MSC08] and in total OpenEHR defines 425 archetypes and 25 templates. Similar to HL7 RIM, the data model of OpenEHR is to be used in combination with reference terminologies such as SNOMED, LOINC and ICD10. In contrast to HL7 Reference Information Model, OpenEHR provides also a model for the electronic health record (EHR) itself [Bea+08].

An ontological analysis of the OpenEHR data model is presented in [BH07]: the authors argue that a successful model making health information systems interoperable has to be created by analysing the process of clinical care delivery as an scientific problem-solving process. Thus, OpenEHR describes clinical data from the data entry perspective to allow efficient capture of data in a structured form. The OpenEHR *Entry* model is based on Clinical Investigation Record ontology (CIR), which is developed with the main class `recorded information` and subclass `care information` (see figure 3.1). Subclasses of `care information` are organized in three branches `history`, `opinion` and `instruction` which represent temporal categories. Subclasses are distinguished as observation-related categories (e.g. `observation`, `diagnosis`, `prognosis`) or intervention categories (e.g. `proposal`, `goal`, `recommendation`, `intervention request`).

In [Gar+07] the impact of the OpenEHR archetype-based approach on healthcare professionals and semantic interoperability are analyzed. Archetypes provide the structure for clinical information and can be bound with a reference terminology such as SNOMED CT to define a set of allowed values (i.e. SNOMED codes) for specific attributes for standard clinical care processes. For instance, the patient position (sitting, lying etc.) during a blood pressure measurement can be specified by using corresponding SNOMED codes. The authors of [Gar+07] argue that the OpenEHR archetype approach allows healthcare professionals (that is the domain experts) to formally define clinical content without technical understanding.

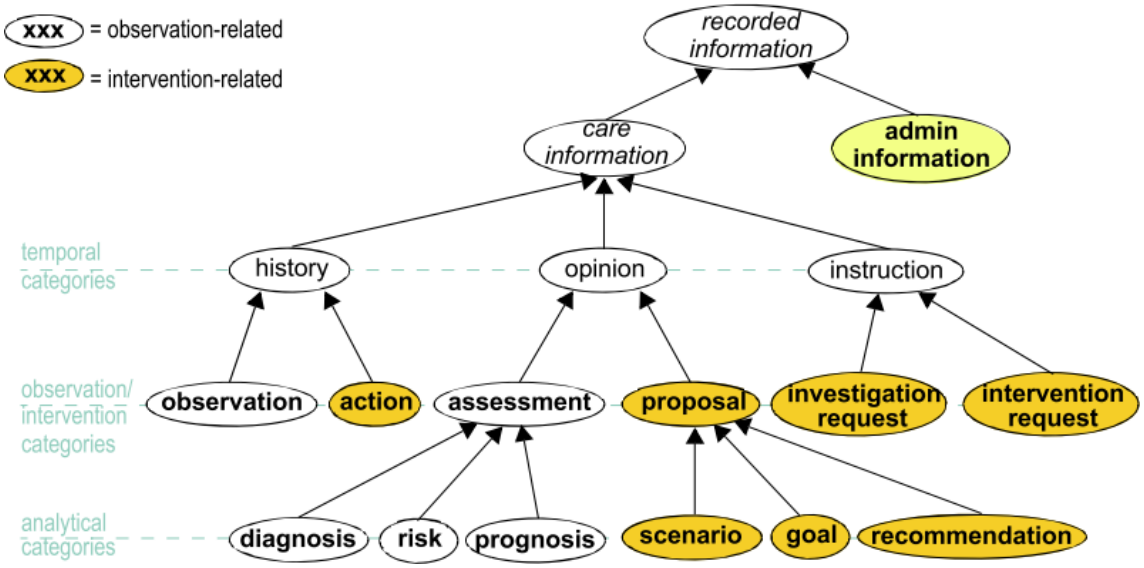


Figure 3.1.: The Clinical Investigator Record (CIR) ontology presented in [BH07; Bea+08].

The authors of [AA11; AAS12] revise the ontological structure of OpenEHR and the role of the data model for representation of clinical data with reference ontologies that commonly implement a so called realist approach. Large parts of the OpenEHR data model for representation of uncertain knowledge such as opinions are not easily expressible with realist ontologies which "require the nodes and edges in an ontology graph to correspond to entities in reality" [CES07]. For example, a *suspected fever* is not a specialization of *fever*. The authors show however that large parts of the OpenEHR archetypes can be represented also with a realist approach which would allow a more consistent combination with existing reference ontologies.

OpenEHR is focused on the *entry*, i.e. documentation clinical care delivery – the development of MCI focuses on providing a general structure for the representation of the *content* of observations and findings. For example, opinions or recommendations are not modelled in MCI. One main disadvantage of OpenEHR is that no official version in RDF is available. This makes a combination with reference ontologies and application of standard reasoning mechanisms more difficult. While for OpenEHR the *entry* of data is the focus, MCI has the purpose to allow *reasoning* over clinical data based on formalized medical knowledge to provide views on the data in different clinical decision making processes. In [QKR07] OpenEHR is taken as an example for a *"especially ambiguous data model"*.

3.1.3. DICOM-Structured Reporting

DICOM structured reporting [Clu00; Hus+04] is a supplement to the Digital Imaging and Communications in Medicine (DICOM) standard [15c] and provides concepts for describing imaging findings as well as the imaging context such as wavelength etc. in a structured form by using specific templates [Hus+04]. Even though the name suggests fully structured reports, DICOM SR also supports free text radiology reports. While the general DICOM defines a standard for the exchange of images and the parameters for presentation, DICOM SR is focused on the *content* of the information [Hus+04].

According to [Ber+08], DICOM SR distinguishes between *anatomy findings*, *lesion findings* and *quality findings*. Furthermore, the properties *finding site*, *finding modifier* and *certainty of finding* can be specified. There are several context specification as e.g. *Anatomic Regions*, *Anatomic Modifiers*, *Image Guided Therapeutic Procedures*, *Interventional Drug*, *Nuclear Medicine Projections* and others which mostly have corresponding concepts in RadLex and SNOMED. The data of DICOM SR documents is serialized in XML and accordingly organized in form of a tree. Detailed information about DICOM Structured Reporting (DICOM SR), the hierarchical tree structure and template specifications can be found in [Clu00]. Even though no official RDF version of DICOM SR is available, the author of [Bru13] developed a transformation of the DICOM meta-data to an RDF representation. An attempt to create a unified representation of radiology findings integrating DICOM SR descriptions and UMLS representations are described in [Ber+08]. The authors argue that a unified representation of findings bears the potential to integrate different knowledge resources (literature on differential diagnosis) with the diagnostic process.

This thesis also shows that uniformly described findings (not only in radiology) are key for integration of clinical data into different decision making processes. In particular the basic representation of clinical findings in MCI is similar to the representation described in [Ber+08]. However, MCI uses only the anatomical entities and some image observations from RadLex, while other entities are taken from OBO ontologies. For example, the modifier (i.e. the qualities) are taken from PATO. Further, MCI provides more categories of clinical findings such as normal/abnormal, increased/decreased, increasing/decreasing etc. which are relevant for realization of a decision support services.

3.1.4. OBO-based Models

MCI is based on ontologies from the Open Biological and Biomedical Ontologies (OBO) library and reuses many classes and properties from there. In this subsection other data models are described which also reuse OBO ontologies and thus share

a similar foundation (i.e. have common classes and properties) with MCI. OBO ontologies which are *reused* by MCI are described later in chapter 4 since they form an integral part of MCI.

The Computer-Based Patient Record ontology

Computer-Based Patient Record ontology (CPR) [11a] is meant to be used in clinical research projects and patient registries [Ogb11]. It is based on BFO1.1, defines 69 classes and contains 136 classes in total. Furthermore, it defines 22 object properties (in total 52) and 5 data properties (14 in total). The ontology attempts to address the terminology needs for patient record systems and is used in combination with the FMA, OWLTime [06c], BioTop [12b], the RO [15o] and SNOMED CT [Int14]. CPR makes use of OWL axioms to formally define its classes. The main classes are *clinical act*, *clinical artefact* (with subclasses such as *clinical finding*, *clinical diagnosis*, *laboratory test finding*, *patient record*, *patient record of a symptom* etc.), *disposition* (with subclasses such as *disease*, *syndrome* etc.), *role* and *object* (with subclasses such as *anatomical structure*, *medical device*, *organism*, *patient*, *physician* etc.). Most of the classes of CPR contain textual and formal definitions, however not all of them are fully consistent. For example, *clinical act* is defined as "a collection of events, when a medical care professional provides a medical service to some human or animal patient(s)" [11a]. The logical definition however does not express the notion of a "collection" of events.

Even though CPR defines a lot of important core terms of a clinical information model, it was easier for us to base MCI on OGMS due to its clarity and illustrative descriptions of concepts [SCS09]. In contrast to CPR, which is focused on clinical research and patient registries, MCI is focused to provide general patterns for clinical findings and measurements which are integrated with medical knowledge. The goal of MCI is to enrich finding descriptions in order to better integrate the data within clinical decision making. Further, the CPR Ontology is still based on BFO1.1 and seems to be not longer actively developed and maintained.

Translational Medicine Ontology

The Translational Medicine Ontology (TMO) is a *"high-level, patient-centric ontology that extends existing domain ontologies to integrate data across aspects of drug discovery and clinical practice"* [12e]. It is based on BFO1.1 and IAO and defines 204 classes (300 classes in total), 5 object properties (in total 34) and 2 data properties (6 in total). The ontology was created by the W3C Semantic Web for Health Care and Life Sciences Interest Group (HLSIG) and attempts to address the terminology needs for patient record systems. It should be used in combination with established domain ontologies. The core classes of TMO are *process* (with subclasses *clinical*

trial and *adverse drug reaction*), *chemical substance* (e.g. *pharmaceutical ingredient*), *molecular entity* (subclasses *DNA*, *protein* etc.) and *biomedical measure* (subclasses *medical measure*, *drug measure*, *genetic measure* etc.). One highlighted use case of TMO is the Alzheimer's Disease since it is influenced by *"a range of genetic, environmental and other factors"* [Luc+11].

In contrast to MCI, TMO is more focused on clinical studies and the representation of molecular and genomic data and lacks classes for a detailed representation of clinical findings other than laboratory measurements. Due to the available patient data of the use cases, TMO was not the right ontology to start with. If MCI is extended to molecular and genomic data an alignment to or reuse of TMO classes is certainly appropriate.

3.1.5. Quantitative Imaging Biomarker Ontology

The Quantitative Imaging Biomarker Ontology (QIBO) [11b] is not based on any upper-level ontology. It defines 618 classes, 56 object properties and 72 data properties to express imaging biomarkers. The term *biomarker "refers to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly"* [ST10]. In general, structured representation of image measurements allows the analysis of large patient cohorts. QIBO was created to support the creation of biomarkers from clinical imaging data with the aim *"to represent, integrate, and harmonize heterogeneous knowledge across the domain of imaging biomarkers"* [Buc+11]. This should improve the access to quantitative imaging data, ease the development of new biomarkers, support annotation of imaging data and further provide a "framework for hypothesis testing". In general, biomarkers are powerful since they are quantified and thus allow comparison and analysis. Within the imaging domain the development of biomarkers however relies on standardization and the integration of data with different contexts. The Radiological Society of North America (RSNA) created Quantitative Imaging Biomarkers Alliance (QIBA) [15n] with the aim to *"improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, patients and time"* [Buc+11]. While the approach of QIBA is to integrate many different stakeholders to standardize biomarkers, the coverage is limited to "most mature biomarkers" [Buc+13]. QIBO, on the other, hand was created to support the creation of biomarkers from clinical imaging data *in general* and thus support the detection of *new* biomarkers based on existing data. QIBO is organized under the following first level classes: *Imaging subject*, *biological intervention*, *imaging agent*, *biological target*, *acquisition device*, *imaging technique*, *post-processing algorithm*, *quantitative imaging biomarker*, *indicated biology* and *biomarker use*.

Even though QIBO is focused on biomarkers, it has a very similar scope as MCI

since both models have a strong focus on image findings and measurements. However, the fact that the ontology is no longer maintained and that many of the classes of QIBO lack (good) textual definitions makes the ontology hard to reuse. Further, QIBO is not aligned with other ontologies.

3.1.6. Other Data Models

There are many more data models such as Integrating the Healthcare Enterprise [15j], Clinical Data Interchanges Standards Consortium (CDISC) [15b] and others which are developed by different standardization organizations. The scope and approach of these models are however significantly different to MCI which uses semantic web technologies and established upper level ontologies. Also classical data ware house providers developed healthcare specific models. For example, the Oracle Healthcare Data Model [Cor10], Teradata Healthcare Logical Data Model [15p] or IBM Unified Data Model for Healthcare [15i]. The underlying motivation behind these models is to support the *management* of a healthcare organization by using of the vendors standard Business Intelligence (BI) and analytics software solutions. This however is not the goal of MCI, which attempts to support *clinicians* in their decision making. In the context of data integration, the i2b2 platform [15h] has to be mentioned which is essentially a data warehouse (DWH) for clinical data. With i2b2 heterogeneous clinical data related to a patient can be made available at one point and structured data can be mapped to standardized coding systems. The problem however is that the i2b2 data schema is too generic to represent more complex data such as findings descriptions which are better represented in graph like structures. In section 7.3, it is described how data can be mapped from i2b2 to MCI.

3.2. Related Work Regarding the Case Studies

The Model for Clinical Information is used to express clinical data and medical knowledge in an integrated manner. The medical knowledge defined in terms of MCI is applied in three case studies: measurement extraction, finding classification and disease ranking. Related work regarding these case studies is presented in the following subsections.

3.2.1. Measurement Extraction and Classification

Extraction of structured measurement information from free text (section 8.2) is a special case of *relation extraction*. This in turn is a specific task of general Information Extraction (IE) which is a very broad research domain for its own. The work presented in this thesis builds upon existing IE techniques such as Named Entity Recognition (NER) and extraction of measurements and provides a knowledge-based approach to infer the *relations* between measurements and the measured entities to obtain structured measurement representations. That is, based on the output of an existing annotator, which is build on the Unstructured Information Management Architecture (UIMA) framework, structured representations are inferred.

In contrast to this very specific task, medical text analytics has been conducted in the context of various use cases. For example, in the cTAKES project [Sav+10], which is also based on the UIMA framework, and by the MedLee system [Fri97]. While cTAKES approach to measurement detection is based on finite state machines, the MedLee system is rule-based. Relations between medical entities like conditions (e.g. diseases or disorders), symptoms and indicated treatment are extracted e.g. in [RGH08] or [BZ11]. The vocabulary of medical ontologies such as those from the Unified Medical Language System is commonly used for the extraction of medical entities (e.g. in [BZ11]). Existing approaches for relation extraction mostly use machine learning (ML) techniques in combination with linguistic parse trees [Sav+10]. Within the i2b2 Relations Challenge in 2010 the supervised ML system described in [RHR11] showed the best results for the extraction of relations between medical problems, treatments and tests with an F-measure of 0.74. The i2b2 Challenge on Temporal Relation extraction showed that ML outperforms other systems in *detection* of events, while hybrid approaches were better at *classification* of temporal relations [SRU13]. Similarly in [BZ11] a hybrid approach (combine relation patterns and ML) yields an F-measure of 0.94 for the extraction of relations between disease and treatment. Since the knowledge-based approach presented in this thesis has a rather specific focus, it is however difficult to compare these results to it. No evaluations of the above mentioned approaches for the extraction of *measurement-entity* relations could be found. However, in [Ang10]

the role of domain knowledge for extraction of information from medical texts in general is emphasized since it *"supports supports the decisions about structuring the extracted text objects into domain statements"*. The thesis follows this idea by using domain knowledge about the normal size of anatomical entities and the generalization hierarchy of the RadLex ontology to infer structured measurement representations based on semantic annotations. That is, by using this domain knowledge, a *set* of annotations is transferred into a *semantic structure*. The authors of [Ang10] refer to these (target) structures as *templates*.

In a subsequent step to extraction of structured measurement relations, this thesis provides a knowledge-based approach to classify the extracted measurement findings as *normal* or *abnormal*. As in the case of measurement extraction, there are numerous NLP techniques with the purpose of classification and semantic labeling of textual entities. It is noted that the knowledge-based approach presented in this thesis is different to to *pure* NLP approaches (such as [BZH13]). On the one hand, the knowledge-based approach is able to classify findings even when *no* interpretation terms such as *normal*, *unremarkable*, etc. occur in the report text (e.g. a sentence like "Spleen 13.8 cm." can be classified). On the other hand, the approach is limited to findings where the measurement is explicitly mentioned in the text. As for the relation extraction, hybrid approaches that take advantage of different methods bear the greatest potential.

3.2.2. Disease Ranking

Diagnosis Systems have a long tradition in Computer Science and Artificial Intelligence, in-particular within the medical domain. There are a variety of formalisms and techniques like set-cover, abductive reasoning, logic approaches, Baysian networks, rule-based systems, case-based reasoning etc. Most of the logical approaches aim to explain the whole set of observations, i.e. provide a complete diagnosis. This however often leads to diagnosis consisting of large fault sets and thus unspecific and redundant diagnosis results, not helping the clinician. Early attempts to formalization of model-based diagnostic knowledge were made in [Rei87]. In [Luc98] evidence-functions are used to encode the relation between defects and findings. Even though these formalisms are more expressive than the disease-symptom model presented in section 8.6, this comes with high computational cost: most diagnosis-algorithms are exponential with the number of possible faults. In the medical domain there are many well known implementations of clinical diagnosis systems (expert systems) from around the 1970th and later like e.g. MYCIN [Sho76], INTERNIST-1 [MPM82], CASNET [KW82], DXplain [Bar+87], CADIAG2[AA89], PATHFINDER [HHN89]. Since Bayes-nets are theoretically best suited for diagnosis, they are successfully implemented in some of the mentioned expert-systems. However, statistical approaches like Bayes-nets require knowledge, which is not broadly available. For example, the a-priori

probabilities for signs and symptoms $P(s)$ are mostly not known. Thus, it is not possible to compute the conditional probability for a disease, given the symptom $P(d|s) = \frac{P(s|d) \cdot P(d)}{P(s)}$ with the help of Bayes-Theorem. Note that $P(s|\neg d)$ is normally not known as well. It is not difficult to see that even though one cannot calculate $P(d_1|s)$ and $P(d_2|s)$ for two diseases d_1 and d_2, one is able to *compare* the values. Thus, without knowing $P(s)$, *ranking* diseases is possible.

In many approaches, as in the one taken by this thesis, a patient is represented through a set of present (and absent) symptoms. For example, in [BSP03], a quality measure for a diagnosis is defined. However, the factors influencing the ranking are different (e.g. they assume to have knowledge about the probability of a finding being caused by a certain disease). The authors of [Mai+11] propose an information retrieval inspired approach to rank likely diseases. Similar to the approach of this thesis, a model represents the disease-symptom relations using fuzzy labels, however the symptoms themselves are not weighted.

Related is also the *set-cover* approach: given a set of symptoms S, diseases are represented as subsets of S. Then one is interested in a minimal set of diseases covering the whole set of present symptoms of S in the context of a particular patient. Set-cover has several optimizations under consideration of so called parsimony criteria like minimal cardinality, irreducibility, relevance, most probable diagnosis and minimal cost [RNW83]. In other words, these systems deliver a *complete diagnosis* explaining all observations or clinical findings under some optimization constraints. This however leads to diagnosis with big disease-sets like "Lymphoma *and* Colorectal Cancer *and* some Infection *and* Diverticulitis can explain the symptoms that were observed at the patient". Clinicians commonly clarify the presence of one specific disease and thus are more interested in a single-fault diagnosis even though a single disease cannot explain *all* but *most* of the symptoms of the patient.

Thus, in contrast to the set cover approaches that aims to cover *all* symptoms, the ranking described in this thesis simply provides a ranking that highlights diseases that are well matching with the patient's symptoms. Further, in contrast to the above mentioned approaches, the algorithm of this thesis includes also information about the hierarchical relationships between symptoms. This information is crucial to augment symptom information before starting the disease ranking algorithm. Firstly, for detection of inconsistencies and secondly for inference of implicit symptom information. Similar to this approach is the work on the Phenomizer [Köh+09] which aims ranks likely diseases based on phenotype information which is represented by reference to the Human Phenotype Ontology (HP) and thus makes use of the hierarchical structure of phenotypes. In the context of radiology, the Gamuts website [15d] describes a network of diseases and symptom (or findings) that is linked to the RadLex Ontology. Gamuts is used in [BLK14] for providing differential diagnosis, similar to the work of this thesis described in section 8.6.

Part II.

THE MODEL FOR CLINICAL INFORMATION

4

Model Foundation

The Model for Clinical Information (MCI) is based on selected upper- and mid-level ontologies from the Open Biological and Biomedical Ontologies (OBO) library and reuses established schemes like the Dublin Core. The respective ontologies define the basic classes and properties and provide the foundation of the model. While a general introduction to the OBO library was given in section 2.2.2, in this chapter it is motivated why ontologies from the OBO library are reused as the main foundation for MCI (section 4.1) and the selected ontologies are described in more detail (section 4.2). For each reused ontology an overview of its size, the general scope and intended role in MCI is given. Then it is described which entities were reused and which adaptations needed to be made to obtain a solid and consistent foundation for MCI. Technically, the reuse of ontologies is realized through so called imports of ontology modules (i.e. subsets of the classes and relations of an ontology). section 4.3 describes the creation of these imports by using the MIREOT Tool.

4.1. Requirements of the Model Foundation

As motivated in the introduction, it is desirable to reuse existing models whenever possible. Existing models were analyzed with respect to the following criteria:

- **Open:**The models should be freely available and reusable so that one is not limited in regarding extensions.
- **Well defined:** The basic classes should be well defined and proved to be well suited to represent clinical data. The class hierarchy and textual definitions of classes provide the basis for extensions.
- **Coverage:** Core information objects such as diagnosis, processes, examinations, reports etc. should be already defined so that one has to extend the model only with use case specific classes.
- **Active Community:** The model has to be based on ontologies which are still maintained so that questions can be addressed. Understanding a model precisely only through written documentation is difficult and likely to produce wrong understanding.
- **Stable:** The foundation of MCI should be as stable as possible. Changes to the reused model affect the structure of MCI and thus need to be aligned. Thus, the reused model should not change fundamentally.
- **Established:** Used as a foundation by other projects in healthcare domain so that one can benefit from analysis of related work.
- **Standard:** If possible, the model should be build upon standard solutions.
- **Modular reuse:** Existing models are sometimes very large. Flexible reuse of parts of a model are important to avoid inclusion of unnecessary entities which are not relevant for the use cases.

Obviously there is no ideal framework or model which fulfills all requirements. As mentioned in chapter 3, the HL7 Reference Information Model, the OpenEHR and several OBO ontologies are important candidates to provide the foundation of MCI . Thus, they are compared along the above mentioned criteria (see table 4.1). HL7 RIM and OpenEHR models can be considered as the most promising models to become a standard in the future and the coverage of both models is good for healthcare related entities. However, both models are not based on upper level ontologies and thus lack a proved well defined ontological basis (for detailed arguments see chapter 3). In the case of HL7 RIM there are even inconsistencies (for details see 3.1.1). For instance, HL7 RIM contains four classes for *patient*. From

Table 4.1.: Comparison of HL7 RIM, OpenEHR and ontologies of the OBO library.

	HL7 RIM	OpenEHR	OBO library
open	+	++	++
well defined	–	-	++
coverage	++	++	++
active community	o	+	+
stable	o	+	+
established	o	++	o
standard	+	++	-
modular reuse	-	o	++

the coverage point of view all models provide a good basis. The ontologies of the OBO library are broadly used within the biomedical domain (genetics, phenotypes etc.), however still not *established* in healthcare.

In summary, OpenEHR and ontologies of the OBO library provide a good basis for MCI . However, because of the fact that OBO ontologies are well defined and that modular reuse easily accomplished, it was decided to base the model on ontologies of the OBO library. Due to the orthogonal development approach taken by OBO one can easily create and combine modules from different ontologies - exactly fitting ones modeling needs. Thus, one avoids the creation of a large model where many unnecessary and duplicate classes and relations are defined. Further there is an active community working on the different OBO ontologies. Thus, modeling questions can be discussed and one can get valuable feedback to extensions to existing ontologies. A disadvantage of OpenEHR is, that the model is formalized in a specific format called 'Archetype Definition Language' ADL), and that there is no official RDF version which makes it more difficult to use with standard semantic tools.

Regarding well defined upper level ontologies there are several options: For example the Basic Formal Ontology [GSG04], BioTop Ontology (BT) [Bei+08], General Formal Ontology (GFO) [Her10], Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [06b] or Suggested Upper Merged Ontology (SUMO) [10b]. For a detailed comparison of these ontologies (except BioTop Ontology (BT)) it is referred to [Mas+06]. Since BFO is used by ontologies of the OBO library it can be considered as the most widely used within the biomedical domain and it was not considered to base the model on one of the other upper level ontologies.

4.2. Reused Ontologies

In this section for each ontology imported by MCI a short summary is provided before the classes and properties that were reused are described. The main imports are the Basic Formal Ontology version 2 (BFO-2), the Ontology for General Medical Science (OGMS), the Information Artefact Ontology (IAO), the Ontology for Biomedical Investigations (OBI), the Phenotypic Quality Ontology (PATO) and the Units Ontology (UO). To give an impression of the size of the described ontology, the number of classes and if relevant also the number of properties are listed. Appendix A.3 provides a detailed explanation about how corresponding numbers are calculated.

4.2.1. Basic Formal Ontology

The Basic Formal Ontology version 2 (BFO-2) [12c] defines 36 classes, 36 object properties, 0 data properties and 2 annotation properties. It is an upper-level ontology which is completely domain independent and provides the basis for almost all ontologies of the OBO library. There are three main distinctions modelled in BFO [BS03; BS04; GSG04]:

- **continuants vs. occurrents:** Continuants exist over time (e.g. objects) while occurrents come to existence and leave existence (e.g. processes).
- **dependent vs. independent:** Continuants are further distinguished as independent and dependent continuants. While independent continuants can exist for their own (e.g. material and immaterial entities), dependent continuants exist only in the context of some other entity (e.g. qualities). For example an anatomical entity is independent while the quality size is dependent on a bearer (e.g. an anatomical entity or another object).
- **classes vs. instances:** Classes (or universals) represent the abstract concepts while instances the particular objects. This is the common approach of ontologies in OWL.

These distinctions are reflected by the BFO class hierarchy shown in figure 4.1: Under the root class `entity`, there are two subclasses `continuant` and `occurrent` and further the distinction of independent and dependent continuants. Independent continuants are e.g. material and immaterial entities. Dependent continuants are distinguished as *generically* or *specifically* dependent continuants. The main distinction between a specifically and a generically dependent, is that specifically dependent continuants S depend on a some (specific) independent continuant during the whole course of existence of S while generically dependent continuants

depend on one or more other entities [12c]. Specifically dependent continuants are e.g. qualities (such as a color or a size). Examples of generically dependent continuants are all information artefacts (such as files, documents, measurements etc.)

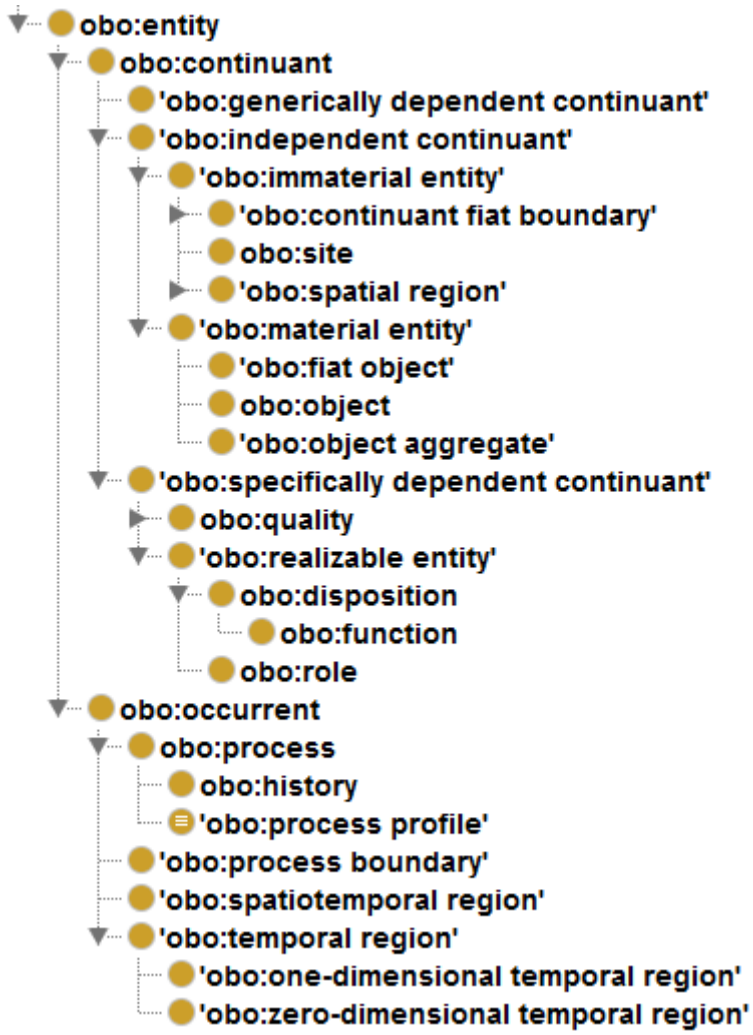


Figure 4.1.: The main part of the class hierarchy of the Basic Formal Ontology version 2 (BFO-2) shows the distinction between continuants and occurrents and between dependent and independent entities.

Reuse of BFO On the first glance the distinction of dependent and independent continuants defined by BFO seem to have small relevance for practical applications. However, since the class hierarchy of the BFO is the backbone of other ontologies of the OBO library – and thus provides the basis for the integration of different ontology modules listed below – all classes of BFO-2 are reused. However, as described in 2.2.2 the temporal relations defined by BFO-2 are still controversially discussed in the community and cannot be considered to be stable. Further, they would make the model more complicated and less easy to use. Thus, it

was decided to follow the OBI approach in using only the BFO-2 classes, but no temporalized relations. For this purpose, MCI imports only the official "classes only" OWL file [12d]. Additionally 14 out of 36 object properties from BFO-2 that are not temoporalised.

4.2.2. Relation Ontology

The Relation Ontology (RO) defines 24 object properties which are very general and used by most of the OBO ontologies. For instance, RO contains the properties `has part`, `has quality`, `located in` and `participates in`. Most of the properties defined by RO have corresponding inverse properties. [Smi+07]. RO is designed to be used with BFO and the range of several object properties is specified by BFO classes.

Reuse of RO: All 24 object properties defined in the RO core version [15o] are reused.

4.2.3. Phenotypic Quality Ontology

The Phenotypic Quality Ontology (PATO) [14n] defines 1570 classes with the root class `quality` as well as 22 object properties and 10 annotation properties. A phenotype is defined as a *"collection of characteristics that arise through the expression of the genes of an organism, in an environment"* [Mun+07] such as color, shape or size. In other words, a quality is a *characteristic* of some entity and PATO contains corresponding classes for these qualities which can be used to describe these phenotypes. In PATO a `quality` is defined as *"a dependent entity that inheres in a bearer by virtue of how the bearer is related to other entities"* [14n]. PATO is meant to be used together with reference ontologies such as the FMA, Uber-Anatomy Ontology (UBERON) or Gene Ontology (GO). As explained in [Was+09], a *"description of an individual phenotypic character can be recorded using a bipartite 'EQ' (Entity + Quality) method, where a bearer entity (such as an anatomical part, cellular process, etc.) is described by a quality (such as small, increased temperature, round, reduced length, etc.)."* That is, a description of a phenotype is constructed by combining (anatomical) entities with at least one quality. For instance, the Human Phenotype Ontology (HP) uses PATO in combination with the UBERON to describe abnormal anatomical structures for humans such as an `hp:enlarged spleen` by combining `pato:increased size` and `uberon:spleen`. Since PATO is a *cross-species* phenotypic quality ontology it provides the basis to integrate phenotype information from different species [Mun+10], which shows the advantage of a clear separation of qualities and the bearer entities. However, since the qualities defined by

PATO are very general, they can be used to describe clinical findings as well. Under the root class `quality`, PATO distinguishes between `qualitative` quality, `process quality`, `physical object quality` as well as `increased quality` and `decreased quality`. Subclasses of `physical object quality` are for example `morphology` (with subclasses `shape`, `size`, length etc.), `physical quality` (with subclasses `mass`, `energy`, `color` etc.) and `functionality` (with subclasses `active`, `inactive` etc.). The qualitative qualities are of special interest since they can be used to classify clinical findings. For instance `normal` and `abnormal` (with subclass `pathological`) and logically defined classes for `decreased quality` and `increased quality` provide terms to describe *deviation* from `normal`. Other qualitative qualities are e.g. `count`, `magnitude` or `intensity`. The class hierarchy of increased and decreased qualities is inferred by using logical definitions for these qualities. For instance `increased length` is a subclass of `increased size` due to the fact that `length` is a subclass of `size`.

Reuse of PATO: Since the qualities defined by PATO are very general, they can be used to describe not only phenotypes, but also findings. As shown in figure 4.3, PATO is used in combination with anatomical entities to describe clinical findings. Even though many of the qualities defined in PATO could be relevant in general, the import is kept as small as possible and only those qualities are included which are actually needed to model the data of the use cases. So only a small subset of the PATO qualities shown in figure 4.2 was reused. In particular, physical object qualities for describing morphologies such as size, mass, color or temperature were reused. These qualities are often used in finding descriptions. Especially the different size qualities are used in the radiology and pathology domain. The classes `increased quality` and `decreased quality` and corresponding logical axioms were also reused. In PATO, the subclasses of increased and decreased qualities repeats almost the entire class hierarchy under `physical object quality` and thus violates the DRY principle (see the notes on ontology design patterns in section 2.1). The object properties used in the logical definitions of the reused qualities, namely`different in magnitude relative to` with sub-properties `increased in magnitude relative to` and `decreased in magnitude relative to` are included. These properties are used in MCI to related particular findings of one patient to normal specifications, i.e. it provides the link between patient data and medical knowledge. All other properties are *not* imported because they are either not related to the classes that were reused or similar ones were defined in RO and thus are not specific for PATO. In total 36 classes and 3 properties were reused from PATO as well as the axioms for increased and decreased qualities.

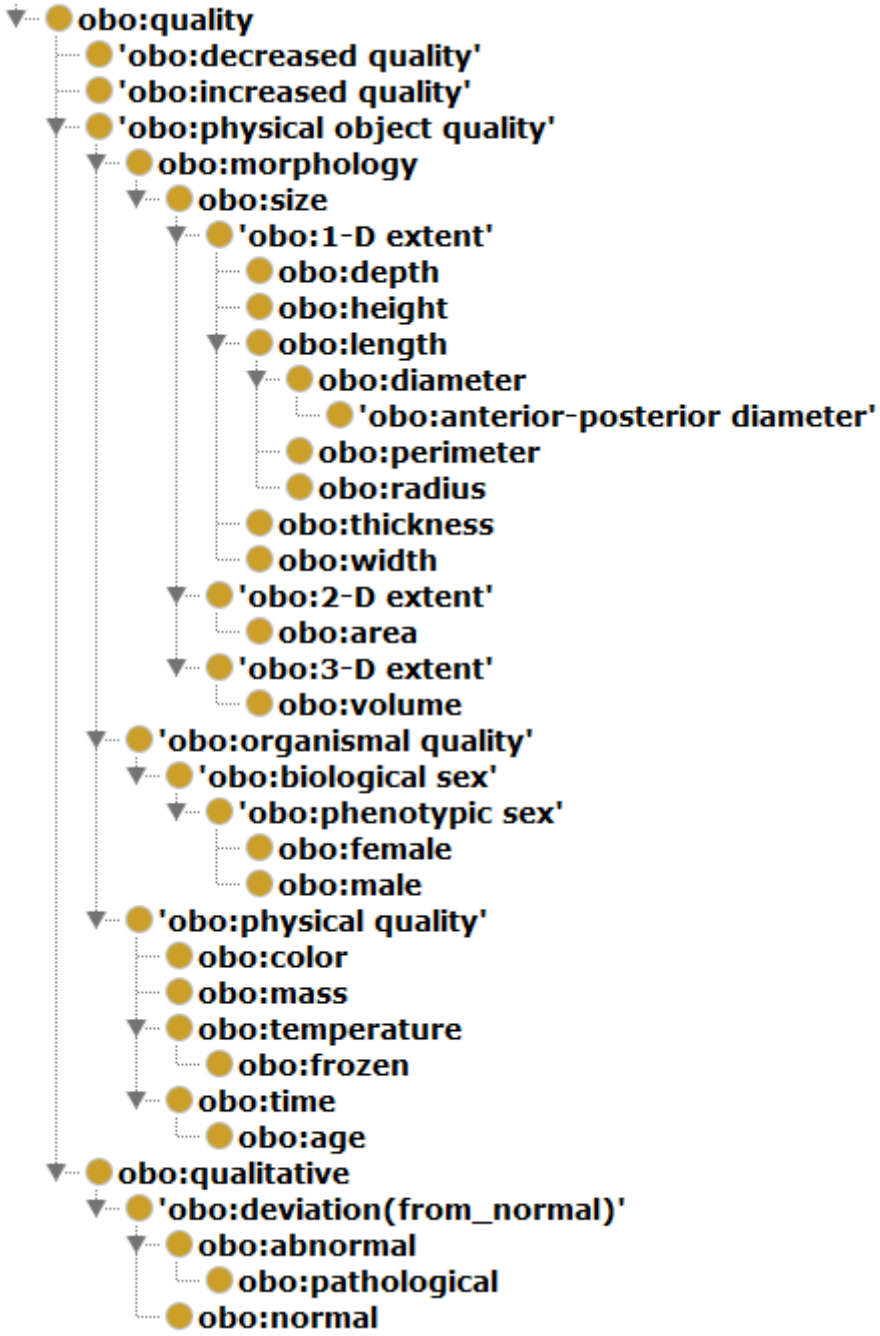


Figure 4.2.: Classes imported from PATO.

4.2.4. Units Ontology

The Units Ontology (UO) [14v] defines 331 classes and contains 379 classes in total. It is a *"comprehensive ontology for the standardization of units of measurements in the biomedical domain"*, which was developed in the context of PATO to describe*"qualitative and quantitative observations in biology"* [GSH12]. UO defines classes for *units* like `gram`, `ampere`, `micrometer` etc. as well as for common unit

Table 4.2.: Reused entities from the Units Ontology.

Class	Instances
concentration unit	gram per liter, mass volume percentage, ...
frequency unit	hertz
length unit	meter, centimeter, millimeter, micrometer, nanometer, ...
mass unit	kilogram, gram, milligram, microgram, nanogram, picogram
temperature unit	degree Celsius, degree Fahrenheit, kelvin
time unit	year, month, week, day, hour, minute, second
volume unit	cubic decimeter, cubic centimeter liter, milliliter, microliter, nanoliter, ...
volumetric flow rate unit	microliters per minute

prefixes like centi, milli etc. Units have a textual definition, a label and sometimes a synonym or abbreviation. The units are hierarchically structured into unit categories: For example, the *unit* gram per mole is a subclass of molar mass unit which is a *category* of units and a subclass of mass unit. UO contains exactly one object property is_unit_of, which is used in logical definitions (object restrictions) of units by the corresponding PATO qualities they describe. For instance, length unit has subclasses meter, centimetre, millimetre etc. and is defined as

uo:length unit rdfs:subClassOf (uo:is_unit_of SOME pato:1-D extent)

Reuse of UO: In OWL, object properties relate instances and since measurements need to be related with corresponding units, these units have to be instances – not classes. UO entities for size, mass, concentration and ratio are reused. Their subclasses were transformed to instances so that they can be used for instances of measurements. That is, centimetre is not a subclass of length unit but an instance of it. Here the example of developers of the Ontology for Biomedical Investigations is followed. The logical definitions for classes such as length unit are preserved, so that one can check that the units are correctly used to describe PATO qualities. table 4.2 gives an overview of the reused units from UO. In total, 74 classes and 1 object property from UO were reused. Since the OntoFox [14j] web service was used to create the module based on the standard (class-based) UO-version, 63 classes needed to be transformed to instances in a post-processing step. Thus, the unit categories are represented as classes (under the IAO class measurement unit label) while the units itself are instances of the corresponding classes.

4.2.5. Information Artefact Ontology

The Information Artefact Ontology (IAO) [11d] is based on BFO-2, defines 109 classes and contains 180 classes in total. It defines 16 object properties (in total 49), 4 data properties (4 in total) and 19 annotation properties (66 in total). The IAO models the domain of *information entities*, i.e. entities that are *about* other entities. The most important class of the IAO is the `information content entity` which is defined as *"an entity that is generically dependent on some artefact and stands in aboutness to some entity"* [11d]. That is, an instance of `information content entity` has to be related to some instance of `bfo:entity` by the `iao:is about` relation. For example a report *is about* some specific patient. The class `information content entity` has 102 subclasses beyond which e.g. `data item`, `measurement datum`, `document` and `document part`, `report` etc. Further IAO defines classes under `obi:planned process` such as `documenting`. Three data properties are defined to describe three dimensional coordinates (`has x coordinate value` etc.) and one data property `has measurement value` is defined to relate measurements to the respective values.

Ontology-Metadata Besides these classes which are contained in the main IAO file, annotation properties are defined in a separate Ontology-Metadata file [11c]. Ontology-Metadata defines 6 classes (in total 16), 28 annotation properties (in total 30) and 18 instances. The annotation properties can be used to further describe ontology resources. For example, the properties `definition`, `definition source` or `example of usage` are used to provide textual definitions and examples. The properties are widely used by ontologies of the OBO library.

Reuse of IAO: Since IAO is fully in scope of MCI, it is completely reused - discarding only obsoleted or deprecated classes. More specifically, the latest IAO release from March 2015 was reused, which is based on the BFO-2 class hierarchy. MCI is an information model, i.e. many classes of its extension are about other entities and thus will be subclasses of `information content entity`.

4.2.6. Ontology for General Medical Science

The Ontology for General Medical Science (OGMS) [14s] defines 76 classes and contains 126 classes in total. The latest OGMS version (from June 2014) is based on the BFO-2 and all object properties are imported from BFO-2. It does not contain any data properties and all 52 annotation properties contained in OGMS are imported from other ontologies. OGMS defines classes for basic clinical entities such as `diagnosis` defined as *"the representation of a conclusion of a diagnostic process"*,

`symptom` defined as *"a quality of a patient that is observed by the patient or a processual entity experienced by the patient, either of which is hypothesized by the patient to be a realization of a disease"*,`sign` defined as *"a quality of a patient, a material entity that is part of a patient, or a processual entity that a patient participates in, any one of which is observed in a physical examination and is deemed by the clinician to be of clinical significance"* and others [14s]. Detailed definitions for basic OGMS concepts such as disorder and disease can be found in [SCS09]. Note that even though OGMS is based on BFO-2 the do not make use of the BFO-2 properties in logical definitions of OGMS classes. OGMS contains seven classes from IAO such as `iao:information content entity` and `iao:data item`. In particular `clinical finding` defined as *"a representation that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof"* is a subclass of data item.

Reuse of OGMS: Even though there are no logical definitions for the classes defined by OGMS the class hierarchy and the textual definitions provide a very useful starting point for MCI . Since OGMS defines many useful classes with high relevance for the use case data, it was the initial reason to base MCI on OBO ontologies. OGMS is fully in scope of MCI , so all classes were reused. The general pattern of clinical findings is shown in figure 4.3.

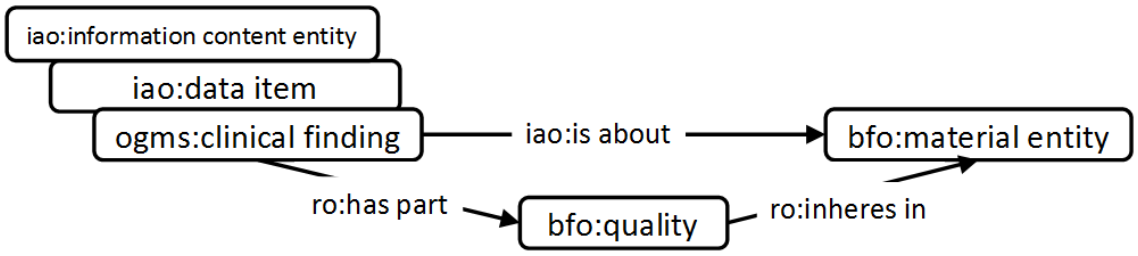


Figure 4.3.: The basic pattern of the representation of a clinical finding: A clinical finding is a data item which is about some material entity and describes some a quality that inheres in that entity.

4.2.7. Ontology for Biomedical Investigations

The Ontology for Biomedical Investigations (OBI) [14l] defines 2216 classes (in total OBI contains 2739 classes), 23 object properties (in total 88), 2 data properties (in total 7) and 4 annotation properties (in total 50). It is an integrated ontology for the description of life-science and clinical investigations [14r]. OBI is actively developed within "a cross-community effort and provides an integrative framework" [Bri+10]. It is based on BFO and can be considered as one of the core ontologies within the OBO library. The main focus of OBI lies on modeling

investigations as processes: About the half of all classes defined by OBI are subclasses of `bfo:process` – about 630 of them define subclasses of `assay` which is defined as *"planned process with the objective to produce information about the material entity"*. Further, 620 classes are defined under `bfo:material entity`, mostly as subclasses of `processed material` (e.g. `molecular-labeled material` or `blood plasma specimen`) which allows capturing the different changes of materials along an investigation process. OBI also defines about 100 roles such as `donor`, `antigen role` or `manufacturer role` to precisely express the participants of a biomedical investigation and 70 classes under `iao:objective specification`. Most important for us is the definition of `value specification` described in the next paragraph.

OBI Value Specifications The main motivation to introduce the class `value specification` is that value-unit pairs appear in very different contexts [Ove13]. For instance "1.6 cm" might denote the following (adapted from [Ove13]):

- The size of a lymph node of some patient at some time.
- The maximal size of some lymph node over time.
- The upper bound used for classifying of inguinal lymph nodes as 'normal'.
- The predicted measurement of some lymph node in the future.
- The length of a stenosis.

OBI defines `value specification` as *"an information content entity that specifies a value within a classification scheme or on a quantitative scale"* [14l]. This provides a shared structure for all these different cases listed above. The class `value specification` has subclasses `scalar value specification` which is *"a value specification that consists of two parts: a numeral and a unit label"* and a subclass `categorical value specification` which is *"a value specification that is specifies one category out of a fixed number of nominal categories"* [14l].

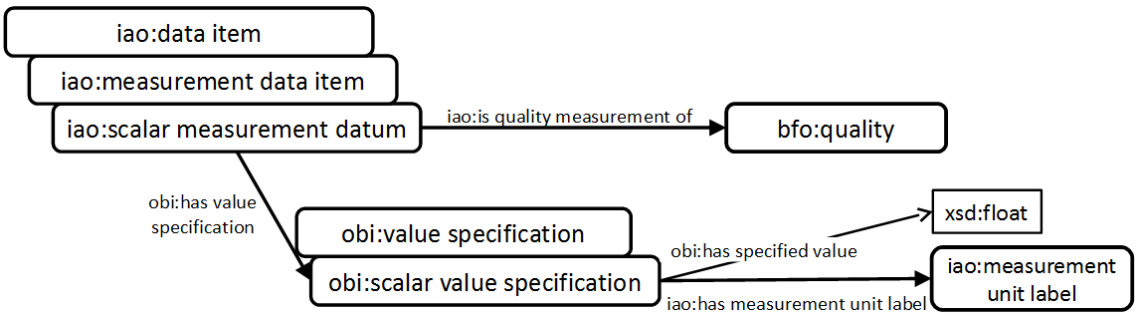


Figure 4.4.: The pattern of a scalar measurement datum expressed by a scalar value specification and a corresponding quality.

Reuse of OBI: The separation of the value-unit pair from its context and semantics (measurement, average value, upper bound etc.) leads to more precise modeling and allows easier comparison. The class `value specification` is used e.g., to represent measurements. The basic pattern for measurements with value specification is shown in figure 4.4, but also in a knowledge model where the typical size of anatomical entities is stored (section 6.1). Even though the representation of a simple measurement involves one intermediate entity, the benefits of this separation showed to be useful in implementations.

4.2.8. Foundational Model of Anatomy

The Foundational Model of Anatomy (FMA) is the most comprehensive reference ontology for anatomy. It contains 83,319 classes (version 3.1, released 03/03/2010) for material and immaterial anatomical entities and relations between them. For details on the FMA it is referred to [RJ07].

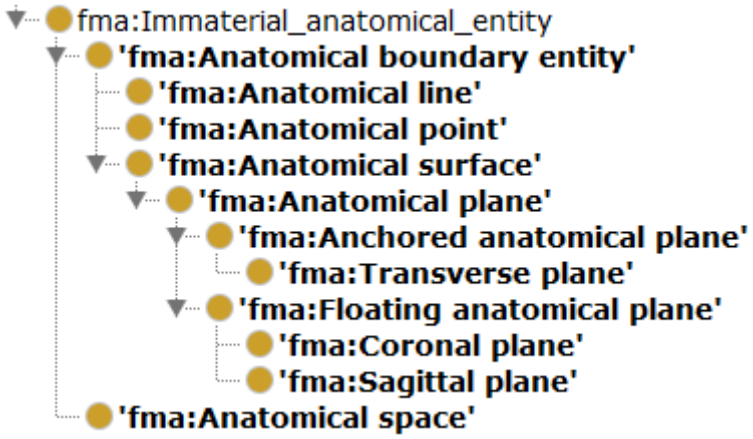


Figure 4.5.: Imported classes from the Foundational Model of Anatomy (FMA).

Reuse of FMA: Only very few high-level classes are imported to represent location and orientation of findings within the body. More precisely, the class `Anatomical boundary entity` with its subclasses `Anatomical plane`, `Anatomical line` and `Anatomical point` as shown in figure 4.5 are imported. In total the FMA contains 8059 subclasses of `Anatomical boundary entity`. However, only the high-level classes are needed. For instance there are 79 subclasses of `Anatomical plane`, but only the class for anatomical plane and the three main body planes `Transverse plane`, `Sagittal plane`, `Coronal plane` are needed for specifications of size measurements. Annotation properties of FMA are mapped to existing properties if corresponding ones exist already. For example the annotation property `fma:definition` is mapped to `iao:definition` so that textual definitions a represented in a uniform way in MCI.

4.2.9. Ontology of Medically Related Social Entities

The Ontology of Medically Related Social Entities (OMRSE) [13b] is based on BFO-1.1 and defines 44 classes (in total 94) and one object property (in total 8). OMRSE does not contain any data properties and does not define own annotation properties. It is meant to be related with OGMS and defines classes for organizations and for various clinical roles. For instance an organization is defined as *"a continuant entity which can play roles, has members, and has a set of organization rules ..."* [13b] and has subclasses such as governmental organization or sub-national entity. The main contribution of OMRSE however is the definition of *roles*: Under the root class `role in human social processes` it defines e.g. `human healthcare role` as *"a role that is realized by health care processes such as seeking or providing treatment for disease and injury, diagnosing disease and injury, or undergoing diagnosis"* [13b]. Subclasses of this role (see figure 4.6) are e.g. `health care provider role` defined as *"role inhering in an organization or human being that is realized by a process of providing health care services to an organism"* and `patient role` defined as *"a role that inheres in an organism as the recipient of a health care service"*.

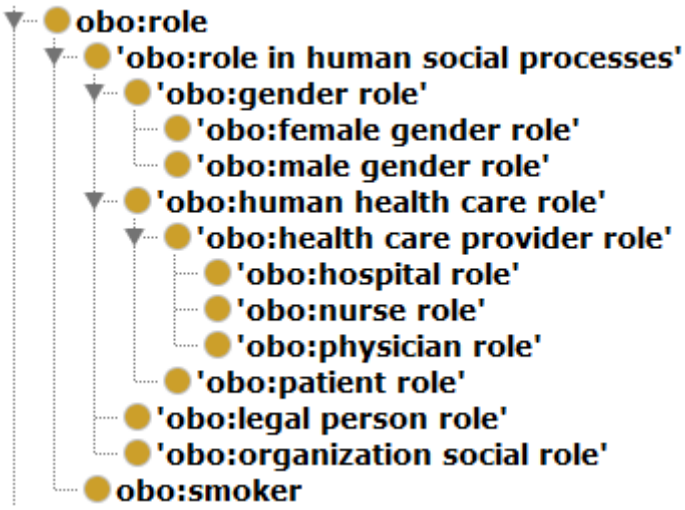


Figure 4.6.: OMRSE subclasses of *human health care role*.

Reuse of OMRSE: For the used use case data, the most important classes are those under `role in human social process` which is a subclass of `bfo:role`. Further OMRSE imports the complete subclass path from National Center for Biotechnology Information (NCBI) to include `ncbi:Homo sapiens`. This class is included by simply making `ncbi:Homo sapiens` a direct subclass of `obi:organism`, skipping all 28 intermediate NCBI classes such as `ncbi:Eukaryota`, `ncbi:Metazoa`, `ncbi:Mammalia` etc., since they are of no relevance for data of the use cases. In total, 16 classes and one object property were reused from OMRSE and aligned to the BFO-2 class hierarchy.

4.2.10. Other Reused Ontologies

Besides the ontologies listed above, a few single classes from other ontologies of the OBO library are indirectly reused. Namely from NCBI (homo sapiens, bacteria, viruses etc.), Vaccine Ontology (vaccine), Gene Ontology (biological process), Chemical Entities of Biological Interest (molecular entity, nucleic acid etc.) and Cell Ontology (cell). From RadLex the *Hounsfield unit* was reused.

Some properties from other established schemes were reused. For example,, Dublin Core (dc-terms [14e]) defines relations to express meta-data of resources. Protégé-dc [03] is a subset of the dc-terms [14e] and defines the relations of the /elements/1.1/ namespace [14e] through annotation properties. Thus, one can express meta-data of any resources in the ontology with dc-terms. Namely by using the following 15 annotation properties: `contributor`, `coverage`, `creator`, `date`, `description`, `format`, `identifier`, `language`, `publisher`, `relation`, `rights`, `source`, `subject`, `title` and `type`. Since Protégé-dc is widely used within the OBO library, all 15 annotation properties were included. Additionally several classes from the NLP Interchange Format (NIF) ontology [14i] are reused to represent textual entities such as a `Sentence` or `Word`.

The Open Annotation data model (OA) [13a] is used to represent annotations to clinical data in a coherent way. Since the annotation pattern of OA is very general, one can have basically the same schema for text and image annotations. Details on the role of OA for integration of annotations is given in chapter 7. The basic pattern of an annotation is shown in figure 4.7.

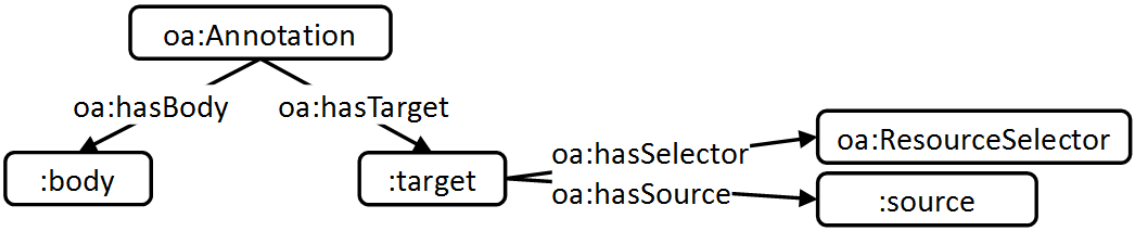


Figure 4.7.: The basic pattern of annotations using classes and properties from the Open Annotation data model.

4.2.11. Summary of Reused Ontologies

In the previous sections the main imports of MCI were described. To obtain an impression of the impact of the different ontologies to the foundation of MCI, the number of reused classes and properties are listed in table 4.3. Note that for UO some classes were transformed to individuals as described in section 4.2.4. In total MCI reuses 447 classes, 83 object properties, 15 data properties, 75 annotation properties and 124 named individuals from imported ontologies.

Table 4.3.: The foundation of MCI, analyzed by number of reused classes, properties and individuals for the main ontology imports.

ontology	classes	object properties	data properties	annotation properties	individuals
BFO-2	36 (36)	14 (36)	0 (0)	2 (2)	0 (0)
PATO	35 (1570)	2 (22)	0(0)	0 (10)	0 (0)
UO	11 (371)	1 (1)	0(0)	0 (12)	63 (0)
IAO	109 (102)	14 (16)	4(4)	29 (19)	18 (19)
OGMS	76 (76)	0 (0)	0 (0)	0 (0)	0 (0)
OBI	114 (2216)	7 (23)	1 (2)	2 (4)	0 (0)
OMRSE	16 (44)	1 (1)	0 (0)	0 (0)	0 (0)
OA	11 (23)	7 (17)	4 (9)	0 (0)	(13) 13
others	39	37	6	42	43

Despite the orthogonality approach of the OBO library classes were defined in different ontologies. For these classes mappings were used to align these entities. For instance, the following mappings were defined:

`obi:performing a diagnosis` = `ogms:diagnostic process`,
`obi:disease` = `ogms:disease`
`obi:healthcare provider role` = `omrse:healthcare provider role`
`obi:patient role` = `omrse:patient role`

4.3. MIREOTing

It is often the case that one is only interested in a part of some ontology. For instance to avoid very large imports, or import only stable parts of ontologies which are under development. It could also be the case that one agrees only with a certain subset of some ontology but not with all defined classes or properties. Further two imported ontologies might be created by different design principles or based on different upper level ontologies. If these ontologies are not aligned, importing both might *"lead to inconsistencies or unintended inferences"*[Cou+09]. Since OWL allows only the import of a complete ontology file, a mechanism to reuse parts of ontologies is needed. The Minimal Information to Reference an External Ontology Term (MIREOT) [Cou+09] is a set of guidelines to support reuse of parts of other ontologies: To consistently include entities from an external ontology one needs at least the *source ontology URI*, the *source term URI* and the *target direct superclass*[Cou+09]. OntoFox [Xia+10][14j] is a web-service that implements the MIREOT guidelines so that one can automatically create ontology modules. Based on a set of entities and properties one can extract modules of different scope from the ontology. For instance one can get all subclasses of some given entity, together with all their annotation properties such as labels, comments and textual definitions. But one could also recursively include all axioms mentioning the respective entity. The OntoFox web-service was used to create imports for all reused ontologies from the OBO library.

Note that the focus of MIREOTing is the reuse of *terms*. Even though one can specify to which extent axioms should be included in the module one has to accept that inference might be not complete when using only the extracted module instead of the complete ontology. There are other modularization approaches which guaranty that the same meaning and inferences are obtained using the module or the complete ontology. For instance [Cue+07] offer an approach to extract modules with minimal size. In the context of ontologies of the OBO library however MIREOTing is accepted practice and for us the reuse of *terms* is most important.

5

The Model for Clinical Information

The Model for Clinical Information (MCI) provides the basis for data integration and knowledge exploration, i.e. the structural concepts for the representation of clinical data. This is, at first place meta-information about the patient characteristics like diagnoses and findings (e.g. target lesions that were determined for a cancer patient) as well as provided examinations, procedures and therapies. MCI defines classes and properties and provides the basis to infer the changes of the patient's health status over time (i.e. the change of diagnoses, findings, observations, symptoms and signs). These changes might then be analyzed in the context of provided examinations and procedures e.g. in order to measure their effectiveness. To make this possible MCI integrates various types of information. The following *four dimensions* of medical data were identified, that need to be integrated:

1. **Administrative data about clinical processes:** treatments, procedures, examinations etc.

2. **Data about the health status:** diagnoses, findings, observations, symptoms and signs, allergies, basic demographic information such as age, gender etc.

3. **Longitudinal integrated data:** to evaluate the development of the health status, data needs to be integrated along the time axis. For instance findings from consecutive examinations.

4. **Medical knowledge:** Firstly, knowledge to classify and interpret clinical data, such as classification of findings as normal or abnormal. Secondly, diseases information, such as statistics, relation to symptoms and clinical findings etc. need to be linked.

This chapter presents the Model for Clinical Information, which covers all four dimensions for the data of the use cases. The knowledge models are described in detail in chapter 6. In the previous chapter, the foundation of MCI was described. Many useful classes and properties are thus already defined. The foundation consists of 447 classes, 83 object properties, 15 data properties, 75 annotation properties and 124 named individuals. Even though the most basic terms are

defined by the ontologies forming the foundation of the model, the representation of clinical data needs more detailed entities. On the one hand, *classes* need to be added, to provide more granular types and on the other hand *properties* for relating instance data. MCI defines 103 classes, 23 object properties, 18 data properties, 12 annotation properties and 5 named individuals. Together with the model foundation MCI contains in total 550 classes, 106 object properties, 33 data properties, 87 annotation properties and 129 named individuals.

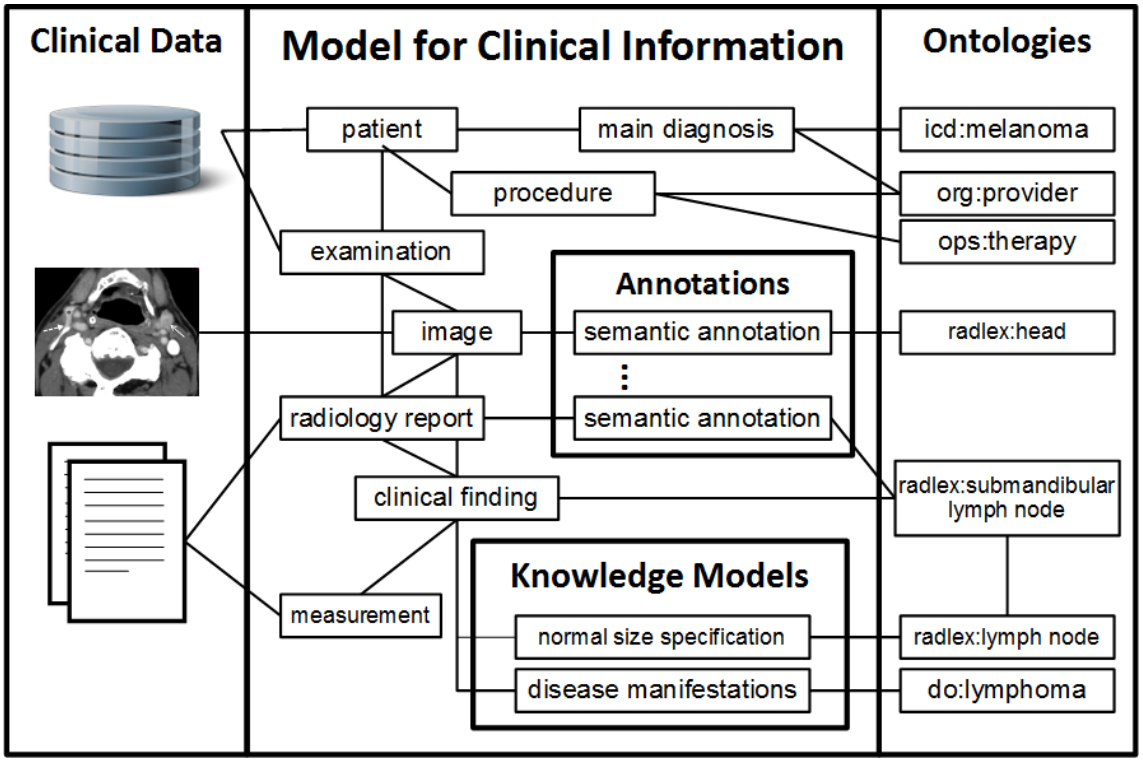


Figure 5.1.: The high level classes of MCI and their relations.

In the following, describe the general modeling decisions of the approach are described in section 5.1. MCI is an *information model*, i.e. it provides the classes and properties to *structure* clinical data. The vocabulary however is taken from reference terminologies. In section 5.2, it is described how these reference terminologies are integrated with MCI. Then, the classes, properties and individuals defined by MCI are presented by listing their label and textual definition. Several example representations are provided to allow a better understanding of the usage of the defined classes and properties.

5.1. General Modelling Decisions

In this subsection the general modeling decisions and principles are described.

5.1.1. Definitions

According to best practices, for each entity at least one human readable label and textual definition should be given. Textual definitions of classes have an Aristotelian form "an A is a B that C", where A is a subclass of B, distinguished by C. For example a `mci:size measurement` has the following textual definition: "A size measurement is a measurement that specifies a size quality". Logical definitions are specified if possible. To allow a better understanding, examples of their usage are additionally given. For properties which relate specific classes, corresponding domain and range axioms are defined.

5.1.2. Don't Repeat Yourself Principle

The Don't Repeat Yourself (DRY) principle commonly found in software engineering is also applicable to semantic modeling. It generally means that one should avoid duplication of code. In the context of semantic modeling it means to avoid duplication of class hierarchies RadLex for example contains more than 250 subclasses of *lymph node*. The finding *lymphadenopathy* (i.e. enlarged lymph node) has however only one subclass. That is the subclass hierarchy under *lymph node* is not repeated. Sometimes duplication is however convenient regarding later querying: For example, PATO defines a class hierarchy of qualities such as weight, size, length, diameter, craniocaudal diameter etc. (see previous figure 4.2). Measurements specify these qualities and there are classes for measurements of specific qualities: For instance, IAO defines subclasses of `obo:measurement datum` such as `obo:length measurement datum` or `obo:mass measurement datum` – with corresponding logical definitions so that measurements are automatically classified. To retrieve all length measurements one can thus simply query for

```
?lmd a 'obo:length measurement datum' .
```

insted of

```
?lmd a 'obo:measurement datum';
     'obo:is quality measurement of'/a obo:length .
```

This duplication is convenient to allow *shorter* queries, however one should avoid the creation of specific measurements for *all* qualities of PATO. By defining, additionally to length measurement, a *diameter measurement*, *craniocaudal diameter measurement* etc. – one would duplicate the entire class hierarchy of PATO.

5.1.3. Focus on Class Hierarchies

In many ontologies the *class hierarchies* are the most important structure – simply because the subclass property is used more than any other property. These detailed class hierarchies allow to precisely define the type of individuals. In OWL it is further possible to define also *hierarchies for properties*. That is, one can define sub-properties of some given property. For instance a property *has examination modality* (domain: examination; range: examination modality) could have a sub-property *has imaging modality* (domain: radiology examination; range: imaging modality). Usage of sub-properties allows more specific domain and range definitions for properties and also more precise representation of relations between particular entities (see figure 5.2): One can easily retrieve one specific examination modality by using the subproperty.

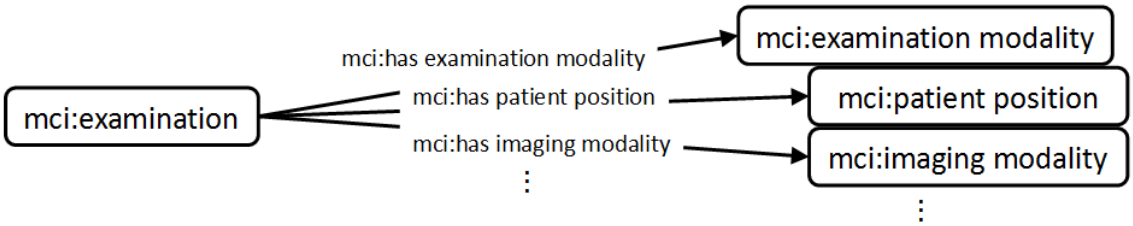


Figure 5.2.: Representation variant of examination modalities with sub-properties.

However, many properties and especially property hierarchies make an ontology more complex, more difficult to modify. For instance, modification of domain and range for a sub-property has to respect the domain and range specification of the super-property. Further, as shown in figure 5.2, sub-properties tend to *repeat* the class hierarchy – violating the DRY principle described above. Further, many similar properties can be confusing: A user definitely has problems to select the right property between *part of*, *constitutional part of*, *member of*, *contained in*, *located in*, *part of at all times* etc.!

The main difference to a more simple representation – without sub-properties as shown in figure 5.3 – however concerns querying and required reasoning level: Retrieval all examination modalities can be done in both cases by the following query:

```
SELECT ?examination ?modality
WHERE { ?examination 'mci:has examination modality'  ?modality. }
```

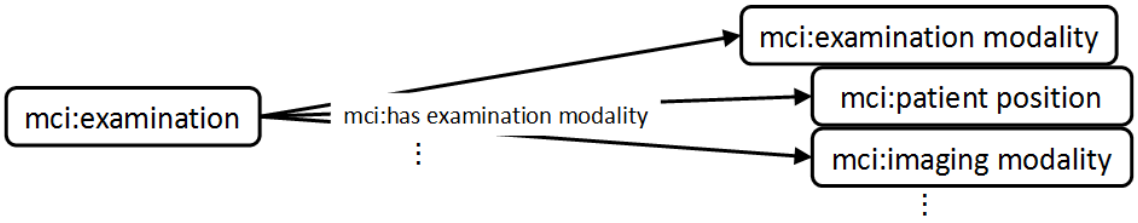


Figure 5.3.: Representation variant of examination modalities without sub-properties.

The difference, is that *with sub-properties* one needs a running reasoner while for the simpler representation no reasoner is needed. On the other hand – in case of the simple representation – retrieval of one specific modality requires one extra triple in the query that specifies the type of `?modality`. Due to the enhanced complexity of sub-properties, in MCI the focus is set on class hierarchies and the definition of sub-properties are kept to a minimum.

5.2. Reference to Terminologies

The representation of heterogeneous clinical data using terminologies has to be realized through links from instance data (individuals) to terminologies, which often represented through class hierarchies. These links can be established either using `rdf:type`, a data property or an annotation property. Data properties have the disadvantage that the referenced class is represented by a string for the URI or the ID. Annotation properties are very convenient, since they can be used to relate any resources, however they are disregarded during the reasoning process. Thus, `rdf:type` is the preferred way, to refer to terminologies if one intends to run inference for automatic classifications of individuals. If there is a class in the terminology exactly representing the required type, then one `rdf:type` relation to that class is sufficient (see figure 5.4).

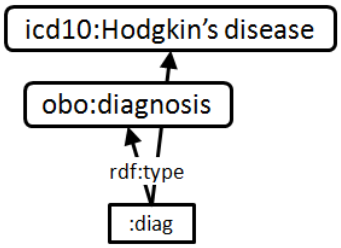


Figure 5.4.: Representation of a diagnosis with a reference to the International Classification of Diseases (ICD).

The code for that data is the following:

```
:diag rdf:type obo:OGMS_0000073 , icd10:C81 .
```

In combination with the reference terminology such as ICD reasoning mechanisms can be easily applied. For instance it could be inferred by a standard RDFS reasoner that the diagnosis represents a malignant neoplasm:

```
:diag rdf:type 'icd10:Malignant neoplasms'.
```

5.2.1. Post-Coordination

If a statement is more complex it might be the case that there is no single corresponding class in the terminology. Then it is necessary to post-coordinate classes from the terminology, e.g. for further specification of the intensity, severity or other qualities of some entity. The main challenge in post-coordination is that the coordinated entities can be correctly classified in the referenced terminology. For instance the Human Phenotype Ontology (HP) contains a class `hp:Lymph node hypoplasia` with the following definition:

```
'hp:Lymph node hypoplasia' = 'obo:has part' SOME (obo:hypoplastic AND
                             ('obo:inheres in' SOME 'obo:lymph node'))
```

`hp:Lymph node hypoplasia` has no subclasses. Expression of clinical findings such as "abdominal lymph node hypoplasia" requires to follow the pattern of the definition above, so that the findings gets correctly classified by HP.

```
:finding1 a 'obo:clincal finding' ;
          'obo:has part' [
              a obo:hypoplastic ;
              'obo:inheres in'/a 'obo:abdominal lymph node'
              ].
```

Since `obo:abdominal lymph node` is a subclass of `obo:lymph node`, it is automatically inferred that `:finding1` is of type `hp:Lymph node hypoplasia` and also of all super-classes such as `hp:Abnormality of the lymph nodes`. In some cases one has a coordinated representation even when one single class is available. For instance HP contains the class hp:Splenomegaly (enlarged spleen) – however this clinical finding is often represented by referencing the `obo:increased size` and `obo:spleen`. The classification of this finding as a `hp:Splenomegaly` is a useful inference one can make when one uses the post-coordination pattern of HP. In RadLex there is also a corresponding class `radlex:enlarged spleen` however without logical axioms. Thus, a finding would not be classified as a `radlex:enlarged spleen` directly. Correct classification of post-coordinated entities depends on the availability of logical definitions in the reference ontology and the use of the same coordination pattern. Since MCI is based on OBO ontologies and HP is used to classify findings, the corresponding patterns shown above are followed.

5.2.2. Bindings to Terminologies

Consistent reuse of reference such as ICD, RadLex, FMA, HP and others needs to be ensured using ontology *bindings*. The different binding options below will be explained along the following example, that a radiology examination has a certain examination modality like some specific patient position during the examination. MCI defines a class `patient position` but no concrete patient positions since these entities are already available in existing ontologies – and reused. For instance RadLex defines the class `radlex:patient position` and more than 40 different subclasses. As in the above example, the `rdf:type` property is used to reference one of these classes (figure 5.5).

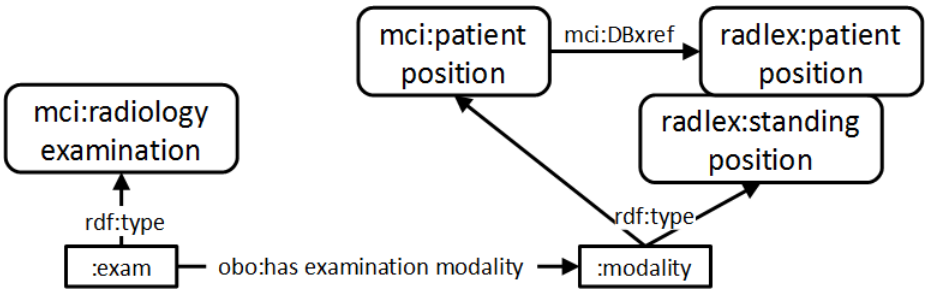


Figure 5.5.: Representation of a specific patient position during an examination by referencing a corresponding RadLex class.

To ensure that an instance of `mci:patient position` has a type reference only to RadLex classes which are subclasses of `radlex:patient position` (such as `radlex:standing position`), but not to other RadLex classes, one has to *bind* the MCI classes to the corresponding classes of reference ontologies. That is a binding of `mci:patient position` to `radlex:patient position` is needed. It is realized by the following `oboInOwl:dbxref` annotation property. In general the binding serves two purposes: consistency verification and restriction of selectable entities in structured reporting. The verification is done by an SPARQL ASK query that checks for each referenced ontology that the referenced entities are subclasses of the binding of the MCI class. For example,, that all references to RadLex from instances of mci:patient position (such as `radlex:standing position`) are subclasses of `radlex:patient position`. MCI has bindings to ICD, German procedure classification (Operationen- und Prozedurenschlüssel) (OPS), the FMA and RadLex. In the case of ICD the class `ogms:diagnosis` is bound to the root class of the entire ICD. In summary, MCI is aligned to the different reference ontologies, but lets them provide the terminology to express clinical information.

5.3. Classes and Properties

In this section, the classes and properties defined by MCI are described, example are provided to illustrate their usage and the defined bindings and axioms are listed (if available). At the beginning, definition of general properties are given, before roles, diagnoses, examinations and reports are described. Then, more specific classes and properties are presented to describe images and image regions, qualities and units, range specifications, measurements as well as clinical findings. The representation of location information is described towards the end of this section.

5.3.1. General Properties

The different type of properties available are *object properties* which are used to relate individuals, *data properties* which are used to relate individuals with literals (such as a string or a date value) and *annotation properties* which can relate any resource with other resource or literal. Data properties are mainly used for representation of temporal information, a reference to identifiers and specification of numeric values, e.g. in a coordinate system. For instance, `has time stamp` is a generic relation that can be used for any entity, the other properties are more specific (as shown in figure 5.6. All of them have range `xsd:dateTime`.

mci:has timestamp A relation to timestamp entities.

mci:date of birth A relation between a human and the date time of his or her birth.

mci:start date time Relates a process with a date time that represents the start of the process.

mci:end date time Relates a process with a date time that represents the end of the process.

mci:creation date time The date time when an information content entity was created.

The following property is used to store a textual representation of some individual. For instance this property is used to link a *report section* to the corresponding text represented as a string.

mci:has textual representation Has textual representation relates some entity with a human readable string representation.

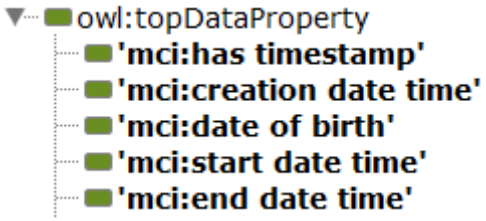


Figure 5.6.: Data properties defined by MCI to represent different date values.

Identifiers for specific entities are listed as sub-properties under `skos:notation` as shown in figure 5.7. In general it is avoided to define several properties for reports, images etc. and MCI contains only one such relation for unique identifiers.

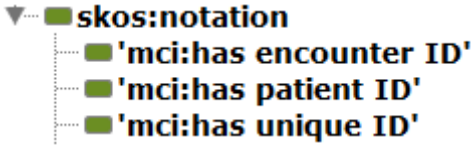


Figure 5.7.: Data properties defined by MCI to represent different identifiers.

5.3.2. Roles

OMRSE defines `obo:patient role` as *"A role borne by an organism being as the recipient of a health care service"*, but the foundation has no class for patient! To retrieve all patients one would need to query for `?patient 'obo:has role'/a 'obo:patient role'`. To allow more direct access to individuals of type *patient* a corresponding class with logical axioms is defined. Then all humans with a patient role are automatically classified as patients and one can retrieve all patients by simply querying for all individuals of type patient.

mci:patient A homo sapiens (human) who has a patient role.

```
mci:patient = 'obo:Homo sapiens'
              AND ('obo:has role' SOME 'obo:patient role')
```

OMRSE defines `obo:healthcare provider role` as *"a human health care role inhering in an organization or human being that is realized by a process of providing health care services to an organism"* [13b]. For capturing the administrative role of admission and discharge of patients, the following subclasses are defined additionally:

mci:admission role An admission role is a healthcare provider role that is realized by the admission of some patient to an inpatient hospital stay.

mci:discharge role An admission role is a healthcare provider role that is realized by the discharge of some patient.

mci:age at admission Relation between a patient and the patient's age at admission to inpatient stay some healthcare organization.
Range: `xsd:integer`

The age at admission is unique for each clinical encounter. This is possible since triples are stored in different named graphs for different clinical encounters as described in chapter 7.

5.3.3. Diagnosis

OGMS defines the class `obo:diagnosis` as *"the representation of a conclusion of a diagnostic process"* [14s]. To fully capture diagnosis information contained in the use case data, subclasses `main diagnosis`, `secondary diagnosis` and `working diagnosis` are required. The distinction between main and secondary diagnosis is officially defined by the German Coding Guidelines [14d]:

mci:main diagnosis: According to the German Coding Guidelines ("Deutsche Kodierrichtlinien") [14d] the diagnosis, which was established after analysis as that, which is mainly responsible for the inducement of the inpatient hospital stay is referred to as main diagnosis. The term "after analysis" refers to the evaluation of findings at the end of an inpatient stay. It has to be reported according to the 10. revision of the international statistical classification of diseases and related health problems ICD-10 GM.
Alternative terms: principal diagnosis, primary diagnosis.

mci:secondary diagnosis Diseases or complaints, which either exist from the beginning with the main diagnosis or develop during the stay in hospital are referred to as relevant secondary diagnosis (comorbidity and complication). A requirement for this is a diagnostic measure (operation and/or procedure) or a therapeutic measure or an increased need for care and/or monitoring.

mci:working diagnosis A working diagnosis is a diagnosis which has not been confirmed.

The classes `main diagnosis`, `secondary diagnosis` and `working diagnosis` are pairwise disjoint. An example for the representation of diagnosis information is given in figure 5.8. Since ICD is used for representation of the different types of diagnoses, the superclass `obo:diagnosis` is bound to the root class of the ICD10 class hierarchy.

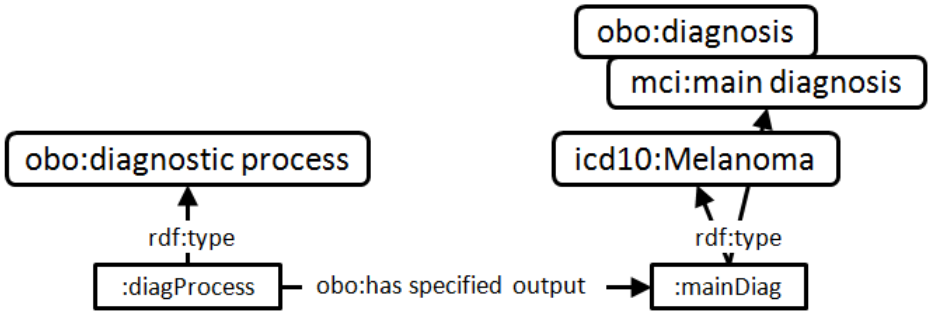


Figure 5.8.: A main diagnosis (here: melanoma) is the output of a diagnosis process.

5.3.4. Examinations

Examinations are one of the most important processes in healthcare. The model foundation provides the class `obo:process` with subclasses `obo:clinical history taking` (anamnesis), `obo:diagnostic process`, `obo:physical examination` and `obo:laboratory test`. To ease the retrieval of examination information the class `examination` is defined as an intermediate class and further subclasses for specific examinations are added (see figure 5.9).

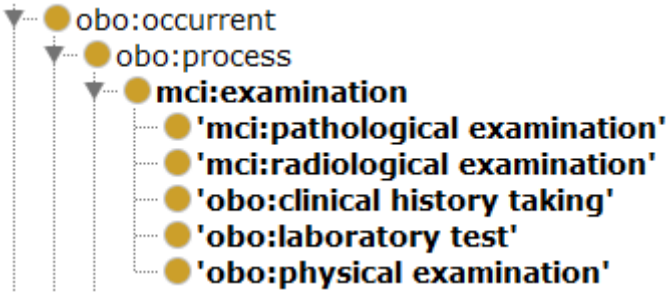


Figure 5.9.: Examination and subclasses.

mci:examination An examination is a healthcare process which has the objective to gather information about the health status of the patient.

mci:radiology examination A radiology examination is an examination with an associated radiology imaging modality.

mci:pathological examination A pathological examination is an examination based on some extracted tissue or body substance from the patient.

Processes – and thus also examinations – have a start and end time For many examinations however only one timestamp is provided, in which case the start and end date reference the same value. Even though the temporal sequence of examinations can be inferred from the above listed properties at query time, an object property for linking consecutive examinations explicitly is defined:

mci:is consecutive examination of Relates an examination to the directly previous examination of the same type, e.g. the consecutive examinations in a series of radiology examinations.

The participants of a process and thus also of an examination are described using the object property `obo:has participant` which relates the process with continuants that are involved in the process. Participants are for example the healthcare provider, the patient or some material entity such as specimen or a device.

mci:imaging device An imaging device is a device that can be used to produce images.
Binding: `radlex:imaging device`.

Besides participants, examinations might have an associated modality that specifies how an examination was provided. The modalities are important, since they are needed for correct interpretation of examination outputs.

mci:examination modality An examination modality is a data item that describes the context of an examination.

mci:imaging modality An imaging modality is an examination modality that describes the technique of the image acquisition.
Binding: `radlex:imaging modality`.

mci:patient position A patient position is an examination modality describing the bodily posture of the patient during the examination.

Examinations are linked to corresponding modalities by the object property `has examination modality`. An example of a computed tomography with different imaging modalities and participants is given in figure 5.10.

mci:has examination modality Relates an examination with and examination modality that further specifies the examination.

5.3.5. Reports

The foundation provides a class `obo:report` which is defined as a *"document assembled by an author for the purpose of providing information"* and which is *"the output of a documenting process"*. The input of a documenting process can be an image but also other data items. The pattern of the radiology examination and documentation process is given in figure 5.11.

Since the use case data does not always make distinctions between the process

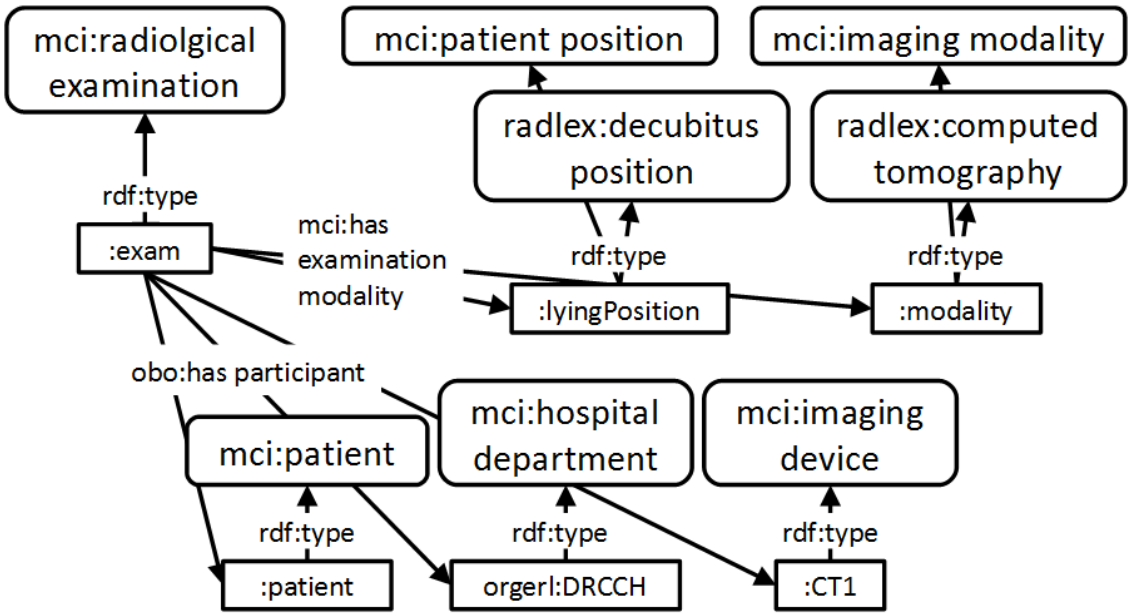


Figure 5.10.: A radiology examination with imaging modality computed tomography.

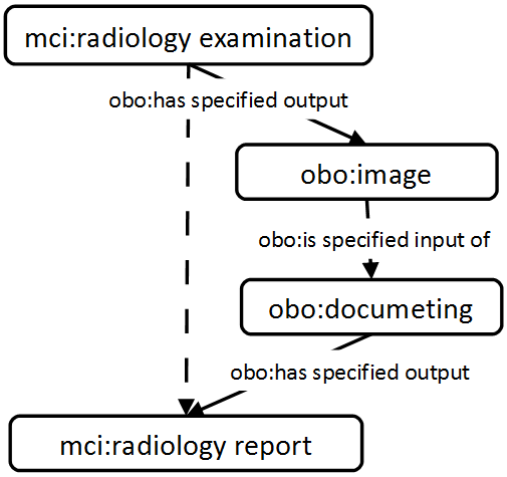


Figure 5.11.: The basic pattern of the examination and documentation process with corresponding inputs and outputs.

of examination and documenting, the clinical report is also related as a direct output of the corresponding examinations. The point in time when an information content entity (e.g. a report) was created, is expressed by using the data property *creation date time*. Further, classes for specific types of reports are defined:

mci:radiology report A radiology report is an report which is the specified output of some radiology examination.

mci:pathology report A pathology report is an report which is the specified

output of some pathological examination.

mci:laboratory report A laboratory report is an report which is the specified output of some laboratory examination.

mci:medication report A medication report is an report which is the specified output of some laboratory examination.

The foundation further contains classes for document parts (such as abstract or footnote) and textual entities (such as document title). Additionally, classes for specific parts of clinical reports are required:

mci:findings section The findings section is the part of a report where clinical findings are described.

mci:findings subsection A findings subsection a part of the findings section which groups clinical statements according to some rational. For example, findings of radiology reports are often grouped by anatomical regions such as head, neck, thorax or abdomen.

mci:assessment section An assessment section is the report part where the clinician gives an interpretation about the findings described in the findings section.

To group these report parts a corresponding class is defined:

mci:report part A document part that is part of a report.

```
‘mci:report part’ = ‘obo:document part’
                     AND (‘obo:part of’ SOME obo:report’)
```

Report parts can be further split into sentences. The foundation provides the class `nif:Sentence`, however in clinical reports one often has specific types of sentences that compare findings from consecutive examinations. This distinction is important for later retrieval and linking of finding data.

mci:comparison sentence A sentence containing a comparison of findings from consecutive examinations.

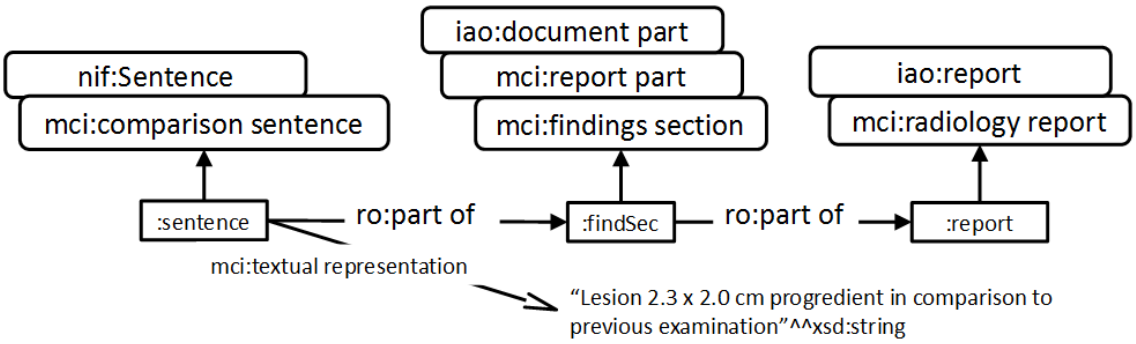


Figure 5.12.: A sentence of a finding section of a radiology report describing a progredient lesion.

5.3.6. Images and Image Regions

The foundation defines `obo:image` as *"an affine projection to a two dimensional surface, of measurements of some quality of an entity or entities repeated at regular intervals across a spatial range, where the measurements are represented as color and luminosity on the projected on surface"* [11d]. It has one subclass `obo:photograph`, but additional classes to represent the use case data is needed.

mci:3D-image A 3D-image is an information content entity which consists of voxels.

Note that the definition of image does not include 3D-images and thus 3D-image is not a subclass of image.

mci:image slice An image slice is an image which is part of a 3D-image and specified by some plane of that 3D-image.

In the context of DICOM, multiple images are part of an an `dicom image series`.

mci:dicom image series A dicom image series is part of some dicom study and consists of one or more images.

Multiple image series are part of a dicom study:

mci:dicom study The output of a radiology examination that consists of one or more dicom image series.

The analysis and corresponding reporting on images often requires the creation of image regions to point to some object or to base some calculation (e.g. average intensity) on a subset of the pixels. The foundation does not provide any classes for image regions – thus MCI defines:

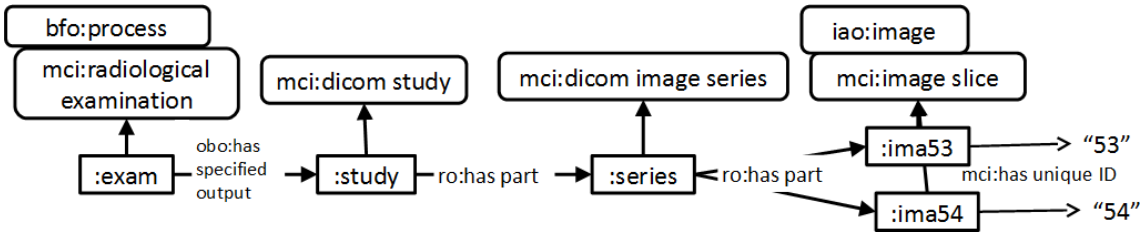


Figure 5.13.: Radiology examination and associated radiological image series.

mci:image region An image region is a data item representing a subset of pixels of some image or a subset of voxels of some 3D-image.

Note that an image region does not necessarily be connected.

mci:region of interest A region of interest (ROI) is an image region identified for a particular purpose.

In order specify which pixels belong to an image region, imaging tools commonly provide different geometric shapes such as an ellipse or a rectangle to define the boundary of an image region. Thus, the following classes are defined:

mci:geometric shape A geometric shape is the geometric information which remains when location, scale, orientation and reflection are removed from the description of a geometric object.[1]

The following geometrical shapes were used in the THESEUS MEDICO project [Sei+10] and transferred to MCI: rectangle, ellipse, box, landmark, line, curve and mesh. The geometrical shapes are used as `oa:AreaSelector` and linked from the image region by the property `oa:hasSelector` as shown in figure 5.14.

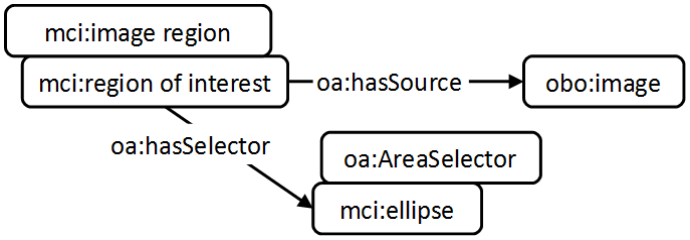


Figure 5.14.: A region of interest (ROI) with relations to selector and source.

5.3.7. Qualities and Units

Several qualities are imported from PATO, however some extension need to be made to capture the qualities described by findings and measurements more

[1]Definition source: en.wikipedia.org/wiki/Geometric_shape

precisely. For instance, `obo:1-D extent` (a subclass of `obo:size`) has subclasses `length`, `width` and `height` and also `diameter`. In the description of solid tumors and lymph nodes however, the size is specified by the longest 1-D extension and the longest perpendicular 1-D extension within the transverse plane. Thus, MCI defines:

mci:longest axis The longest axis is a 1-D extent describing the longest straight line (where all points of the line are within the respective object).

mci:short axis The short axis a 1-D extent describing the longest straight line of an object in orthogonal direction to the longest axis of that object.

In contrast to these specific 1-D extensions there are qualities describing the extension in parallel to the main body axes. The foundation provides the class `anterior-posterior diameter` (a subclass of `diameter`) however one additionally need classes for the other two main body axes:

mci:craniocaudal diameter A diameter that is along the craniocaudal axis.

mci:left right diameter A diameter that is along the left-right axis.

To express orthogonality between lines and also between lines and planes MCI defines the following properties:

mci:orthogonal to The relation between two orthogonal subspaces.

This property is symmetric and can be used between any orthogonal subspaces. Due to the symmetry domain and range of that property have to be the same. In a two dimensional space orthogonality is defined between lines. In a three dimensional space orthogonality can occur between lines and planes or between two lines. To distinguish between general orthogonality and the orthogonal complement in a three dimensional space the following object property is defined:

mci:orthogonal complement in 3D Relates orthogonal complements in a three dimensional space.

5.3.8. Anatomical structures

An abnormal anatomical structure might be pathological or non-pathological. The foundation contains a class *pathological structure* – a corresponding class non-pathological structure is required as well. For instance in reports, one finds assertions like “lymph node not pathologically enlarged”, i.e. the node does not

bear a pathological process. To capture this information, the following classes are defined:

mci:anatomical structure A material entity which is a part of the human body.

mci:non-pathological anatomical structure An anatomical structure not bearing any pathological process.

This definition is analogue to the definition of `ogms:pathological anatomical structure` [14s].

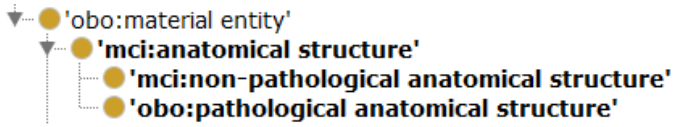


Figure 5.15.: Pathological and non-pathological anatomical structures.

5.3.9. Scalar Range Specifications

The foundation provides classes which can be used to represent scalar measurements. There are however cases where representation of a value *range* instead of a scalar value is needed. *Scalar range specifications* are intervals, upper bounds, lower bounds and imprecise values with deviation (as defined below). There are two distinct cases where such range specifications are required: Firstly, the measurement values in reports are not precise but described by a range: e.g. by an upper bound as in *"mediastinal and hilar nodes smaller than 1 cm"*. Secondly medical knowledge describes the typical value of qualities such as size in ranges: e.g. *"diameter of ureter normally up 9 mm"* or *"anterior-posterior diameter of liver normally 10-13 cm"*.

mci:scalar range specification A scalar range specification is an information content entity which specifies a range of some quality.

mci:lower bound specification A lower bound specification is a scalar range specification which specifies a range of some quality by providing a minimal value for that quality.

mci:upper bound specification An upper specification is a scalar range specification which specifies a range of some quality by providing a maximal value for that quality.

mci:interval specification An interval specification is a scalar range specification which specifies the interval of some quality by a lower and an upper bound specification of that quality.

5.3.10. Measurements

In the previous chapter about the model foundation it was shown how measurements are represented using classes and properties from ontologies of the OBO library (see e.g. figure 4.4). Some extensions are needed to represent measurements derived from images. There are many different contexts in which measurements occur. Here, measurements are described, which specify a quality which *inheres in* some material entity. This pattern is shown in figure 5.16. In the context of images there are cases where this simple representation needs to be extended to allow expression of different size, density and ratio measurements.

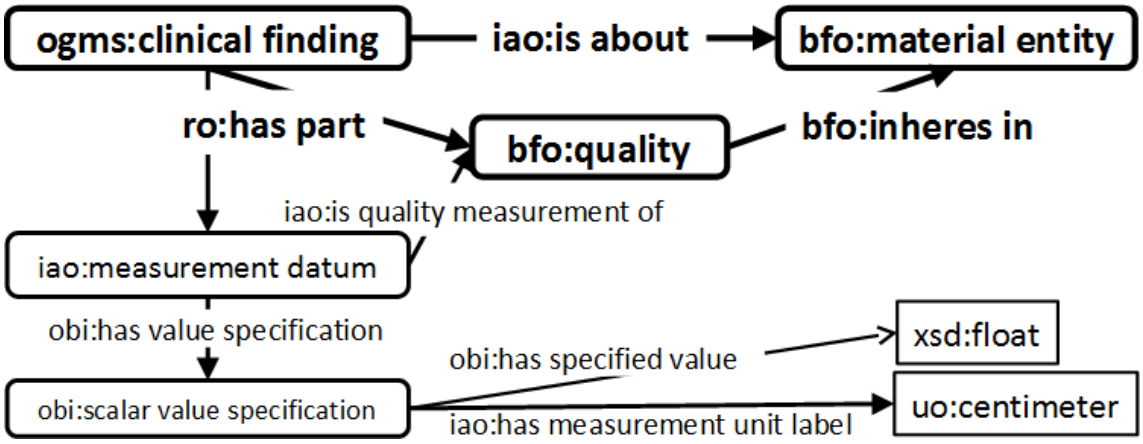


Figure 5.16.: The basic pattern of the representation of a clinical finding with a measurement datum using OBO classes and relations.

Size Measurements

In radiology reports there are mainly measurements specifying the size of anatomical entities in terms of volume, area or length. One commonly finds assertions such as "splenomegaly with 477.2 cm^3", "unremarkable diameter of ascending aorta with 3.2 cm" or simply "lymph node 1.3 cm". The foundation provides `obo:length measurement datum`. Additionally needed:

mci:area measurement datum A scalar measurement datum that is the result of measurement of area quality.

mci:volume measurement datum A scalar measurement datum that is the result of measurement of volume quality.

Analysis of radiology reports yielded that *length* measurements are the most frequent, however they are used in very different contexts. In the simple case it describes the extension of an anatomical entity or structure into one dimension.

For instance “mediastinal lymph node with diameter 21 mm”, “wall of gallbladder 12 mm”, “hepatic duct dilated up to 1 cm.”, “craniocaudal diameter of liver 14.0 cm” or “liver with anterior-posterior diameter of 15.5 cm”. The size of an implanted devices is also mentioned in reports. For example, *“placement of a 6 cm long, 16 mm wide wall stent into left V. brachiocephalica and V. cava superior.”* These size measurements comprise the following three components:

- **(Anatomical) entity:** lymph node, liver, wall of gall bladder, lesion, stent etc.
- **Measured quality:** length, width, diameter, anterior-posterior diameter, height, thickness, volume etc.
- **Value specification:** 21 mm, 1.1 cm, 477.2 cm^3, 6 cm etc.

The measured quality might not be precisely specified in reports. For instance, in*“lymph node 3 cm”* the quality would be a `obo:length` even though it is not specified, whether 3 cm denotes the long or short axis of the corresponding lymph node. Two or three length measurements might be grouped together to describe the extension of a certain entity along orthogonal axes: e.g. “lesion in segment 7/8 with 1.4 x 1.1 cm” or “spleen with 3.8 x 9 x 10.5 cm not enlarged”. These length measurements refer to the spatial extent along different axes. These types are referred to as *2D* or *3D length measurements*. For organs these measurements are sometimes taken in parallel to the main body axes to specify height, width or depth. For smaller entities the axes are mostly defined by the form of the entity itself: For the evaluation of tumors or metastatic lesions in computed tomography (CT) or magnetic resonance imaging (MRI) the radiologist firstly measures the longest diameter in the axial slices and then the longest perpendicular extension [Eis+09].

The actual value specification of a measurement can refer to multiple entities for which a range or a bound is given: For instance one commonly finds a assertions like “axillary lymph nodes smaller than 1 cm” or “Nodular densities at head of pancreas with diameters up to 1 cm”. In all cases 1 cm refers an upper bound of the size of the corresponding entities. An example representation for these type of measurements is given in figure 5.17.

Density Measurements

The density of anatomical structures is measured by radiologist in order distinguish for example cysts from solid tumors. Each pixel of some radiology image represents a density measurement for its own. In the context of radiology the density is given in *Houndsfield scale*. To analyze the density of a part of the image, (e.g. an anatomical entity) the radiologist creates a region of interest (ROI). All

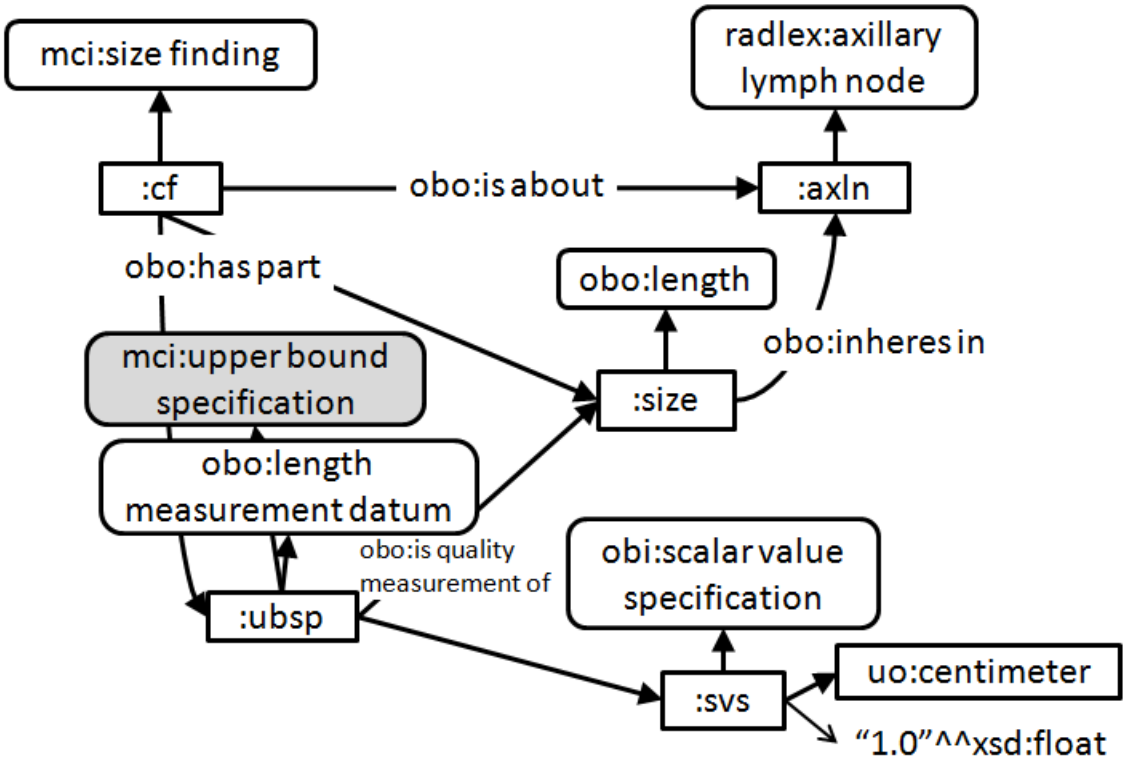


Figure 5.17.: An upper bound used to represent the size of axillary lymph nodes. In reports assertions like "unspecific axillary lymph nodes smaller than 1 cm" are commonly found.

pixels within this ROI are the input for the calculation of the minimal, average and maximal density values for that ROI. As shown in figure 5.18 the ROI is drawn within the suspicious structure. The average density value of 8 HU is close to the density value of water (0 HU) which is typical for a cyst.

Ratio Measurements

Ratio measurement are important in the context of laboratory test or pathological examinations. In the context of cancer, *staining* is used to measure the amount of active tumor cells (see figure 5.20). The protein Ki-67 for instance increases in cells when they prepare to divide. Thus, the percentage of Ki-67 positive tumor cells is a measure for the aggressiveness of the tumor. For example, in breast cancer, a rate of 10% Ki-67 positive is considered low while 10-20% intermediate/borderline, and 20% and more is considered high.

Ratio measurements are expressed in the same way as other measurement, i.e. with a scalar measurement datum specifying some quality. The quality then is e.g. the prevalence of Ki67+ cells within a tumor as shown in figure 5.21. This has the

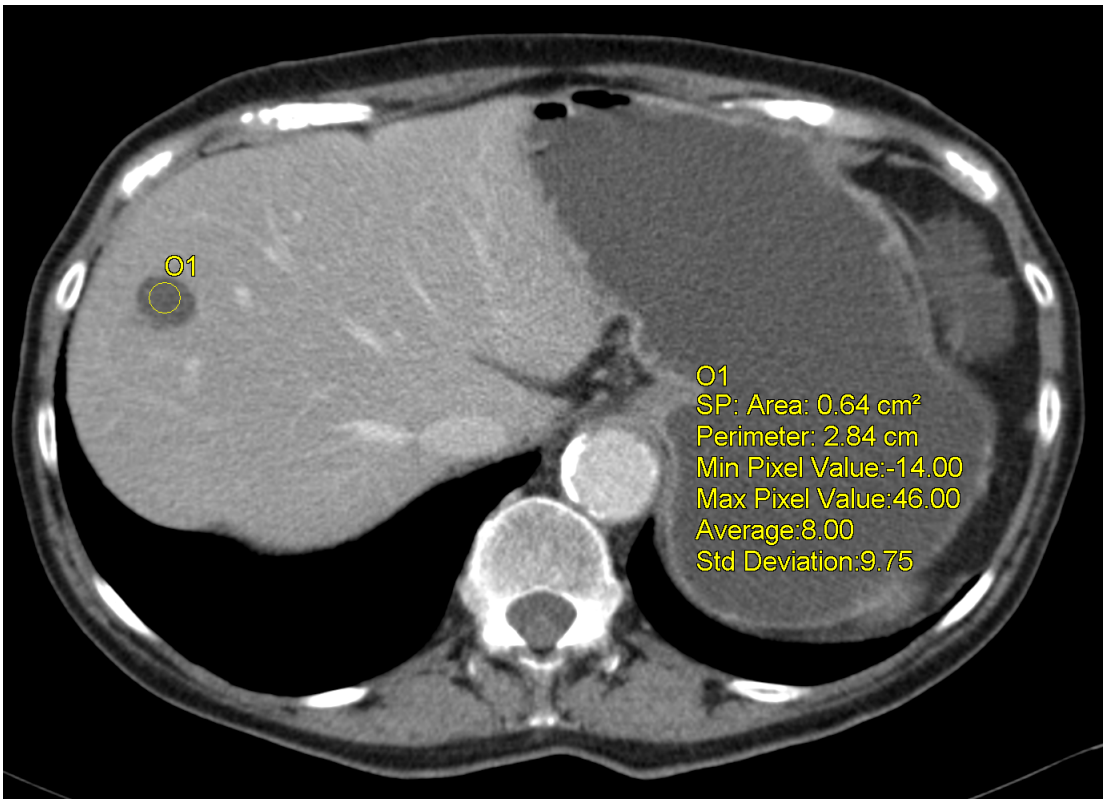


Figure 5.18.: The density is evaluated based on a circular ROI which is part of the liver. The image shows a liver cyst.

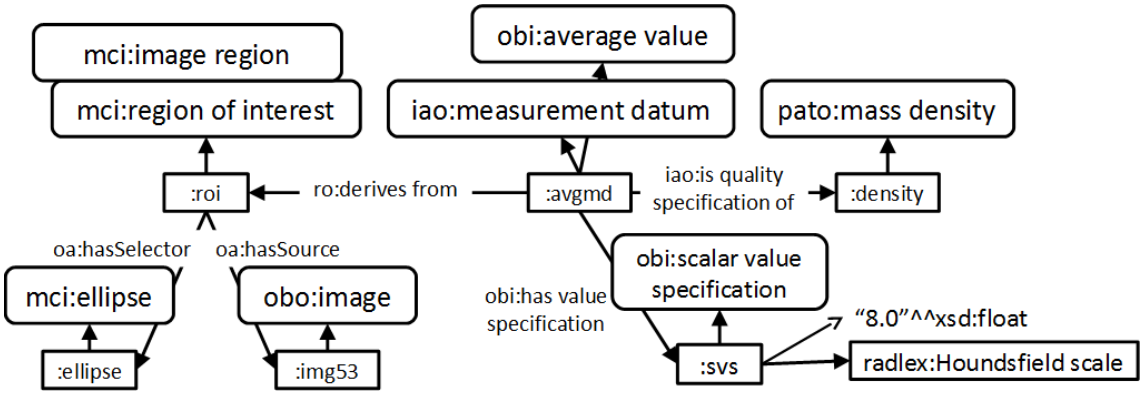


Figure 5.19.: Representation of a measurement of an average density based on an image region.

advantage that dimensional and dimensionless ratios can be represented by the same pattern. In order to be able to compare ratios reasonably the denominator and numerator of ratios have to be made explicit.

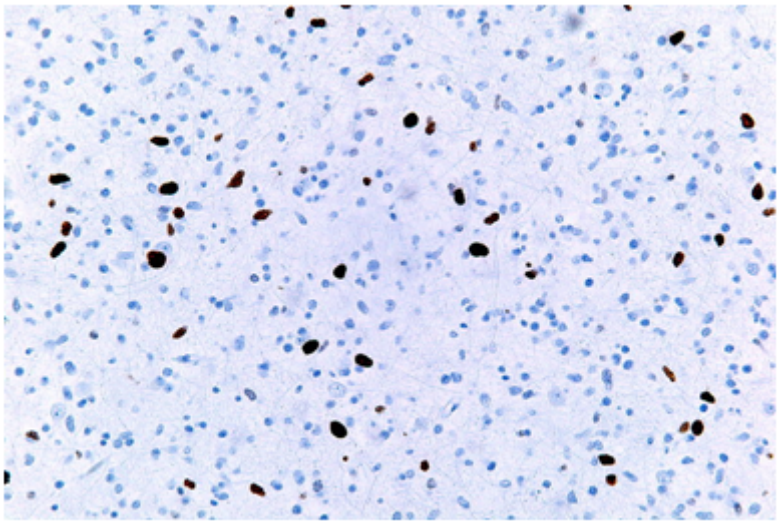


Figure 5.20.: Ki-67 positive brain tumor cells. Image source: [10a]

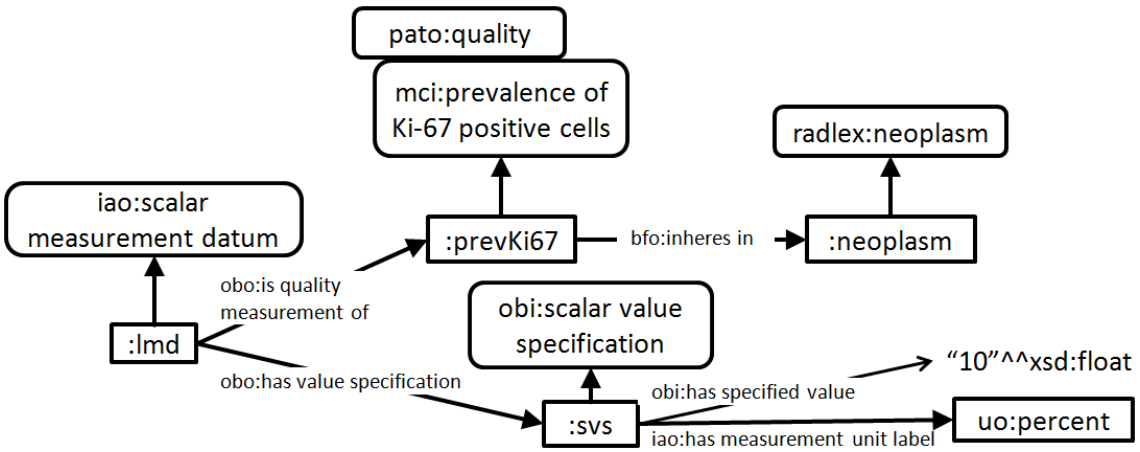


Figure 5.21.: Representation of a measurement of the amount of Ki67-positive tumor cells.

5.3.11. Clinical Findings

Clinical findings are central information objects beyond all patient data, because they are the main input for clinical decision making, such as stating a diagnosis or treatment evaluation. OGMS defines clinical finding as *"a representation that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof"*. The distinction of clinical findings and symptoms is defined by the observer: While findings are observed by clinicians, symptoms are observed by the patient itself. For instance pain is a symptom but not a clinical finding. Data about clinical findings is important because it forms the basis for clinical decisions and thus require detailed representations. The model foundation provides subclasses of clinical finding such as clinical history, image finding, laboratory finding and physical examination finding. These are important classes

to classify finding data by *origin*. To allow efficient clinical decision making, additional types to enhance retrieval of relevant findings are needed. In practice, clinical findings are distinguished by clinicians mainly according to two aspects: Firstly, the main distinction is made between *normal* and *abnormal* clinical findings. Secondly, findings are classified with respect to *change*, i.e. regarding the temporal aspect. The complete subclass hierarchy defined under clinical finding is shown in figure 5.22. The corresponding classes are described in the following subsections, before measurement findings are discussed. The representation of location specifications in clinical findings is presented towards the end of this subsection.

Figure 5.22.: The class hierarchy under clinical finding.

Normal and Abnormal Clinical Findings

In general clinicians look for *abnormalities*, however *normal findings* are also important since they allow excluding certain diseases. For instance, lab values are tagged as being *low*, *normal* or *high* for some given patient. Similarly, a lymph node presenting certain characteristics, such as a short axis diameter larger than 1 cm (or 1.5 cm, depending on its location) is categorized by radiologists as *abnormal*, while smaller lymph nodes are classified as *normal*.

mci:normal finding: An normal finding is a clinical finding describing some normal quality that inheres in some anatomical structure.

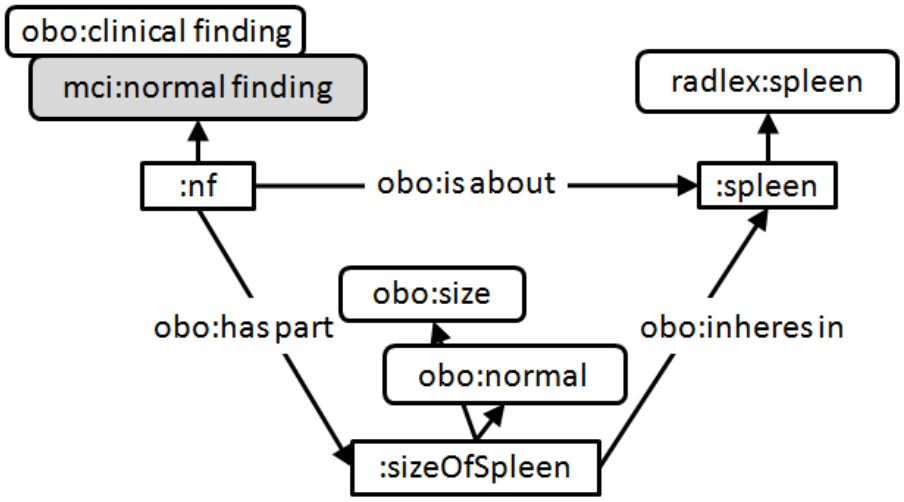


Figure 5.23.: Representation of a normal finding that describes the size of the spleen.

mci:abnormal finding An abnormal finding is a clinical finding describing some abnormal quality that inheres in some anatomical structure.

mci:pathological finding A pathological finding is an abnormal clinical finding which describes some pathological quality which inheres in some anatomical structure.

Further, normal and abnormal findings that describe qualities that are based on measurements are distinguished:

mci:normal value finding A normal value finding is a normal finding where the quality described is specified by some scalar measurement datum.

mci:increased quality finding An increased quality finding is an abnormal finding describing an increased quality.

mci:decreased quality finding A decreased quality finding is an abnormal finding describing an decreased quality.

An example of such a finding is given in figure 5.25.

Change of Clinical Findings over Time

Important for evaluation of the health status is the detection and tracking of changes of findings. For example, lymph nodes might be *progressive* or *regressive* in size, the body temperature might *increase* or *decrease* and the synopsis of this information can be crucial for the physician to come up with the correct diagnosis. The following definitions make use of the object property `has preceding finding` which relates clinical findings from consecutive examinations which describe the same anatomical entities.

mci:has preceding finding A relation between clinical findings from consecutive examinations describing the same quality of the same anatomical entity.

For example the sentence "lesion in segment 7/8 with 1.4 cm x 1.1 cm (Ima 19, pre-investigation 1.4 cm x 1.1 cm)" contains two preceding findings. It is however often the case that findings from consecutive examinations describe not exactly the same quality or not exactly the same entity. For instance, two findings from consecutive examinations, one describing the spleen volume and the other the craniocaudal diameter of the spleen are *related findings*.

mci:has related finding A relation between two different clinical findings either similar qualities of the same entity or the same quality of similar entities.

mci:unchanged finding An unchanged finding is a clinical finding that has a preceding finding with the same qualitative quality description.

For instance, a clinical finding describing the size of the liver as normal, which has a preceding finding describing the volume (and thus the size) of the liver as normal.

mci:changed finding A changed finding is a clinical finding that has a preceding finding with with different qualitative quality descriptions.

To be classified as a changed finding the corresponding described quality needs to be different in comparison to the quality description of the preceding finding. For instance a finding about the size of the spleen as *enlarged* (i.e. abnormal), where the preceding finding described the spleen as *normal*, is a changed finding (see figure 5.24). Similarly a clinical finding from a follow up examinations with

measurement datum specifying the longest 1-D extension of an axillary lymph node, which was measured also in the preceding examination however with different measurement values, is a changed finding.

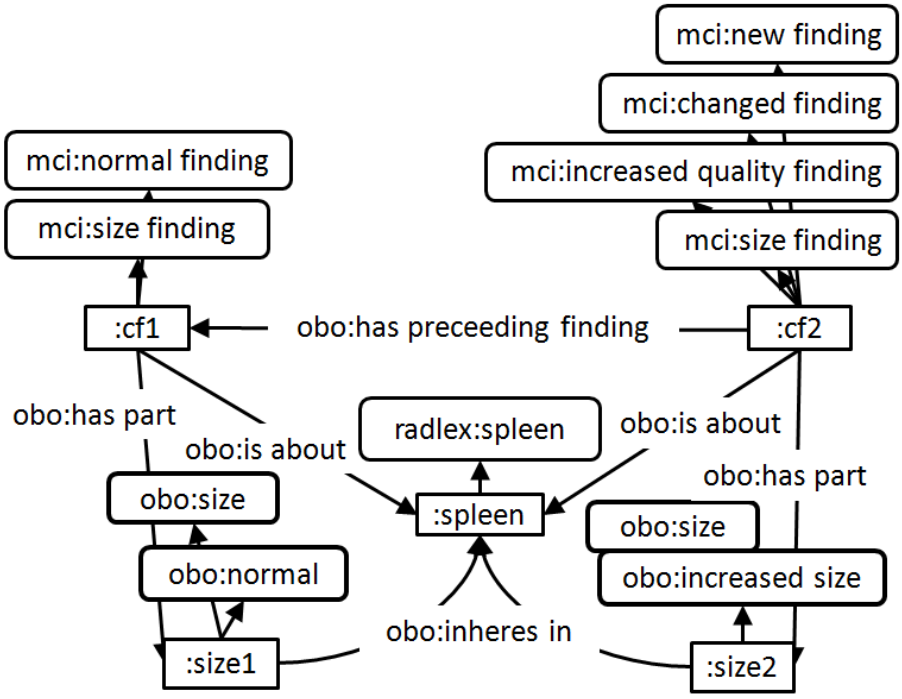


Figure 5.24.: Representation of a changed finding that describes a newly appeared splenomegaly.

mci:new finding: The first occurrence of abnormal findings about some entity in the context of some clinical encounter.

Obviously, not only the fact of a change but also the tendency of the change should be qualified if possible. For instance a changed finding with a scalar value specification can be classified as a `decrease finding` if the value is falling or as an `increase finding`) if the value is getting higher:

mci:decrease finding A decrease finding is a changed finding describing a quantifiable quality whose value is getting lower.

mci:increase finding An increase finding is a changed finding describing a quantifiable quality whose value is getting higher.

Note that these classes are different to `decreased quality finding` or `increased quality finding`, which describe the status at some point in time - not the change as such. As shown in Figure 5.25 a finding can be and `decrease finding` and at the same time an `increased quality finding`.

Increase and decrease can be further specified by the type of quality described by

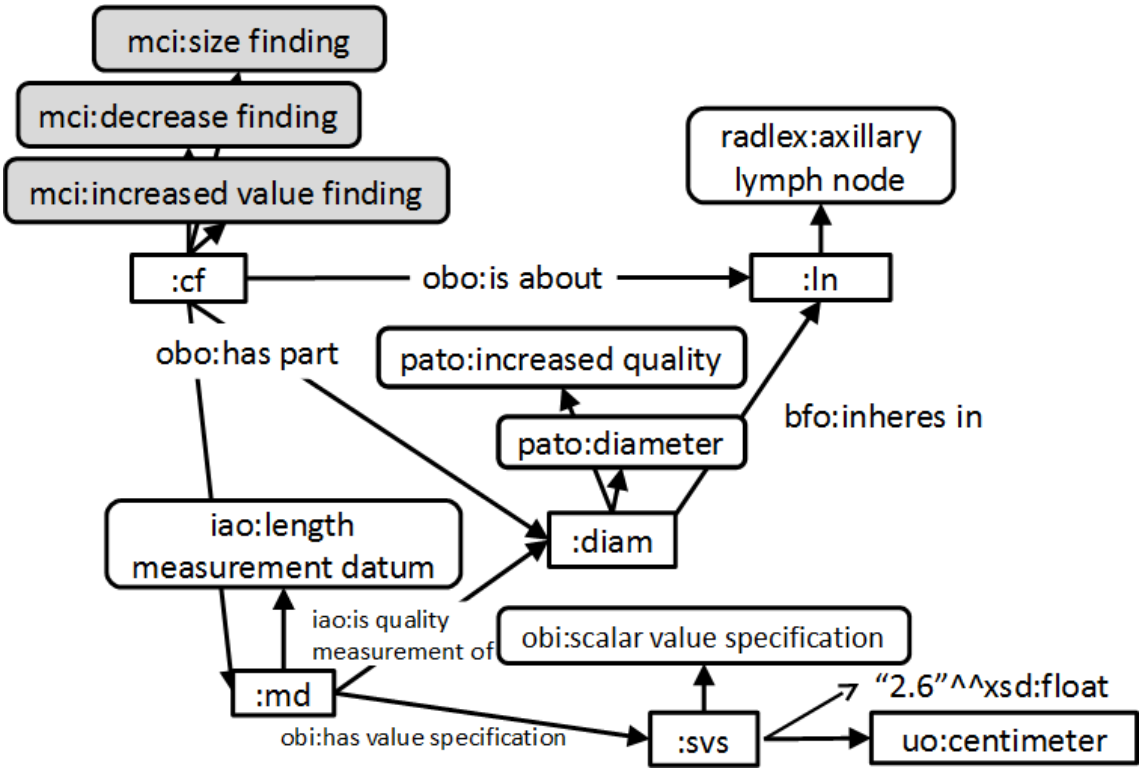


Figure 5.25.: Representation of a size finding describing the diameter of some axillary lymph node. The measured value for the diameter indicated an increased quality. In comparison to the previous examination (not shown) the value however decreased. The types of the findings express this information.

the finding:

mci:regressive finding A regressive finding is a decrease finding describing an decrease of some size quality or some quality related to size.

mci:progressive finding A progressive finding is an increase finding describing an increase of some size quality or some quality related to size.

mci:improvement finding An improvement finding is a changed finding which quality specification changed towards normal homoeostasis.

mci:worsening finding A worsening finding is a changed finding where the corresponding quality specification gets worse with respect to normal homoeostasis.

Measurement Findings

The representation of measurements, in particular image measurements was described above. In the context of images, it is often the case, that two or more measurements are related to each other and that only together the corresponding finding is correctly represented. Before corresponding relations are described, measurement finding and subclasses are defined.

mci:measurement finding A measurement finding is a clinical finding, which has a measurement datum as a part.

mci:weight finding A weight finding is a clinical finding, which has a mass measurement datum as a part.

mci:size finding A size finding is a measurement finding which has at least one measurement datum as a part which specifies some size quality.

For instance the finding shown in figure 5.25 is a size finding.

mci:one directional size finding A one directional size finding is a size finding, which has a measurement datum as a part, specifying some 1-D extent.

For example, a size finding about the kidney with one measurement datum, specifying the length of the kidney along the cranio-caudal axis.

mci:Two directional size finding A two directional size finding is a size finding which has exactly two measurement data items as parts which specify different 1-D extents which stand in relation of orthogonality to each other.

For example, a size finding about the spleen with two measurement data items – one specifying the width the other the length of the spleen as shown in figure 5.26

Similarly, measurements of lymph nodes and solid tumors according to RECIST (response evaluation criteria in solid tumours) guidelines are represented as two directional size findings: in computed tomography (CT) or magnetic resonance imaging (MRI) the radiologist firstly measures the longest diameter in the axial slices and then the longest perpendicular extension [Eis+09] as shown in figure 5.27.

Its representation is shown in figure 5.28. This form of standardized measuring procedure allows comparing measurements from consecutive examinations.

mci:Three directional size finding A three directional size finding is a size finding which has exactly three measurement data items as parts which specify

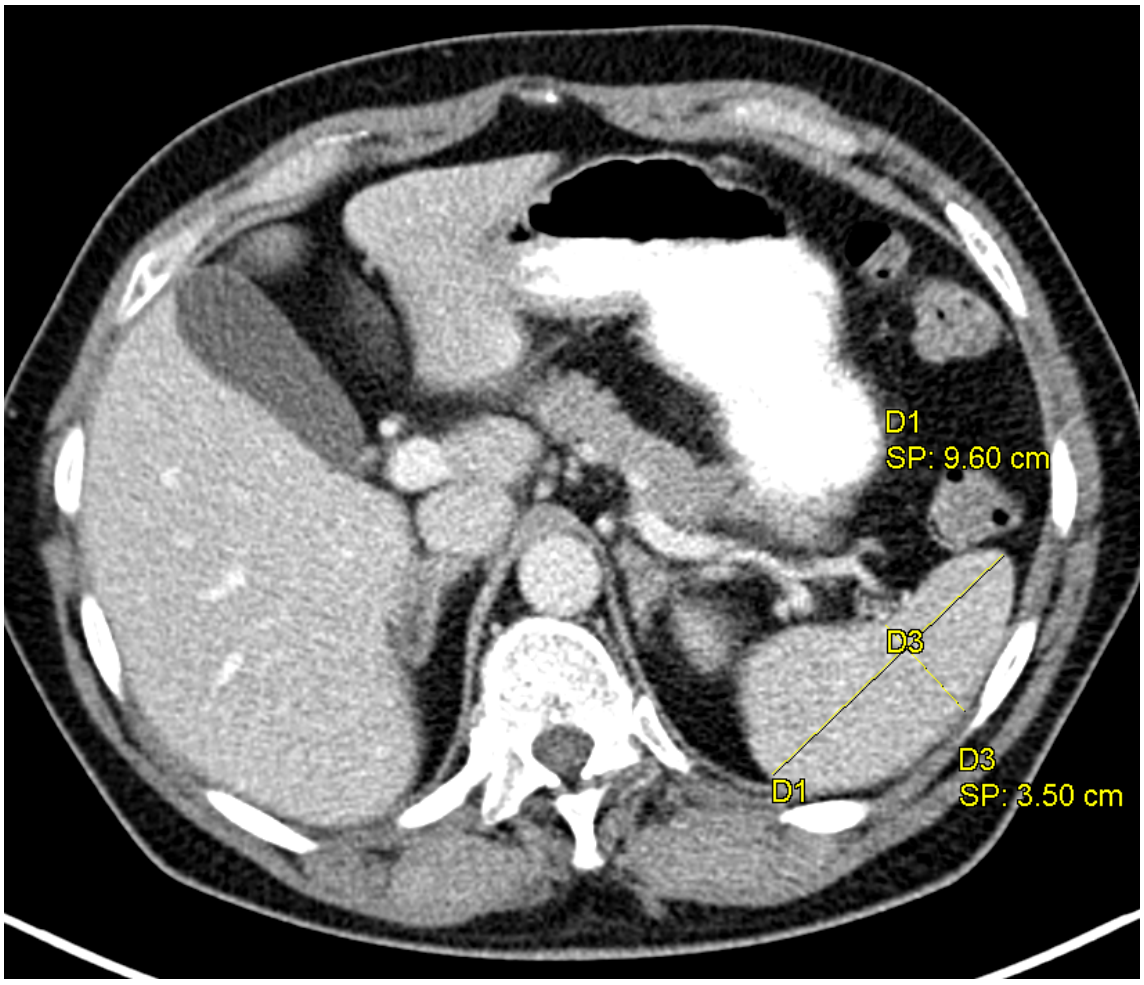


Figure 5.26.: A transversal slice of a contrast enhanced computed tomography examination of the upper abdomen. The size of the spleen in transversal plane is specified by measuring the length and width at the hilum.

different 1-D extents and which stand in relation of orthogonality to each other.

A size finding about the spleen with three different measurement data items - one specifying the width, one the depth and a third the height of the spleen. For example, “spleen with 3.8 x 9 x 10.5 cm not enlarged”.

The Context of Findings

Most of the findings mentioned above describe what the clinician observes – e.g. “an axillary lymph node with a diameter of 2.65 cm”. As shown, MCI provides the required classes and properties to capture these *positive* assertions. There are however also many negative assertion such as “no fever”, “spleen not enlarged”, “liver without lesions” etc. Most of them can be mapped to positive

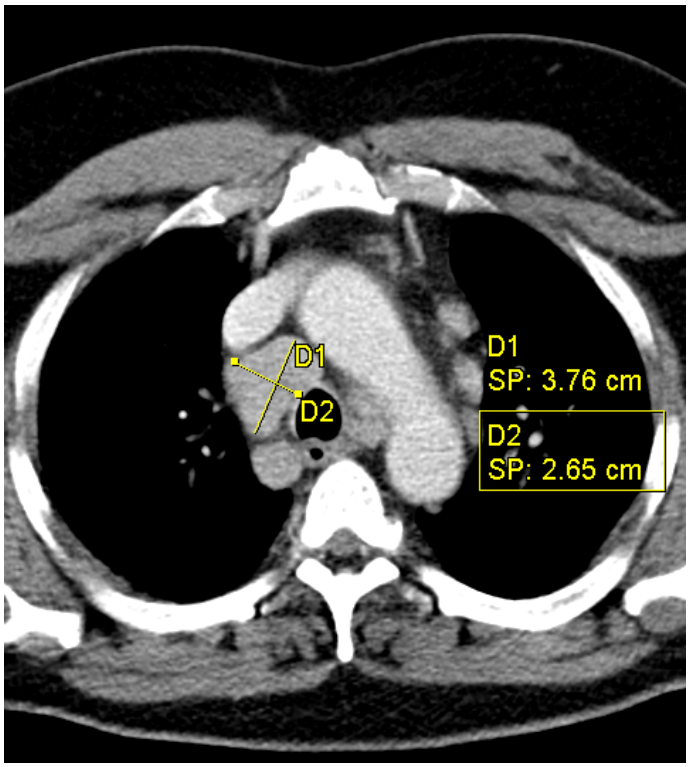


Figure 5.27.: Measurement of a lymph node according to the RECIST guidelines.

assertions "normal body temperature", "normal spleen size", "structure of liver homogeneous" etc. In the context of diagnosis however, clinicians look for specific signs and symptoms whose status they attempt to clarify. MCI defines four types of finding contexts:

mci:known absent A finding context specifying that the finding was checked for and negative evidence was found.

mci:known present A finding context specifying that the finding was checked for and positive evidence was found.

mci:unknown A finding context specifying that the finding was either still not checked or no positive or negative evidence could be found.

mci:suspected A finding context specifying that the finding was checked for and preliminary positive evidence was found.

The relation of finding instance data to the finding context is realized through the following annotation property:

mci:has qualifier Relates a clinical finding with additional context information.

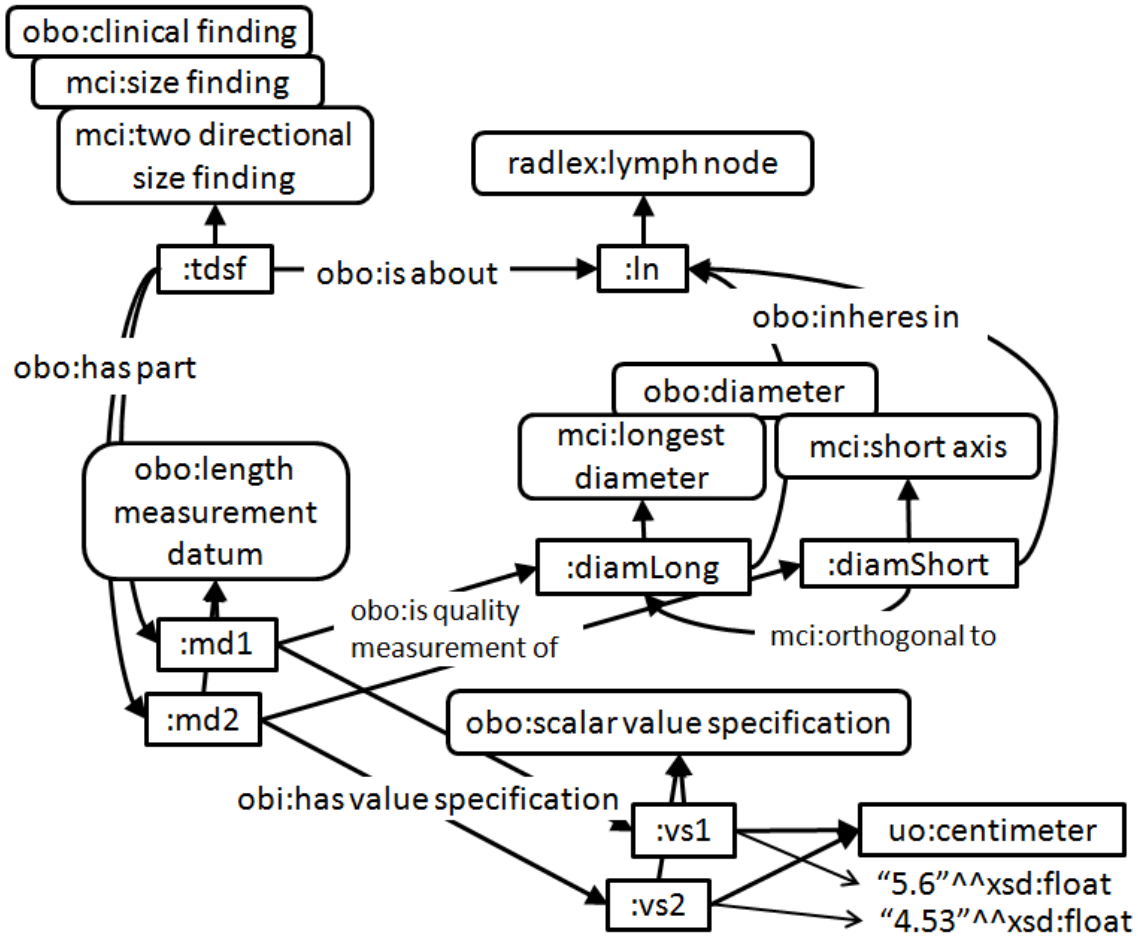


Figure 5.28.: A two directional size finding with long and short axis according to the RECIST guidelines.

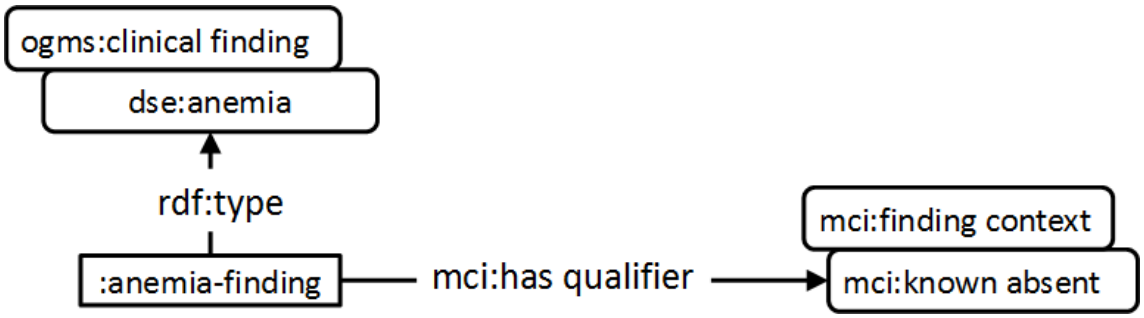


Figure 5.29.: Representation of the finding context in diagnosis.

Location

Earlier it was described how the *orientation* of measurements can be specified by qualities. Here representation of *location* is described. In the context of size findings, two types of location information are distinguished: the location of the described entity and location of the described quality.

The location of the described entity: For example, "lesion in liver segment II", "at terminal ileum on a length of 6 cm thickening of wall up to 1.2 cm." or "next to hilum of spleen there is a small hyperdense lesion with diameter 1.0 cm.". This location information is represented `ro:located in` (see figure 5.30.

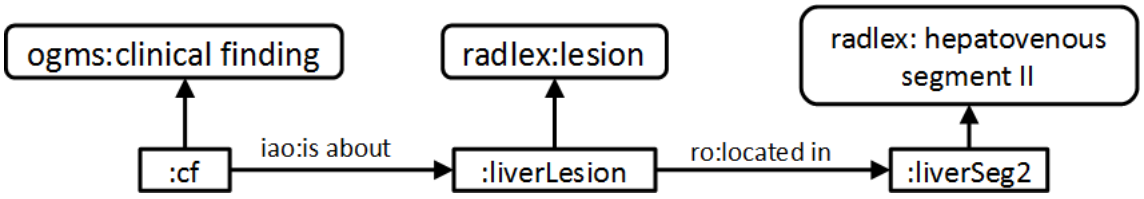


Figure 5.30.: The location of a lesion in a specific segment of the liver.

Further the location information might be the *main content* of the finding: For instance, *"5 mm shift relative to central axis"* describes a deviation from the normal location, i.e. a distance. In these case the quality (length) does not *inhere in* the (anatomical) entity but describes the distance of the entity to some other entity or location.

mci:location finding A location finding is a clinical finding that describes the location of some entity within the body.

5.3.12. RECIST Target Lesions

The RECIST (response evaluation criteria in solid tumors) guideline allows a standardized assessment of tumor burden and tumor response during therapy. Here, a sum of the diameters (short axis for lymph nodes, longest diameter for the remaining lesions) for all target lesions is calculated and reported as the baseline for follow-up sum and indicates therapy response/failure. At baseline, tumor lesions (longest diameter ≥ 10 mm)/lymph nodes (short axis diameter ≥ 15 mm) can be considered and selected as target lesions. When more than one lesion fulfill the described criteria, up to a maximum of five lesions (maximally two lesions per organ) that are representative of all involved organs should be identified as target lesions and measured. In addition, the selected target lesions should allow reproducible repeated measurements. The sum of the diameters (short axis for lymph nodes, longest diameter for the remaining lesions) of all target lesions is calculated and reported at baseline and during follow-up. The following criteria are used to determine objective tumor response. Complete Response: Disappearance of all target lesions (any pathological lymph nodes, whether target or non-target must have reduction in short axis to <10 mm). Partial Response: At least a 30% decrease in the sum of diameters of target lesions (reference = baseline sum diameters). Progressive Disease: At least a 20% increase in the sum of diameters of target lesions (reference = the smallest sum on study; this includes the baseline sum if that is the smallest on study). For Progressive Disease, the sum

must also exhibit an absolute increase of ≥ 5 mm. The appearance of one or more new lesions is also considered as Progressive Disease. Stable Disease: Neither sufficient decrease to qualify for Complete/Partial Response nor sufficient increase to qualify for Progressive Disease. To express this information MCI defines the following classes:

mci:target lesion A target lesion is a lesion that was selected to be tracked in order to measure the response to treatment. Depending on the imaging modality a RECIST defines the minimum size of target lesions.

mci:non-target lesion A non-target lesion is a lesion that was not specifically selected to be tracked in order to measure the response to treatment.

mci:RECIST sum calculation The RECIST sum calculation takes target lesions as input and outputs the sum of the length of their short axes (for lymph nodes) and longest diameter (for all other lesions).

mci:RECIST sum The RECIST sum is the output of a RECIST sum calculation process.

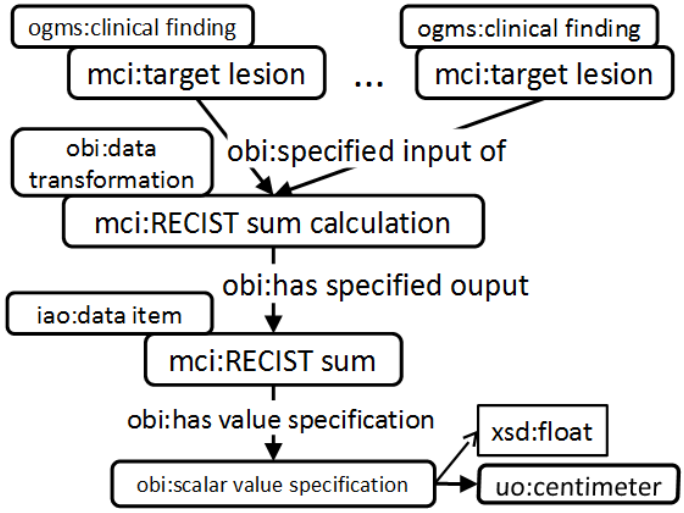


Figure 5.31.: The RECIST sum is calculated based on a set of target lesions. Adjacent items represent by subclass relationships, i.e. the RECIST sum calculation is a subclass of data transformation.

Annotations

The representation of annotations is based on the OA. The usage of OA classes and relations to other classes from MCI, text and image annotations are shown in the following figures.

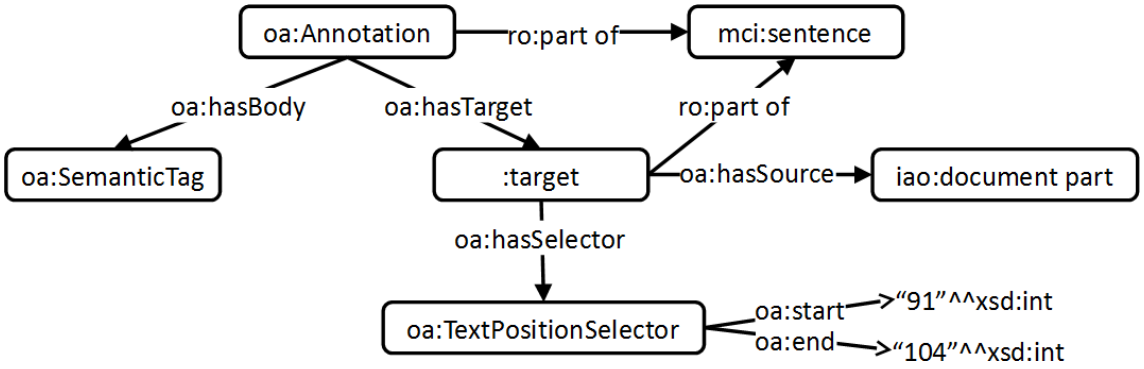


Figure 5.32.: The basic pattern of a text annotation. The target is specified by a text position selector, which defines the position within the corresponding document part (such as a finding section of a report). To ease later retrieval, the target as well as the annotation are linked to the sentence.

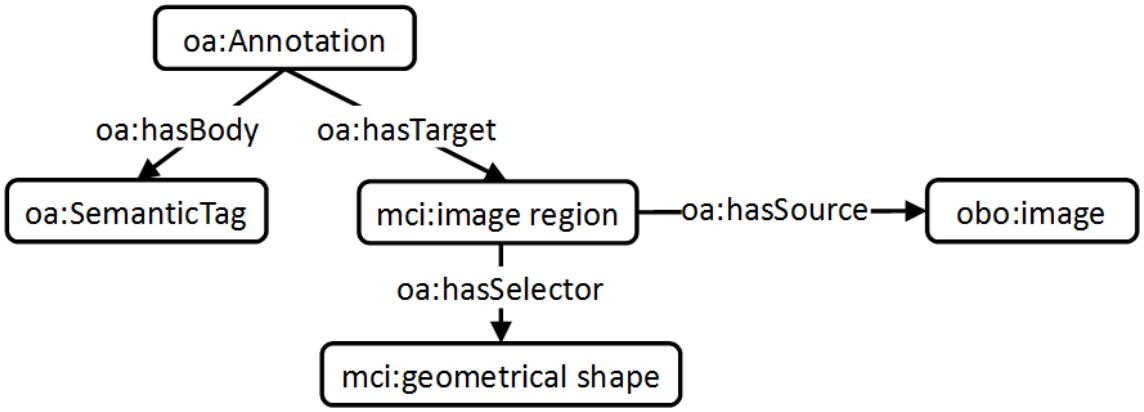


Figure 5.33.: The basic pattern of an image annotation. Here, the target is represented by some image region that is specified by some geometrical shape such as a rectangle or ellipse.

6

Knowledge Models

In the previous chapter the Model for Clinical Information (MCI) was described and its usage for representation of clinical data, i.e. data about patients, was shown along several examples. Another important resource for more efficient decision making by clinicians is *formalized medical knowledge*, which can be used to enrich patient data. For instance, the importance of granular types for representation of clinical findings and especially for normal and abnormal findings was highlighted in the previous chapter. MCI provides corresponding classes, however the distinction is often not explicit in the data. For instance, in a radiology report, the size of the spleen might be specified as "spleen 5 x 9 x 12 cm". Formalized medical knowledge can be used to classify this measurement finding as a *splenomegaly*, i.e. an enlarged and thus *abnormal* spleen. Further, in a diagnostic process, finding data needs to be interpreted with respect to diseases. For example, a splenomegaly is an abnormality of the immune system and related to specific types of cancer such as lymphoma. In this chapter it is shown, how the Model for Clinical Information can be used to represent medical knowledge about the normal and abnormal size of anatomical entities commonly described in radiology examinations (section 6.1). Further, a knowledge model about the relations between diseases, clinical findings and examinations is presented (section 6.2). Applications of both models are described in chapter 8.

6.1. Range Specifications

For quantitative qualities a value range can be specified in many different contexts: For instance to represent a normal range or typical size of anatomical entities, to represent the minimum and maximum values of some measured value over a specific time period etc. In the next subsection, it is described the different types of normal and abnormal range specifications typically found in the medical literature. Then, the general representation pattern and a patient-specific representations are presented before the coverage of the created knowledge model about normal findings is described.

6.1.1. Range Specifications About Normal and Abnormal Anatomical Entities

The medical literature contains much information about the *normal* qualities of anatomical entities as well as descriptions of typical *abnormal* or *pathological* structures like, e.g., cysts, lesions or enlarged lymph nodes. The main function of these specifications is to define, which manifestation of a quality of some anatomical entity is considered to be normal and which not. Clinicians refer to these specifications from medical literature, when they classify observed findings as normal, abnormal (i.e. not normal or outside the normal range) or pathological (abnormal with underlying disease process). For instance, a lymph node with a short axis of 0.7 cm is considered to be normal (since "normal lymph nodes are $\leq$ 1 cm in short axis") while one with 2.4 cm is classified by clinicians as abnormal. The motivation of formalizing knowledge about normal and abnormal qualities, is to automate this kind of classification. The advantages of automatic classification are twofold: Firstly, adding missing information about the finding type, enriches descriptions of clinical findings and thus enhances the data quality. Secondly, automatic classification can be used in combination with automatic segmentation algorithms and thus point the radiologist to suspicious anatomical entities, which enhances the reporting work-flow.

In this section, it is demonstrated that normal size specifications can be represented in MCI. The usage of them for automatic classification of size findings commonly found in radiology is demonstrated in Chapter 8. Even though the provided examples describe normal size qualities, it is emphasized that this approach is suitable to classify any findings describing *measurable qualities*, i.e qualities which can be described by a scalar value specification. table 6.1 lists the three different types of range specifications which are commonly used in the medical literature to describe size qualities. The table lists the *type* of range specification and corresponding examples.

Table 6.1.: Commonly used range specifications to describe the typical size of anatomical entities.

Type	Examples
Interval	Cranio-caudal diameter of kidney normally 8-13 cm
	Thickness of wall of gallbladder normally 0.1 -0.3 cm
	normal width of ureter 0.4-0.7cm
Upper bound	Normal lymph node < 1 cm
Lower bound	Normal aorta diameter > 4 cm at root
	Enlarged lymph node > 1 cm

The basic form consists of three components: An anatomical entity, a quality description, and a scalar range specification. This is similar to the pattern of clinical findings - the main difference is the specification of a *range* instead of a single scalar value and that the quality is additionally typed as normal or abnormal. These specifications are referred to as *normal specifications* (e.g. "Normal lymph node ≤ 1 cm" in short axis), as *abnormal specifications* (e.g. "lymph node with short axis > 1 cm are considered abnormal") or sometimes simply as *range specifications*. Further, a value x is said to be *in the range* of some size specification if x is contained within the respective interval $[a, b]$. Here, upper bounds (where only value b is given) are interpreted as the interval $[0, b]$ and lower bounds (where only value a is given) as the interval $[a, \infty)$. The actual definition of which size of particular anatomical entities is considered to be normal depends on guidelines. There are for example different guidelines on how to measure lymph nodes (in short or long axis) and correspondingly the definition of normal and abnormal is different. However, it is emphasized, that the *model* presented here is able to represent the necessary *types* of specifications (upper bound, lower bound and interval) and can be adapted to the guideline preferred by the user.

Representation of Normal Range Specifications

The knowledge model of range specifications is expressed in terms of the Model for Clinical Information, i.e. by using classes and properties described in the previous chapter. Anatomical entities and structures are included through links to reference ontologies like the FMA or RadLex. A normal range specification has three components as shown in figure 6.1 and figure 6.2: the anatomical entity, the quality described and a corresponding range specification.

As for measurement findings, RadLex [14p] and the FMA [RJ07] are used to

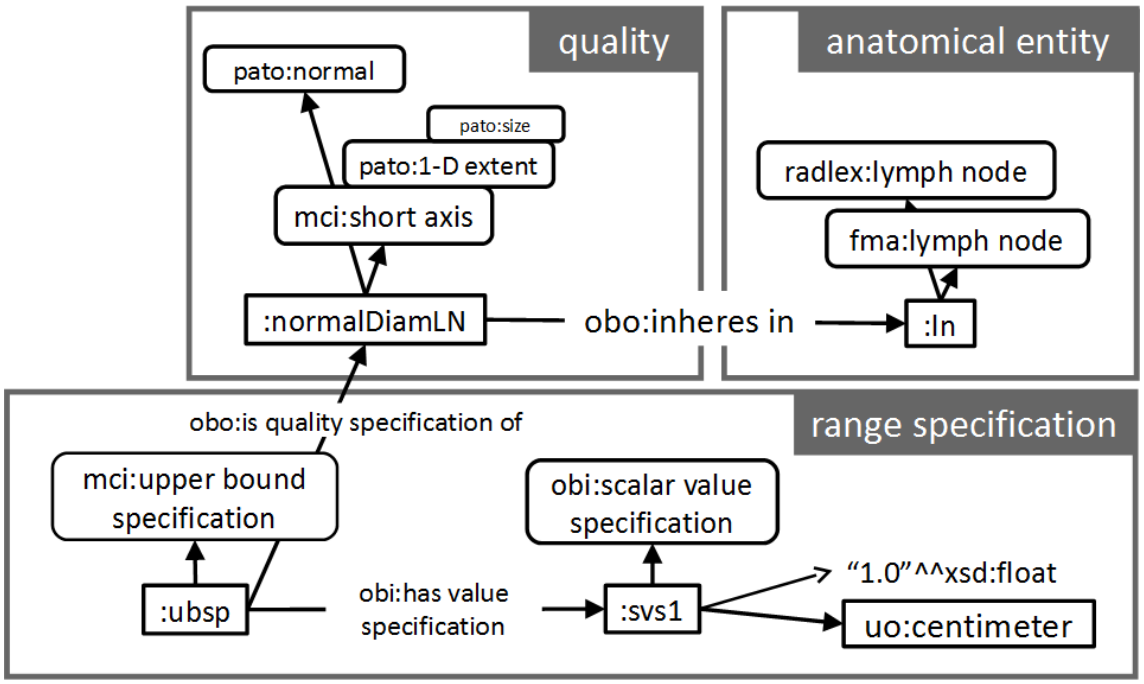


Figure 6.1.: Representation of the normal upper bound specification "Lymph nodes are normally at most 1 cm in short axis" with anatomical entity *lymph node*, quality *normal short axis* and range specification.

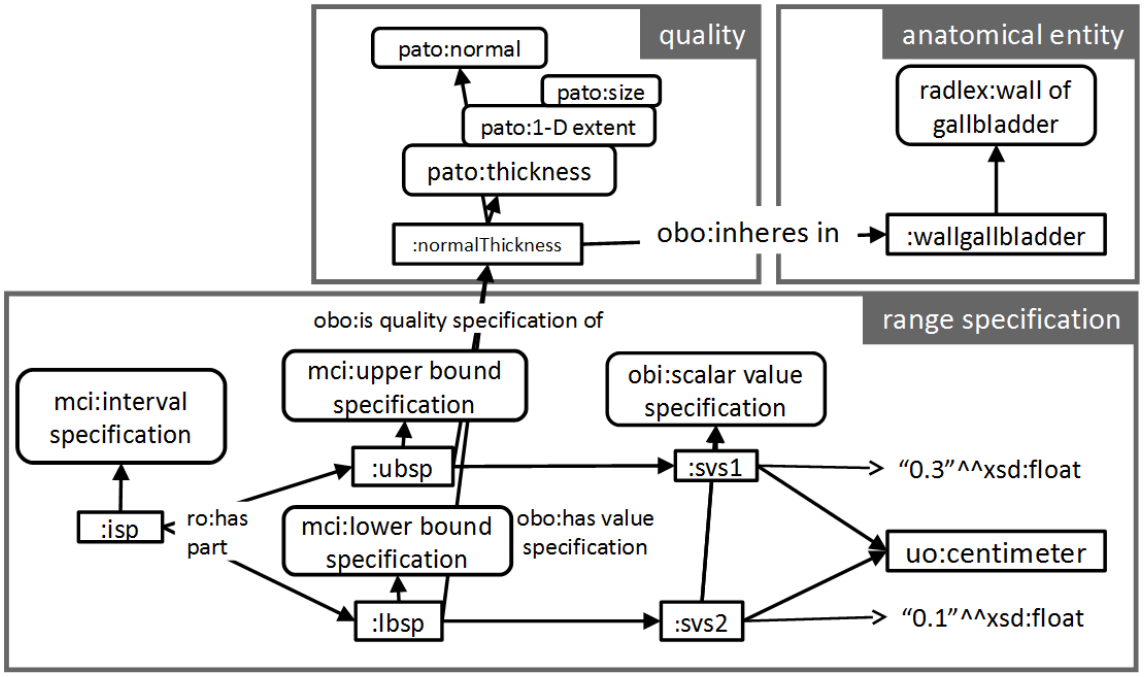


Figure 6.2.: The thickness of the wall of the gallbladder is normally in the range of 1-3 mm. Here the anatomical entity is the *wall of the gallbladder* and the quality is the *normal thickness*.

reference anatomical entities, qualities (such as diameter, length, thickness etc.) are represented by PATO [14n], and units are taken from UO (see e.g. figure 6.1 or figure 6.2).

6.1.2. Range Specifications of Clinical Findings

For entities which normally do *not exist* in the healthy human body, such as lesions or cysts, the structure of the representation has to be slightly adapted as shown in figure 6.3. Besides the size of the findings subject (such as a mass) the location is relevant. In comparison to the previously described normal range specification, here, the criteria for a finding are expressed: For instance, a *mass* within the lung has to be at least 3 cm to be documented as a *lung mass* (see figure 6.3).

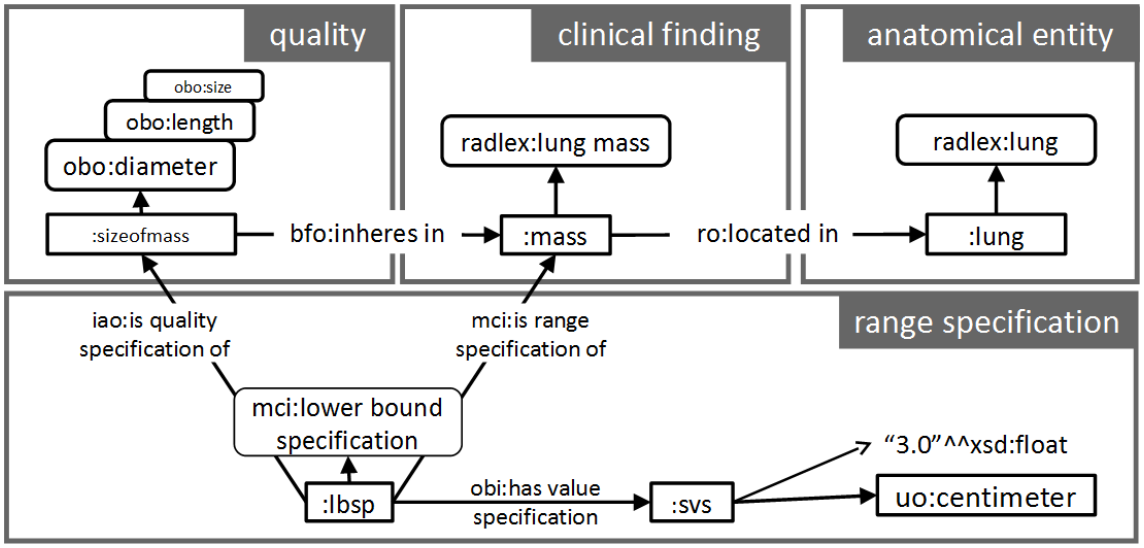


Figure 6.3.: The size specification for a lung mass.

Similarly, the lymph nodes and lesions have to have a minimal size to be selected as target lesions for treatment evaluation according to RECIST guidelines.

6.1.3. Patient-Specific Range Specifications

Some size specifications depend on patient characteristics such as the body height and weight or the age and gender. For instance, the size of the pancreas (subdivided into pancreas head, body and tail) are age-dependent [MR98]. To capture the patient context of a normal or abnormal specification, described above, instances of reference populations are defined which are then linked to the corresponding (ab-)normal range specifications as shown in figure 6.5. A reference population is characterized by at least one quality (such as age), for which a range specification

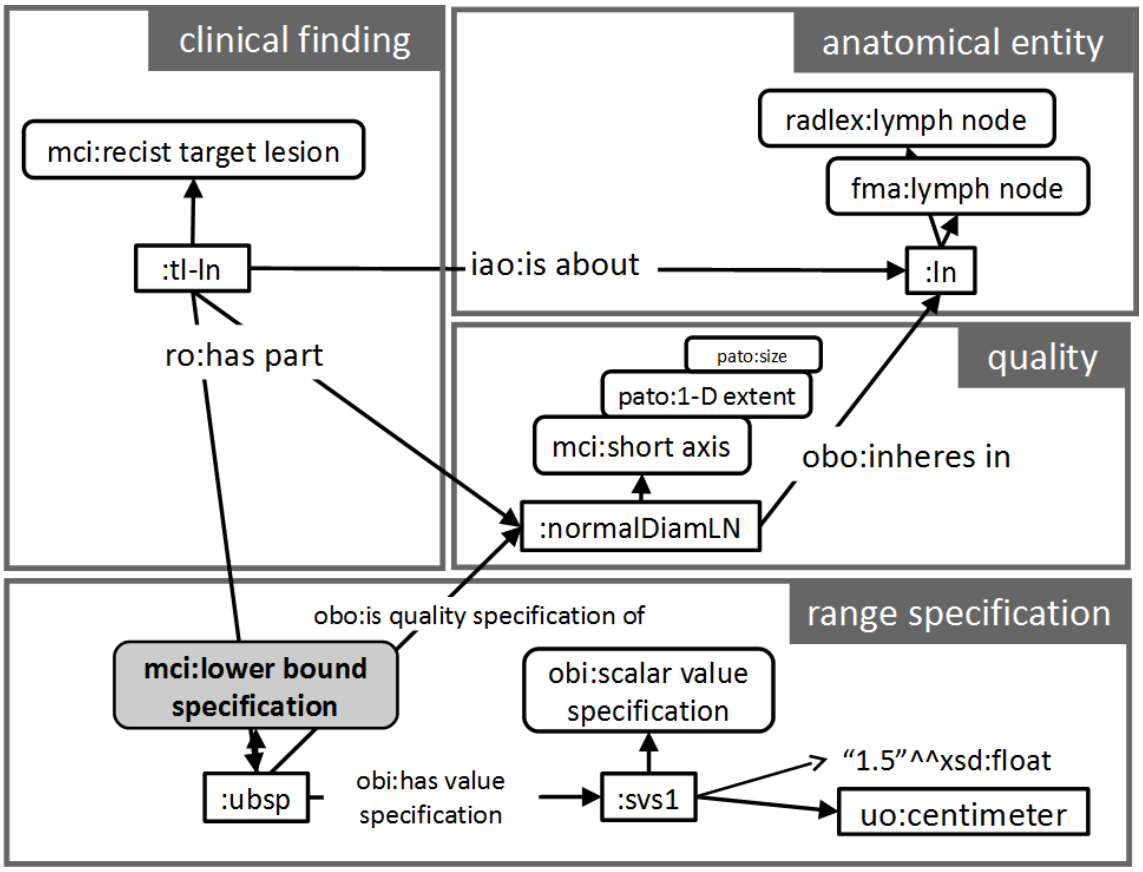


Figure 6.4.: The short axis of a lymph node has to be at least 1.5 cm in order to fit the criteria for RECIST target lesions.

is defined. This is basically the same representation as for normal specifications with an additional link to the corresponding reference population. Note that the definition of a reference population for a range specification is optional and that most of the range specifications are *not* patient-specific.

6.1.4. Coverage

The semantic structure of range specification was described above and illustrated along examples. To evaluate the knowledge model, instance data were created manually using a clinical book about normal findings [MR98] and information from Radiopedia [14o]. The knowledge model on range specifications contains 50 (size) range specifications about 38 different anatomical entities of clinical interest, which are typically measured in radiology examinations of the head, thorax and abdomen area. For instance, the model contains size descriptions for spleen, kidney, gallbladder, pancreas, lymph nodes, aorta, lesion, bile ducts etc. Using the Entity-Quality methodology [Was+09], one range specification potentially applies for many anatomical entities. For instance, the range specification for the normal

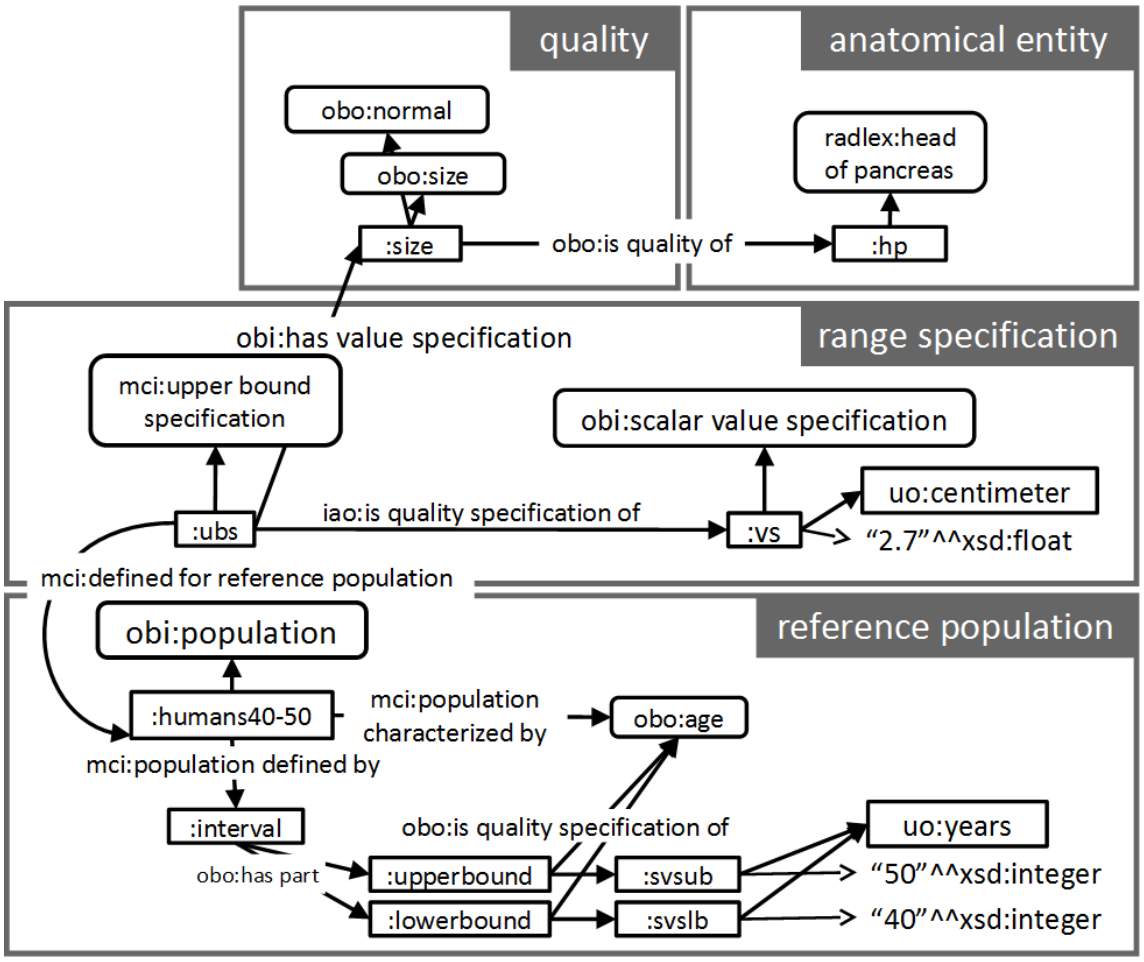


Figure 6.5.: Representation of a patient specific range specification: "The head head of pancreas is normally not larger than 2.7 cm for people between 40 and 50 years".

size of lymph nodes applies more than 200 specific lymph nodes defined in FMA or RadLex. The usage of the model for extraction and classification of size findings from radiology examinations is described in chapter 8.

6.2. Disease-Symptom-Examination Ontology

This section is based on the paper *"Interpreting patient data using medical background knowledge"* [Obe+12a] – however several adaptations of the model were made since then. The main purpose of the Model for Clinical Information is to store patient data, in particular finding descriptions, in a structured and semantic form. However, in the differential diagnosis process as well as in examination planning, available finding information needs to be *interpreted* in the context of diseases and examinations. For instance, a clinician wants to retrieve all *cancer related* findings from all available finding and symptom information about a particular patient. Or, a clinician wants to get a ranking of likely diseases which is based on documented observations. To satisfy this information need one has to capture the relations between diseases and corresponding manifestations in clinical findings and symptoms. Currently a clinician has to have a broad knowledge of diseases and their relation to symptoms and clinical findings. For a better understanding of the relevance of disease-symptom relations (and symptom examination relations) the following paragraph provides a short overview of the work flow of a differential diagnosis:

Differential Diagnosis

1. Based on unspecific symptoms, of a patient the clinician has to map those to most likely diseases which have a strong correlation with these symptoms – so called *leading symptoms*.

2. To differentiate between the different diseases, the clinician has to search for further symptoms of the initial list of diseases which have not been clarified so far.

3. To obtain this information certain examinations have to be performed. Here, the clinician needs to decide which examinations could confirm or exclude diseases. That is, additionally to knowledge about disease symptom relations the relations between examinations and symptoms is needed.

4. The result of examinations provides a more complete picture of the patient. Again, the (larger) set of symptoms needs to be matched with the typical symptoms of diseases and if necessary additional examinations have to be performed.

In the current situation a clinician has to search for all symptoms and findings of a certain disease separately. Further, he has to match the obtained clinical picture to the typical symptoms of *many different diseases* to state a diagnosis. The problem however is, that many diseases are rare and a clinician cannot know

the typical symptoms and clinical findings for all of them. Thus, automatically matching the patient's symptoms to diseases can enhance the quality and efficiency of the diagnostic process. Diseases information relevant within the diagnostic process can be found in medical literature e.g. in [Her11]. First of all, this is the relationship between diseases and their manifestations in clinical findings or symptoms. Further, relations between findings and examinations are needed for integration of finding information in examination planning.

For that purpose, a Disease-Symptom-Examination ontology (DSE) based on MCI was created, that links clinical findings and symptoms with diseases and examinations. Even though existing medical ontologies provide a detailed description of the medical domain this link is not sufficiently represented [Obe+15a]. The main classes required for the knowledge model are the classes *disease*, *clinical finding*, *symptom*, *sign* and *examination*. The representation of corresponding subclasses is described in the following subsections. An overview of the relations between these classes is given in figure 6.6. Since DSE is an ontology – and not an information model as MCI – a new namespace is used for all classes defined by DSE.

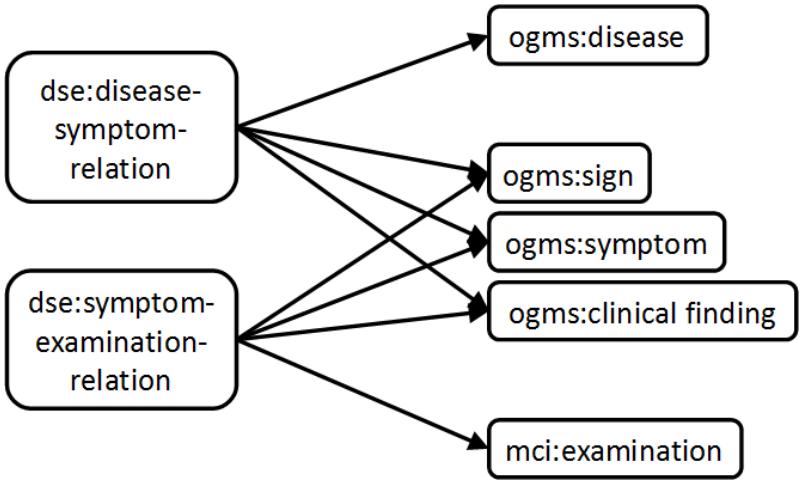


Figure 6.6.: The relation between diseases and their manifestations in signs, symptoms and clinical findings and the corresponding examinations that can be used to check their manifestations.

Within the following subsections, diseases, findings and symptoms, examinations and their relations are described in detail.

6.2.1. Diseases

Instead of defining new classes for diseases, the classes of the Human Disease Ontology (DOID) are reused. The root class `doid:disease` is set as an equivalent class to `ogms:disease` that was defined in MCI. The DOID is part of the OBO

library and contains about 8945 classes for *"descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts"* [Sch+12]. Further, DOID contains many links to other ontologies such as ICD, NCIT, SNOMED CT, Medical Subject Headings (MESH) and thus provides a bridge between the OBO ontologies and UMLS. The OntoFox web service [14j] was used to create a small ontology module of DOID exactly suited for the usecase. More precisely, DSE contains five disease classes (diverticulitis, colorectal cancer, lymphadenopathy Hodgkin's lymphoma and non-Hodgkin lymphoma) and their superclasses as shown in figure 6.7. All annotation properties defined in DOID are also included to preserve labels, synonyms, textual definitions as well as cross references to other ontologies that were defined in DOID.

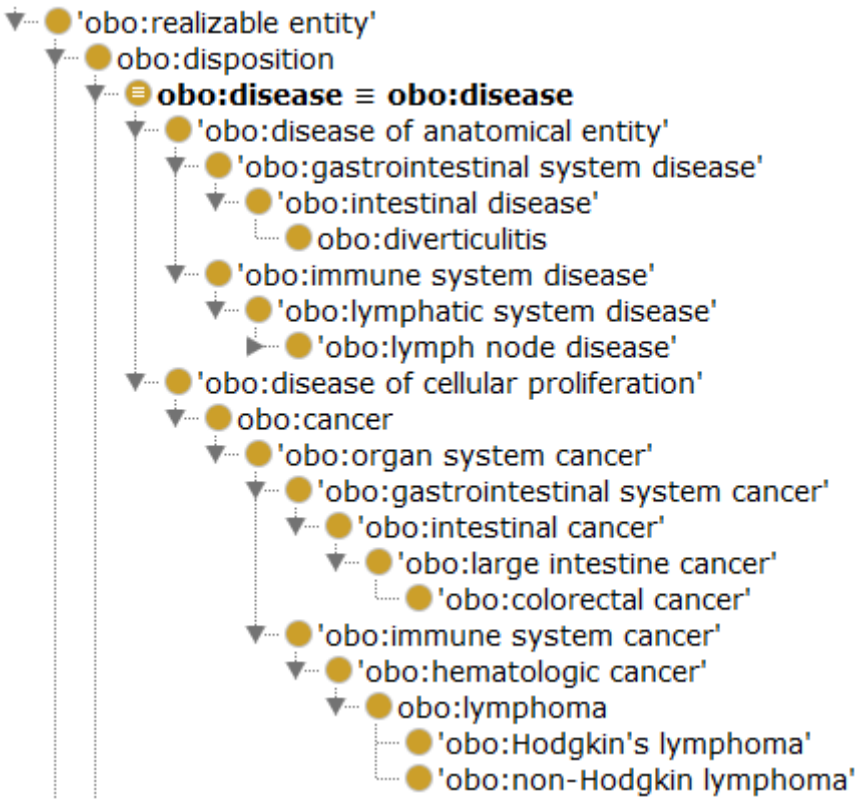


Figure 6.7.: Imported classes from the Human Disease Ontology.

The risk age for a disease is represented through a range specification (upper bound, interval, lower bound) as shown in figure 6.8. Since the specification is references from a class, *has associated risk age* is defined as an annotation property.

6.2.2. Clinical Findings and Symptoms

For clinical findings and symptoms there is no single ontology that could be reused like in the case of diseases. Even though there is a Symptom Ontology

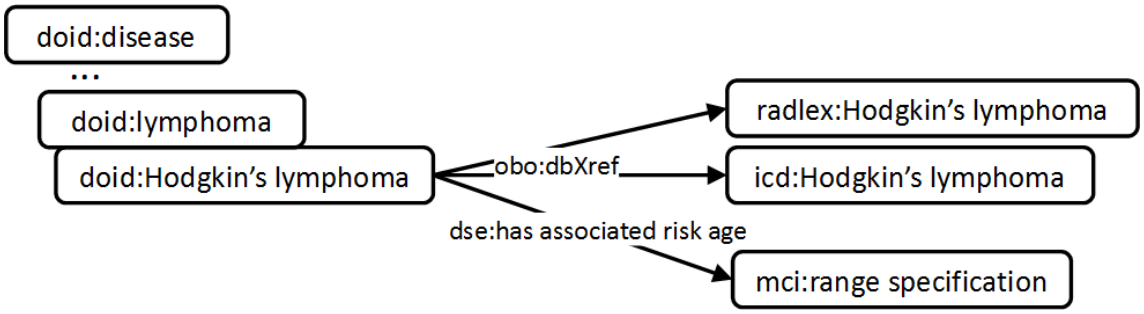


Figure 6.8.: Representation of the disease Hodgkin's lymphoma.

(SYMP), the Human Phenotype Ontology (HP) and other ontologies, none of them contains all symptoms and clinical findings required for the use case data. This is mainly because their description is more complex and involves coordination of classes from different ontologies. For instance while many ontologies include the finding "enlarged lymph nodes", more specific findings such as "enlarged colic lymph nodes" – strongly related to colorectal cancer – are missing. The purpose of DSE however is to link diseases with clinical findings and symptoms. Thus, pre-coordinated classes for corresponding findings and symptoms are required. DSE defines classes for each symptom, sign and finding (see figure 6.9) and provides a label, if possible, logical axioms and also cross references to other ontologies.

Clinical findings and symptoms are represented by the same pattern used by HP. That is, classes are defined with logical definitions and the EQ methodology is applied: For example, the class `enlarged mediastinal lymph node` is defined by the axiom:

```
'dse:enlarged mediastinal lymph node'
      = 'obo:clinical finding'
AND ('obo:is about' SOME fma:mediastinal lymph node)
AND ('obo:has part' SOME ('obo:size' AND'obo:increased quality'))
```

Since the DSE is to be used for the interpretation of finding descriptions specified in MCI, it has to follow the patterns of MCI for clinical findings, i.e. represent findings according to the Entity-Quality methodology [Was+09] that is also used by HP. The FMA is used to refer to anatomical entities and PATO for qualities.

The logical definitions and cross references have two purposes: Firstly, the subclass hierarchy of findings is automatically created along the hierarchy of the anatomical entities . For instance, `enlarged colic lymph nodes` is a subclass of `enlarged abdominal lymph nodes` since colic lymph nodes are a subclass of abdominal lymph nodes in FMA. Secondly, the findings are automatically aligned with HP and other referenced ontologies. For instance, from HP one gets that `enlarged spleen` is an `hp:abnormality of the immune system`.

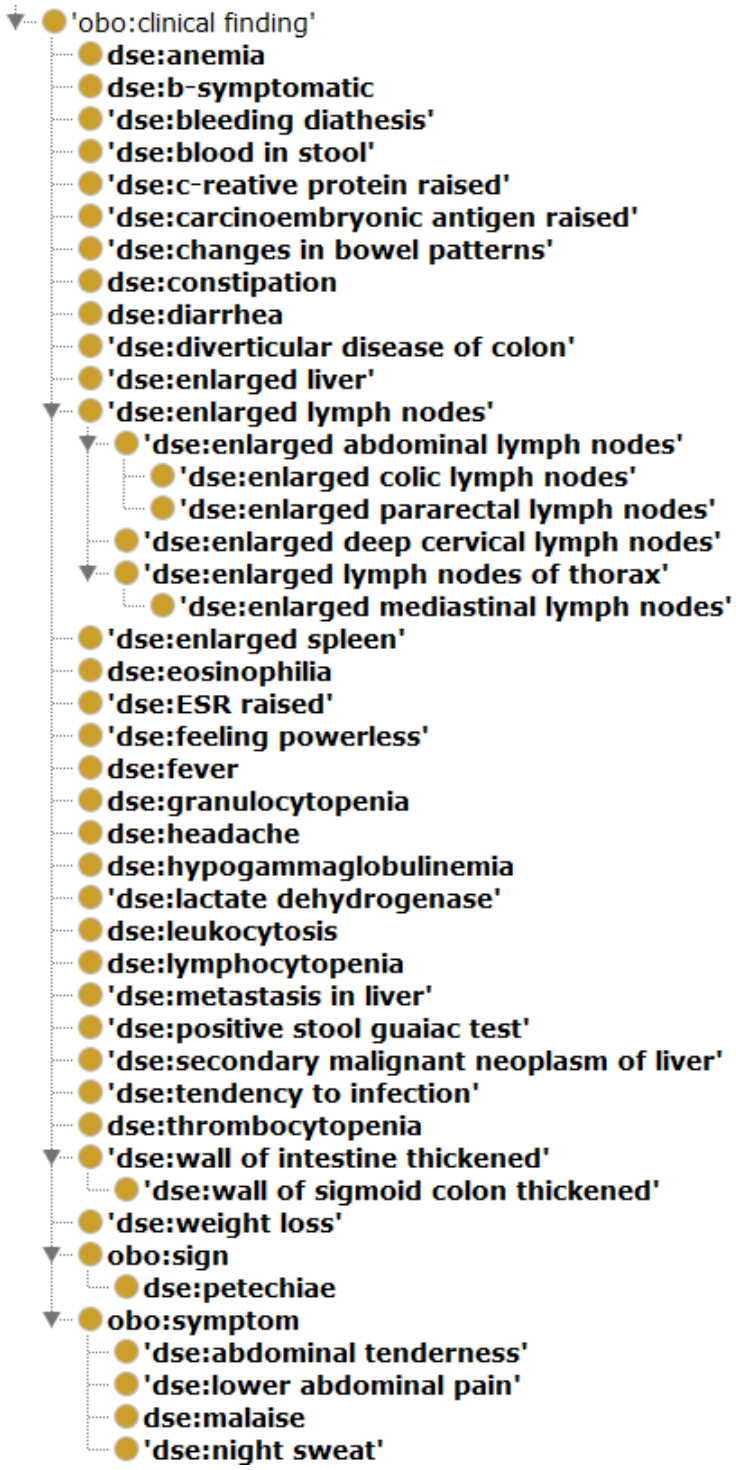


Figure 6.9.: The hierarchy of clinical findings in DSE.

```
‘dse:carcinoembryonic antigen raised’ = ‘obo:clinical finding’
     AND (‘obo:is about’ SOME ‘loinc:Carcinoembryonic Ag’)
     AND (‘obo:has part’ SOME ‘obo:increased quality’)
```

Additionally to the creation of a class hierarchy of findings, further information for single symptom classes useful in the diagnostic process is added. For instance, findings are not equally relevant or *severe*: in general blood in stool is more relevant than a headache, i.e. it is more relevant to clarify the cause blood in stool than the cause of a headache. Additionally the annotation property `has severity` is defined to assign severity (high, moderate, low) to findings and symptoms.

6.2.3. Examinations

Examinations are covered by the Model for Clinical Information, so the classes were used from there. Additionally, classes for stool guaiac test and colonoscopy are added. The classes for examinations are shown in figure 6.10. DSE further provides properties for relating information about the examination such as the cost, duration or risks to the examination classes.

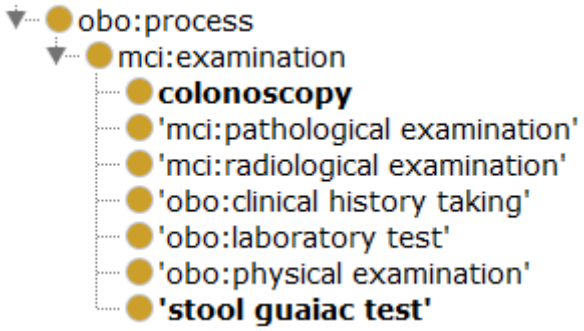


Figure 6.10.: Examinations defined in DSE.

6.2.4. Relations

The general pattern of the disease symptom relation as well as the examination symptom relation was shown in figure 6.6. The relation between diseases and findings/symptoms as well as the relation between examinations and findings/symptoms needs to be represented by an extra RDF node for each relation to allow assignment of extra information to relations. For instance, the frequency, that some symptom occurs with in the context of some disease or the precision of an examination to clarify the status of a certain symptom can be assigned to corresponding relation nodes. An example of a concrete disease-symptom relation is given in figure 6.11

Disease Symptom Relation

Even though DSE distinguishes between clinical findings, symtpoms and signs only one type of relation to diseases is needed. In differential diagnosis, the clinicians commonly refer to them as disease-symptom relations even though not all of the corresponding disease manifestations are symptoms (e.g. results of a blood test are not symptoms). Besides the general disease-symptom relation in differential diagnosis, the notion of *leading symptoms* is important. Certain symptoms are *leading* indicators for specific diseases. For example, a thickened wall of the sigmoid colon is a *leading symptom* for diverticulitis. To express this information, a subclass `disease-leading-symptom-relation` (shown in figure 6.11) of the general disease-symptom relation is defined. Thus, a leading symptom of some disease is also a symptom of a disease. Note that the type of a symptom relation (leading symptom, general symptom) is only defined in the context of some particular disease.

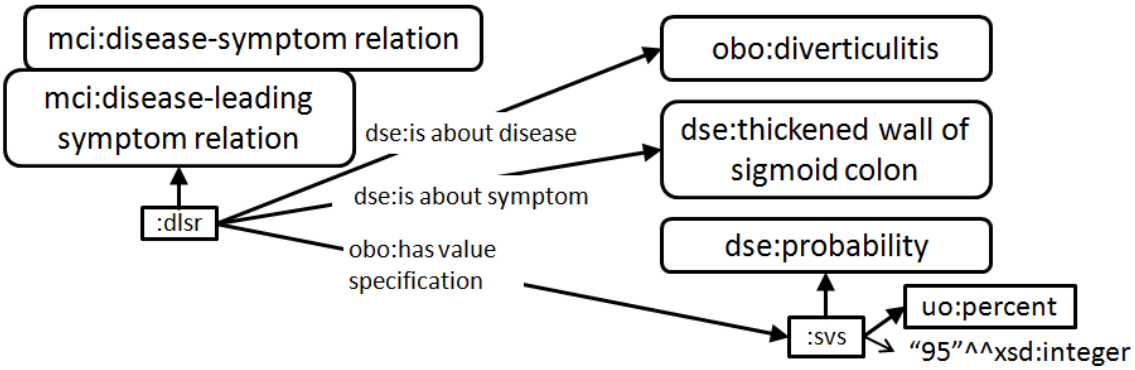


Figure 6.11.: Example of a relation between a disease and a corresponding leading symptom.

Disease symptom relations defined by DSE are not to be understood as "disease always has symptom X" as this is not true in the medical domain. Here, the semantics of this relation simply means that a disease is *correlated* with a symptom. If it is possible, extra information about the concrete probabilities as shown in figure 6.11 is defined. Using OWL it is difficult to express uncertainty such as that a disease *may* be caused by something or that a disease *may* show up by some symptom. For a detailed analysis of that problem it is referred to [RSD08].

Symptom Examination Relations

The representation of relations between clinical findings and symptoms and corresponding examinations that are suitable to clarify the symptom status are quite simple in DSE. However, the same pattern as for disease symptom relations is

used in order to be able to add additional information to the relation such as the precision of an examination to clarify the status.

Part III.

USING MCI: STORAGE, INFERENCE AND ACCESS

7

Architecture

This chapter describes the architecture of realized implementations. That is, an overview of the different technical components and their interaction is provided to allow a better understanding for the general role of those components. Details of specific *realizations* of concrete components, such as the inference component or data access components, are given in later chapters. The first section of this chapter provides a schematic overview of the components before they are described in more detailed in the subsequent sections.

7.1. Overview

The implementation relies on the following main components (see figure 7.1):

- **Source data:** The patient data is provided in different formats. Most of the radiology reports are provided as excel sheets or as structured data in relational databases. For images only the annotations that were created manually by clinicians are available in form of RDF triples.

- **Annotator:** The annotator adds structured information to unstructured data such as images or free text reports. In this work, the annotator is an NLP pipeline for automatic annotation of medical texts. Image annotations were created in a separate manual image annotation tool.

- **Mapper:** The mapper transforms source data and annotations to the schema of MCI which is then loaded into the triple store. While unstructured patient data (i.e. free text reports) is annotated first, structured patient data such as meta-descriptions of reports or data from relational databases are directly mapped to MCI.

- **Triple store:** The triple store is the main storage component. It contains all mapped patient data as well as an ontology repository which provides all used reference ontologies. The data in the triple store is contained in different data sets for the MCI (/mci), patient data (/pdata) and the different reference ontologies (/radlex etc). The triple store can be accessed by a SPARQL endpoint through HTTP.

- **Inference component:** While standard RDFS inference is directly performed by the reasoner of triple store more complex inference algorithms such as the measurement-entity resolution, the normality classification of image findings or the ranking of likely diseases are performed by the inference component.

- **Jetty server:** The jetty server is responsible for request handling for specific applications. HTTP requests are translated to corresponding SPARQL queries and responses are transformed to XML or HTML to be displayed in a user interface. That is, the Jetty server provides APIs for applications such as the ReportViewer or the DiseaseAnalyzer.

- **Applications:** Applications can access the data contained in the triple store either through the Jetty server or by directly querying the triple store. For instance, different demo user interfaces realize the visualization of the data for clinicians.

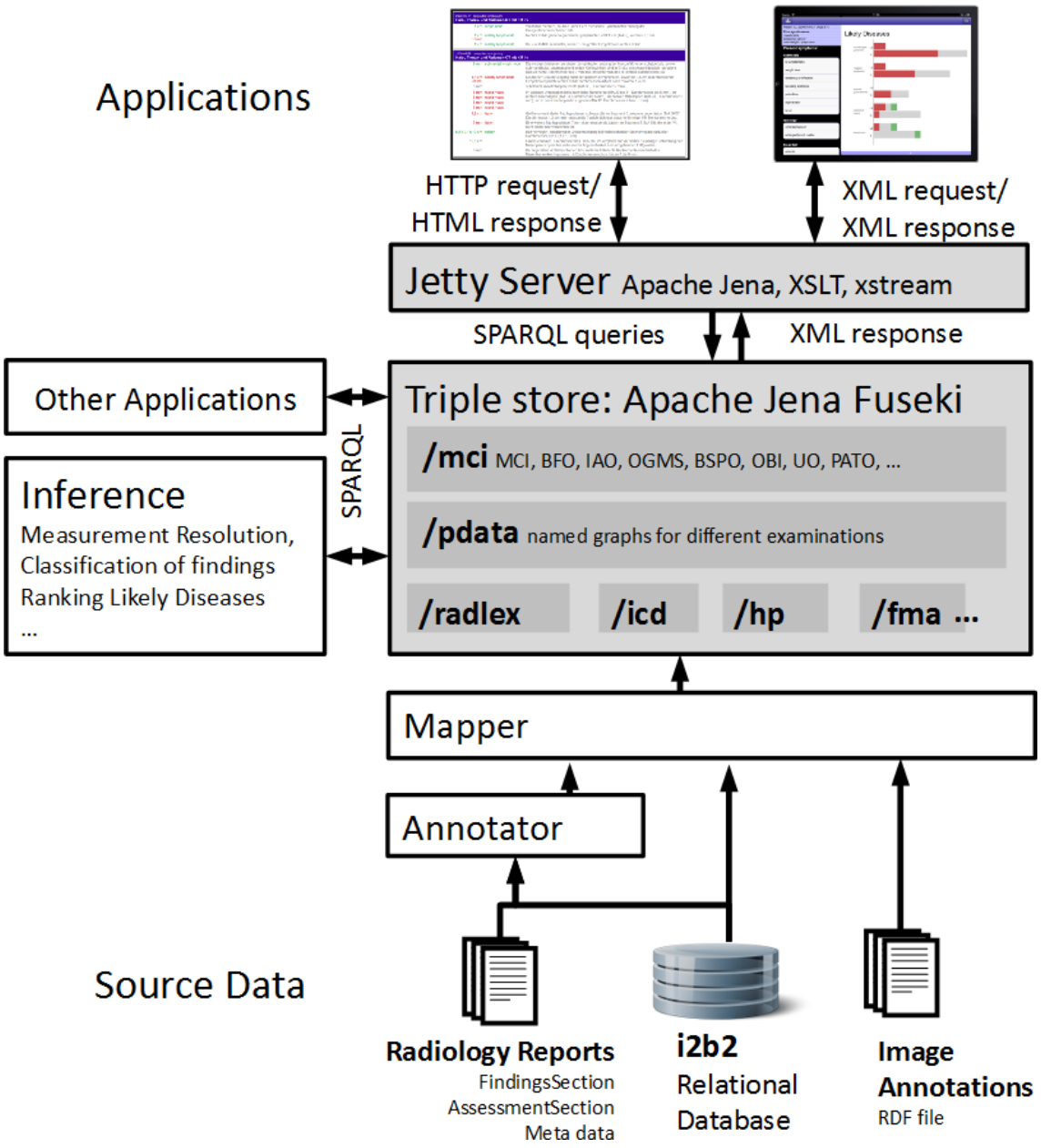


Figure 7.1.: Architecture of the MCI implementation.

7.2. Annotator

Large parts of clinically relevant information is contained in unstructured formats such as images or free text. To integrate this crucial information in a structured form an *annotator* is commonly used which assigns structured information to unstructured data or specific parts of it. The output data are called (semantic) *annotations* and, roughly speaking, they provide a structured representation of the entities that occur in the corresponding unstructured data. For instance, a clinician detects some suspicious anatomical structure in one image. Then, he can annotate the corresponding image region, e.g., with a comment or with entities from some controlled vocabulary. Similarly, a substring of a free text report such as "23 mm" can be automatically annotated as a *measurement* with value "23" and unit "millimetre". In general, the capabilities of annotators range from fully automatic (e.g., image segmentation, NLP) to fully manual. The annotator, i.e. the component that creates the annotations, is considered as a black-box. The contribution of this work regarding annotations is to integrate image and text annotations into a common schema using the Open Annotation data model (OA) which is part of the model foundation of MCI as described in chapter 4. Annotations created by NLP pipelines, image segmentation algorithms or manual image annotations are mapped to this common schema. In the following subsections, the image and text annotations used in this work are briefly described.

7.2.1. Image Annotation

Using the MEDICO annotation tool, semantic image annotations are created semi-automatically as described in [Sei+10]: important organs are segmented automatically and corresponding image regions are annotated with respective classes from the FMA. The clinician then validates the automatic annotations and might add additional annotations manually. For instance, the clinician might add to an organ, e.g., the spleen, a modifier for the size, e.g., *enlarged*. The clinician can also view the 3D scan and mark specific regions and selects ontology concepts for annotation independently of any image segmentation algorithm. The annotations created by the MEDICO annotation tool are stored in the MEDICO Annotation Ontology (MANO), which is described in [MRS09; Sei+10; Sei+11]. The MANO reuses classes from FMA to represent anatomical entities *"whereas concepts for the visual manifestation of an anatomical entity on an image are derived from the modifier and imaging observation characteristics sub-tree of RadLex"* [Sei+10]. Additionally, free text descriptions might be used to annotate image regions. The basic schema of the MEDICO annotation is given in Figure 7.2: There are four types of annotations (spacial, anatomical, describing and measurement) which are linked to a finding. The finding is part of a study of some patient and, optionally, has additional references to an image region or some text region.

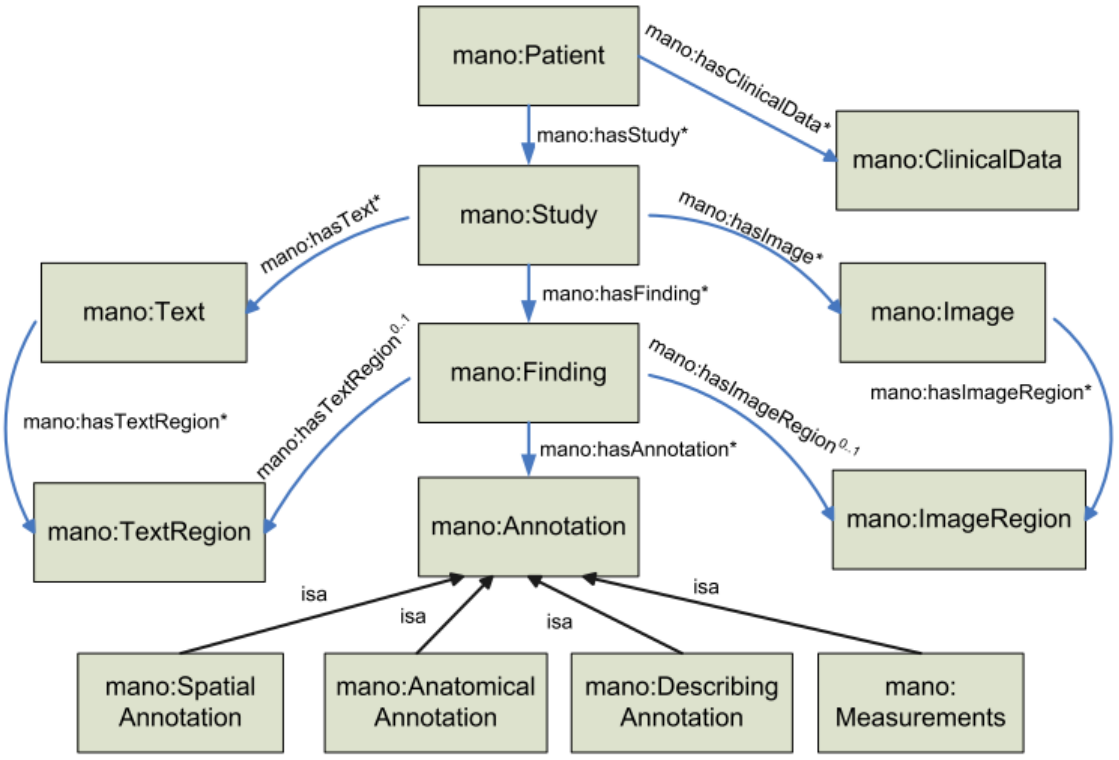


Figure 7.2.: Schema of MEDICO Annotation Ontology (MANO) [Sei+11].

7.2.2. Text Annotation

The description of the text annotator is based on [Obe+14]. The component that takes care of the extraction of relevant information from unstructured clinical texts is the *text annotator*. Information Extraction (IE), as a task of NLP, is a technique to find important pieces of information in unstructured texts and to extract them as structured information [Mey+08]. For instance, IE is used to detect semantic entities such as date values, names or measurements in texts. Furthermore, the annotator can detect entities of a provided dictionary, e.g., the controlled vocabulary of an ontology. This task is also referred to as Named Entity Recognition (NER) or *semantic annotation*. There are freely available annotators like, e.g., the annotator of the BioPortal [Whe+11], for annotation of text with concepts from biomedical ontologies. However, a more specific annotator tailored for German medical texts is needed, so the annotator described in [BOZ15] is used.

Functional Scope of the Text Annotator

The used annotator extracts two main pieces of information. Firstly, measurements, date values, abbreviations etc., i.e. predefined information objects. These entities

are extracted using pattern-based techniques (i.e., through detection of utterances by using a set of regular expressions that express predefined combinations of measurement values and units). Secondly, the annotator can detect entities of a provided ontology. There are three features that the annotator supports regarding this semantic annotation:

- It detects multi-word terms independently of the ordering of the individual tokens.
- The annotator respects the sentence boundaries and maps multi-word terms only when they occur within the same sentence.
- It recognizes inflected forms of ontological concepts in the text such as plural form or other grammatical inflections based on stemmed forms.
- The annotator is able to split compound terms into their parts. This is especially useful regarding annotation of German texts since the parts of compounds would not be recognized otherwise.

The implementation of the annotation pipeline builds on the Unstructured Information Management Architecture (UIMA) framework [14a]. The annotator itself is an adapted version of the UIMA Concept Mapper, which annotates texts that are pre-processed by a special medical text pre-processing pipeline including sentence splitting and tokenisation. The output of the annotator is represented in RDF which allows easy integration with MCI . Details of the functional scope of the annotator are described in [BOZ15]. An example of a measurement annotation of the sentence "Axillary lymph node 56 x 45 mm" is shown in figure 7.3.

```
:AnnotationType-1772554695
  a               oa:Annotation ;
  rdfs:label      "56 x 45 mm"^^xsd:string ;
  :annotationID   "AnnotationType-1772554695" ;
  oa:hasBody      :MeasurementAnnotation-2013480429 ;
  oa:hasTarget    :Token-1078343793 .
:MeasurementAnnotation-2013480429
  rdfs:label       "56 x 45 mm"^^xsd:string ;
  :annotationID    "MeasurementAnnotation-2013480429" ;
  :dimensions      2 ;
  :measures        "45.0"^^xsd:float , "56.0"^^xsd:float ;
  :uimaType        "de.reports.types.MeasurementAnnotation"^^xsd:string ;
  :unit            "mm"^^xsd:string .
:Token-1078343793  a          nif:Word ;
  rdfs:label       "56 x 45 mm"^^xsd:string ;
  nif:beginIndex   20 ;
  nif:endIndex     30 ;
  nif:sentence     :Sentence-1887026300 ;
  nif:stem         "56 x 45 mm"^^xsd:string ;
  :text            "56 x 45 mm"^^xsd:string ;
  :uimaType        "de.reports.tokenizer.Token"^^xsd:string .
```

Figure 7.3.: Measurement annotation created by the UIMA text annotator for the sentence “Axillary lymph node 56 x 45 mm.”

7.3. Mapper

The mapper transforms all relevant structured data, i.e. annotations as well as data from relational data bases, to the schema of the MCI. As described above, the textual annotator takes a string as input and outputs annotations serialized in RDF. Similarly, the MEDICO image annotations are provided in RDF and stored in the MEDICO Annotation Ontology. The mapper fulfills the following tasks:

- Integration of annotations from different annotators to the schema of OA.
- Filtering and structuring of annotations since not all annotations are mapped to MCI.
- Normalization of annotations, e.g., dates, measurements, units from strings to URIs.
- Representation of the context of the annotations. Since only parts of the source data are annotated (e.g. the finding or assessment section of a report) annotations need to be related to corresponding entities in MCI such as the report section instance.
- Creation of consistent URIs for entities: A string pattern for the creation of URIs based on identifiers from source data sets is necessary in order to preserve references between entities.
- Transformation reference codes to URIs of classes from ontologies. For example an ICD10 code (C00-D48.9) to the corresponding class URI from the ICD10 *ontology*.

Annotations which are already in RDF format can be simply transformed by using SPARQL queries. Structured data from relational databases needs to be queried with SQL and, based on that, triples need to be generated and loaded to the triple store.

7.3.1. Mapping of Image and Text Annotations

To map image and text annotations, the corresponding RDF file containing the annotations is loaded. Then, a set of SPARQL queries is used to retrieve specific triple patterns which are then transformed to the schema of MCI and uploaded to the triple store. One important task thereby is to consistently create URIs for the created entities so that they are correctly referenced by different queries. Furthermore transformation of triple patterns, involves the transformation of entities,

properties and the creation of additional entities. The concrete transformation of data and annotations from MANO (image and text annotations) and UIMA (text annotations) are described along examples in the following subsections.

Mapping of MANO Image and Text Annotations

The MANO defines several different data properties to represent unique IDs, e.g., patientUID, findingUID, studyUID, radiologicalReportUID etc. The instances however are modelled as blank nodes in MEDICO. To be able to consistently refer to instances, URIs need to be created for the blank nodes. As shown in figure 7.2, the MEDICO schema contains classes for patient and study which are directly linked. In MCI, a DICOM study is the output of a radiology examination which has the patient as a participant. The transformation of this pattern is shown in figure 7.4. These transformations were created manually.

```
CONSTRUCT {
 ?patient  a :MCI_0000150 ;              # a patient
           :MCI_0000090 ?patientUID .    # with unique ID
 ?radExam  a :MCI_0000003 ;              # a radiology examination
           obo:BFO_0000057 ?patient ;    # has participant
           obo:OBI_0000299 ?dicomStudy .# has specified output
 ?dicomStudy a :MCI_0000052 ;            # a dicom study
           :MCI_0000090 ?studyUID .      # with unique ID
 } WHERE {
 ?p        a mano:Patient ;
           mano:patientUID ?patientUID ;
           mano:hasStudy/mano:studyUID ?studyUID .
 BIND(CONCAT("http://.../mci.owl/patient",?patientUID)) AS ?patient)
 BIND(CONCAT("http://.../mci.owl/exam",?studyUID)) AS ?radExam)
 BIND(CONCAT("http://.../mci.owl/dicomStudy",?studyUID)) AS ?dicomStudy)
}
```

Figure 7.4.: SPARQL query to map structured patient information from MANO to MCI.

The mapping of image annotations is more complicated. In MEDICO, a finding has 1 - 9 annotations of various types. In MCI, annotations are represented using classes and properties of OA, where a composite construct is used to bind multiple annotations as shown in figure 7.5. These annotations are then the basis to create instance of clinical findings using MCI classes and properties.

Mapping of UIMA Text Annotations

The text annotations produced by the annotator are described in section 7.2. Since the annotations created by the UIMA text annotator are very detailed and include

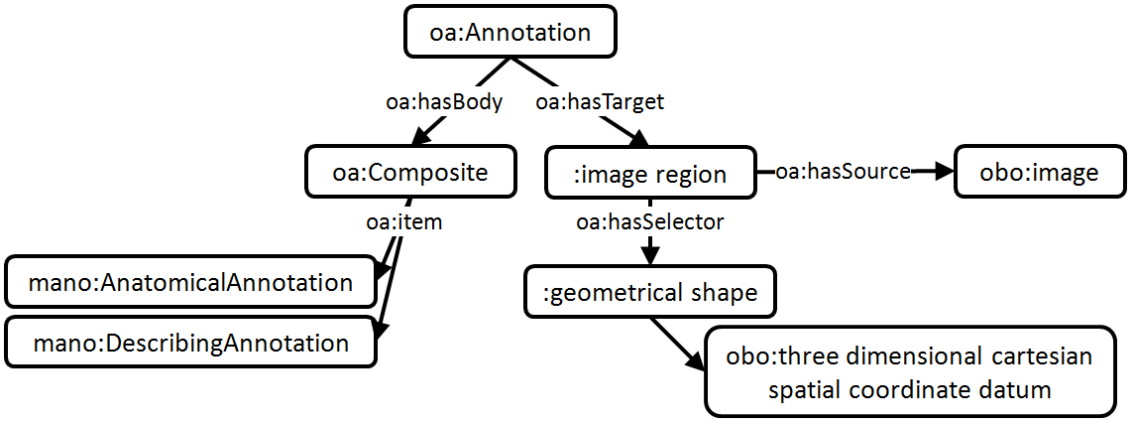


Figure 7.5.: Target representation of an image annotation transformed from the MEDICO Annotation Ontology (MANO).

all kind of linguistic annotations, the mapper needs to filter and normalize these annotations for their representation in MCI. For instance, information about the stemmed form or the UIMA type are filtered out in this step. The measurement annotation shown in figure 7.3 is directly transformed to a size finding which is contained in a specific sentence of the corresponding report section. Furthermore, two instances of length measurements are created which describe length qualities that are orthogonal to each other. The unit (represented by the annotator with a string) is mapped to a normalized entity of the Units Ontology (UO). In the example, the length would be normalized and represented in centimetres. This is depicted in figure figure 5.28, though the relation to the lymph node is omitted. The relation between measurements and corresponding anatomical entities is not provided by the annotator and needs to be inferred in subsequent steps.

7.3.2. Mapping Data from Relational Data Bases

There are different possibilities to integrate data from relational data base (RDB) into semantic web applications. Firstly, one can transform the data from the RDB into an RDF representation, which is referred to as *triplification*. Secondly, one can rewrite SPARQL queries to SQL queries in order to access an RDB by semantic queries. In cases where the underlying data is very large or when the RDB is frequently updated, the latter approach is more appropriate. Since the data from the RDBs considered in this work is supposed to be rather static and several manipulations and mappings are necessary for the applications, the triplification of relational data was chosen.

i2b2 Transformation

The relational data base i2b2[15h] is a data warehouse, specifically designed to integrate clinical data from various hospital departments. The main purpose of i2b2 is to identify patient cohorts based on fundamental patient characteristics such as the main diagnosis, the gender or the age. To accomplish this, i2b2 integrates different classification systems such as the International Classification of Diseases (ICD), German procedure classification (OPS), Anatomical Therapeutic Chemical (ATC) for medication and Logical Observation Identifiers Names and Codes (LOINC) for lab values.

Structured and coded data elements such as diagnosis information are directly mapped to MCI. Unstructured data such as finding descriptions from radiology reports are additionally annotated (via the annotator described above) and then mapped together with the annotations to MCI. The mapping of annotations is described below.

Technically, mapping of structured data from i2b2 to MCI is realized through several SQL queries which retrieve data elements from the relational i2b2 data base. Then, the data is mapped to MCI using JAVA and uploaded to the triple store using SPARQL. In the transformation (or triplification) process, consistent reference to URIs for the different entities is required. Here, the IDs of i2b2 are used to create new URIs, if necessary by appending existing identifiers. For instance, while a radiology report has an ID in i2b2, the different sections do not have predefined IDs. Further, for all entities that are created additionally to entities defined in i2b2, new URIs need to be created. For a patient with patient_num 1000000006 the following URI is created:

http://www.siemens.com/ontologies/mci.owl/i2b2-patient-1000000006

or short

mci:i2b2-patient-1000000006

The basic patient information is then represented as

```
mci:i2b2-patient-1000000006 a mci:patient ;
   mci:date of birth "1946-01-01T00:00:00"^^xsd:dateTime ;
   mci:has patient ID "1000000006"^^xsd:string .
```

7.4. Triple Store

The Apache Jena triple store [15a] contains all reference ontologies, the MCI and the patient data represented in MCI. In general, there are two technical possibilities to separate data in a triple store. Firstly, it is possible to create different data sets. For each data set, it can be specified which services (query, update, load etc.) are provided by the corresponding SPARQL endpoint. *Federated queries* are used to query over two or more data sets. Secondly, within each data set, named graphs can be created that group triples additionally. Technically this is realized by using quadruples where (in contrast to classical triples) one additional entry specifies the graph the triple belongs to. The separation in data sets and named graphs is employed in the triple store as shown in figure 7.6. The main advantage of this kind of separation is that one can more easily maintain and modify data. For instance, if a new version of some ontology is available one can simply delete the old content of one data set and upload the new version. Since data sets are separated this also leads to better query performance when one queries only one or two data sets. Additionally, the separation allows us to have different reasoning levels for the different data sets. MCI is held with OWL-reasoning while the patient data and the referenced ontologies without any reasoning. Furthermore, named graphs allow to group patient data triples with respect to the context of clinical encounters. The separation of triples belonging to different clinical encounters is necessary since, e.g., some clinical department might realize different roles within the context of different encounters (e.g., admission or discharge role). For instance, the patient's age at admission makes sense only within the context of some clinical encounter. The triple store does not support any user or transaction management.

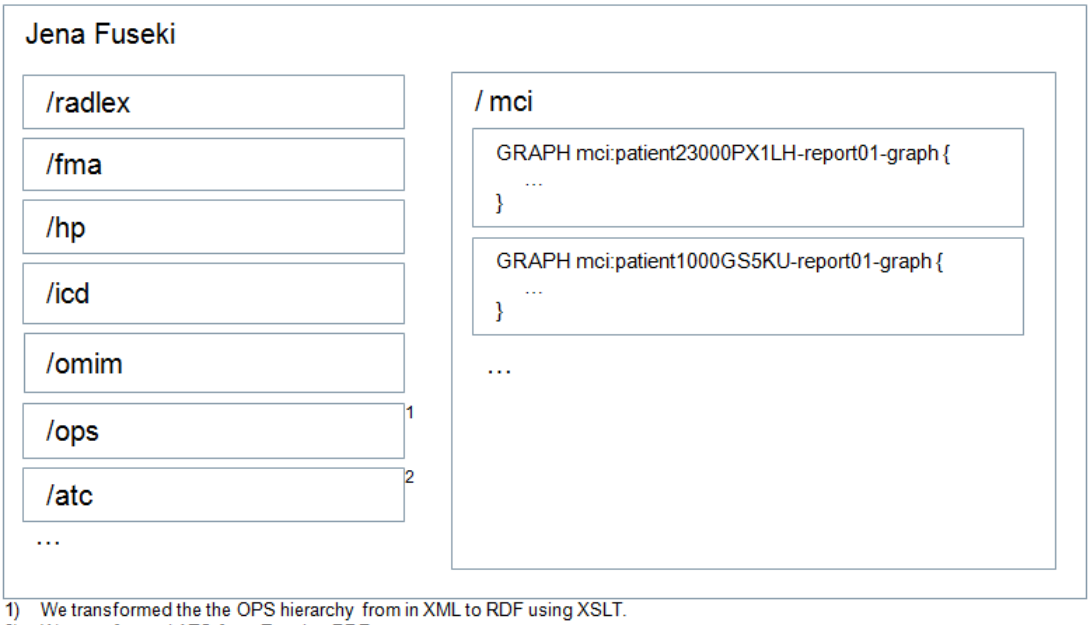


Figure 7.6.: Triple store with different data sets indicated by '/...' and named graphs.

7.5. Inference Component

The role of the inference component is to enrich patient data stored in MCI using the knowledge models described in the previous chapter. For instance, the relations between measurements and corresponding anatomical entities is inferred by this component since this relation is not provided by the annotator. Similarly, the classification of size findings based on comparison with normal values or the interlinking of findings from consecutive examinations is done by the inference component. The implementation of this component is realized as follows (for details it is referred to chapter 8):

1. **All triples necessary for inference are retrieved.** For instance, all size findings without a classification as normal or abnormal.

2. **Based on these triples, knowledge models are used to infer new relations.** For instance, the measurement values of size findings are compared to normal size specifications defined in the knowledge model to classify them as normal or abnormal.

3. **New relations are added to the triple store.** For example, the resulting finding type is added to the triple store.

Some inference procedures involve multiple SPARQL queries and a complex algorithm with several steps, e.g., the measurement entity resolution described in section 8.2. However, some procedures are realized by single SPARQL update queries such as the linkage of consecutive examinations.

8

Inference

This chapter describes different inference techniques and their role for enriched data representation – beginning with an overview in the first subsection. It is described how standard RDFS and OWL reasoning is performed and how the knowledge models presented in chapter 6 are applied technically. Afterwards, specific inference procedures requiring extended algorithms are described in detail: Based on semantic annotations of radiology reports the knowledge model on normal size specifications (section 6.1) is used to resolve the relation between measurements and corresponding anatomical entities. After this resolution, complete structured representations of measurement findings have been extracted from their previous representations in free text reports. Based on these structured representations a subsequent inference step is performed to classify these findings as *normal* or *abnormal*. Furthermore, their deviation in comparison to previous examinations is highlighted. Finally, the Disease-Symptom-Examination ontology (DSE) is used to enable a disease-centric access to the finding descriptions: By matching the patient's symptoms to the typical symptomatology of different diseases, a ranked list of likely diseases is inferred and subsequent examinations are proposed.

8.1. Overview

There are different possibilities to infer additional information from the patient data and annotations, i.e., using logical definitions from reference ontologies, MCI and the knowledge models. The basics of knowledge representation and reasoning are described in section 2.1. Here, it is focused on the description of how inference techniques and procedures are used to enhance data of the case studies presented later in chapter 10. There are several ways to realize inference: Firstly, the reasoner of the triple store can be used for inference based on logical axioms and definitions defined in reference ontologies, MCI and the knowledge models. For instance, RDFS reasoning is used to automatically infer the transitive closure of subclass relations for reference ontologies such as the RadLex or the ICD. Secondly, more complex algorithms are commonly realized in external applications. Similarly, when dealing with higher inference levels such as OWL or OWL2, it is appropriate to first extract a subset into an in-memory ontology model such as a Jena Inference Model to perform inference within that model – instead of running the inference on the whole data set of the triple store. This strategy is also suitable to keep the inference level of most of large data sets of the triple store low. Thirdly, SPARQL update queries are used to infer information based on *rules*. This technique is mainly used for data normalization and cleaning, however it can be also used to perform inference that would be difficult or impossible to realize by using DL. For instance it is not possible to write an axiom such as the following in DL:

$$hasParent(x,y) \wedge hasParent(x,z) \wedge married(y,z) \rightarrow C(x)$$

which defines a "class C of children whose parents are married" [KMK11]. Also, comparison of data values is better done by using SPARQL than by logical axioms.

8.1.1. Inference with RDFS and OWL

A very important role of inference in the context of description logic is the classification of instances. The most important knowledge contained in reference ontologies is the generalization hierarchy of their classes. By specifying an RDFS reasoner of the triple store for a corresponding data set such as those containing the FMA, RadLex or ICD, one can easily include this knowledge at query time. For example, a diagnosis of type `icd10:Hodgkin's disease` is also classified as type `icd10:Malignant neoplasms` and `icd10:Neoplasms` due to the subclass hierarchy of ICD. Further domain and range restrictions of properties are used to classify instances related by corresponding properties. Similarly, logical definition of classes are commonly used by reference ontologies. Firstly, to allow classifications of instances as in the above example and secondly to automatically infer the subclass

hierarchy of the ontology. For instance, the Human Phenotype Ontology (HP) defines classes of phenotypes by logical axioms such as:

```
hp:splenomegly = obo:has part SOME (obo:increased size
                 AND (obo:inheres in SOME obo:spleen))
```

A clinical finding describing an increased size (quality) which inheres in some spleen is thus automatically classified as an `hp:splenomegaly`. By applying the HP subclass hierarchy, the finding gets further classified as `hp:Abnormality of the spleen` and also as `hp:Abnormality of the lymphatic system` etc. As explained in section 6.2, the Disease-Symptom-Examination ontology (DSE) follows this pattern and defines clinical findings such as *enlarged lymph nodes*, or *raised carcinoembryonic antigen* accordingly.

Using the Structure of Reference Ontologies

The relations between different classes or instances defined in existing reference ontologies are another valuable knowledge resource. For instance, the aggregation relations such as `has part` are widely used. Often these aggregation relations are defined as *transitive* properties. Inference about transitive properties however requires OWL reasoning level. Thus, by specifying an OWL reasoner, one can retrieve all parts of some given entity. However, since OWL reasoning is computationally expensive, is not feasible as the default reasoning level in the context of very large ontologies or data sets. An alternative is to use specific SPARQL constructs to get the transitive closure at query time. For instance, a query such as

```
SELECT ?entity ?part
WHERE {
 ?part fma:part-of+ ?entity
}
```

can be used to get the transitive closure of all parts for each entity without requiring a running reasoner.

8.2. Resolution of Measurement-Entity Relations

A large percentage of clinically relevant radiologic patient information is represented in unstructured formats such as free text reports. Measurements represent important information documented in reports. On the one hand clinicians measure only things of importance, on the other hand measurements are comparable and thus provide valuable insights into the change of the patient's health status over time. In radiology, there are mainly size measurements describing the spatial extent of anatomical entities. For instance, radiologists measure the size of tumors and metastatic lesions (characteristic changes of parenchyma in different organs, enlarged lymph nodes) in consecutive examinations to evaluate response to treatment. Currently, radiologists and clinicians need to manually collect measurements from different reports in order to compare the respective values. Sometimes they even need to go back to the original image and measure the entities again. Since measurement information is only useful when it is exactly known what the measurement *is about*, a mechanism to resolve *measurement-entity relations* automatically from annotated texts was developed. The results of the inference described in this section lead to a complete and structured representation of measurement findings and thus provides the basis for their classification and linkage. The overall aim is to facilitate and speed up the comparison of measurements from consecutive reports. To illustrate the challenge of extracting measurement-entity relations, consider the *example sentence*:

Example Sentence: "Enlarged lymph node right paraaortal below the renal pedicle now 23 mm."[1]

The resolution algorithm has to extract that *23 mm* specifies the size of a *lymph node* (or better a paraaortal lymph node), i.e. the *relation* between the measurement and the measured anatomical entity. This is especially challenging in long sentences, where many different entities occur with the measurement and the measurement is not close to the entity described – in the example sentence above, the entity *lymph node* is at the beginning while the *measurement* is at the end of the sentences. The annotator (described in section 7.2) is used to detect and extract the measurement value *23* and unit *mm* as well as ontology concepts. For the example sentence the following set of RadLex annotations are obtained: `lateral aortic lymph node`, `right`, `lymphadenopathy`, `enlarged`, `lymph node`, `inferior para-aortic lymph node`, `renal pedicle`, `inferior`, `paraaortic` and `kidney`. The resolution algorithm has to extract the correct measurement-entity relation based on this set of annotations.

[1] Original sentence in German "Vergrößerter Lymphknoten rechts paraaortal unterhalb des Nierenstiels jetzt 23 mm."

8.2.1. Overview of Resolution Algorithm

For each sentence containing at least one measurement, the set of annotations is analyzed to infer the anatomical entity the measurement *is about*. The resolution of measurement-entity relations is based on a combination of three ranking algorithms (see figure 8.1):

- **Knowledge-based Resolution:** The knowledge model about normal size specifications for anatomical entities (see section 6.1) and typical size specifications for other findings as well as the structure of a reference ontology is used.
- **Statistics-based Resolution:** Based on 1000 manually resolved sentences a statistic about the frequency of commonly measured anatomical entities is used.
- **Distance-based Resolution:** The distance within the sentence between the measurement and annotations is used for the resolution.

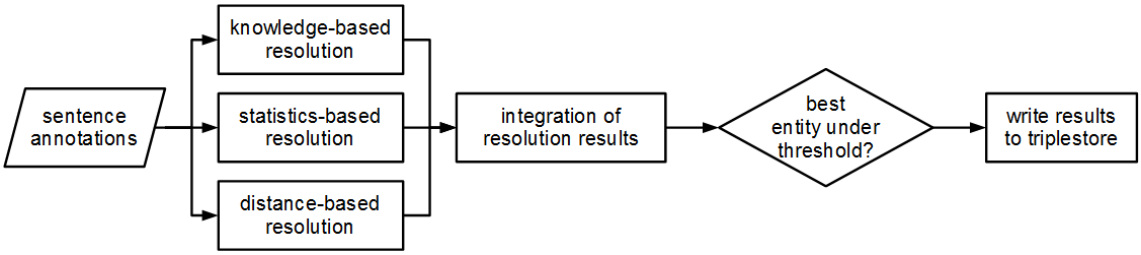


Figure 8.1.: Overview of the measurement-entity resolution algorithm that combines a knowledge-based approach with a statistical and a sentence distance measure.

Most important here is the *knowledge-based* algorithm described in detail in the following subsection (see also [Obe+14]). The statistics-based algorithm as well as a simple distance-based algorithm refine the results and help to avoid false resolutions of the pure knowledge-based algorithm described in [Obe+14]. The refinement is realized by an ensemble algorithm that integrates all three rankings with weights and *the best entity is selected* in dependence on some threshold criteria. For the example sentence described above the entity `inferior para-aortic lymph node` is selected for the resulting relation to the measurement *23 mm*. Using the threshold, selection of wrong entities is avoided in some cases and thus enhances precision. Not selecting any entity can have two reasons: Either the knowledge model does not contain information about the annotated entities or all the set of annotation is of low quality and the correct entity is not annotated.

8.2.2. Knowledge-based Resolution Algorithm

This subsection is based on the publication [Obe+14]. The knowledge-based resolution is based on the annotations, the knowledge model and the structure of RadLex. The idea of the knowledge-based approach is to abstract from the sentence structure and use a knowledge model containing information about the typical size of anatomical entities in combination with hierarchical information of a medical ontology (in this case RadLex). The knowledge-based algorithm relies on six steps: First, the set of annotations is *filtered* (1) and *extended* (2) using the ontology structure of RadLex. Then, a *spanning tree* (3) covering the annotations is created, *corresponding size specifications are attached* (4) from the knowledge model, *measurement values are compared to them* (5) and a *ranking of entities* (6) is computed. These steps are explained in detail in the following paragraphs before limitations are analyzed.

1. Filter Annotations Not all generated annotations are good candidates for the measurement-entity resolution: For instance, while `lymph node` is a good candidate, `inferior` is not. In RadLex, good candidates are subclasses of `anatomical entity`, `imaging observation` and `clinical finding`. Annotations under these classes are referred to as *relevant annotations*. All other annotations, e.g. those under `imaging modality`, `procedure`, `medical device` or `Radlex descriptor` are filtered out. This removes `right`, `enlarged`, `inferior`, and `paraaortic` from the list of annotations of the example sentence above.

2. Extend Annotations In RadLex, some clinical findings are linked to anatomical entities by the property `radlex:Anatomical_Site`. For example, the finding `radlex:hepatomegaly` is linked to `radlex:liver` by this property. The initial set of annotations is extended by using this property. That is, for each initial annotation RadLex is queried for related anatomical entities which are then added to the set of annotations. Thus, annotations for anatomical concepts are added, which could not be detected directly by the annotator.

3. Create Spanning Tree Using the set of relevant annotations, a minimal spanning tree is created from the RadLex subclass hierarchy. The spanning tree for the annotations of the example sentence is shown in figure 8.2. The spanning tree is represented in RDF, like the knowledge model on normal size specifications.

4. Attach Size Specifications For all entities of the spanning tree, the knowledge model is checked for size information about the respective entity and attached to the spanning tree. Regarding the example case, size specifications for

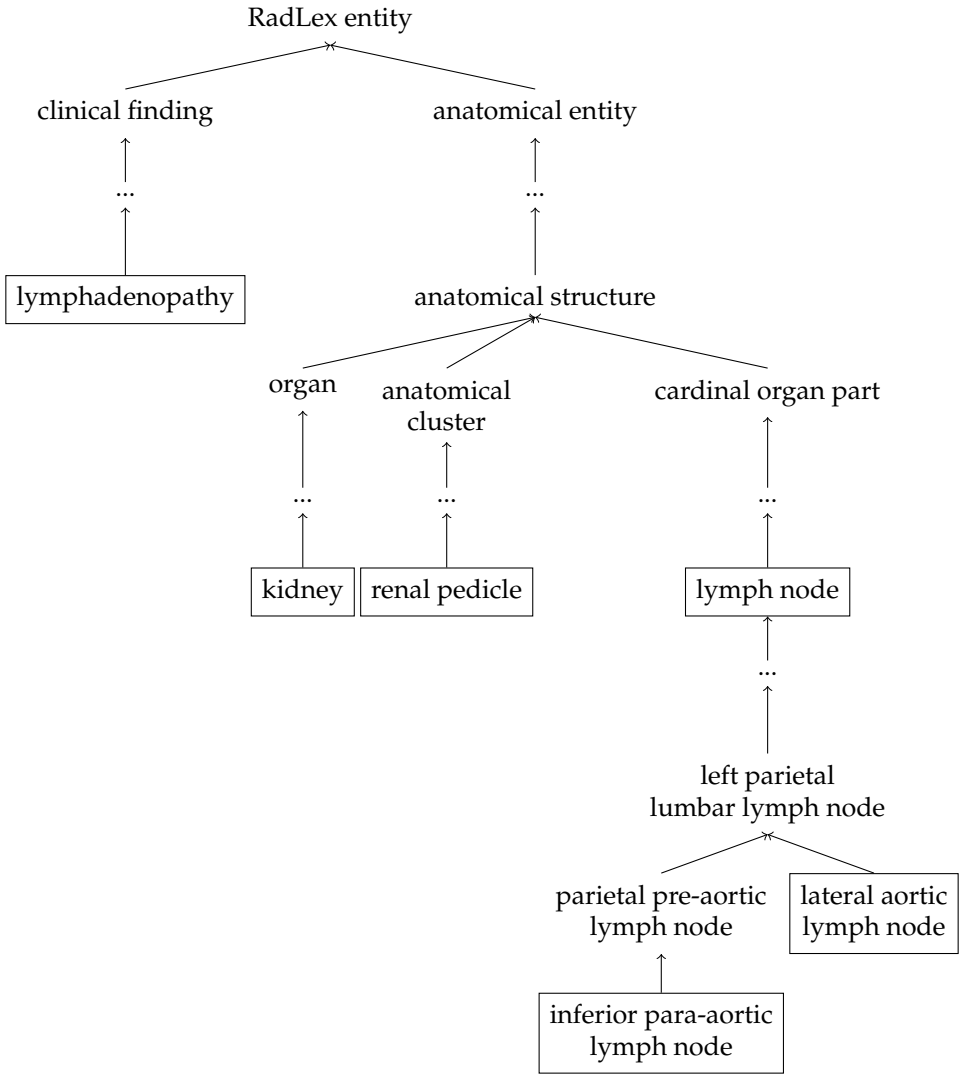


Figure 8.2.: Minimal spanning tree for RadLex annotations relevant for relation resolution for the sentence "Enlarged lymph node right paraaortal below the renal pedicle now 23 mm". Boxed entities represent annotations of the sentence, arrows a subclass relationship and dots indicate that a subclass path is omitted.

`radlex:kidney` and `radlex:lymph node` are found. Using the hierarchy of the ontology in form of the spanning tree has two advantages: Firstly, size information is propagated down to subclasses. Thus, as shown in the next step, each size assertion in the knowledge model implicitly applies to *many* concepts. Secondly, the spanning tree with subclass paths enhances the chance to find a matching concept in the knowledge model.

5. Compare Size Specifications For all entity e of the spanning tree with size specifications a comparison value is computed which indicates how well the

measurement value x fits into typical size range $[m, M]$ of e:

$$compValue(e, x) := \begin{cases} \frac{m-x}{x} & , x < m. \\ 0 & , x \in [m, M]. \\ \frac{x-M}{M} & , x > M. \end{cases} \tag{8.1}$$

That is, if the measurement value is within the range, a comparison value 0 is assigned, otherwise a value > 0 is computed. If there are multiple attached size specifications, the lowest (best) comparison value is saved. Other entities of the spanning tree then get the comparison value of the closest superclass assigned if available. For example, `lymph node` has comparison values 1.3 (“normal lymph nodes are $[0, 1]$ cm”) and 0 (“enlarged lymph nodes are $[1, 5]$ cm") so the value 0 is assigned. `kidney` gets comparison value 0.73 (“anterior-posterior diameter of kidney is normally 4 cm”). Then information is propagated: `lateral aortic lymph node` and `inferior para-aortic lymph node` get comparison value 0 assigned since `lymph node` is their closest superclass with size specification. The knowledge model does not cover `lymphadenopathy` and `renal pedicle`, so these entities get a default comparison value assigned.

6. Compute Ranking The final ranking value of the knowledge-based algorithm includes also the position of the entity e within the RadLex class hierarchy:

$$rankValue(e) := compValue(e, x) + \frac{1}{depth(e)} \tag{8.2}$$

Thus, in case of equal comparison values more special concepts (deeper in the hierarchy) are preferred. This again shows the advantage of using the ontology hierarchy.

8.3. Normality Classification

Normality classification of findings is very important for clinicians. During the diagnosis and examination process their attention is mainly focused on abnormalities. For blood test results this classification is standard: The measured values are compared to normal intervals and then flagged as low, normal or high. This kind of normality classification for *image findings* however is missing since their finding descriptions are more complex. Here, the normality classification of size findings, based on the model about normal size specifications (see section 6.1) is described. Given some size finding where anatomical entity as well as the measured quality and the measured value are defined, the most specific size specification is retrieved from the knowledge model (if available) and the corresponding normal range is compared to the actual measurement value. Most specific means that there is no other specification defined in the knowledge model that has a more specific quality description or which is about a more specific anatomical entity. Since the classification of measurement findings is based on value comparisons it is not realized by defining logical axioms. Instead, SPARQL is used to retrieve relevant data from the triple store. Then findings are classified by comparison of their measurement values to normal size specifications of the knowledge model. For a size finding that is about an *inguinal lymph node* with as measurement of the length with value specification `1.4 cm`, the knowledge model contains two size specifications: "lymph node, normal short axis, ≤ 1 cm" and "inguinal lymph node, normal short axis, ≤ 1.5 cm". Since the second description is more specific, the value `1.4` is compared to it and the finding is consequently classified as a *normal size finding*. More precisely, depending on the comparison the *quality* described by the size finding is classified as `increased size`, `decreased size` (which are both subclasses of `deviation from normal`) or as `normal`. Then the corresponding finding is classified as an `mci:increased size finding`, `decreased size finding` (both subclasses of `mci:abnormal size finding`) or as an `mci:normal finding`.

In the case of a multidimensional size finding the situation is slightly more complicated: Here it has to be ensured, that *all* specified values are in the normal range. For instance, a measurement `8 x 14 mm` about a mediastinal lymph node is classified as normal since not both values are `> 1 cm`. A measurement "spleen 10 x 4.8 cm" is classified as normal since the values fit into the range specifications of the knowledge model: "spleen, normal width, [4,6] cm", "spleen, normal length, [7,10] cm" and "spleen, normal height, [11,15] cm". A measurement "spleen 12 x 13 x 5 cm" would be classified as abnormal: Even though for each value one could find *some* fitting normal range, the three values have to be mapped to three *different* range specifications. Since no value is within the normal range of the length of the spleen, the finding is classified as abnormal. The evaluation of the normality classification is given in section 11.2.

8.3.1. Patient-specific Normality Classification

Additionally to the selection of the most specific size specification from the knowledge model, the patient context is also taken into account. As described in Section 6.1 some normal size specifications are age or gender dependent. For example, the normal size of the pancreas is age-dependent. During selection of appropriate size specification that are to be used for comparison it is ensured, that only specifications are retrieved that match the patient context. That is, for a finding "width of head of pancreas, 2.8 cm" and a patient with age between 51 and 60 a normal range 21-27 mm is retrieved while for another patient with age between 41 and 50 a normal range 22-29 mm is retrieved. Accordingly the classifications are different for a measurement value of 2.8 cm for patients with an age in the different ranges.

8.4. Linking Findings

Besides classification of findings as *normal* or *abnormal* (described in the previous section), information of the *change* of findings over time is important as well. Especially a comparison to the previously performed examination is required in many clinical processes such as diagnosis, monitoring or treatment evaluation. For instance, the clinician wants to know which findings are *progressive*, *regressive* or *new*. To allow corresponding classifications, findings describing the *same* anatomical entity at different points in time need to be linked. For findings about anatomical entities which are unique in the human body this can be achieved by matching their types – for findings about entities that occur multiple times, like nymph nodes or lesions, this is problematic.

8.4.1. Unique Anatomical Entities

Findings about an anatomical entity which exists only once in the human body (such as the liver, spleen etc.) can be linked: Here, it is required, that the anatomical entities match *and* that there does not exist a more specific type for the anatomical entity in RadLex. Then an `owl:sameAs` relation is created between the instances. For instance, there is no subclass of `radlex:spleen` and two findings about the spleen are linked (see figure 8.3). However two findings about `radlex:kidney` are not linked since there subclasses `radlex:right kidney` and `radlex:left kidney`. These findings are only linked if both are about the right kidney (or left kidney respectively). Thus, it is ensured, that the linking is at least as accurate as the used reference ontology. However, as described in the next subsection this is not sufficient for lymph nodes and lesions.

In a subsequent step, measurements of linked findings can be directly compared (as with normal value specifications) and the corresponding type `increasing finding` or `decreasing finding` assigned.

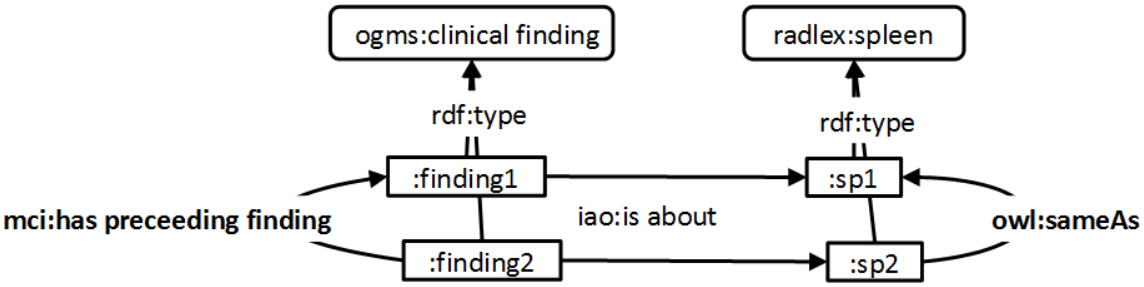


Figure 8.3.: Linking findings from consecutive examinations by anatomical entity.

8.4.2. Lymph Nodes and Lesions

Linking findings about lymph nodes or lesions from consecutive examinations is problematic: There are more than 250 different lymph nodes and lesions can occur in almost all locations in the human body. Often corresponding finding descriptions found in radiology reports are not precise enough to allow automatically linking these findings (see details below). Thus, these findings are *not* linked directly. Instead, a user interface (ReportViewer, see section 9.2) was developed, that allows to review these findings in a longitudinal fashion. That is, the structured representation of clinical findings with MCI is used to provide an integrated view for the clinician.

Lymph Nodes

Since the human body contains *many* lymph nodes, it is not enough, that two findings are about *some* lymph node. Even though RadLex defines a very detailed class hierarchy for lymph nodes, linking findings about lymph nodes, following the above approach of "most specific types", does not guarantee accurate enough results. For instance, `radlex:para-aortic thoracic lymph node` has no subclass – and thus is a *most specific type*. However, depending on the patient there might be *multiple* para-aortic thoracic lymph nodes. Consequently a linkage would not be accurate enough to justify a value comparison.

Lesions

As mentioned above, lesions can occur in any location of the body. Since RadLex does not contain a very detailed subclass hierarchy of lesions, the *location* of lesions has to be respected when corresponding findings should be linked. One could try to establish links between lesions, found at the same location, requiring that there is no more specific location description in RadLex. Then a location specification "liver" would not be sufficient: At least a specific liver segment would be required. However, even requiring a "most specific location" would not guarantee, that two findings indeed describe the *same* lesion: For example, there are numerous cases where multiple lesions are found in the same liver segment.

8.5. Propagation of Finding Information

In the context of some particular patient, each finding and symptom of Disease-Symptom-Examination ontology (DSE) can be classified into one of the following three categories:

Present finding (abnormal finding) A finding which is observed at a patient. Present findings are important, since they indicate the *presence* of some diseases and require monitoring.

Absent finding (normal finding) A finding which was under examination but did not show up (e.g. "no enlarged lymph nodes in neck area"). Absent findings are of importance, since they allow to *exclude* certain diseases.

Unknown finding A finding for which no information is available for the patient in the current situation. These findings have not been examined yet and need to be targeted next.

The status of a finding is meaningful only within the context of some particular patient at some point in time. Thus, all descriptions below always refer to findings and symptoms of one patient. After initial mapping of finding information specified in MCI to the findings and symptoms defined in DSE, a patient-specific *finding tree* is created and the consistency of available finding information is checked.

8.5.1. Mapping

For some given finding of DSE, such as `dse:splenomegaly` the status is determined by querying the patient data represented in MCI: All clinical findings are retrieved together with all their types, the corresponding qualities (if specified) and the date of the corresponding examination. Then each finding is mapped to the most specific class of DSE, and a status and date is attached. For example:

```
mci:radFinding123
   rdf:type          'dse:enlarged colic lymph node' ;
   dse:has_status    'dse:known present' ;
   dc:date            "2006-04-07" .
```

By referencing RadLex the coverage of DSE is extended: Even though DSE does *not* contain a finding, `dse:enlarged facial lymph node`, a corresponding finding defined in MCI will be captured at least as the finding `dse:enlarged lymph node`,

since lymph node is a superclass of facial lymph node in RadLex. However this is done only for findings that have status *known present*. That is the *absence* of enlarged facial lymph nodes does not map to "enlarged lymph nodes absent". The rational behind this is described in more detail below.

8.5.2. Łukasiewicz Three Value Logic

Łukasiewicz logic is used for three truth values (*true*, *unknown*, *false*) – here described as *present*, *unknown*, *absent*. The truth-table of Łukasiewicz's logic is shown in table 8.1, where the values denote the truth value for the validity of the implication $A \rightarrow B$.

Table 8.1.: Truth table of three-value logic (T - true, U - unknown, F - false) for the implication relation. Adapted from [Mal07] to the subsumption relation between findings (finding A is a subclass of finding B) regarding their status (present, unknown, false).

	$A \rightarrow B$	B F (absent)	U (unknown)	T (present)
A	F (absent)	T	T	T
	U (unknown)	U	T	T
	T (present)	F	U	T

Regarding the class hierarchy of clinical findings defined in DSE, the subclass relation between two clinical findings can be translated to a logical assertion. That is, an axiom denoting a subclass relation such as

```
dse:finding1 rdfs:subClassOf dse:finding2
```

is interpreted as the logical assertion

```
dse:finding1 → dse:finding2
```

since each instance of dse:finding1 is also an instance of dse:finding2. For example an instance of `dse:enlarged abdominal lymph nodes` is also an instance of `dse:enlarged lymph nodes` and thus

```
enlarged abdominal lymph node → enlarged lymph node
```

Since the subclass axiom is considered to be always true, only finding descriptions (i.e. status descriptions) are consistent that are marked as "T" (true) in table 8.1. For

example, it is not consistent when *A enlarged abdominal lymph nodes* is true (present) and *B enlarged lymph nodes* is false (absent) since this violates the implication $A \rightarrow B$. More precisely: if finding1 is true (present) then finding2 must be true (present) as well; if finding1 is unknown, then finding2 is either unknown or false (absent) – but it cannot be true (present), since this would contradict the logical assertion above; if finding1 is false (absent), no assertion about finding2 can be inferred. Similarly, if finding2 is true (present), no assertion about finding1 can be inferred. However if finding2 is unknown than finding1 must be unknown as well or true. If finding2 is false (absent) then finding1 must be false (absent) as well.

Basically, information about present findings propagates to *super*classes while absent findings propagates to *sub*classes.

Definition: finding status tree For one patient at one point in time a *finding status tree* is defined as the class hierarchy of findings of DSE, where for each class at most one status *known present* or *known absent* is specified.

Definition: consistent/inconsistent finding status tree A finding status tree is called *inconsistent* if and only if there exist a pair of classes A and B, connected by a subclass path from A to B, for which the truth-table is evaluated to *F (false)* – otherwise the status tree it is called *consistent*.

That is, the status tree is inconsistent if and only if A has status *known present* and B has status *known absent*, where A is a subclass of B. For instance, it is not consistent to state "known present, enlarged abdominal lymph nodes" and at the same time "known absent, enlarged lymph nodes".

8.5.3. Getting to a Consistent State of the Patient's Findings Status

As mentioned above, a finding status tree might be inconsistent. Two rules are used to resolve inconsistencies arising from an initial set of findings with status and date: Firstly, newer findings are more important than older findings. Secondly, findings with status *known present* are more important than those with status *known absent*.

To obtain a consistent status tree based on an initial set of findings with status and date the procedure is as follows:

1. Sort clinical findings: clinical findings are sorted descending by date, i.e. newest at the top of the list, and secondly by status (first: known present;

then: known absent).

2. Iterate over the sorted list and add a class C to the status tree (ST) if one of the following conditions are met:

   a) C has status present and there is no superclass B of C in ST with status known absent .

   b) C has status absent and there is no subclass A of C in ST with status known present .

After those steps one has the complete finding status for the patient. For example, as shown in figure 8.4.

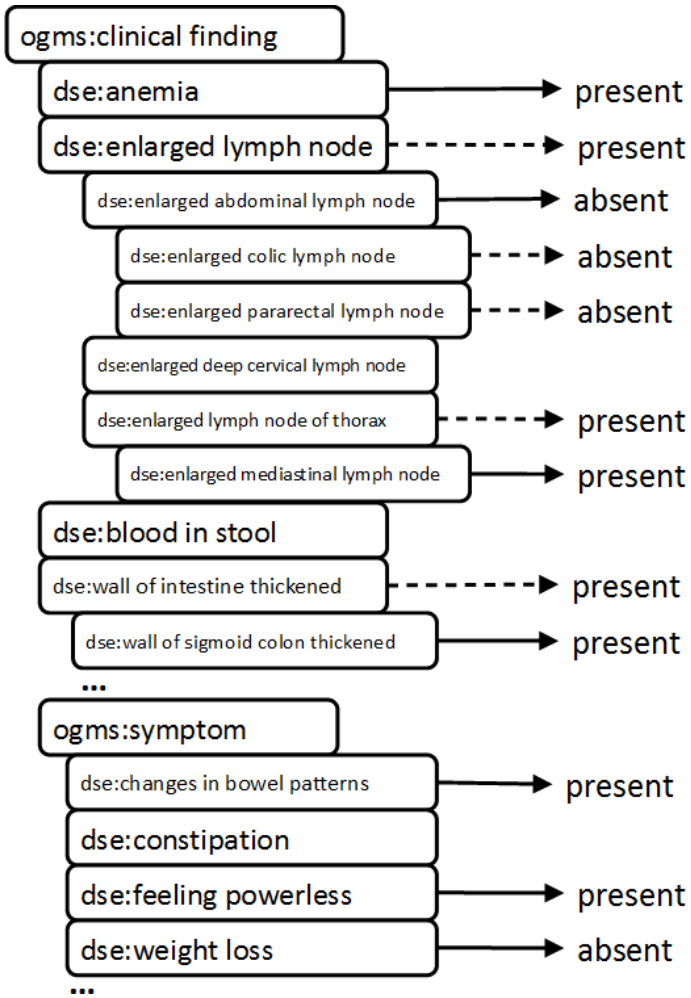


Figure 8.4.: From patient data specified in MCI to a comprehensive overview of the patient's current findings. Dashed arrows express inferred information about the corresponding finding status.

8.5.4. Determining the Status of Findings

With a consistent finding status tree one can query the status for any finding of the tree. If the status is explicitly represented, it is directly retrieved. If not, then the status of a finding class *C* can be inferred in some cases (examples illustrated in figure 8.4):

- If there is an entry about a subclass of *C* with status *known present* then this status is inferred also for *C*. For example since `wall of sigmoid colon thickened` is present, the finding `wall of intestine thickened` is present as well.

- If there is an entry about a superclass with *absent* annotation then this finding is absent as well. For example, since `enlarged abdominal lymph node` is absent, the findings `enlarged colic lymph nodes` and `enlarged pararectal lymph nodes` are regarded as absent as well.

Regarding the status tree shown in figure 8.4, one cannot infer any information about the status of `enlarged deep cervical lymph node` or `blood in stool`. The inferred information is not materialized in the status tree in order to keep a distinction between original and inferred information. This is important in the situation where new information is entered to the status tree as explained in the next subsection.

8.5.5. Changing the Finding Status

The status of a finding might change over time: Firstly, as a result of a recently performed examination (status *known present* or *known absent*). Secondly, because some status is not considered valid by a clinician, thus has to be set to *unknown*. For example, the presence of a headache from several weeks ago might be removed or discarded in diagnosis.

If the status of some finding changes, then the finding status tree is adapted to avoid inconsistency. As before, the newer entry is considered more important than older entries: For example (again referring to figure 8.4), if the status of *enlarged colic lymph nodes* is set to *known present*, then the status of *enlarged abdominal lymph nodes* is set to *known absent* is in conflict with the new entry and thus need to be discarded. Since the inference described above (dashed arrows in figure 8.4) is not materialized, the status of *enlarged pararectal lymph nodes* is now *unknown*.

If the status of a finding class *C* is changed to *unknown*, then the following adaptations need to be made: Firstly, the status *known present* of subclasses of *C* needs

to be removed. For example, if a clinician considers the finding status of *wall of intestine thickened* to be unknown, then status *know present* of the subclass *wall of sigmoid colon thickened* needs to be removed. Secondly, the status *known absent* of super-classes of C needs to be removed. For example if a clinician considers the finding status of *enlarged colic lymph nodes* to be unknown, then status *know absent* of the superclass *enlarged abdominal lymph nodes* needs to be removed.

8.6. Ranking Likely Diseases

Given a patient with an initial set of symptoms (explicitly represented within the patient's findings status tree described in the previous section, a ranked list of likely diseases is inferred. A relatively simple disease-centric view on patient data as shown in figure 8.5 is suitable to assist the clinician during diagnosis.

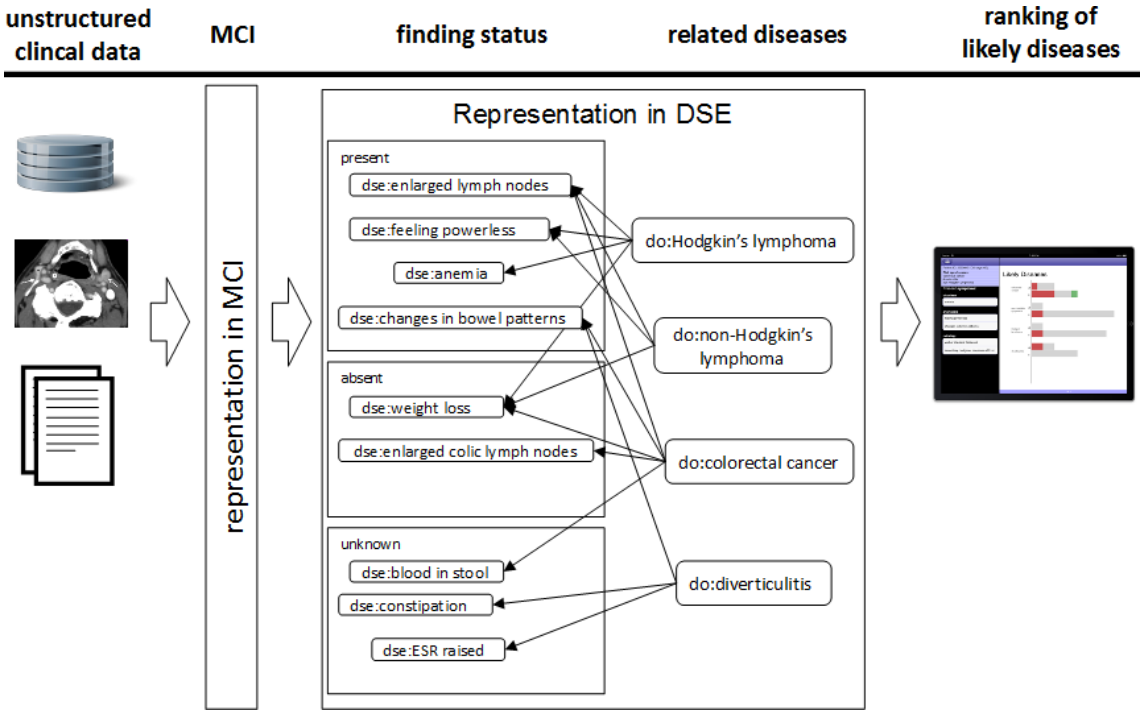


Figure 8.5.: From annotations to a ranking of likely diseases. Matching the typical symptomatology of a disease with the patient's symptom information.

The basic idea is to measure how well the patient's symptom information matches the typical symptomatology of some given disease. Weighting the patient's symptoms and then figuring out the *best match* under the set of diseases is very similar to the decision-making process of clinicians. So it was decided to base the ranking of likely diseases on an adapted measure of precision and recall. The purpose of such a system is to prevent that clinicians miss a disease they are rarely confronted with and for which they do not know all correlated symptoms. For that purpose, the clinician is provided with a graphical overview about likely diseases (see next Chapter).

Factors with Influence on the Ranking

The factors influencing the ranking were identified in interviews with clinicians. In general, all patient data and all existing medical knowledge is relevant for stating

the correct diagnosis. Here, it was concentrated on the following three aspects:

- **Medical knowledge** about diseases, their manifestations in findings and symptoms distinguishing between leading symptoms other findings and also risk age of the disease. This knowledge is available in medical literature (e.g. [Her11]) and modelled in DSE.

- **Patient-specific data** like status of findings and symptoms (represented in a finding status tree described above) as well as the patient's age.

- **Clinical knowledge** on the relative importance of clinical findings. For instance blood in stool is more important to track than an unspecific headache.

The ranking of likely disease is described in detail in [Obe+12b]. Based on the *finding status tree* and the relations to diseases (see figure 8.5), for each disease a ranking value is calculated. It is measured how well the overall patient's finding status (present and absent) matches the symptomatology of a given disease. In an adapted measure of precision and recall it is calculated which amount of findings with status *known present* can be explained by the disease and further how many findings with status *known absent* indicate the absence of the corresponding disease. Additionally an uncertainty factor is competed for each disease, which measures the relative amount of findings related to the disease for which the status is unknown. For instance (referring to figure 8.5), the disease *colorectal cancer* has two matching findings with status *known present*, one with status *known absent* and one for which the status is unknown. This information is weighted during the calculation of the ranking factor: In the described example, the absent symptom *weight loss* is a leading symptom for colorectal cancer and thus weighted stronger than other present symptoms such as *changes in bowel patterns*. Thus, it is less likely that the patient has colorectal cancer but more likely to have lymphoma.

9

Integrated Data Views

The previous chapter explained how knowledge models and reference ontologies are used to enrich the clinical data represented in MCI. In this chapter they are used to provide an integrated view on data, i.e. they are used during retrieval and presentation of the data. The structured and semantic representation of clinical findings and their examination context allows to formulate integrated queries that could not be posed to the original data. Two types of views are presented: Firstly, a disease-centric view is presented which can be obtained by using the knowledge model about disease-symptom relations presented in section 6.2. This provides a *snapshot*-view on the finding data for one patient at a specific point in time. The role of this view is to assist the clinician in the differential diagnosis and examination process by highlighting the most likely diseases together with the status of their typical symptoms. Further the resulting information is used to point to missing finding information, i.e. it provides directions for the follow-up examination planning. Secondly, a *longitudinal* view on patient data is presented. Here, the different dimensions of clinical findings specified in MCI are used to realize different perspectives and filters on the data. This eases the comparison between findings from consecutive examinations. The scenarios are explained along demo applications.

9.1. Disease-centric Snapshot View

The process of differential diagnosis was explained in section 6.2. Basically, a clinician starts with a small set of initial symptoms and – based on leading symptom relations – gets a first list of possible diagnoses. In a subsequent step of differential diagnosis, examinations are planned to obtain information about symptom that have not been covered so far. With the additional symptom information the list of possible diagnoses is refined. In this section it is shown, how data represented in MCI and the knowledge model about disease-symptom relations are used to enhance the differential diagnosis process.

9.1.1. Findings-Status for one Disease

A search for findings related to a specific disease is currently not possible since information about disease and symptom interrelations is not covered by existing medical ontologies. Thus, a search for cancer-indicating findings is only possible through a search for specific finding as e.g. enlarged mediastinal lymph nodes or enlarged spleen assuming that the clinician is informed about likely symptoms of the disease. However, clinicians are usually experts in one particular domain, leading to a lack of prior knowledge about the interrelations of symptoms and diseases – in case certain diseases are no longer in the scope of their expertise. In other words, there is a clear danger that the information about the relevance of identified symptoms remains overlooked or misinterpreted, leading to wrong or not appropriate treatments. The disease-symptom knowledge model described in section 6.2 is used to query patient data for finding information of a specific disease as shown in figure 8.5.

9.1.2. Ranking of Likely Diseases

Allowing the clinician to get all finding information for one specific disease is only the first step. A natural extension is then, to create a *ranking* of likely diseases. The prototype presented here, employs disease-symptoms relations defined in DSE, to make likely diseases of the patient explicit. The corresponding ranking algorithm is described in section 8.6 and more detailed in [Obe+12b]. Similar to the cognitive decision process of clinicians, who rely on their experience and expertise, medical background knowledge is used to automatically interpret the findings and symptoms for integration in aforementioned clinical processes. Basically, the patient's symptoms are matched with the typical symptoms of diseases, where present symptoms are counted as pro-evidence and absent symptoms as contra-evidence for the related diseases. This is shown in figure 9.1, where the status

of symptoms is indicated by color (red: symptom is present, green: symptom is absent, gray: symptom has not been checked). The better the patient's symptoms match with the typical symptoms of a disease, the higher the corresponding disease is ranked. Further, it is distinguished between leading symptoms (LS) and other symptoms (S) for each disease. In figure 9.1, one can see that colorectal cancer has more associated present symptoms than the other diseases. However, many symptoms have not been checked so far, which is indicated by light gray bars in the ranking chart. A ranking of likely diseases improves the diagnosis process significantly as the clinician will not miss any findings and symptoms of diseases that might be out of his (main) expertise. Knowing likely diseases makes it easier to plan next examinations. Further, the clinician gets an overview of what symptoms of certain disease are present without going through all images and reports. Especially this helps the clinician not to miss any symptoms in the diagnostic process. Thus, the relevance-based highlighting of information about clinical observations in the context of likely diseases supports clinicians to improve their treatment decisions.

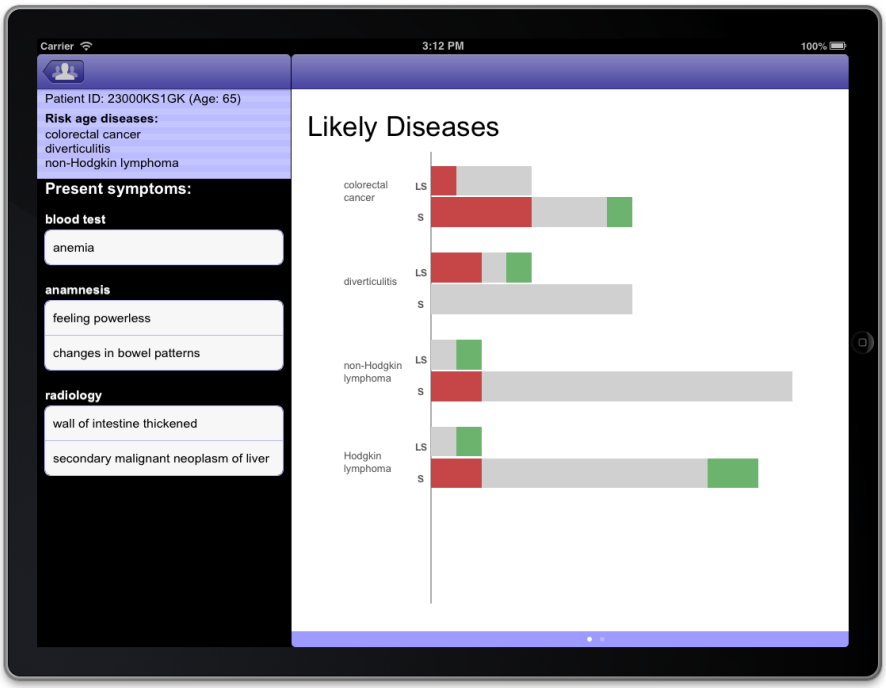


Figure 9.1.: Ranking of likely diseases based on information about present and absent clinical findings and symptoms.

9.1.3. Examination Planning

After the automatic detection of likely diseases based on an initial set of findings, this information can be used in differential diagnosis, where the clinician intends

to either exclude or strengthen particular diseases. So the first list of symptoms and likely diseases in turn indicates to check for further symptoms. For example, *enlarged lymph nodes* indicates to check for B-symptomatic i.e. weight loss, fever and night sweat). Similarly, certain blood tests could be performed. In order to help the clinician in planning the next examinations, unchecked symptoms and clinical findings are ranked by relevance, which is defined through the relations to the diseases of the ranked list of likely diseases. A finding has a higher relevance to be checked when it is related to (many) top ranked likely diseases, than when it is related only to (few) less likely diseases. Then the ranking values for particular findings/symptoms are aggregated on examination level and a ranking of examinations is proposed (see figure 9.2).

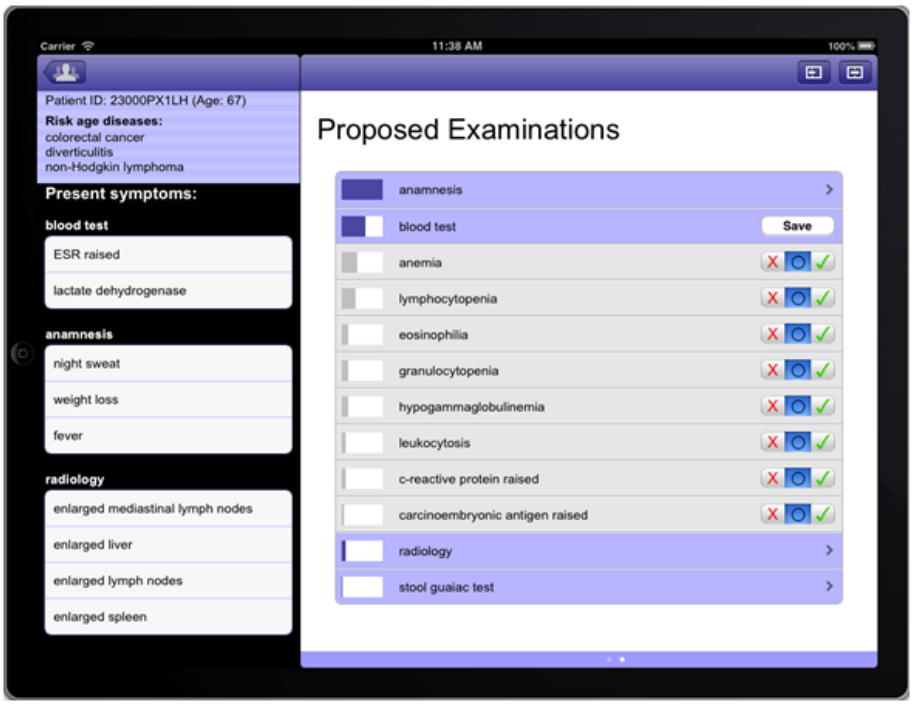


Figure 9.2.: Ranking of proposed examinations that are suitable to efficiently check further symptoms. In this case anamnesis and blood test are proposed next examinations.

9.2. Longitudinal Finding-centric View

A longitudinal view on the data of one patient is important since it allows the clinician to get an impression of the *change* of the patient's health status. For instance, clinicians need to know whether the health status is improving or getting worse. Besides this high-level information, even more important is the tracking of *individual* clinical findings. For instance, clinicians compare the size of lesions over time to evaluate treatment success. Similarly, findings about specific anatomical entities such as enlarged lymph nodes or abnormalities of organs are commonly compared between consecutive examinations. Having structured semantic representations of patient data, these types of comparisons can be realized through simple queries. As shown in figure 9.3 a clinical finding has four different dimensions: type, anatomy, quality and location. These dimensions of finding descriptions are used to retrieve longitudinal patient data in different views as described in the following subsections. Additionally to direct links between findings from consecutive examinations, reference ontologies are used to get a more comprehensive overview of corresponding information over time. Findings can always be sorted by their date and to provide a longitudinal view on the patient.

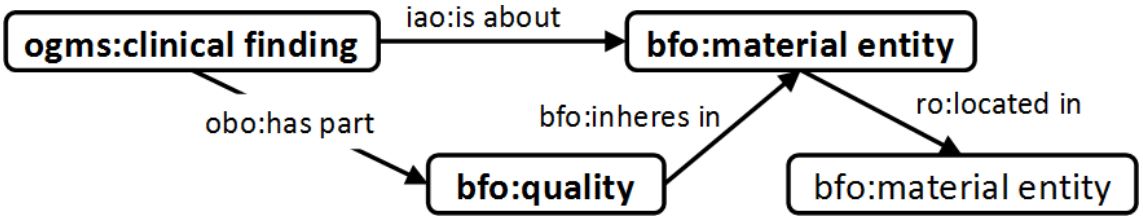


Figure 9.3.: The main dimensions of a clinical finding: the type of the finding, the anatomical entity, its quality and location.

9.2.1. Type of Clinical Finding

As presented in chapter 5, MCI provides a detailed class hierarchy for clinical findings (see e.g. figure 6.9). This classification is now used in the finding-centric view on the patient data. For instance, to retrieve all *abnormal findings* or all *new appeared* or *changed findings* for one specific examination. figure 9.4 shows a view on *measurement* findings from consecutive examinations. The type of a clinical finding is not restricted to the class hierarchy in MCI. For instance, the RadLex or HP contain detailed hierarchies of clinical findings (or phenotypes). These hierarchies are used again as background knowledge for retrieving certain type of findings based on text annotations. For instance, one can retrieve all lesions such as a lung mass or liver lesions using the RadLex class hierarchy. Similarly, one can retrieve all abnormalities of the immune system using the HP class hierarchy which returns findings such as splenomegaly or enlarged lymph nodes.

2006-05-31 computed tomography **Hals-, Thorax- und Abdomen-CT mit KM i.v**	
1 cm lymph node	Weiterhin mehrere, deutlich unter 1 cm messende Lymphknoten entlang der Halsgefäßnervenscheiden bds..
1 cm (1,3 cm) axillary lymph node	Rechts axillär größenregredienter Lymphknoten jetzt 1 cm (IMA 9), vormals 1,3 cm.
1 cm axillary lymph node	Bei Z.n. B-NHL residualer, kleiner 1 cm großer Lymphknoten rechts axillär.

2006-02-28 computed tomography **Hals-, Thorax- und Abdomen-CT mit KM i.v**	
8 mm submental lymph node	Die vorbeschriebenen zervikalen Lymphknoten entlang der Halsgefäßnervenscheide bds. sowie submandibulär, submental und in den Kieferwinkeln sind in Größe und Anzahl deutlich regredient (aktuell noch Durchmesser bis 8 mm IMA 39 li unterhalb des M. sternocleidomastoideus).
1,4 cm (3,5 cm) axillary lymph node	Deutlicher Größenrückgang auch der axillären Lymphknoten, zuvor bis 3,5 cm durchmessende Lymphknotenpakete rechts axillär durchmessen aktuell noch maximal 1,4 cm.
5 mm	subpleural im Mittellappen rechts (IMA 35, Durchmesser 5 mm).
5 mm round mass 5 mm round mass 5 mm round mass 6 mm round mass	Im apikalen Unterlappen links zahlreiche Rundherde (IMA 20 bis 31, Durchmesser bis 6 mm), im rechten Unterlappen (IMA 34, Durchmesser 5 mm), im basalen Mittellappen (IMA 47, Durchmesser 5 mm) und in den Oberlappen bds. (jeweils IMA 12, Durchmesser 4 bzw. 5 mm).
1,2 cm lesion	Größenunveränderte, flau hypodense, subkapsulär im Segment 5 gelegene Leberläsion (IMA 66/67, Durchmesser 1,2 cm) mit randständig fraglich diskreter stäbchenförmiger KM-Mehranreicherung.
7 mm lesion	Eine weitere flau hypodense, 7 mm durchmessende Läsion im Segment 5 (IMA 65), die in der VA nicht sicher nachvollziehbar ist.
8,2 x 9,1 x 13 cm spleen	Milz homogen, ausgeprägter Größenrückgang bei vorbestehender Splenomegalie (aktueller Durchmesser 8,2 x 9,1 x 13 cm).
10,5 cm	Cranio-caudaler Durchmesser bds. 10,5 cm, im Vergleich zur VU hierbei rückläufige Schwellung des Nierenparenchyms bei verbesserter Abgrenzbarkeit zum umgebenden Fettgewebe.
8 mm	Bei liegenden Harnleiterschienen bds. mehrere kalkdichte Konkremente innerhalb des Nierenbeckenkelchsystems mit Durchmessern, bds. bis zu 7 bis 8 mm.

Figure 9.4.: Measurement findings from consecutive examinations with classification as normal and abnormal.

9.2.2. Anatomical Entity

Another important dimension of a finding is the described anatomical entity (or material entity). The anatomy-centric view retrieves all findings – regardless of their type – that are about some anatomical entity in questions. Thus, the clinician is able to track the status of specific entities such as organs over time (see example for spleen in figure 9.5). Here the reference ontologies are used to retrieve not only findings about exactly the same entity but also closely related entities. For instance a finding about the left kidney is related to findings about the right kidney. Similarly, findings describing lymph nodes are related. Here, the distance within the subclass hierarchy of the reference ontology is taken into account. Text and image annotations are additionally used to retrieve further observations.

9.2.3. Quality

It is also possible to retrieve only findings describing specific qualities such as size, color, thickness or their modifications such as `increased size`. Some of these qualities are also reflected in the class hierarchy of clinical findings, however the hierarchy of qualities contains even more classes.

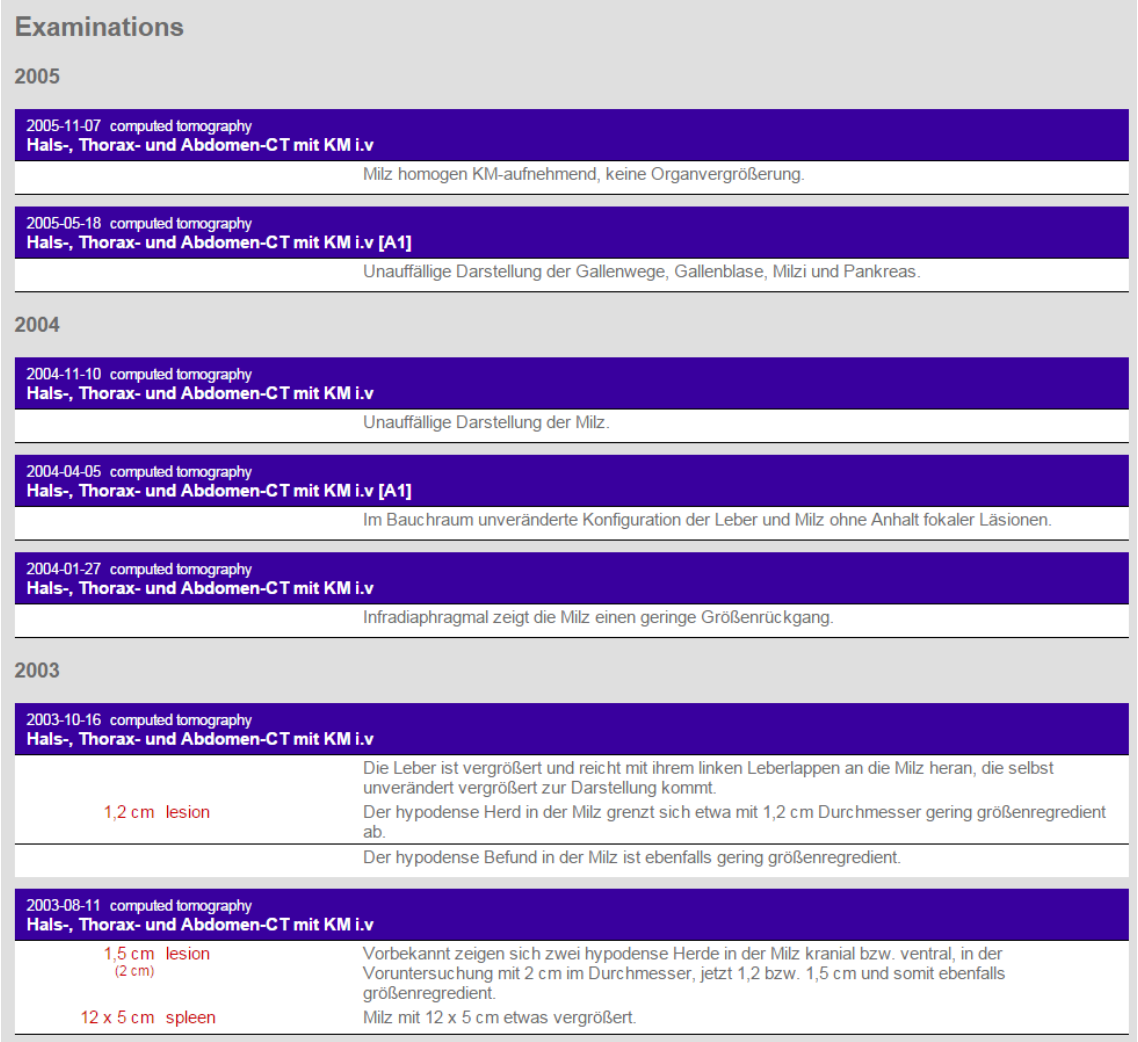


Figure 9.5.: Findings from consecutive examinations about the spleen: while in 2003 lesions were found in the spleen, since 2004 the spleen is unspecific.

9.2.4. Location

The location is another important dimension of a clinical finding. Not only in the context of lesions, one might want to retrieve all (abnormal) findings describing entities that are contained in a specific region such as the *thorax* (see figure 9.6). Here, the partonomy of the reference ontology is employed. That is, all findings that are about an anatomical entity that is part of a certain region are retrieved.

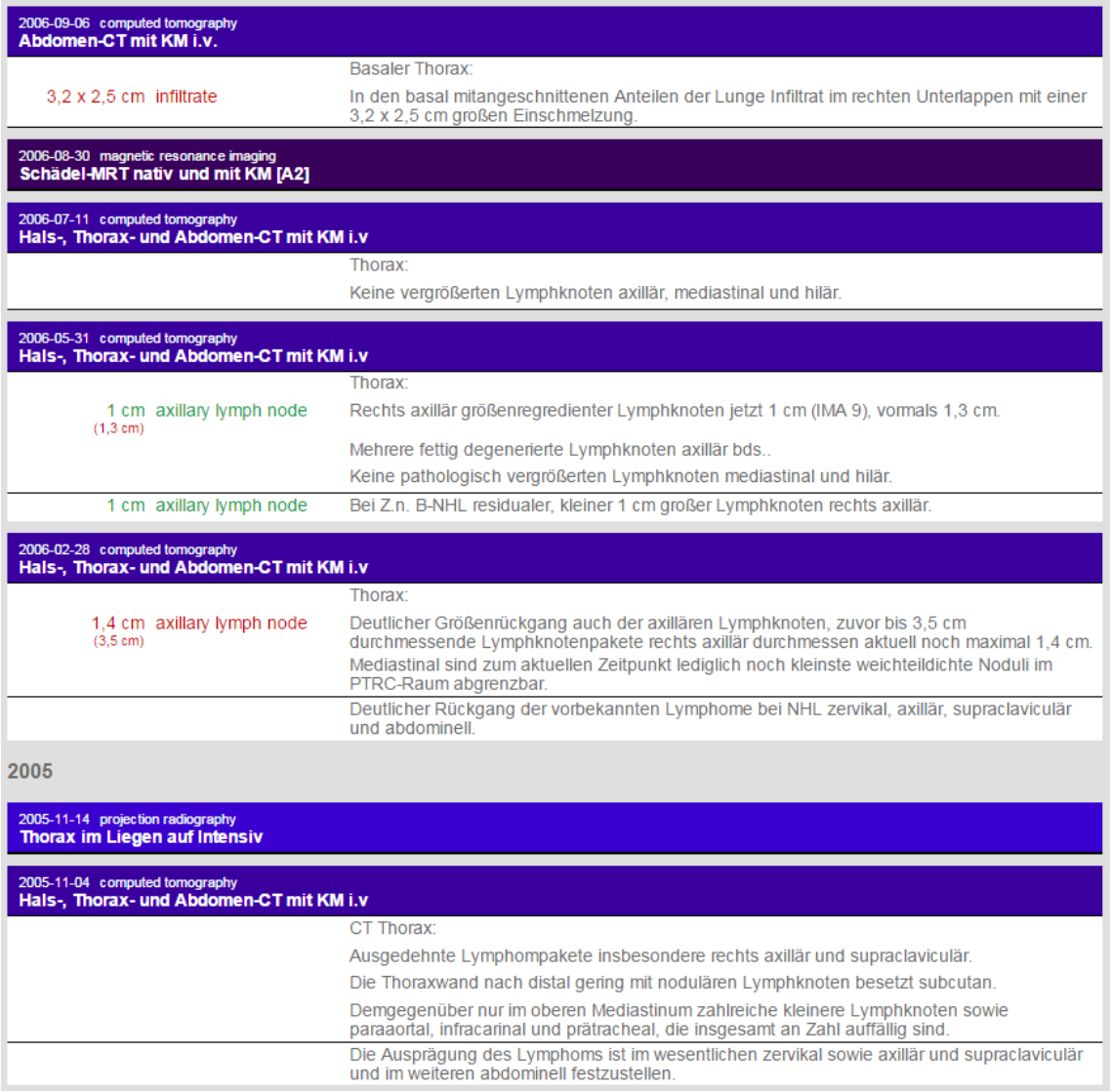


Figure 9.6.: Clinical findings from consecutive examinations within a specific anatomical region of interest – in this case the *thorax*.

Part IV.

CASE STUDIES

10

Data Sets

The three data sets described in this chapter were used during the development of the semantic Model for Clinical Information (MCI) and for the evaluation of the inference algorithms of chapter 8. Detailed descriptions and example data for all three data sets are provided within the three sections of this chapter. Firstly, the *base line data set* that contains data of 6 melanoma patients with complete information for a period of three years. This data set contains structured information about the demographic background of the patients, diagnoses, procedures, drug administration and laboratory values as well as unstructured German reports from radiology and pathology. Secondly, the *data set of free text radiology reports* that consists of three corpora: (1) A corpus of about 2700 German radiology reports for 377 *lymphoma* patients with different imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (US). Since this data set contains many reports from consecutive radiology examinations for each patient, it is well suited for the longitudinal analysis of corresponding clinical findings. (2) A corpus of 6007 German radiology reports from CT examinations of *internistic* patients with diverse disease background. This data set has no patient identifiers and thus is mainly used for evaluation of the algorithms that extract and classify measurement findings. (3) A corpus of 3000 English radiology reports from examinations with diverse imaging modality and patient background (disease, age, gender etc.). The advantage of this data set is that English texts are better covered by labels of ontology entities from RadLex, leading to a better overall annotation coverage. The data set is used for evaluation of the measurement entity resolution algorithm. Thirdly, the *data set of 40 patients with structured finding information* for which the age, an associated main disease (that was assigned by a clinical expert) and an initial set of 3 - 15 known present findings or symptoms and up to 5 known absent findings or symptoms are specified. This data set is used for the evaluation of the ranking of likely diseases and the corresponding disease-centric data access.

The data sets described below were all handled according to the German privacy regulations “Datenschutzgesetz”. The data from the university hospital in Erlangen was provided in de-identified form. The de-identification of free text data was done manually by the clinical partners.

10.1. Baseline Data Set on Melanoma Patients

In the context of the funded projects with the university hospital Erlangen "Data Intelligence for Clinical Solutions" and "Klinische Datenintelligenz" [15k] a *baseline data set* of melanoma patients was provided. The data set contains the complete and timestamped data of 6 patients with main diagnosis melanoma for a period of three years from 2009 up to 2012. Melanoma patients were selected, since they create data in many different clinical domains (diagnosis, laboratory analysis, pathology, radiology etc.). Thus, the data set provides a good basis for the development of the semantic data model. The data was provided in a relational data base called i2b2 [15h], which has a star schema as typical for data warehouses. The central table is the OBSERVATION_FACT table with the schema shown in table 10.1.

Table 10.1.: The columns of the i2b2 OBSERVATION_FACT table with corresponding data types.

column	data type
ENCOUNTER_NUM	NUMBER(38,0)
PATIENT_NUM	NUMBER(38,0)
CONCEPT_CD	VARCHAR2(50 BYTE)
PROVIDER_ID	VARCHAR2(50 BYTE)
START_DATE	DATE
MODIFIER_CD	VARCHAR2(100 BYTE)
INSTANCE_NUM	NUMBER(18,0)
VALTYPE_CD	VARCHAR2(50 BYTE)
TVAL_CHAR	VARCHAR2(255 BYTE)
NVAL_NUM	NUMBER(18,5)
VALUEFLAG_CD	VARCHAR2(50 BYTE)
QUANTITY_NUM	NUMBER(18,5)
UNITS_CD	VARCHAR2(50 BYTE)
END_DATE	DATE
LOCATION_CD	VARCHAR2(50 BYTE)
OBSERVATION_BLOB	CLOB
CONFIDENCE_NUM	NUMBER(18,5)
UPDATE_DATE	DATE
DOWNLOAD_DATE	DATE
IMPORT_DATE	DATE
SOURCESYSTEM_CD	VARCHAR2(50 BYTE)
UPLOAD_ID	NUMBER(38,0)

Further i2b2 contains several *dimension* tables:

- PATIENT_DIMENSION
- PROVIDER_DIMENSION

- CONCEPT_DIMENSION
- MODIFIER_DIMENSION
- VISIT_DIMENSION

Every entry of the OBSERVATION_FACT table is timestamped with a start date, and optionally, an end date. Each entry contains IDs for the *clinical encounter* (a reference to VISIT_DIMENSION), *patient* (a reference to PATIENT_DIMENSION) and *provider* (a reference to PROVIDER_DIMENSION). The columns CONCEPT_CD and MODIFIER_CD specify the type of information represented in the row. For example, an ICD code (concept) representing a main diagnosis (modifier), or the findings section (modifier) of a radiology report (concept). The other columns are used for the actual data values. The column VALUE_FLAG is used to specify whether lab values are within a *normal range*, *low* or *high*. i2b2 uses ontologies to code data in the OBSERVATION_FACT table: In particular, the base-line data set contains codes for diagnosis (ICD), procedures (OPS), medication (ATC), laboratory tests (LOINC) and a hierarchy for the provider (ORG-ID). In the following subsections the provided structured and unstructured data is described in detail.

10.1.1. Structured Data

The data set contains structured data about the demographic patient background, diagnosis, provided examinations, procedures and drug treatment as well as results of laboratory tests.

Demographics

The basic patient information is provided in the i2b2 dimension table called PATIENT_DIMENSION. It contains information about the patient's vital status, the date of birth and death, the gender and a ZIP code. Columns for specification of the language, race, martial status, religion and income are not filled (see table 10.2).

Diagnosis

Diagnosis information is contained in the OBSERVATION_FACT table. For each diagnosis entry the patient ID, the ICD code, the provider, the date as well as the diagnosis type (main or secondary diagnosis) are specified. An example is given in table 10.3. In total the data set contains 306 diagnosis entries - 60 of type main diagnosis and 246 of type secondary diagnosis. The ICD code is a

Table 10.2.: The main columns of the i2b2 PATIENT_DIMENSION table with example data.

column	data type	example entry
PATIENT_NUM	NUMBER(38,0)	1000000001
VITAL_STATUS_CD	VARCHAR2(50 BYTE)	Y
BIRTH_DATE	DATE	01.01.46
DEATH_DATE	DATE	(null)
SEX_CD	VARCHAR2(50 BYTE)	M
AGE_IN_YEARS_NUM	NUMBER(38,0)	66
LANGUAGE_CD	VARCHAR2(50 BYTE)	(null)
RACE_CD	VARCHAR2(50 BYTE)	(null)
MARITAL_STATUS_CD	VARCHAR2(50 BYTE)	(null)
RELIGION_CD	VARCHAR2(50 BYTE)	(null)
ZIP_CD	VARCHAR2(10 BYTE)	97618
STATECITYZIP_PATH	VARCHAR2(700 BYTE)	(null)
INCOME_CD	VARCHAR2(50 BYTE)	(null)
PATIENT_BLOB	CLOB	(null)
UPDATE_DATE	DATE	15.10.12
SOURCESYSTEM_CD	VARCHAR2(50 BYTE)	ISH

reference to the CONCEPT_DIMENSION table where the concept path (position in the ICD hierarchy) and a human readable label is specified. For instance, the code ICD10:K66.0 stands for *"Peritoneale Adhäsionen"*. In total 77 distinct ICD10 codes are used within the data set. The provider ID is a reference to the PROVIDER_DIMENSION table where, for example, ORG:CH_CHOP is listed as *"CH OP Allgem. Chirurgie"*.

Table 10.3.: Columns of the OBSERVATION_FACT table relevant for diagnosis with example data.

column	example entry
ENCOUNTER_NUM	158065970
PATIENT_NUM	1000000006
CONCEPT_CD	ICD10:K66.0
PROVIDER_ID	ORG:CH_CHOP
START_DATE	27.11.09
MODIFIER_CD	DIAGNOSEART:ND
INSTANCE_NUM	158065970
END_DATE	27.11.09

Procedures

Information about provided procedures is contained in the OBSERVATION_FACT table. For each entry the encounter ID, the patient ID, the OPS code (Operationen-

und Prozedurenschlüssel), the provider and the date are specified. Even though the table schema (see table 10.1) has columns for start and end date, for each procedure, the corresponding values are identical. An example is given in table 10.4. In total the data set contains 457 procedure entries with references to 110 distinct OPS codes. The OPS code is a reference to the CONCEPT_DIMENSION table where the concept path (position in the OPS hierarchy) and a human readable label are specified. For instance, the code OPS:8-522.d0 stands for *"Hochvoltstrahlentherapie: Linearbeschleuniger mehr als 6 MeV Photonen oder schnelle Elektronen, 3D-geplante Bestrahlung: Ohne bildgestützte Einstellung"*. Again, the provider ID refers to the dimension table PROVIDER_DIMENSION where ORG:ST_STLIN is listed as *"ST Linearbeschleuniger (STLIN)"* which is part of ORG:ST *"ST (Strahlenheilkunde)"*.

Table 10.4.: Columns of the OBSERVATION_FACT table relevant for procedures with example data.

column	example entry
ENCOUNTER_NUM	158065988
PATIENT_NUM	1000000006
CONCEPT_CD	OPS:8-522.d0
PROVIDER_ID	ORG:ST_STLIN
START_DATE	07.11.11
END_DATE	07.11.11

Drug Therapy

Information about therapies with drugs is contained in the OBSERVATION_FACT table. For each procedure entry the encounter ID, patient ID, the ATC code, the provider and the date is specified. No information about dosage or frequency is available in the provided data set. An example is given in table 10.5. In total the data set contains 47 therapy entries with references to 9 distinct ATC codes. The CONCEPT_DIMENSION table does not contain entries for ATC codes. Again, the provider ID refers to the PROVIDER_DIMENSION table where ORG:DO_DE2 is listed as *"DE Station D2 (DE2)"* which is part of ORG:DO *"DO (Dermatologische Onkologie)"*.

Table 10.5.: Columns of the OBSERVATION_FACT table relevant for drug therapy with example data.

column	example entry
ENCOUNTER_NUM	158065862
PATIENT_NUM	1000000003
CONCEPT_CD	THER|ATC:B05BB11
PROVIDER_ID	ORG:DO_DE2
START_DATE	02.12.11

Laboratory Values

Results of laboratory tests are contained in the OBSERVATION_FACT table. For each entry, a encounter ID, a patient ID, a LOINC code, a provider, a date, a instance ID, a value type (numeric or qualitative), a value, a unit and a value flag (normal, low, high) are specified. An example is given in table 10.6. In total the data set contains 2033 lab entries for distinct 151 laboratory tests. The LOINC code is a reference to the CONCEPT_DIMENSION table where the concept path (position in the LOINC hierarchy) and a human readable label is specified. For instance, the code LAB:6CRP stands for *"C-Reaktives Protein (CRP)"*. The provider ID is a reference to the PROVIDER_DIMENSION table where ORG:PM_PMS1 is listed as *"PM Station 1 (PMS1)"* which is part of ORG:PM *"PM (Palliativmedizin)"*.

Table 10.6.: Columns of the OBSERVATION_FACT table relevant for laboratory tests with example data.

column	example entry
ENCOUNTER_NUM	158065981
PATIENT_NUM	1000000003
CONCEPT_CD	LAB:6CRP
PROVIDER_ID	ORG:PM_PMS1
START_DATE	10.12.11
INSTANCE_NUM	1546329820111210
VALTYPE_CD	N
NVAL_NUM	17
TVAL_CHAR	(null)
VALUEFLAG_CD	H
UNITS_CD	mg/l

10.1.2. Unstructured Data

The baseline data set contains unstructured data in form of free text clinical reports from radiology and pathology. This data is also contained in the OBSERVATION_FACT table. Both types of reports are described in the following subsections.

Radiology Reports

In total the data set contains 60 free text radiology reports with sections for clinical question (FRA - Fragestellung), title (UNTMET - Untersuchungsmethode), findings (BEF - Befund), assessment (BEURT - Beurteilung) and optional clinical information (KA - klinische Angabe). For each of these sections a separate row

is specified (the section type is defined by the MODIFIER_CD column) and the text is stored in the OBSERVATION_BLOB column. An example is given in table 10.7. The different sections of one report are identified through a common instance ID. The standard columns for encounter, patient, provider, date etc. are also provided.

Pathology Reports

In total the data set contains 14 free text pathology reports. Pathology reports are not split into sections as the radiology reports described above. The standard columns for encounter, patient, provider, date etc. are provided. An example is given in table 10.8.

Table 10.7.: Information for a radiology report of the baseline data set. The data is stored in i2b2 in multiple rows of the OBSERVATION_FACT table.

field	example entry
ENCOUNTER_NUM	158065985
PATIENT_NUM	1000000001
CONCEPT_CD	RAD:FREI
PROVIDER_ID	ORG:DR_DRFB
START_DATE	18.03.11
INSTANCE_NUM	7932157
UNTMET	Hals-, Thorax- und Abdomen-CT mit KM i.v
KA	GSK Studie, BRAF+ Patient, Studiennummer: 001487
FRA	Staging
BEF	Voraufnahmen in Teilen vom 14.01.2011 zum Vergleich Hals: Halsweichteile annähernd symmetrisch und ohne fokale Kontrastmittelaufnahme. Zervikale Lymphknoten z.B. entlang der Halsnervengefäßscheide beidseits <1 cm im kurzen Durchmesser. Schilddrüse homogen kontrastiert, nicht vergrößert. NNH regelrecht angelegt und frei belüftet. Mastoidzellen und Tympanon beidseits pneumatisiert. Halswirbelsäule mit degenerativen Veränderungen. Thorax: Keine vergrößerten Lymphknoten in der rechten Axilla. Links axillär streifig narbige Verdichtungen mit mehreren vergrößerten, teils inhomogen kontrastierten Lymphknoten entlang der Nervengefäßscheide (Bilder 11-18) der größte = Targetläsion 1 mit 2,5 x 2,2 cm (Bild 16; VU Bild 18; 2,0 x 1,7 cm). Keine vergrößerten mediastinalen oder hilären Lymphknoten. Kein Perikarderguss. Kein Pleuraerguss. Zahlreiche Koronarstents. Intrapulmonal mehrere rundliche Verdichtungen <1 cm und größenkonstant zur VU z.B. im lateralen rechten Oberlappen (Bild 25) oder im ventralen Mittellappen (Bild 44). Keine Infiltrate. Abdomen: Leber normgroß, glatt berandet und mit multiplen größenprogredienten, teilweise neuaufgetretenen flau hypodensen Läsionen. Zielläsion 2 im zentralen S IVa mit 4,4 x 4,4 cm (Bild 62; VU Bild 65; 1,4 x 1,2 cm) und Zielläsionen 3 im S V mit 3,1 x 2,9 cm (Bild 67; VU Bild 74; 1,1 x 0,7 cm). Gallenblase gefüllt und mit Kontrastmittelaufnehmender Wand sowie Flüssigkeit im Gallenblasenbett. Keine Erweiterung der intra- und extrahepatischen Gallengänge…
BEURT	Bei Melanom Stadium IV größenprogrediente Lymphknotenmetastasen links axillär und an der Leberpforte. An Zahl und an Größe deutlich progrediente Lebermetastasen. Diss. ossäre Filiae, derzeit nicht stabilitätsgefährdend. Größenkonstante intrapulmonale Verdichtungen unklarer Dignität. RECIST Summe 10,0 (VU 4,5) .

Table 10.8.: Example of a pathology report from the baseline data set.

field	example entry
ENCOUNTER_NUM	158065842
PATIENT_NUM	1000000002
CONCEPT_CD	PATHO:FREITEXT
PROVIDER_ID	ORG:DR_DRFB
START_DATE	06.10.10
INSTANCE_NUM	24604
OBSERVATION_BLOB	Eingangsnummer: E/10/24604 Makroskopischer Befund: Klinische Angabe: V.a. Metastasen malignes Melanom beidseits. Zusendung: I. Keilresektat rechter Unterlappen: Lungenresektat von 3 x 1,5 x 0,9 cm. Auf lamellierenden Schnitten ein grauer Knoten von 0,9 cm im Durchmesser. Abstand zur Absetzungsfläche 0,2 cm. Kapsel 1 Schnellschnitt, Kapsel 2 Rest. II. Keilresektat rechter Unterlappen: Lungenteilresektat von 3,6 x 1,7 x 1,2 cm. Klammernahtverschlossen auf 2,8 und 2,5 cm. Auf lamellierenden Schnitten durch das Lungenparenchym ein weißlicher Herd, unscharf begrenzt, von maximal 0,3 cm im Durchmesser. Abstand zur Absetzungsfläche (blau getuscht) 0,9 cm. Im übrigen Lungenparenchym kein Herdbefund. Kapsel 1 Lungenparenchym mit Knoten, Kapsel 2+3 übriges Lungenparenchym. III. Keilresektat rechter Unterlappen: Lungenresektat von 2,3 x 1,7 x 0,8 cm. Pleura glatt. Klammernaht auf 2,2 cm. Auf lamellierenden Schnitten durch das Lungenparenchym ein Knoten von hellbrauner Schnittfläche, scharf begrenzt, Durchmesser 0,4 cm. Abstand zur tuschemarkierten Absetzungsfläche 0,2 cm…

10.2. Corpora of Radiology Reports

Three different corpora of radiology reports are described in the following subsections.

10.2.1. German Radiology Reports on Lymphoma Patients

The *lymphoma data set* consists of 2584 German radiology reports on 377 lymphoma patients. The imaging modality is mainly computed tomography (CT), but also magnetic resonance imaging (MRI) and ultrasound (US). The reports are from 27 different readers and the inspected body regions are mainly abdomen, thorax and head, but include also various other regions from the whole body. The data is provided in form of an excel sheet. The data schema and example entries are shown in table A.1 in appendix A.2. Since the data set contains multiple reports from consecutive examinations for each patient, it is suitable for longitudinal analysis of clinical findings. The reports contain many measurements and thus they are also used during the development of the knowledge model of normal size specifications. Since the reports are about patients diagnosed with lymphoma disease, the data set contains many measurements of lymph nodes in various body regions and also lesions and tumors.

10.2.2. German Radiology Reports on Internistic Patients

The *internistic data set* consists of 6007 German radiology reports on internistic patients from 27 different readers, where imaging modality was computed tomography (CT). The examined patients have a diverse disease background. The data set does not contain patient identifiers or date values and thus cannot be used for longitudinal analysis of clinical findings. However, the advantage of this data set is that the coverage of patients is *diverse* and that the data set contains many measurements (e.g. more than 20 thousand length measurements). Thus, the data set is used for the evaluation of the measurement entity resolution and finding classification algorithms. The data schema and example entries are shown in table A.2 in appendix A.2.

10.2.3. English Radiology Reports

The data set consists of English radiology reports with diverse imaging modality and patient background (disease, age, gender etc.). The data set does not contain

patient identifiers or date values and thus cannot be used for longitudinal analysis of clinical findings. However, the advantage of this data set is that English texts are better covered by ontologies such as RadLex leading to a better annotation coverage. Thus, the data set is used for the evaluation of the measurement entity resolution algorithm.

10.3. Patients with Structured Finding Information

The data set on patients with structured finding information is provided in form of excel sheets and covers 40 patients. For these patients, the age, an associated main disease (that was assigned by a clinical expert) and a set of present and absent findings and symptoms are specified as shown for two patients in table 10.9. For each patient, 3 - 15 known present and up to 5 known absent findings or symptoms were specified by a clinical expert. The age of the patients ranges from 50 to 87 years. For each of the diseases (lymphoma, reactive lymphadenitis, diverticulitis and colorectal cancer) the data set contains 10 patients with an expert associated main diagnosis. The set of present and absent findings express typical symptomatology of the associated main disease with some deviations commonly found within the diagnosis process. The knowledge model about diseases and their manifestation in clinical findings and symptoms is described in section 6.2. It is used to infer the likely diseases for the patients based on the structured finding information. In total the model contains 5 diseases (Hodgkin lymphoma, non-Hodgkin lymphoma, reactive lymphadenitis, colorectal carcinoma and diverticulitis) and 41 clinical findings or symptoms. The main diseases of the patients are covered by the knowledge model.

Table 10.9.: The initial set of known present (+) and absent (-) findings and symptoms for a patient with main disease colorectal cancer and for a patient with main disease lymphoma.

patient	patient 100000ABC	patient 100000XYZ
age in years	50	51
main disease	colorectal cancer	lymphoma
anamnesis	feeling powerless (+) blood in stool (+) changes in bowel patterns (+) fever (-) malaise (+)	b-symptomatic (+) feeling powerless (+) night sweats (+) weight loss (+) tendency to infection
laboratory	anemia (+) positive stool guaiac test (+) carcinoembryonic antigen raised (+)	c-reative protein raised (+) ESR raised (-)
radiology	wall of intestine thickened (+) enl. pararectal lymph nodes (+) wall of intestine thickened (+) diverticular disease of colon (-)	enlarged lymph nodes (+) enl. mediastinal lymph nodes (+) enl. lymph nodes of thorax (+) enl. abdominal lymph nodes (+) enl. deep cervical lymph nodes (+) enlarged spleen (+)
other signs		petechiae (+)

11

Evaluation

This chapter presents the evaluation of the algorithms described in chapter 8 using the data sets from the previous chapter. Firstly, the knowledge-based extraction of measurement-entity relations based on annotated radiology reports is evaluated on all three data sets of radiology reports. For the two corpora of German reports, an accuracy above 80% and for English reports an accuracy of above 74% is achieved. The difference is mostly due to the fact, that the English reports are significantly more diverse and thus less well covered by the knowledge model (cf. section 6.1). Secondly, the classification of measurement findings is evaluated. Here, an average accuracy of 90% is achieved with only very few false classifications. Thirdly, the ranking of likely diseases based on structured finding information is evaluated. In 35 of 40 cases the most likely disease from the ranking algorithm was also the main disease associated by a clinical expert. In the remaining 5 cases, the main disease was ranked at position 2 or 3. The inference of change is not evaluated in terms of accuracy, since it was realized through the ReportViewer user interface (cf. section 9.2).

11.1. Knowledge-based Extraction of Measurement-Entity Relations

The evaluation of the knowledge-based extraction of measurement entity relations described in section 8.2 is done on the data sets of German radiology reports on lymphoma patients and on internistic patients and further on the data set of English radiology reports. The initial version of the resolution algorithm was presented in [Obe+14]. The algorithm described in 8.2 is an extension to the initial version encompassing the following adaptations:

No restriction on specific sentences: While the initial version was restricted to sentences with one or two measurements, the adapted algorithm does not have any constraints on the number of measurements in a sentence. That is, all report sentences with measurements are in scope for the evaluation.

Not pure knowledge-based: While the initial approach was purely knowledge-based, the algorithm described in section 8.2 involves also a statistical component and the sentence distance.

Radlex adaptation: One result of the evaluation of the initial version of the resolution algorithm in [Obe+14] was that many false resolutions occurred due to *wrong* or *missing* annotations. In previous work [Bre+14], 539 German labels were added to 433 distinct RadLex entities by semi-automatic translation process to enhance annotation coverage regarding the lymphoma data set. For this evaluation RadLex was adapted a second time: In particular, several synonyms of lesion classes that led to *wrong* annotations were removed. For instance, within the original RadLex version, the entity 'radlex:breast mass' has synonyms 'mass', 'nodule', 'lesion', 'nodular enhancement' and 'area of enhancement'. Thus, each time a 'mass' or 'lesion' is mentioned in a report the annotator assigned 'radlex:breast mass' and then the resolution algorithm *falsely* resolved to 'breast mass'. By removal of these synonyms from RadLex less false annotations were obtained. Furthermore, another 106 German labels were manually added to 35 distinct RadLex entities to enhance annotation coverage. For instance, German labels were added for renal cyst, all liver segments, nodule and others entities.

11.1.1. Evaluation Schema

For each data set a list containing all sentences and the resolved entity, if available, was generated. An example is given below in table 11.1. Evaluation was done

according to the following schema:

- **correct:** The entity resolved is exactly what the measurement of the sentence is about. The radiologist cannot find a better entity by reading the sentence.
- **(correct):** The entity resolved is correct however it could be more specific. For example, “lymph node in jaw angle with 1 cm" with resolution to ‘lymph node’. Here, a radiologist can name a better entity.
- **unresolvable:** The sentence does not allow a resolution (e.g. “The biggest is now 2.7 cm.") or the measurement does not represent a size description of an anatomical entity (e.g. “Tracheal tube 4 cm above the carina.") *and* the algorithm did not resolve to a false entity.
- **false:** The resolved entity is false or the entity was not resolved, but the radiologist can identify the correct entity within the sentence.

Table 11.1.: Example row of the evaluation of the measurement entity resolution. The example is evaluated as ‘(correct)’, because the expert can name a better entity (the *abdominal* aorta) by reading the sentence.

sentence	measurement	resolved entity	evaluation
The maximum sagittal diameter of the aorta is 1.6 cm in the midabdomen.	1.6 cm	aorta	(correct)

11.1.2. Summary of Results

For all three corpora the findings and assessment section of the radiology reports were annotated and taken as the input for the resolution algorithm. Then, for 500 different and randomly selected sentences one resolved measurement-entity relation was evaluated by a radiologist. If a sentence contained more than one measurement, one measurement was randomly selected for evaluation. As shown in table 11.2, the algorithm resolves measurement-entity relations with an accuracy of above 80% for German reports and above 74% for English reports. The results for English reports are slightly worse than those for German reports. This overall result for German reports is similar to the results of the initial pure knowledge-based approach [Obe+14], however in comparison to the evaluation of the initial work – where about 8% of the sentences were excluded – now *all sentences* are in scope.

As shown in table 11.2, for the German corpora, for about 24% of all resolutions, the clinical expert cannot name a better entity than the algorithm by reading

Table 11.2.: Evaluation results for 500 randomly selected sentences for each data set.

data set	lymphoma reports	internistic reports	english reports
evaluated	500	500	500
correct	135	119	97
(correct)	249	267	179
unresolvable	20	17	96
false (not resolved)	58	56	60
false (incorrect)	38	41	68
accuracy	80.8%	80.6%	74.4%

the sentence. About 50% of the resolutions are only '(correct)', i.e. for these sentences the clinician can name a better entity. As explained below, most of these cases are measurements of *lesions*, where the location aspect is missing in the resolved entity. Furthermore, the evaluation shows that more than half of all *false* resolutions arise because no entity could be resolved at all. Both problems are mainly due to low annotation quality. Even though the annotator was able to annotate multi-word terms and recognize inflected forms, the quality of the resulting annotations strongly depends on the quality of the provided vocabulary of the ontology (i.e. RadLex). Firstly, only about 25% of all RadLex concepts have German labels (including the extension). Thus, the correct entity gets not always annotated. Secondly, there are not always detailed entities in RadLex that could precisely describe the measured entity. For example, *nodule* or *metastasis* do not have subclasses in RadLex. Consequently, a resolution of measurements of some metastasis cannot be as precise as a clinical expert. The location aspect can only be captured by reference to a second annotation. This is described below.

In comparison the algorithm has better results for German reports than for English reports. Since Radlex has only for about 25% of its entities German labels, one would expect better results for English reports. However, even though the overall coverage of the annotations was better, there are two reasons, why this is not reflected in the results. Firstly, the data set of English reports was significantly more diverse and contained reports on many different examinations. Secondly, the annotator is specific for the German language. For example, the annotator had problems with sentence splitting and also detection of plural forms (e.g., *cysts*) for English reports. The annotator creates 0-12 relevant annotations per sentence for the sentences of the data set of German radiology reports and 0-15 for the data set of English reports respectively.

11.1.3. Detailed Evaluation for German Radiology Reports

Since the overall results shown in table 11.2 are similar for both data sets of German reports, detailed analysis regarding resolved anatomical entities is done together. In summary, the algorithm resolves relations to 68 different anatomical entities. That is, by using the RadLex hierarchy, the coverage of the knowledge model is extended to more entities than defined by the knowledge model explicitly. In the following paragraphs the evaluation results for different anatomical entities (lymph nodes, lesions and cysts etc.) are presented, before causes for *false* resolutions are analyzed.

Lymph Nodes: As shown in table 11.3, good results for resolution of lymph nodes are achieved, which is due to the fact that RadLex has a very detailed subclass hierarchy for lymph nodes. In total, the algorithms resolves to 24 different lymph node entities and in 50% of these resolutions the radiologist cannot name a better entity than the algorithm. The problem however is, that not all of the subclasses of 'lymph node' have German labels. Thus, in many cases the annotator detects only 'lymph node' even though the sentence contains a more precise description. Only in 2 of 101 cases the clinical expert could select a better entity from the set of annotations. Similarly, there are several cases where the resolution to 'mediastinal lymph node' is not precise enough. For example, in *"Mulitple lymph nodes mediastinal slightly regressive, e.g. infracarinal 1.9 x 1.7 cm ..."*[1], the clinical expert can name the measured entity more precisely as 'infracarinal lymph node'. False resolutions of lymph node entities are mainly due to very long sentences where *multiple* measurements and different lymph nodes are mentioned: For example in *"New suspectly enlarged lymph nodes e.g. retrocaval 3.1 x 1.1 cm (image 81) or right ventral of the musculus psoas with 1.1 x 0.9 cm (image 98) "*[2] the resolution "1.1 x 0.9 cm, retrocaval lymph node" is false.

Lesions and Cysts: The resolution results for lesions and cysts are good, i.e. there are only few false resolutions. However, the location of lesions and cysts is not resolved sufficiently by the standard algorithm, which resolves to *one* entity from RadLex. Since the RadLex class hierarchy for lesions is not as detailed as for lymph nodes, the resolution is not precise enough for lesions. For example, "Unchanged hypodense structure in gallbladder with 0.4 cm (IMA 33).)"[3] is resolved to 'lesion'. Similarly, resolution to *renal cyst* is mostly not precise enough since the cyst might be on the right or left kidney. However, RadLex does not define more classes

[1] Original in German: Mehrere Lymphknoten mediastinal, gering größenregredient, z.B. infrakarinal mit 1,9 x 1,7 cm, VU 2,0 x 1,8 cm (Bild 31, VU 38).

[2] Neue suspekt vergrößerte Lymphknoten , z.B. retrokaval mit 3, 1 x 1,1 cm (Bild 81) oder rechts ventral des Musculus psoas mit 1,1 x 0,9 cm (Bild 98).

[3] Unveränderte hyperdense Struktur in der Gallenblase mit 0,4 cm (IMA 30).

Table 11.3.: Evaluation results for lymph nodes in German reports.

anatomical entity	correct	(correct)	false
aortopulmonary lymph node	5	0	0
axillary lymph node	30	1	0
bronchopulmonary lymph node	6	0	0
common iliac lymph node	3	4	1
esophageal lymph node	1	0	0
inguinal lymph node	19	2	0
jugulodigastric node	1	0	0
lateral aortic lymph node	15	8	1
left bronchopulmonary lymph node	5	0	0
lymph node	10	101	1
lymph node of abdomen proper	1	0	0
mastoid lymph node	0	1	0
mediastinal lymph node	31	27	1
mesenteric lymph node	13	0	0
paracardiac lymph node	1	1	0
parietal lumbar lymph node	0	0	1
parietal lymph node of abdomen proper	5	1	0
retrocaval lymph node	1	0	1
retrocrural lymph node	3	0	0
right bronchopulmonary lymph node	9	0	0
right paraesophageal lymph node	5	0	1
submandibular lymph node	7	4	1
submental lymph node	3	1	0
supraclavicular lymph node	1	8	0
all lymph nodes	175	167	8

for these specific cysts. Thus, the location information needs to be resolved in a subsequent resolution step which is described in the next paragraph. Note, however, that through removal of RadLex synonyms (described above), many false resolutions were avoided in comparison to the initial version described in [Obe+14].

Resolution of the Location of Lesions and Cysts: As shown in table 11.4, resolution of measurements of lesions and cysts is often not precise enough. This is mainly because of the fact that RadLex does not contain detailed enough entities. For example, RadLex contains the entity 'lesion' but no entities like 'lesion is liver segment 1' etc. Correspondingly, annotations are not precise enough and the

Table 11.4.: Evaluation results for lesions and cysts in German reports.

anatomical entity	correct	(correct)	false
carcinoma	0	1	0
cyst	0	12	0
enhancement	0	0	8
fluid	0	2	0
focus	0	6	0
lesion	12	165	4
lymphoma	0	8	0
mass	0	61	4
metastasis	1	6	0
neoplasm	1	7	0
nodule	0	48	1
ovarian cyst	2	0	0
portion of tissue	0	1	0
renal cyst	1	7	0
round mass	0	16	0
Total Result	14	309	17

location aspect of lesions and cysts has to be resolved in a subsequent step.

The resolution of the location aspect was performed on a data set of sentences where the measurement was resolved to one of the entities listed in table 11.4. The location resolution algorithm is fairly simple and should only demonstrate that the location aspect can be resolved by a similar approach as for the measured entity itself. The algorithm simply takes entities from a predefined set of possible locations and selects from the set of annotations the one that is mostly close to the measurement and the measured entity. From the data set of German radiology reports on internistic patients as well as from the data set of German radiology reports on lymphoma patients 100 sentences were analyzed. In total in 52.5 % sentences the location could be correctly resolved. In 71.4 % of these cases the resolution was as precise as manual resolution by the clinical expert. The detailed results by anatomical entity are shown in table 11.5.

In total the algorithm resolved to the following locations (number of resolutions in brackets): aorta (1), aortic arch (1), area X (2), breast (1), caudate lobe of liver (1), caudomedial auditory cortex (1), cerebellum (2), head of pancreas (3), hepatovenous segment II (5), hepatovenous segment III (1), hepatovenous segment IV (4), hepatovenous segment IVa (5), hepatovenous segment IVb (4), hepatovenous segment V (8), hepatovenous segment VII (10), hepatovenous segment VIII (2), kidney (1), L1 vertebral body (1), L4 vertebral body (1), left adrenal gland (4), left kidney (2), liver (7), lower lobe of left lung (3), lower lobe of right lung (11),

Table 11.5.: Evaluation results for resolution of the *location* of lesions and cysts in German reports.

anatomical entity	correct	(correct)	false
cyst	5	1	1
enhancement	0	3	2
focus	1	0	2
lesion	52	10	36
mass	7	8	20
metastasis	1	2	4
neoplasm	0	1	2
nodule	3	3	22
round mass	6	1	6
Total Result	75	30	95

mediastinum (1), occipital brain region (1), pleura (1), porta Hepatis (1), right kidney (2), spleen (4), T10 vertebral body (1), upper lobe of left lung (1) and upper lobe of right lung (1).

The false location resolutions (47.5 %) are mainly due to missing annotations for the location aspect and thus *not resolved locations*: 85 of the 95 false resolutions (i.e. 89.5 %) are due to unresolved locations.

Other Entities: As shown in table 11.6 the algorithm correctly resolves to many more entities than only lymph nodes and lesions. Only for 8 out of these 37 entities there are false resolutions at all and the false resolutions mainly concern aorta, gallbladder and liver. In the case of the one of the false resolutions to liver, the measurement described a very large lesion (18.0 x 10.0 cm), which is as large as a typical liver.

Analysis of False Resolutions: In total there are 193 cases of false resolution for both data sets of German reports together. In 114 of those cases the algorithm did not come up with a resolution and 79 times the resolution was incorrect. For all false resolutions the clinical expert reviewed the available annotations and tried to find a suitable resolution. Only in 41 of all 193 false resolutions (i.e. in 21.2%) the expert could find a correct entity beyond the annotations. In all other cases (78.8%) the correct entity was simply not annotated and the resolution algorithm had no chance to pick the right entity.

There are entities, where the algorithm always falsely resolved to, mainly due to the lack of alternative annotations. Thus, these entities were falsely selected

Table 11.6.: Evaluation results for other entities in German reports.

anatomical entity	correct	(correct)	false
aorta	0	2	5
ascending aorta	1	0	0
bile duct	1	0	0
colon	0	1	0
common bile duct	1	0	0
dehiscence	0	1	0
gallbladder	1	0	4
head of pancreas	0	1	0
heart	1	0	0
hematoma	1	0	0
hilum	1	0	2
lipoma	0	1	0
liver	4	0	4
osteolysis	0	2	0
pericardial effusion	2	0	0
pleura	0	1	1
pleural effusion	5	0	1
pneumothorax	0	2	0
prostate	6	0	0
pulmonary artery	0	1	1
rib	0	1	0
right adrenal gland	2	1	0
scar	0	1	0
spleen	32	0	0
stenosis	0	1	0
tail of pancreas	0	1	1
thymus	1	0	0
ventricular septal defect	1	0	0
wall of gallbladder	2	0	0

because they were close to the measurement within the sentence. The following entities had *only* false evaluation results (frequency given in brackets): abdominal aorta (1), amyotrophic lateral sclerosis (2), anus (2), aortopulmonary window (1), area X (1), breast (1), carina (1), chest wall (1), dilation (1), enhancement (8), grade I chondromalacia (1), gut (2), head (2), hepatovenous segment II (1), hepatovenous segment IVa (1), ileus (1), infiltrate (3), lower abdomen (1), lungs (1), pancreatic duct (1), parietal lumbar lymph node (1), porta Hepatis (1), renal vein (1), retroperitoneum (1), right atrium (1), sacral promontory (1), segmental

enhancement (1), skull (1), T6 vertebral body (1), tuber (1) and ventral anterior nucleus (1). By creating a list of a subset of these entities that are commonly *not* measured in radiology examinations these false resolutions could be possibly avoided.

11.1.4. Detailed Evaluation for Data Set of English Reports

The overall results of the relation extraction for the data set of English reports are shown in table 11.2. In this subsection the results are analyzed in detail in the same manner as in the previous subsection for German reports. In summary, the algorithm resolves relations to 90 different anatomical entities. In the following paragraphs the evaluation results for different anatomical entities (lymph nodes, lesions and cysts, etc.) are presented, before *false* resolutions and *unresolvable* sentences are analyzed.

Lymph Nodes: As shown in table 11.7, good results for resolution of lymph nodes are achieved, which is due to the fact that RadLex has a relatively detailed subclass hierarchy for lymph nodes. In total, the algorithms resolves to 8 different lymph node entities and in 42% of these resolutions the radiologist cannot name a better entity. However, in 8 sentences the annotator detects only 'lymph node' even though the sentence contains a more precise description. Thus, in none of these cases the clinical expert could select a better entity from the set of annotations. For example, in *"A 19 x 13 mm lymph node is seen in Station 2 L anterior to the left mainstem bronchus."* has only annotation 'lymph node', 'bronchus', 'main bronchus', 'left main bronchus', 'lung' etc. however no entity such as 'bronchial lymph node'. In comparison to the German corpora, the English reports contain far less lymph node measurements.

Table 11.7.: Evaluation results for lymph nodes in English reports.

anatomical entity	correct	(correct)	false
aortopulmonary lymph node	1	0	0
axillary lymph node	1	1	0
common iliac lymph node	2	0	0
hepatic lymph node	1	0	0
inguinal lymph node	0	1	0
left bronchopulmonary lymph node	0	0	1
lymph node	1	8	0
subcarinal lymph node	2	0	0
all lymph nodes	8	10	1

Lesions and Cysts: The resolution results for lesions and cysts are good, i.e. there are only few false resolutions (cf. table 11.8). However, one important aspect of lesions and cysts – the location – is not resolved. This is because the algorithm resolves to *one* entity. Since the RadLex class hierarchy for lesions is not as detailed as for lymph nodes, resolution is not precise enough. Thus, the algorithm often resolves to *cyst*, *lesion*, *mass* or *nodule*. For example in "A 1 mm lesion is noted posteriorly in the left acetabular roof." is resolved to 'lesion'. Thus, the location information needs to be resolved in a subsequent resolution step. In the previous subsection on German reports it was shown that it is possible to resolve the location aspect in a subsequent processing step.

Table 11.8.: Evaluation results for lesions and cysts in English reports.

anatomical entity	correct	(correct)	false
breast mass	1	2	0
chest mass	0	1	0
cyst	1	29	0
enhancement	0	2	1
fluid	0	8	0
focus	0	2	2
lesion	0	24	0
lobular mass	0	1	0
mass	4	16	5
mass in or on skin	1	0	0
meniscal cyst	1	0	0
neoplasm	0	1	0
nodule	4	45	3
ovarian cyst	0	4	0
parapelvic cyst	0	1	0
renal cyst	1	2	0
round mass	0	1	0
synovial cyst	1	0	0
Total Result	14	139	11

Other Entities: As shown in table 11.9 the algorithm correctly resolves to many more entities than only lymph nodes and lesions. From these 32 entities there are false resolutions only for two entities (*fragmentation* and *ovary*) while for 14 entities the resolution is always as good as the clinical expert (*abdominal aorta*, *common bile duct*, *dilation*, *intraosseous hemangioma*, *isthmus of thyroid gland*, *leiomyoma*, *lipoma*, *liver*, *nasal septum deviation*, *right kidney*, *scar*, *spermatocele*, *spleen* and *wall of gallbladder*).

Table 11.9.: Evaluation results for other entities in English reports.

anatomical entity	correct	(correct)	false
abdominal aorta	2	0	0
aorta	0	5	0
calcification	0	4	0
cervix of uterus	1	0	0
common bile duct	10	0	0
common duct	5	1	0
compression	0	2	0
consolidation	1	1	0
dilation	1	0	0
fat	0	1	0
fragmentation	0	1	2
ganglion	0	1	0
granuloma	0	2	0
hernia	0	1	0
infiltrate	0	2	0
intraosseous hemangioma	1	0	0
isthmus of thyroid gland	1	0	0
left kidney	14	1	0
leiomyoma	1	0	0
lipoma	1	0	0
liver	5	0	0
lobe	0	2	0
nasal septum deviation	1	0	0
ovary	1	2	2
pleural effusion	1	2	0
right kidney	12	0	0
scar	1	0	0
set of testicles	0	1	0
spermatocele	1	0	0
spleen	14	0	0
tear	0	1	0
wall of gallbladder	1	0	0

Analysis of False Resolutions: In total there are 128 cases of false resolution for the data set of English reports. In 60 of those cases the algorithm did not come up with a resolution and 68 times the resolution was incorrect. For all false resolutions the clinical expert reviewed the available annotations and tried to find a suitable resolution. In 63 of all 128 false resolutions (i.e. in 49.2%) the expert could find a

better entity beyond the annotations. In all other cases (50.8%) the correct entity was not annotated and, consequently, the resolution algorithm had no chance to pick the right entity.

There are entities, where the algorithm always falsely resolved to, in many cases, due to the lack of alternative annotations. Thus, these entities were falsely selected because they were close to the measurement within the sentence. For example, the algorithm falsely resolved *tendon sheath* 5 times. In none of these cases there was an alternative annotation. In the case of *endometritis* (4 false resolutions) the correct entity (*endometrium*) was annotated. This false resolution could be corrected by excluding diseases and disorders from the set of possible resolutions (which would exclude *endometritis*). The following entities had *only* false evaluation results (frequency given in brackets): acoustic tubercle (1), adenoma (1), anus (2), apex (1), area X (6), body (1), carina (1), clavicle (1), dens (1), endometritis (4), gallbladder (2), greater sac (1), iliacus muscle (1), kidney (1), left bronchopulmonary lymph node (1), left vertebral artery (2), lower lobe of right lung (2), lumen (1), lungs (2), origin (1), pelvis (3), pubis (1), right nipple (1), right ureter (1), spine (2), stenosis (1), submucosa (1), tendon (1), tendon sheath (5), thigh (1), ureter (1), ureteropelvic junction (1) and vertebra (1). Here, creation of a list of a subset of these entities that are commonly *not* measured in radiology examinations could possibly avoid these false resolutions.

Analysis of Unresolvable Sentences: In contrast to the data sets of German radiology reports, the English sentences were significantly more often classified by the clinical expert as *unresolvable* (96 of 500 sentences). This is due to the higher variability of the English reports, where many measurements are actually not about the *size* of some anatomical entity. For instance, some measurements describe a *position*: For example, "... the sheath tip was placed 2 cm from the saphenofemoral junction". In many cases sentences are unresolvable due to a wrong annotation of the *type* of the measurement: For example, there were many cases such as "Peak systolic velocity in the left ICA is 90 cm/sec and ...", where the annotator falsely detected a size measurement (here *90 cm*)! Furthermore, some measurements were about some medical device or another entity. For example, "The patient swallowed a 13 mm barium tablet without difficulty". Also, technical details of the imaging modality were under these measurements. For example, "Spiral CT imaging through the neck was performed at 3 mm slice ...". In all of these cases the measurements are not in scope of the resolution because they do not represent a size of some anatomical entity. As for German reports, several measurements are also not resolvable within sentence boundaries. For example, "Maximum AP diameter is 2.4 cm." or "Many of these measure 1 to 2 mm in size.". For all 96 sentences classified as *unresolvable* the resolution algorithm did not resolve to a false entity. Thus, it is not a false resolution.

11.2. Normality Classification

The lymphoma data set and the internistic data set were used for the evaluation of the normal classification of size findings described in section 8.3. Based on the extracted and evaluated structured finding representations from the previous section a test set was created: Firstly, all *correctly* extracted findings were added to the test set and, secondly, all findings for which the clinician could find a correct or more specific anatomical entity within the set of annotations were added. That is the test set contains the best possible finding representation that can be manually selected from reports based on the available annotations.

11.2.1. Evaluation Schema

For each data set the clinical expert was provided with an excel sheet where for each classified finding the sentence, the measurement, the anatomical entity and the classifier result (*normal*, *abnormal* or *unclassified*) are given. For example, the sentence "Lymph nodes within aortopulmonary window and pretracheal unchanged with a size of up to 1.6 x 1.2 cm."[4], the measurement "1.6 x 1.2 cm", the anatomical entity "aortopulmonary lymph node" with classification *abnormal*. Then the clinical expert evaluated the classified finding as *true* (i.e. correct) or *false*. All *unclassified* findings are evaluated as *false*.

11.2.2. Results

In table 11.10 the evaluation of the classification of size findings as normal or abnormal is shown for the *lymphoma data set* as well as for the *internistic data set*. For both data sets the number of *false normal* and *false abnormal* classified findings is very small. However, both data sets have a significant amount of *unclassified* findings which occur when the anatomical entities are not covered by the knowledge model. For instance, the size findings about "portion of soft tissue" are never classified due to that reason (6 times for the lymphoma data set and 9 times for the internistic data set). A portion of soft tissue however is too unspecific to be classified as normal or abnormal. Similarly, a *fluid*, a *duct* or a *wall* are too unspecific for a classification by size. Furthermore, there are anatomical entities which are simply not yet covered by the knowledge model. For instance, the finding "2.8 x 0.7 cm, rib" is not classified because the knowledge model does not contain a normal size specification for the *rib*. Similarly, findings about *pleura*, *trachea*, *right ventricle*, *bulla*, *infrarenal aorta*, *mucosa*, *L2 vertebral body*, *T10 vertebral*

[4]Original in German: Die LK im aortopulmonalen Fenster und prätracheal größenunverändert mit einer Größe von bis zu 1,6 x 1,2 cm.

body, *L4 vertebral body*, *aortic arch*, *crus of diaphragm* and *set of biliary ducts* were not classified. The *rate* of correctly classified findings for the lymphoma data set is significantly higher than for the internistic data set. This is due to the fact that for the internistic data set more findings did not get classified (25 out of 209). The underlying reason for this difference is the coverage of the knowledge model regarding the findings contained in the data sets. In particular, the lymphoma data set contains more measurements of lymph nodes.

Table 11.10.: Evaluation results for the classification of size findings as normal or abnormal.

	lymphoma data set	internistic data set
total number of findings	250	209
true normal	115	86
true abnormal	120	95
false normal	2	2
false abnormal	2	1
unclassified	11	25
rate of correctly classified	94.0%	86.6%

Analysis of False Classifications: Most of the false classifications are spleen measurements. For instance, the finding "14.8 x 8.5 cm, spleen" was falsely classified as *normal*. This is because the classification algorithm mapped 14.8 cm to the height (normally 11-15 cm) and the length (normally 7-10 cm). However the size was measured in the transverse plane and thus describes the length and *width* (which is normally 4-6 cm). Accordingly, the finding describes a splenomegaly, i.e. an *abnormal* spleen. Furthermore, the finding "9 x 3.5 x 6.5 cm, spleen" was falsely classified as *abnormal* since the algorithm detected that the value 3.5 cm is *below* the normal width of 4-6 cm. However, according to the clinician, this deviation is not considered to be abnormal. Here, the algorithm needs to be adapted accordingly. Furthermore, the finding "4 cm, ascending aorta" was falsely classified as *normal*. In this case the size was compared to the normal size of the aorta at the root (normally at least 4 cm). However, the 'ascending aorta' is normally slightly smaller and thus 4 cm should have been classified as *abnormal* since it represents an ectasia (i.e., a widening) of the ascending aorta. This type of fault can be avoided by extension of the knowledge model with more granular size descriptions.

11.3. Disease-centric Data Access

The evaluation of the disease-centric data access, described in chapter 9, is done on the data sets of 40 patients, for which a clinical expert manually created structured finding information (cf. section 10.3). Firstly, it is evaluated, whether the ranking of likely diseases described in section 8.6 coincides with the associate main disease that was assigned by a clinical expert. Secondly, it is described, to which extent the corresponding prototype implementation, described in chapter 9, helps the clinician within differential diagnosis. An initial version of the approach and the ranking was described in [Obe+12a; Obe+12b].

Evaluation of the Ranking: Before the ranking is calculated the set of present findings or symptoms is augmented as described in section 8.5. For instance, for patient 100000ABC (cf. section 10.3), the finding 'enlarged lymph nodes' is not explicitly defined. However, since there are 'enlarged pararectal lymph nodes', the following findings are inferred: 'enlarged abdominal lymph nodes' and 'enlarged lymph nodes'. After this inference step the ranking is calculated. As shown in table 11.11, for 35 out of 40 patients the inferred most likely disease from the ranking algorithm corresponds to the main disease associated by a clinical expert. In the remaining 5 cases the main disease was ranked at position 2 or 3. In the case of patients with associated main disease lymphoma, non-Hodgkin lymphoma (a subclass of lymphoma) was always ranked at position one and Hodgkin lymphoma (also a subclass of lymphoma) was always ranked at position two. This is due to the fact that non-Hodgkin lymphoma is about 3 times more frequent in clinical practice than Hodgkin lymphoma. Since the expert associated main disease *lymphoma* subsumes both Hodgkin and non-Hodgkin lymphoma, these cases were considered to be correct. The average ranking position of the disease that was associated as the main disease by the clinical expert is for all diseases better than the average position of the other diseases (cf. table 11.11).

For lymphadenopathy, the difference of the average ranking value is close to those of non-Hodgkin and Hodgkin lymphoma. However, for colorectal cancer and lymphoma the algorithm clearly identified the main disease in all cases. This is also shown in table 11.12, where the exact ranking values are listed. For patient 100000ABC, the gap between ranking values of colorectal cancer and lymphadenopathy is significant. For patient 100000XYZ, the gap is between the two lymphoma diseases (Hodgkin and non-Hodgkin lymphoma) and the third ranked disease lymphadenopathy.

Evaluation of Prototype Implementation: The prototype implementation for visualization and interaction with the ranking of likely diseases is based on a stacked bar diagram to show the absolute and relative amount of present and

Table 11.11.: Ranking results for the disease ranking.

expert diagnosis	average rank position of diseases
colorectal cancer	1.0 colorectal cancer
	2.7 non-Hodgkin lymphoma
	2.8 lymphadenopathy
	3.5 Hodgkin lymphoma
	4.0 diverticulitis
lymphoma	1.0 non-Hodgkin lymphoma
	2.0 Hodgkin lymphoma
	3.5 lymphadenopathy
	3.7 colorectal cancer
	4.8 diverticulitis
lymphadenopathy (LYM)	1.6 lymphadenopathy
	2.2 non-Hodgkin lymphoma
	2.6 Hodgkin lymphoma
	4.0 diverticulitis
	4.6 colorectal cancer
diverticulitis (DIV)	1.2 diverticulitis
	1.8 lymphadenopathy
	3.6 colorectal cancer
	3.6 non-Hodgkin lymphoma
	4.8 Hodgkin lymphoma

Table 11.12.: Ranking results for the two patients listed in table 10.9. Raking values from algorithm are given in brackets. Higher values indicate more likely diseases.

patient	patient 100000ABC	patient 100000XYZ
expert diag.	colorectal cancer	lymphoma
rank 1	colorectal cancer (0.706)	non-Hodgkin lymphoma (0.700)
rank 2	lymphadenopathy (0.286)	Hodgkin lymphoma (0.632)
rank 3	non-Hodgkin lymphoma (0.286)	lymphadenopathy (0.267)
rank 4	Hodgkin lymphoma (0.286)	colorectal cancer (0.154)
rank 5	diverticulitis (0.154)	diverticulitis (0.154)

absent findings or symptoms (see figures in section 9.1). Leading symptoms (denoted by 'LS') are distinguished from other findings or symptoms (denoted by 'S'). Information about risk age and adapted incidence proportion is given additionally. By clicking on the chart the clinician can see lists of the present, open

and absent symptoms for a selected disease.

In interviews with clinicians a positive feedback was obtained: inferring likely diseases shows the usefulness of annotations for clinical decision support. Through the bar chart, the mapping of symptom information to likely diseases gets transparent, i.e. the clinician can easily see, why a certain disease is top ranked. Furthermore, the graphical visualization and especially the possibility to get an overview of open symptoms helps to plan further examinations. By changing the status of symptoms the clinician can make what-if scenarios. For example, for the patient 100000XYZ from above selection of the finding 'weight loss' (a leading symptom for colorectal cancer) as present results in 'B-symptomatic' (a leading symptom for lymphoma) being present as well, since this the combination of fever, night sweats and weight loss. Based on the prototype implementation the clinicians could explain what questions should be addressed next, i.e. which findings should be checked next. A ranked list of open findings or symptoms by examination helps the clinician to plan according next examinations.

It would be of high relevance to include information about time-sequences of symptoms in order to detect their development. This aspect was realized for radiology findings in the prototype implementation of the ReportViewer described in section 9.1. Further, the clinicians see a need in representing the intensity of present symptoms and e.g. the amount of enlarged lymph nodes (1, 2, many). Additionally, the clinicians pointed out that it would be extremely helpful to create a connection between medication and symptoms since some findings or symptoms might represent simply side effects of certain drugs and thus do not represent manifestations of some underlying disease process. Consequently, there should be an option to ignore or exclude them during the diagnosis process.

Part V.

CONCLUSION

12

Conclusion

Today the increase in available clinical data means mainly a *data overload* for clinicians. The potential of data for more efficient and better treatment of patients is not realized mainly due to the following three problems: Firstly, clinical data is stored case-centric and distributed across different systems. It is also not longitudinal integrated. Secondly, only small amount of the data is structured while high percentage of clinically relevant data is unstructured. Thirdly, existing data is not sufficiently aligned to medical knowledge and thus not on the appropriate level of detail for decision support systems. As a result of these problems most of the available data is simply not used in their full strength for better treatment. The contribution of this thesis is three-fold: Firstly, it demonstrates that the created semantic Model for Clinical Information (MCI), which is based on established upper ontologies from the Open Biological and Biomedical Ontologies library, is suitable to capture important clinical data in a well structured way, overcoming the case-centric representation. It is also feasible to use the model to formalize medical knowledge about normal and abnormal clinical findings and their relations to diseases. Secondly, the thesis demonstrates how this medical knowledge can be used to extract structured representations of measurement findings from free text radiology reports. Thirdly, the thesis demonstrates how formalized medical knowledge can be used to enrich clinical data in a way that allows realization of different *views* on the data (disease-centric, longitudinal etc.). Thus, the results of this thesis allow to provide clinicians with data on the appropriate level of detail, needed for more efficient decision making. In the following, the strengthens and limitations of these three contributions are described and afterwards an outlook is given.

12.1. Contribution

The first contribution is the creation of a semantic Model for Clinical Information (MCI). By reusing classes and properties from nine ontologies of the Open Biological and Biomedical Ontologies (OBO) library, the work demonstrates that modular reuse of existing ontologies is feasible in this framework to create a semantic model for *clinical information*. MCI is used in combination with reference terminologies and covers basic patient data, meta data about provided examinations and diagnosis as well as clinical findings and their integration over consecutive examinations. Since clinical findings are central information objects, the model is focused on them and provides a general pattern expressing the anatomical entity, quality, the finding type (static and longitudinal) and also their relation to diseases. This pattern facilitates the representation of medical knowledge that is necessary to enrich finding descriptions automatically. The thesis demonstrates that it is possible to capture this knowledge even in a patient-specific manner. Further, the structured representation of clinical findings with MCI allows to realize different views on the original data that could not be realized before. Since the use cases described in this thesis are mostly from the radiology domain, the limitations of the model are mainly its coverage. However, by reusing established upper ontologies, MCI provides a good basis for further extensions to capture data from other clinical domains at a similar level of detail as demonstrated for the radiology domain.

In the second contribution, the thesis demonstrates that it is feasible to use a knowledge-based approach to extract structured representations of size measurements from free text radiology reports. Existing semantic annotation tools can automatically extract *sets* of named entities from text or images. The problem with simple annotation of unstructured reports and images is that a *set* of annotations does not represent the *content* in a form that allows effective integration into clinical decision processes. An additional step is necessary to get a *structure* for these annotations that fully represent the content of clinical findings. Without this step the relations between entities and thus the value of the annotations remains locked. This work demonstrates that knowledge-based strategies are suitable to resolve measurement entity relation and thus obtain a full semantic representation of clinical findings from radiology reports. Even though the use case presented in this thesis concentrated on structured representations of *size measurements* from radiology reports, the approach for extraction and also classification of measurement findings is applicable to other clinical domains such as pathology as well. For instance, measurements of density, temperature, weight, concentrations etc. could be extracted and classified using a similar normal quality representation as for size qualities. For qualities such as color which are commonly not measured on an absolute scale, this approach however will not work. The limitations of a pure knowledge-based approach are the following: With respect to other relation extraction algorithms the approach is very specific and cannot be used to extract

e.g. relations between medication and dosage or targeted symptoms. Further, the model is defined for a certain set of anatomical entities of the neck, thorax and abdomen region and needs to be extended for others. So far the algorithm does not cover negations, which are however not that often in radiology reports since measurements express what is *present*. More commonly there are cases where the measurement is not explicitly stated but expressed by bounds. For example in the sentence "No lymph nodes bigger than 1 cm". Here, the algorithm would resolve that the lymph node is of size 1 cm. Since 1 cm is exactly the upper bound for lymph nodes to be considered *normal*, the subsequent classification as normal however would be correct! The contribution of the thesis here is also to evaluate the strengthens and limitations of a knowledge-based approach which can serve as a good component in a more comprehensive and general approach for relation extraction.

In the third contribution, the thesis demonstrates how formalized medical knowledge can be used to enrich clinical data in a way that allows realization of different *views* on the data (disease-centric, longitudinal, abnormality focused etc.) – needed by clinicians for more efficient decision making. This is possible since the clinical data and the medical knowledge are integrated in one semantic model. Firstly, the classification of normal and abnormal findings from radiology measurements is demonstrated within this work. The presented algorithm is patient specific and depends on age and gender. Here, extensions to the patient's body mass index, or previous examinations would be interesting contexts which have not been explored by this thesis. In a subsequent step (after classification) the class hierarchy of findings is used to infer implicit information on symptoms and finding. Here the thesis presents an approach that is based on Łukasiewicz's three-value logic applied to the three status types of clinical findings (present, open and absent). The different dimensions of clinical findings are expressed in a well structured form so that the access to patient data gets possible from different perspectives such as anatomical entity, quality, body region or the type of the finding. For example, a longitudinal view of radiology finding data is realized through a prototype implementation of a report viewer which relies on medical knowledge about the human anatomy, the semantic representation of finding descriptions and the meta-data of the original radiology reports.
Regarding the linkage of findings with diseases the thesis has two contributions: providing a disease-centric view on available symptom information and further a subsequent ranking of likely diseases. By matching symptom information to diseases a disease-centric view can be provided. This view got very positive user response by clinicians since it allows to efficiently obtain an overview within the differential diagnosis work flow. The ranking of diseases is new in the sense that it does not try to give a diagnosis for all occurred symptoms and findings, but rather ranks disease which are best matching the symptoms of the patient. That is, the clinician is provided with a better *view* on the data to come to a diagnosis more accurately and efficiently. Even though the ranking showed good results, it was tested on a rather small disease symptom model and thus the strengthens and

limitations of the *ranking algorithm* need further exploration. While the disease-symptom model presented in this thesis was created manually, in recent work a larger disease-symptom cluster graph was created based on existing ontologies [Obe+15a].

In contrast to clinical *guidelines* that define certain rules of how patients should be treated under different conditions, the work presented in this thesis is focussed on the semantic representation of the patient data to provide clinicians with the data they need in *their* decision making processes. Only the classification of findings as normal or abnormal can be considered as very low level guidelines since it defines which findings are paid attention to and consequently treated.

12.2. Outlook

As argued in [Zil+14] the *"healthcare industry faces tremendous productivity challenges"* some of which are potentially addressed by Big Data technologies. Also the authors of [14c] (page 76) argue that the *"availability and use of large data sets is becoming ever more important"* in research and innovation in the context of university medicine. While in other industries, successful application of these technologies has been demonstrated, the healthcare domain still poses several challenges that need to be overcome to get Big Data *ready*. Besides data digitalization, semantic annotation, data privacy and security challenges, the creation of a standardized (semantic) data model for clinical information that allows structured representation and automatic processing of clinical findings is a part of the solution. As argued in [Zil+14] *"the highest impact of Big Data applications is expected when data from various healthcare areas, such as clinical, administrative, financial, or outcome data, can be integrated."* Structured representation and usage of standardized reference vocabularies (coding systems) are the first essential requirement to allow integration of data from different stakeholders and repositories. The developed model does exactly this for clinical data from the radiology domain. In a subsequent step, treatment can be comparatively analyzed and also biomarker can be developed based on large patient cohorts.

The above mentioned problems and requirements hold for data with *clinical context*, i.e. data created and stored at hospitals. Recent innovations such as wearables and tracking applications provide new data sources. Here, people share their personal health data with the vendors of the corresponding devices or apps with the goal to optimize their lifestyle (e.g. walking a certain minimal distance each day). Since data are continuously collected from a large set of people, the obtained data are big and suitable for various machine learning and data mining algorithms. The main difference between data from wearables and clinical data is the level of detail and corresponding complexity of the data. While it is not difficult to measure and store the walking distance and blood pressure, it is more difficult to store clinical findings from pathology, radiology etc. in a structured form.

Another important aspect for application of semantic models is Linked Data, which provides the potential to include many different data resources in clinical decision making. For example, a clinician could be provided with latest clinical studies that are related to the problems of the patient. Here, the usage of established reference terminologies and availability, of ontology mappings is important. Then, Linked Data can be also used to create application specific knowledge models such as the disease symptom graph presented in [Obe+15a].

Part VI.

Annex

Bibliography

Printed Sources

[14b] *Code of Federal Regulations 45 CFR 170.207 Vocabulary standards for representing electronic health information*. 2014.

[14c] *Comission of Experts for Research and Innovation: Research, Innovation and Technological Performance in Germany*. Tech. rep. 2014.

[AA11] André Q Andrade and Maurício B Almeida. "Realist representation of the medical practice: an ontological and epistemological analysis". In: *Proceedings of the 4th Ontobras*. 2011, pp. 73–84.

[AA89] Klaus-Peter Adlassnig and Mohammad Akhavan-Heidari. "CADIAG-2/GALL: An experimental expert system for the diagnosis of gallbladder and biliary tract diseases". In: *Artificial Intelligence in Medicine* 1.2 (1989), pp. 71–77. ISSN: 0933-3657. DOI: 10.1016/0933-3657(89)90018-3.

[AAS12] André Q Andrade, Maurício B Almeida, and Stefan Schulz. "Revisiting ontological foundations of the OpenEHR Entry Model". In: *3rd International Conference on Biomedical Ontology (ICBO 2012)*. KR-MED Series CEUR, 2012, pp. 1–5.

[AF07] Maristella Agosti and Nicola Ferro. "A Formal Model of Annotations of Digital Content". In: *ACM Transactions on Information Systems* 26.1 (2007).

[Ang10] Galia Angelova. "Use of Domain Knowledge in the Automatic Extraction of Structured Representations from Patient-Related Texts". In: *Conceptual Structures: From Information to Intelligence*. Ed. by M. Croitoru, S. Ferré, and D. Lukose. Springer Berlin Heidelberg, 2010, pp. 14–27. ISBN: 978-3-642-14196-6. DOI: 10.1007/978-3-642-14197-3_6.

[Ash+00] Michael Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." In: *Nature genetics* 25.1 (2000), pp. 25–29. arXiv: 10614036.

[Baa+10] Franz Baader et al., eds. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2010. ISBN: 9780521150118.

[Bar+87] G Octo Barnett et al. "DXplain. An evolving diagnostic decision-support system." In: *JAMA The Journal Of The American Medical Association* 258.1 (1987), pp. 67–74.

[Bea+08] Thomas Beale et al. *EHR Information Model, Release 1.0.2*. Tech. rep. 2008.

[Bei+08] Elena Beisswanger et al. “BIOTOP : An Upper Domain Ontology for the Life Sciences”. In: *Applied Ontology* 3.4 (2008), pp. 205–212. DOI: 10.3233/AO-2008-0057.

[Ben10] Tim Benson. *Principles of Health Interoperability HL7 and SNOMED*. Springer London, 2010. ISBN: 9781848828025. DOI: 10.1007/978-1-84882-803-2.

[Ber+08] Valérie Bertaud et al. “Toward a unified representation of findings in clinical radiology.” In: *International journal of medical informatics* 77.9 (Jan. 2008), pp. 621–629. ISSN: 0926-9630.

[BH07] Thomas Beale and Sam Heard. “An Ontology-based Model of Clinical Information.” In: *Studies in health technology and informatics* 129 (Jan. 2007), pp. 760–764. ISSN: 0926-9630.

[BHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. *The Semantic Web*. 2001. DOI: 10.1038/scientificamerican0501-34. arXiv: 1204.6441.

[BLK14] Joseph J Budovec, Cesar A Lam, and Charles E Kahn. “Informatics in radiology: radiology gamuts ontology: differential diagnosis for the semantic web.” In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 34.1 (2014), pp. 254–64.

[BOZ15] Claudia Bretschneider, Heiner Oberkampf, and Sonja Zillner. “UIMA-2LOD: Integrating UIMA Text Annotations into the Linked Open Data Cloud”. In: *International Conference on Knowledge Engineering and Semantic Web*. Moscow, Russia, 2015.

[Bre+14] Claudia Bretschneider et al. “Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation”. In: *Proceedings of the Third Workshop on Semantic Web and Information Extraction*. Association for Computational Linguistics and Dublin City University, 2014, pp. 1–8.

[Bri+10] Ryan R Brinkman et al. “Modeling biomedical experimental processes with OBI.” In: *Journal of biomedical semantics 2010* 1 Suppl 1 (2010), S7.

[Bru13] Michael Brunnbauer. “DICOM metadata as RDF.” In: *GI-Jahrestagung*. 2013, pp. 1796–1804.

[BS03] Thomas Bittner and Barry Smith. *Formal ontologies of space and time*. 2003.

[BS04] Thomas Bittner and Barry Smith. “Normalizing Medical Ontologies using Basic Formal Ontology”. In: *Medical Informatics* (2004), pp. 199–201.

[BSP03] Joachim Baumeister, Dietmar Seipel, and Frank Puppe. “Incremental Development of Diagnostic Set-Covering Models with Therapy Effects”. In: *International Journal of Uncertainty, Fuzzyness and Knowledge-Based Systems* 11.Nov (2003), pp. 25–49.

[Buc+11] Andrew J Buckler et al. "A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging." In: *Radiology* 258.3 (2011), pp. 906–914.

[Buc+13] Andrew J. Buckler et al. "Quantitative imaging biomarker ontology (QIBO) for knowledge representation of biomedical imaging biomarkers". In: *Journal of Digital Imaging* 26.4 (2013), pp. 630–641.

[BZ11] Asma Ben Abacha and Pierre Zweigenbaum. "Automatic extraction of semantic relations between medical entities: a rule based approach." In: *Journal of biomedical semantics* 2 Suppl 5.Suppl 5 (Jan. 2011), S4. ISSN: 2041-1480. DOI: 10.1186/2041-1480-2-S5-S4.

[BZH13] Claudia Bretschneider, Sonja Zillner, and Matthias Hammon. "Identifying Pathological Findings in German Radiology Reports Using a Syntacto-Semantic Parsing Approach". In: *Proceedings of BioNLP 2013*. BioNLP. 2013, pp. 27–35.

[Cam+98] Keith E Campbell et al. "Representing Thoughts, Words, and Things in the UMLS". In: *Journal of the American Medical Informatics Association : JAMIA* 5.5 (1998), pp. 421–431.

[CES07] Werner Ceusters, Peter Elkin, and Barry Smith. "Negative findings in electronic health records and biomedical ontologies: A realist approach". In: *International Journal of Medical Informatics* 76.SUPPL. 3 (2007).

[Cha+10] David S Channin et al. "The caBIG Annotation and Image Markup Project". In: *Journal of Digital Imaging* 23.2 (2010), pp. 217–225. DOI: 10.1007/s10278-009-9193-9.

[Clu00] David A. Clunie. *DICOM Structured Reporting*. 1st Editio. Vol. 24. PixelMed Publishing, 2000, p. 394. ISBN: 0-9701369-0-0.

[Cor09] Ronald Cornet. "Definitions and qualifiers in SNOMED CT". In: *Methods of Information in Medicine* 48.2 (Jan. 2009), pp. 178–83. ISSN: 0026-1270. DOI: me10.3414/ME9215.

[Cor10] Oracle Corporation. "Oracle ® Healthcare Data Model". In: 2.November (2010).

[Cou+09] Mélanie Courtot et al. "MIREOT: the Minimum Information to Reference an External Ontology Term". In: *Nature Precedings*. 2009.

[Cue+07] Bernardo Cuenca-Grau et al. "Just the Right Amount: Extracting Modules from Ontologies." In: *Proceedings of the Sixteenth International World Wide Web Conference (WWW2007)*. ACM, 2007, pp. 717–726. ISBN: 9781595936547. DOI: 10.1145/1242572.1242669.

[Cun06] Hamish Cunningham. "Information Extraction, Automatic". In: *Encyclopedia of Language & Linguistics*. Vol. 5. 2006, pp. 665–677. ISBN: 9780080448541.

[DSM02] Robert H Dolin, Kent Spackman, and David Markwell. "Selective retrieval of pre- and post-coordinated SNOMED concepts." In: *AMIA 2002 Annual Symposium Proceedings* (Jan. 2002), pp. 210–214. ISSN: 1531-605X.

[DuC11] Bob DuCharme. *Learning SPARQL*. O'Reilly Media, Inc., 2011. ISBN: 1449306594, 9781449306595.

[Eis+09] Elisabeth A. Eisenhauer et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)". In: *European Journal of Cancer* 45.2 (2009), pp. 228–247.

[Fri97] Carol Friedman. "Towards a comprehensive medical language processing system: methods and issues." In: *Proceedings of the AMIA Annual Fall symposium*. American Medical Informatics Association, 1997, p. 595.

[Gar+07] Sebastian Garde et al. "Towards semantic interoperability for electronic health records: Domain Knowledge Governance for openEHR Archetypes". In: *Methods of information in medicine* 46.3 (2007), pp. 332–43. ISSN: 0026-1270. DOI: 10.1160/ME5001.

[GPS11] Rafael S Gonc, Bijan Parsia, and Uli Sattler. "Analysing the Evolution of the NCI Thesaurus". In: *Proc. of CBMS*. 2011.

[Gru93] Thomas R Gruber. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing". In: *International Journal Human-Computer Studies 43* (1993), pp. 907–928.

[GSG04] Pierre Grenon, Barry Smith, and Louis Goldberg. "Biodynamic ontology: Applying BFO in the biomedical domain". In: *Studies in Health Technology and Informatics*. Vol. 102. 2004, pp. 20–38.

[GSH12] Georgios Gkoutos, Paul N Schofield, and Robert Hoehndorf. "The Units Ontology: a tool for integrating units of measurement in science." In: *Database : the journal of biological databases and curation* 2012 (2012), bas033. ISSN: 1758-0463. DOI: 10.1093/database/bas033.

[Har+07] Frank van Harmelen et al. *Handbook of Knowledge Representation*. San Diego, USA: Elsevier Science, 2007. ISBN: 0444522115, 9780444522115.

[Her10] Heinrich Herre. "General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling". In: *Theory and Applications of Ontology: Computer Applications*. Ed. by Roberto Poli, Michael Healy, and Achilles Kameas. Springer Netherlands, 2010, pp. 297–345. ISBN: 978-90-481-8846-8. DOI: 10.1007/978-90-481-8847-5_14.

[Her11] Gerd Herold. *Herold: Internal Medicine*. Gerd Herold, Köln, 2011.

[HHN89] David E Heckerman, Eric J Horvitz, and Bharat N Nathwani. "Update on the Pathfinder Project". In: *Proc Annu Symp Comput Appl Med Care.* 1989, pp. 203–207.

[Hor+06] Matthew Horridge et al. “The manchester OWL syntax”. In: *CEUR Workshop Proceedings*. Vol. 216. 2006.

[Hus+04] Rada Hussein et al. “DICOM Structured Reporting: Part 1. Overview and Characteristics”. In: *Radiographics* (2004), pp. 891–896.

[JSM09] Clement Jonquet, Nigam H Shah, and Mark A Musen. “The Open Biomedical Annotator”. In: *Summit on translational bioinformatics* (2009), pp. 56–60.

[KBS11] Daniel Karlsson, Martin Berzell, and Stefan Schulz. “Information Models and Ontologies for Representing the Electronic Health Record”. In: *ICBO2011*. Vol. 6. Buffalo, NY, USA, 2011, pp. 6–10.

[KMK11] Markus R Krötzsch, Frederick Maier, and Adila A Krisnadhi. “A Better Uncle For OWL: Nominal Schemas for Integrating Rules and Ontologies”. In: *Proceedings of the 20th International Conference on World Wide Web (WWW2011)*. ACM, 2011, pp. 645–654.

[Köh+09] Sebastian Köhler et al. “Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies”. In: *American Journal of Human Genetics* 85.4 (2009), pp. 457–464.

[KW82] Casimir A Kulikowski and Sholom M Weiss. “Representation of expert knowledge for consultation: The CASNET and EXPERT projects”. In: *Artificial Intelligence in Medicine*. Ed. by Peter Szolovits. Westview Press, 1982, pp. 21–55.

[Lan+13] Melanie Langermeier et al. “Management of Variability in Modular Ontology Development”. In: *9th International Workshop on Semantic Web Enabled Software Engineering*. 2013.

[Lan+14] Melanie Langermeier et al. “Change and Version Management in Variability Models for Modular Ontologies”. In: *16th International Conference on Enterprise Information Systems (ICEIS)*. 2014.

[Luc+11] Joanne S Luciano et al. “The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside.” In: *Journal of Biomedical Semantics 2011* 2 Suppl 2.S1 (Jan. 2011). ISSN: 2041-1480. DOI: 10.1186/2041-1480-2-S2-S1.

[Luc98] Peter Lucas. “Analysis of notions of diagnosis”. In: *Artificial Intelligence* 105(1-2) (1998), pp. 293–341.

[Mai+11] Carmen De Maio et al. “Fuzzy Knowledge Approach to Automatic Disease Diagnosis”. In: *IEEE International Conference On Fuzzy Systems* (2011), pp. 2088–2095.

[Mal07] Grzegorz Malinowski. “Many-valued logic and its philosophy”. In: *Handbook of the History of Logic* 8 (2007), pp. 13–94.

[Mas+06] Viviana Mascardi et al. *A Comparison of Upper Ontologies*. Tech. rep. 2006.

[Mey+08] Stéphane M Meystre et al. "Extracting information from textual documents in the electronic health record: a review of recent research." In: *Yearbook of medical informatics* (Jan. 2008), pp. 128–44. ISSN: 0943-4747.

[MPM82] Randolph A Miller, Harry E Pople, and Jack D Myers. "Internist-1: An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine". In: *New England Journal of Medicine* 307.8 (1982), pp. 468–476. ISSN: 00284793. DOI: 10.1056/NEJM198208193070803.

[MR98] Thorsten B Möller and Emil Reif. *CT- und MRT-Normalbefunde*. Thieme, 1998.

[MRS09] Manuel Möller, Sven Regel, and Michael Sintek. "RadSem: Semantic Annotation and Retrieval for Medical Images". In: *ESWC 2009, Heraklion Proceedings of the 6th European Semantic Web Conference*. Ed. by Lora Aroyo et al. Vol. 5554. Lecture Notes in Computer Science 01. Springer Berlin Heidelberg, 2009, pp. 21–35. ISBN: 978-3-642-02120-6. DOI: 10.1007/978-3-642-02121-3_6.

[MSC08] David Markwell, Laura Sato, and Edward Cheetham. "Representing clinical information using SNOMED Clinical Terms with different structural information models". In: *Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)*. Ed. by K Spackman and Ronald Cornet. 2008, pp. 72–79.

[Mue+12] Andreas Mueller et al. "Knowledge Engineering Requirements for Generic Diagnostic Systems". In: *KEOD 2012 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain*. Ed. by Filipe Joaquim and Jan L. G. Dietz. Barcelona, Spain: SciTePress, 2012, pp. 184–189.

[Mun+07] Chris Mungall et al. "Representing Phenotypes in OWL". In: *OWLED 2007*. 2007.

[Mun+10] Christopher Mungall et al. "Integrating phenotype ontologies across multiple species." In: *Genome biology* 11.1 (Jan. 2010), R2. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-1-r2.

[Mun14] Christopher J Mungall. *A critique of temporalized relations*. Tech. rep. Berkeley: Genomics Division, Lawrence Berkeley National Laboratory, 2014, pp. 1–10.

[NLM09] NLM. *UMLS Reference Manual*. Bethesda (MD): National Library of Medicine (US), 2009.

[Obe+12a] Heiner Oberkampf et al. "Interpreting Patient Data using Medical Background Knowledge". In: *Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series*. Ed. by Ronald Cornet and Robert Stevens. Graz, Austria: CEUR-WS.org, 2012, pp. 1–5.

[Obe+12b] Heiner Oberkampf et al. “Towards a Ranking of Likely Diseases in Terms of Precision and Recall”. In: *Proceedings of the 1st International Workshop on Artificial Intelligence and NetMedicine. In conjunction with ECAI 2012.* Montpellier, France, 2012, pp. 11–20.

[Obe+13] Heiner Oberkampf et al. “An OGMS-based Model for Clinical Information (MCI)”. In: *Proceedings of International Conference on Biomedical Ontology 2013*. 2013, pp. 97–100.

[Obe+14] Heiner Oberkampf et al. “Knowledge-based Extraction of Measurement-Entity Relations from German Radiology Reports”. In: *IEEE International Conference on Healtcare Informatics*. Sept. 2014, pp. 149–154. DOI: 10.1109/ICHI.2014.27.

[Obe+15a] Heiner Oberkampf et al. “From Symptoms to Diseases - Creating the Missing Link”. In: *ESWC 2015*. Springer, 2015.

[Obe+15b] Heiner Oberkampf et al. “Semantic Representation of Radiology Findings for Better Decision Making”. In: *(submitted to) BMC Medical Informatics and Decision Making* (2015).

[Ogb11] Chimezie Ogbuji. “A Framework Ontology for Computer-Based Patient Record Systems”. In: *Proceedings of the ICBO: International Conference on Biomedical Ontology*. 2011, pp. 217–223.

[OPS10] Jasmin Opitz, Bijan Parsia, and Ulrike Sattler. “Evaluating Modelling Approaches for Medical Image Annotations”. In: *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences*. CEUR Workshop Proceedings. Berlin, Germany: CEUR-WS.org, 2010.

[Ove13] James A Overton. “OBI Data Prototype”. 2013.

[QKR07] Rahil Qamar, Jay Kola, and Alan Rector. “Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies.” In: *AMIA 2007 Symposium Proceedings* (Jan. 2007), pp. 608–613. ISSN: 1942-597X.

[QR07] Rahil Qamar and Alan Rector. “Semantic Mapping of Clinical Model Data to Biomedical Terminologies to Facilitate Data Interoperability”. In: *Healthcare Computing 2007 Conference* (2007), p. 257.

[Rei87] Raymond Reiter. “A Theory of Diagnosis from First Principles”. In: *Artificial Intelligence* 32.1987 (1987), pp. 57–95.

[RGH08] Angus Roberts, Robert Gaizauskas, and Mark Hepple. “Extracting Clinical Relationships from Patient Narratives”. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. BioNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 10–18. ISBN: 978-1-932432-11-4.

[RHR11] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. “Automatic extraction of relations between medical concepts in clinical texts.” In: *Journal of the American Medical Informatics Association : JAMIA* 18.5 (2011), pp. 594–600. ISSN: 1067-5027.

[RJ07] Cornelius Rosse and L V Mejino Jr. “The Foundational Model of Anatomy Ontology”. In: *Esophagus* 2007 (2007), pp. 59–117. DOI: 10.1007/978-1-84628-885-2_4.

[RNW83] James A Reggia, Dana Nau, and Pearl Wang. “Diagnostic expert systems based on a set covering model”. In: *International Journal of ManMachine Studies* 19.5 (1983), pp. 437–460. ISSN: 00207373. DOI: 10.1016/S0020-7373(83)80065-0.

[Rob+08] Peter N. Robinson et al. “The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease”. In: *American Journal of Human Genetics* 83.5 (2008), pp. 610–615.

[RQM06] Alan Rector, Rahil Qamar, and T Marley. “Binding Ontologies & Coding systems to Electronic Health Records and Messages”. In: *KR-MED 2006*. 2006, pp. 11–19.

[RSD08] Alan Rector, Robert Stevens, and Nick Drummond. “What Causes Pneumonia? The Case for a Standard Semantics for "may" in OWL”. In: *OWLED* (2008).

[Rub+08] Daniel L Rubin et al. “Medical Imaging on the Semantic Web : Annotation and Image Markup”. In: *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*. AAAI, 2008, pp. 93–98.

[Sav+10] Guergana K Savova et al. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.” In: *Journal of the American Medical Informatics Association : JAMIA* 17.5 (Sept. 2010), pp. 507–513. ISSN: 1527-974X. DOI: 10.1136/jamia.2009.001560.

[SC06] Barry Smith and Werner Ceusters. “HL7 RIM: an incoherent standard.” In: *Studies in health technology and informatics* 124 (2006), pp. 133–138.

[Sch+07] Daniel Schober et al. “Towards naming conventions for use in controlled vocabulary and ontology engineering”. In: *In ISMBECCB Special Interest Group SIG Meeting Program Materials BioOntologies*. Ed. by Robert Stevens et al. Vol. SIG Worksh. 2007, p. 4.

[Sch+09] Daniel Schober et al. “Survey-based naming conventions for use in OBO Foundry ontology development.” In: *BMC Bioinformatics* 10 (Jan. 2009). ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-125.

[Sch+10] Stefan Schulz et al. “Bridging the semantics gap between terminologies, ontologies, and information models.” In: *Studies in health technology and informatics* 160.Pt 2 (Jan. 2010), pp. 1000–4. ISSN: 0926-9630.

[Sch+12] Lynn Marie Schriml et al. “Disease Ontology: a backbone for disease semantic integration.” In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D940–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkr972.

[SCS09] Richard H Scheuermann, Werner Ceusters, and Barry Smith. "Toward an Ontological Treatment of Disease and Diagnosis". In: *2009 AMIA Summit on Translational Bioinformatics*. 2009.

[SCV13] Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. "Designing the W3C open annotation data model". In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13* (2013), pp. 366–375. DOI: 10.1145/2464464.2464474.

[Sei+09] Sascha Seifert et al. "Hierarchical Parsing and Semantic Navigation of Full Body CT Data". In: *Proceedings of SPIE Medical Imaging 2009: Image Processing 725902* (2009). DOI: 10.1117/12.812214.

[Sei+10] Sascha Seifert et al. "Semantic Annotation of Medical Images". In: *Proceedings of SPIE Medical Imaging 2010*. San Diego, CA, United States, 2010. DOI: 10.1117/12.844207.

[Sei+11] Sascha Seifert et al. "Combined Semantic and Similarity Search in Medical Image Databases". In: *SPIE Medical Imaging 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications (SPIE), Lake Buena Vista, FL, USA*. Ed. by William W. Boonn and Brent J. Liu. Mar. 2011. DOI: 10.1117/12.878179.

[Sho76] Edward H Shortliffe. *Computer-Based Medical Consultations: MYCIN*. Ed. by Elsevier North Holland. Vol. 85. 6. American Elsevier, 1976, pp. 831–. DOI: 10.1059/0003-4819-85-6-831_1.

[Smi+07] Barry Smith et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." In: *Nature biotechnology* 25.11 (Nov. 2007), pp. 1251–5. ISSN: 1087-0156. DOI: 10.1038/nbt1346.

[SRU13] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." In: *Journal of the American Medical Informatics Association : JAMIA* 20.5 (2013), pp. 806–13. ISSN: 1527-974X. DOI: 10.1136/amiajnl-2013-001628.

[SS10] Stefan Schulz and Barry Smith. "Strengths and limitations of formal ontologies in the biomedical domain". In: 3.1 (2010), pp. 31–45. DOI: 10.3395/reciis.v3i1.241en.Strengths.

[ST10] Kyle Strimbu and Jorge A Tavel. "What are Biomarkers?" In: *Current opinion HIV and AIDS* 5.6 (2010), pp. 463–466. DOI: 10.1097/COH.0b013e32833ed177.

[Ter13] Ken Terry. *Analytics : The Nervous System of IT-Enabled Healthcare*. Tech. rep. Institute for Health Technology Transformation, 2013.

[Was+09] Nicole Washington et al. "Linking human diseases to animal models using ontology-based phenotype annotation." In: *PLoS biology* 7.11 (Dec. 2009), e1000247. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000247.

[Wel07] Katrin Weller. “Folksonomies and Ontologies: Two New Players in Indexing and Knowledge Representation”. In: *Applying Web 2.0: Innovation, Impact and Implementation. Proceedings of the Online Information Conference, London, Great Britain*. Ed. by Helen Jezzard. Vol. 2. Learned Information Europe Ltd, 2007, pp. 108–115. DOI: 10.1128/AEM.67.12.5833.

[Wen+08] Pinar Wennerberg et al. “KEMM: Knowledge Engineering Methodology in the Medical Domain”. In: *In Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS)* June (2008).

[Whe+11] Patricia L Whetzel et al. “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.” In: *Nucleic acids research* 39.Web Server issue (2011), W541–W545.

[Xia+10] Zuoshuang Xiang et al. “OntoFox: web-based support for ontology reuse.” In: *BMC research notes* 3 (2010), p. 175.

[Xia+11] Zuoshuang Xiang et al. “Ontobee: A Linked Data Server and Browser for Ontology Terms”. In: *ICBO*. 2011.

[Zil+14] Sonja Zillner et al. “Towards a Roadmap for Big Data Applications in the Healthcare Domain”. In: *Proceedings of IEEE IRI HI 2014*. 2014.

Online Sources

[03] *Protege-dc: The Dublin Core Element Set v1.1 namespace providing access to its content by means of an OWL DL Ontology*. 2003. URL: http://protege.stanford.edu/plugins/owl/dc/dublincore.owl.

[06a] *Artificial Intelligence and the Semantic Web: AAAI2006 Keynote*. 2006. URL: http://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html (visited on 02/16/2015).

[06b] *Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)*. 2006. URL: http://www.loa-cnr.it/ontologies/DLP_397.owl (visited on 11/03/2014).

[06c] *Time Ontology in OWL - W3C Working Draft*. 2006. URL: http://www.w3.org/TR/owl-time/ (visited on 06/06/2015).

[10a] *High magnification micrograph of an anaplastic astrocytoma. Ki-67 immunostain.* 2010. URL: https://commons.wikimedia.org/wiki/File:Anaplastic_astrocytoma_-_ki67_-_high_mag.jpg (visited on 05/15/2016).

[10b] *Suggested Upper Merged Ontology*. 2010. URL: http://www.adampease.org/OP/SUMO.owl (visited on 06/27/2015).

[11a] *Computer-Based Patient Record Ontology*. 2011. URL: https://code.google.com/p/cpr-ontology/ (visited on 06/06/2015).

[11b] *Quantitative Imaging Biomarker Ontology*. 2011. URL: http://purl.bioontology.org/ontology/QIBO (visited on 06/06/2015).

[11c] *The IAO ontology metadata ontology*. 2011. URL: http://purl.obolibrary.org/obo/iao/ontology-metadata.owl (visited on 09/02/2014).

[11d] *The Information Artifact Ontology (IAO)*. 2011. URL: http://purl.obolibrary.org/obo/iao.owl (visited on 07/31/2014).

[12a] *BFO 2.0 Tutorial at International Conference on Biomedical Ontology*. 2012. URL: http://ontology.buffalo.edu/BFO/Graz-Tutorial-2012/BFO-Smith.ppt (visited on 10/21/2014).

[12b] *BioTop - A Top-Domain Ontology for the Life Sciences*. 2012. URL: http://purl.org/biotop (visited on 06/06/2015).

[12c] *The Basic Formal Ontology (BFO 2.0) Graz*. 2012. URL: http://purl.obolibrary.org/obo/bfo.owl (visited on 09/02/2014).

[12d] *The Basic Formal Ontology (BFO 2.0) Official Classes-only OWL file*. 2012. URL: http://purl.obolibrary.org/obo/bfo/classes-only.owl (visited on 10/21/2014).

[12e] *The Translational Medicine Ontology*. 2012. URL: http://www.w3.org/2001/sw/hcls/ns/transmed/tmo (visited on 06/06/2015).

[13a] *Open Annotation Data Model (OA)*. 2013. URL: http://www.w3.org/ns/oa# (visited on 11/12/2014).

[13b] *The Ontology of Medically Related Social Entities (OMRSE)*. 2013. URL: http://purl.obolibrary.org/obo/omrse.owl (visited on 07/31/2014).

[14a] *Apache UIMA project*. 2014. URL: http://uima.apache.org/ (visited on 11/20/2014).

[14d] *Das Informationssystem der Gesundheitsberichterstattung des Bundes (The Information System of the Federal Health Monitoring)*. 2014. URL: http://www.gbe-bund.de (visited on 10/03/2014).

[14e] *DCMI Metadata Terms*. 2014. URL: http://purl.org/dc/terms/ (visited on 07/31/2014).

[14f] *Gene Ontology Consortium*. 2014. URL: http://www.geneontology.org/ (visited on 07/31/2014).

[14g] *International Classification of Diseases*. 2014. URL: http://www.who.int/classifications/icd/en/ (visited on 07/28/2015).

[14h] *NCBO BioPortal*. 2014. URL: http://bioportal.bioontology.org/ (visited on 07/28/2015).

[14i] *NLP Interchange Format (NIF) 2.0 Core Ontology*. 2014. URL: http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.owl (visited on 10/31/2014).

[14j] *OntoFox*. 2014. URL: http://ontofox.hegroup.org/ (visited on 07/28/2015).

[14k] *Ontology Design Patterns*. 2014. URL: http://ontologydesignpatterns.org (visited on 07/31/2014).

[14l] *Ontology for Biomedical Investigations (OBI)*. 2014. URL: http://purl.obofoundry.org/obo/obi.owl (visited on 07/31/2014).

[14m] *OPS - German Procedure Classification*. 2014. URL: https://www.dimdi.de/static/en/klassi/ops/index.htm (visited on 02/16/2015).

[14n] *Phenotypic quality (PATO)*. 2014. URL: http://purl.obolibrary.org/obo/pato.owl (visited on 07/28/2015).

[14o] *Radiopedia, the wiki-based collaborative Radiology resource*. 2014. URL: http://radiopaedia.org/ (visited on 07/16/2014).

[14p] *RadLex version 3.11*. 2014. URL: http://radlex.org/ (visited on 07/28/2015).

[14q] *The Biological Spatial Ontology (BSPO)*. 2014. URL: http://purl.obolibrary.org/obo/bspo.owl (visited on 07/31/2014).

[14r] *The Ontology for Biomedical Investigations Consortium*. 2014. URL: http://purl.obolibrary.org/obo/obi (visited on 07/31/2014).

[14s] *The Ontology for General Medical Science (OGMS)*. 2014. URL: http://purl.obolibrary.org/obo/ogms.owl (visited on 07/31/2014).

[14t] *The Open Biological and Biomedical Ontologies Foundry*. 2014. URL: http://www.obofoundry.org/ (visited on 07/31/2014).

[14u] *Unified Medical Language System*. 2014. URL: http://www.nlm.nih.gov/research/umls/ (visited on 07/28/2015).

[14v] *Units of Measurement Ontology (UO)*. 2014. URL: http://purl.obolibrary.org/obo/uo.owl (visited on 10/21/2014).

[15a] *Apache Jena*. 2015. URL: https://jena.apache.org/ (visited on 09/12/2015).

[15b] *Clinical Data Interchange Standards Consortium*. 2015. URL: http://www.cdisc.org/ (visited on 06/06/2015).

[15c] *DICOM Homepage*. 2015. URL: http://dicom.nema.org/ (visited on 05/17/2015).

[15d] *Gamuts*. 2015. URL: http://www.gamuts.net/ (visited on 06/06/2015).

[15e] *Health Level Seven FHIR*. 2015. URL: http://hl7.org/implement/standards/fhir/ (visited on 05/10/2015).

[15f] *Health Level Seven International*. 2015. URL: http://www.hl7.org (visited on 05/17/2015).

[15g] *Health Level Seven Reference Implementation Model, Version 3*. 2015. URL: http://purl.bioontology.org/ontology/HL7/ (visited on 04/11/2015).

[15h] *i2b2: Informatics for Integrating Biology and the Bedside*. 2015. URL: https://www.i2b2.org/ (visited on 07/28/2015).

[15i] *IBM Unified Data Model for Healthcare*. 2015. URL: http://www-03.ibm.com/software/products/en/ibm-unified-data-model-for-healthcare (visited on 05/18/2015).

[15j] *IHE - Integrating the Healthcare Enterprise*. 2015. URL: http://www.ihe.net/ (visited on 05/18/2015).

[15k] *Klinische Datenintelligenz*. 2015. URL: http://www.klinische-datenintelligenz.de/ (visited on 03/31/2015).

[15l] *Online Mendelian Inheritance in Man*. 2015. URL: http://omim.org/ (visited on 02/17/2015).

[15m] *OpenEHR Website*. 2015. URL: http://www.openehr.org (visited on 07/28/2015).

[15n] *Quantitative Imaging Biomarkers Alliance*. 2015. URL: https://www.rsna.org/QIBA.aspx (visited on 06/06/2015).

[15o] *Relations Ontology*. 2015. URL: http://purl.obolibrary.org/obo/ro/core.owl (visited on 03/09/2015).

[15p] *Teradata Healthcare Data Model*. 2015. URL: http://www.teradata.de/logical-data-models/healthcare/ (visited on 05/18/2015).

[CS14] G. Carothers and A. Seaborne, eds. *RDF 1.1 N-Triples*. 25 February 2014. URL: http://www.w3.org/TR/n-triples/ (visited on 02/16/2015).

[CWL14] R. Cyganiak, D. Wood, and M. Lanthaler, eds. *RDF 1.1 Concepts and Abstract Syntax*. 25 February 2014. URL: http://www.w3.org/TR/rdf11-concepts/ (visited on 02/16/2015).

[GS14] Fabien Gandon and Guus Schreiber, eds. *RDF 1.1 XML Syntax*. Feb. 2014. URL: http://www.w3.org/TR/rdf-syntax-grammar/ (visited on 02/16/2015).

[Hor+04] Ian Horrocks et al. *SWRL: A Semantic Web Rule Language*. 21 May 2004. URL: http://www.w3.org/Submission/SWRL/ (visited on 02/16/2015).

[HP12] Matthew Horridge and Patel-Schneider, eds. *OWL 2 Web Ontology Language Manchester Syntax (Second Edition)*. 11 December 2012. URL: http://www.w3.org/TR/owl2-manchester-syntax/ (visited on 10/21/2014).

[HS] Steve Harris and Andy Seaborne, eds. *SPARQL 1.1 Query Language*. URL: http://www.w3.org/TR/sparql11-query/.

[Int14] International Health Terminology Standards Development Organisation. *SNOMED CT*. 2014. URL: http://www.ihtsdo.org/snomed-ct/ (visited on 08/09/2014).

[Mot+12] Boris Motik et al., eds. *OWL 2 Web Ontology Language Profiles (Second Edition)*. 11 December 2012. URL: http://www.w3.org/TR/owl2-profiles/ (visited on 07/31/2014).

[MPP12] Boris Motik, Peter F. Patel-Schneider, and Bijan Parsia, eds. *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)*. 11 December 2012. URL: http://www.w3.org/TR/owl2-syntax/ (visited on 07/31/2014).

[PAR14] PAREXEL International Coporation. *RECIST 1.1 - PAREXEL*. 2014. URL: http://www.recist.com/ (visited on 10/28/2014).

[PC14] Eric Prud'hommeaux and Gavin Carothers, eds. *RDF 1.1 Turtle*. Feb. 2014. URL: http://www.w3.org/TR/turtle/ (visited on 02/16/2015).

[Reg14] Regenstrief Institute. *Logical Observation Identifiers Names and Codes (LOINC)*. 2014. URL: http://loinc.org/ (visited on 07/16/2014).

[Spo14] Sporny, M. and Kellogg, G. and Lanthaler, M., ed. *JSON-LD 1.0*. Jan. 2014. URL: http://www.w3.org/TR/turtle/ (visited on 02/16/2015).

Glossary

application ontology An application ontology is created to fulfil the need of a certain application. Its reuse for other applications is mostly not possible without adaptations. 26

classification system In the biomedical domain a classification system defines a set of classes where the classes are pairwise disjoint. Often classification systems are additionally hierarchically structured. In this case sibling classes are pairwise disjoint. 25, 26

coding system In the biomedical domain a coding system is a set of codes represented by letters and numbers. The role of coding systems is to define a common reference for a domain. 25

disease A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism [SCS09]. 67

disorder A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease [SCS09]. 67

domain ontology A domain ontology (also domain-specific ontology) is an ontology which covers the concepts used with a certain subspeciality. For instance RadLex covers the domain of radiology or the Foundational Model of Anatomy (FMA) covers the domain of anatomy. 26

Don't Repeat Yourself (DRY) A design pattern commonly used in software engineering that emphasized to avoid duplication of code. In the context of semantic modelling this design emphasizes to avoid duplication of (parallel) class or property hierarchies that are distinguished only by some single attribute. 27, 77

HL7 Reference Information Model The HL7 RIM version 3 is an information model that is designed to be used in combination with external health terminologies like SNOMED CT to express clinical data. 7, 44–46

Integrating the Healthcare Enterprise IHE is an initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information [15j]. 51

MIREOT Minimal Information to Reference an External Ontology Term (MIREOT) is a set of guidelines to support reuse of parts of other ontologies [Cou+09]. 57, 73

ontology An explicit specification of a conceptualization [Gru93]. 8, 25

OpenEHR The OpenEHR is a set of specifications that define an open standard for electronic health record information. 7

Protégé-dc A commonly used subset of the Dublin Core Metadata Initiative (DCMI) Meta data Terms used to specify meta data of ontology resources such as classes or properties. 71

RECIST Response Evaluation Criteria In Solid Tumors (RECIST) is a set of rules used to determine *"when cancer patients improve ("response"), stay the same ("stable") or worsen ("progression") during treatments"* [PAR14]. 104, 106, 107, 115, 238

reference ontology A reference ontology is an ontology whose main purpose is to serve as a standard reference system which provides a shared vocabulary. 26

taxonomy A hierarchical generalization (is-a or subclass) structure of concepts. 25, 26

terminology A set of (controlled) terms. 25, 26

thesaurus A vocabulary that groups terms by semantic similarity (e.g. synonyms). 25, 26

UMLS Metathesaurus The UMLS Metathesaurus aligns the concepts of different UMLS terminologies and coding systems by assigning concept unique identifiers to them. 33

UMLS Semantic Network The UMLS Semantic Network is a model containing 133 different high level semantic types which are interlinked by semantic relationships. It is used to label the concepts of the UMLS Metathesaurus. 33

upper ontology An upper ontology (or upper-level ontology; some times also referred to as foundational ontology) describes the basic classes and relations of an ontology without being specific to any particular domain. Thus, it

serves as the backbone for different domain ontologies. Examples are the BFO, DOLCE and SUMO. 26

vocabulary A set of terms. 25

Acronyms

ATC Anatomical Therapeutic Chemical classification system. 32, 38, 177

BFO Basic Formal Ontology. 35, 62, 231

BFO-1.1 Basic Formal Ontology version 1.1. 32, 35, 36, 70

BFO-2 Basic Formal Ontology version 2. 32, 36, 60–62, 66, 67, 70, 237

BI Business Intelligence. 51

BSPO Biological Spatial Ontology. 32, 251

BT BioTop Ontology. 32, 59

CARO Common Anatomy Reference Ontology. 32

CDISC Clinical Data Interchanges Standards Consortium. 51

ChEBI Chemical Entities of Biological Interest. 71

CIR Clinical Investigation Record ontology. 46

CL Cell Ontology. 71

CMO Clinical Measurement Ontology. 32

CPR Computer-Based Patient Record ontology. 41, 49

CT computed tomography. 173, 182, 246, 248

CUI Concept Unique Identifier. 33

DCMI Dublin Core Metadata Initiative. 230

DICOM Digital Imaging and Communications in Medicine. 48, 90

DICOM SR DICOM Structured Reporting. 48

DL Description Logic. 23, 24, 27, 144

DOID Human Disease Ontology. 25, 26, 32, 37, 119, 120, 239

DOLCE Descriptive Ontology for Linguistic and Cognitive Engineering. 59, 231

DSE Disease-Symptom-Examination ontology. 119–124, 143, 145, 155–157, 162, 164, 239

DWH data warehouse. 51

EHR electronic health record. 46

FHIR Fast Healthcare Interoperability Resources. 46

FMA Foundational Model of Anatomy. iv, 4, 6, 26, 32, 37, 38, 40, 49, 62, 69, 81, 82, 113, 117, 132, 144, 229, 237, 251

GFO General Formal Ontology. 59

GO Gene Ontology. 6, 37, 62, 71

HP Human Phenotype Ontology. 32, 37, 54, 62, 80, 81, 121, 145

IAO Information Artefact Ontology. iv, 32, 36, 42, 49, 60, 65–67, 77, 225

ICD International Classification of Diseases. 3, 6, 25, 32, 34, 37, 80–82, 85, 120, 144, 237

IDO Infectious Disease Ontology. 26

IE Information Extraction. 8, 40, 52, 133

IRI Internationalized Resource Identifier. 28, 250

LOINC Logical Observation Identifiers Names and Codes. 6, 25, 32, 34, 178

MANO MEDICO Annotation Ontology. 132, 133, 136–138, 239

MCI Model for Clinical Information. 13, 14, 19, 35, 41, 44, 45, 47, 48, 50–52, 57–60, 62, 63, 66, 67, 71, 72, 75, 76, 81–84, 91, 111, 112, 118, 119, 121, 123, 132, 134, 136, 137, 139, 140, 142, 144, 154, 155, 167, 173, 205, 206, 237–239, 241, 244, 251

MESH Medical Subject Headings. 32, 120

MRI magnetic resonance imaging. 173, 182, 246

NCBI National Center for Biotechnology Information. 70, 71

NCIT National Cancer Institute Thesaurus. 25, 32, 34, 37, 120

NER Named Entity Recognition. 8, 40, 52, 133

NIF NLP Interchange Format. 71

NLP Natural Language Processing. 8, 13, 40, 53, 71, 130, 132, 133, 235

OA Open Annotation data model. 71, 109, 132, 136, 137, 237

OBI Ontology for Biomedical Investigations. iv, 32, 36, 37, 60, 62, 65, 67, 68, 250

OBO Open Biological and Biomedical Ontologies. iii, 13, 16, 19, 23, 32–36, 48, 57, 59–62, 67, 71, 72, 94, 119, 205, 206, 238, 250

OGMS Ontology for General Medical Science. iv, 32, 36, 49, 60, 66, 67, 85, 98, 250

OMIM Online Mendelian Inheritance in Man. 32, 34

OMRSE Ontology of Medically Related Social Entities. iv, 32, 70, 84

OPS German procedure classification (Operationen- und Prozedurenschlüssel). 32, 38, 82

OWL Web Ontology Language. 12, 30, 37, 44, 62, 65, 73, 144

PATO Phenotypic Quality Ontology. iii, 32, 37, 48, 60, 62–65, 77, 78, 91, 115, 121, 237

QIBA Quantitative Imaging Biomarkers Alliance. 50

QIBO Quantitative Imaging Biomarker Ontology. iii, 41, 50

RadLex Radiology Lexicon. 6, 16, 17, 25, 26, 32, 38, 53, 54, 81, 82, 113, 117, 132, 144, 148, 173, 186, 189, 190, 229, 237

RDB relational data base. 138

RDF Resource Description Framework. 12, 29, 44, 47, 134, 136

RDFS Resource Description Framework Schema. 29

RO Relation Ontology. iii, 32, 35, 36, 49, 62

ROI region of interest. 95–97, 238

RSNA Radiological Society of North America. 50

SNOMED CT Systematized Nomenclature of Medicine – Clinical Terms. 6, 7, 32–34, 37, 42, 46, 49, 120

SPARQL SPARQL Protocol and RDF Query Language. 12, 31

SUMO Suggested Upper Merged Ontology. 59, 231

SWRL Semantic Web Rule Language. 31

SYMP Symptom Ontology. 38, 120, 121

TMO Translational Medicine Ontology. 32, 41, 49, 50

UBERON Uber-Anatomy Ontology. 62

UIMA Unstructured Information Management Architecture. 52, 134

UMLS Unified Medical Language System. iii, 23, 32, 33, 40

UO Units Ontology. iii, 32, 37, 60, 64, 65, 71, 115, 138

URI Uniform Resource Identifier. 28

URL Uniform Resource Locator. 28

US ultrasound. 173, 182, 246

VO Vaccine Ontology. 71

XML Extensible Markup Language. 28

List of Figures

List of Tables

A

Appendix

A.1. Big Picture of the Model for Clinical Information

The big picture of MCI (see figure at the end of this document) shows the different models that are used to express clinical data and medical knowledge. Since MCI contains 550 classes in total, only high level classes and their relations are shown in the figure. In particular, the class hierarchies under the top-level entities are omitted: the figure shows, e.g., only one class for *clinical finding*, without its 29 subclasses.

A.2. Example Radiology Reports

Three different corpora of radiology reports are described in the following subsections.

German Radiology Reports on Lymphoma Patients

The *lymphoma data set* set consists of 2584 German radiology reports on 377 lymphoma patients. The imaging modality is mainly computed tomography (CT), but also magnetic resonance imaging (MRI) and ultrasound (US) and the reports are from 27 different readers. The inspected body regions were mainly abdomen, thorax and head, but includes also various other regions from the whole body. The data is provided in an excel sheet. Since the data set contains multiple reports from consecutive examinations for each patient it is suitable for longitudinal analysis of clinical findings.

Table A.1.: Data fields provided for each report of the data set of German radiology reports on lymphoma patients.

field	example entry
patient ID	1000000XYZ
date of birth	07.01.1978
gender	W
pat. num	149311
CIS ID	10000001
date	16.10.2004
time	11:39:00
device	CT2C
ref. body	M343
title	Hals-, Thorax- und Abdomen-CT mit KM i.v
findings	Es liegt eine Voruntersuchung vom 11.08.04 zum Vergleich vor. Dem gegenüber keine neue entstandenen pathologisch vergrößerten Lymphknoten zervikal. Unverändert in den Gefäßnervenscheiden kleinere unter 1 cm große Lymphknoten. Der vorbeschriebene vergrößerte Lymphknoten mit 1,6 cm neben der V. jugularis linksseitig ist mit einem Rückgang auf 1,13 cm größenregredient, jedoch sind noch gruppierte kleinere Lymphknoten bds. an der V. jugularis erkennbar. CT Thorax/Abdomen: Vergrößerte Lymphknoten im vorderen Mediastinum und axillär, vor allem linksseitig mit einem Referenzlymphknoten der vormals über 3 cm groß, jetzt mit 2,6 cm zur Darstellung kommt. Auch die übrigen Lymphknoten der linken Achsel sind größenregredient. Im Mediastinum finden sich Lymphome im vorderen Mediastinum, in der Thymusloge sowie im mittleren Mediastinum paratracheal, die ebenfalls im Verlauf größenregredient sind. Aktuell misst ein eingezeichneter Lymphknoten im vorderen Mediastinum 2,3 cm und im Azygoswinkel 1,5 cm. Die Weichteilmasse um die Aorta descendens ist ebenfalls rückläufig. Die Lunge stellt sich unverändert frei von suspekten Lungenrundherden, Erguss oder Infiltraten dar. Die Leber ist vergrößert und reicht mit ihrem linken Leberlappen an die Milz heran, die selbst unverändert vergrößert zur Darstellung kommt. Der hypodense Herd in der Milz grenzt sich etwa mit 1,2 cm Durchmesser gering größenregredient ab. Die Leber ist homogen kontrastiert, suspekte hypodense Läsionen sind nicht erkennbar. In der Gallenblase sind keine kalkdichten Konkremente abgrenzbar. Pankreas, Nieren und Nebennieren kommen unauffällig zur Darstellung. Präaortal infradiaphragmal keine pathologisch vergrößerten Lymphknoten. Inguinal fidnen sich kleinere Lymphknoten die nicht als pathologisch vergrößert zu werten sind.
assessment	Im Verlauf rückläufige Größe der Lymphome zervikal, axillär und mediastinal. Der hypodense Befund in der Milz ist ebenfalls gering größenregredient. Weitere Organmanifestationen sind weiterhin nicht erkennbar.

German Radiology Reports on Internistic Patients

The *internistic data set* consists of 6007 German radiology reports on internistic patients from 27 different readers, where imaging modality was computed tomography (CT). The examined patients have a diverse disease background.

Table A.2.: Data fields provided for each report of the data set of German radiology reports on internistic patients.

field	example entry
title	Abdomen-CT 2 Phasen mit KM
device	CT1N
findings	Voraufnahmen vom 23.11.2012 zum Vergleich. Basaler Thorax: Soweit abgebildet kein Perikarderguss. Kein Pleuraerguss. In den basal miterfassten Lungenabschnitten regelrechte Lungengefäß- und Lungengerüstzeichnung. Keine suspekten intrapulmonalen Läsionen. Abdomen: Neue hypodense Läsion im Lebersegment S8, Durchmesser 1,4 x 1,2 cm (Bild 11). Restl. Leberparenchym homogen KM-aufnehmend, ohne weitere fokale Läsionen. Normale Weite der intra- und extrahepatischen Gallenwege. Gallenblase gefüllt, zartwandig, ohne röntgendichte Konkremente. Pancreas fettig lobuliert, ohne Herdbefund. Kein Gangaufstau. Milz normgroß, homogen KM-aufnehmend. Nebennieren beidseits zart. Nieren beidseits orthotop gelegen, zeitgleich und regelrecht kontrastiert. Nephrostoma beidseits mit den Spitzen in den Nierenbecken. In der Spätphase rechts kein pararenales/-ureterales Kontrastmitteldepot mehr abgrenzbar. Ureteren beidseits regelrecht, nicht erweitert, keine KM-Aufnahme. Einmündung beider Ureteren im Conduit im rechten Unterbauch mit Ausleitung über die Bauchdecke. Größenprogredienter inhomogen KM-aufnehmender Tumor der Neoblase an der Einmündungsstelle der Ureteren beidseits mit 3,7 x 3,1 cm (Bild 63, VU 2,9 x 2,5 cm, Bild 59). Kein Harnstau. Keine vergrößerten Lymphknoten paraaortal, iliakal oder inguinal. Keine freie Flüssigkeit abdominal. Eingeschränkte Beurteilbarkeit im kleinen Becken bei Metallartefakten durch Hüft-TEP rechts. Sigma elongatum. Zustand nach medianer Laparotomie. Rektumwand regelrecht. Im Knochenfenster degenerative Veränderungen der Wirbelsäule. Deckplattenimpression BWK 12, idem. Keine umschriebenen Osteolysen.
assessment	Größenprogredienter Tumor an der Einmündungsstelle der Ureteren an der Neoblase. CT-morphologisch kein H.a. aktive Blutung aus dem Conduit. Dringender V.a. neue Lebermetastase im Segment 8. Nephrostoma beidseits regelrecht. Kein Harnstau. Keine vergrößerten Lymphknoten abdominal. Keine freie Flüssigkeit.

English Radiology Reports

The data set consists of English radiology reports on with diverse imaging modality and patient background (disease, age, gender etc).

Table A.3.: Data fields provided for each report of the data set of English radiology reports.

field	example entry
proc_desc_long	US ABDOMEN
proc_num	5
date	01.04.2011
result	Urinary frequency and decreased renal function. History of lung cancer. Findings: The liver is heterogeneously hyperechoic but not enlarged. No focal hepatic lesion is seen. Small echogenic polyps are noted in the otherwise normal looking gallbladder. There is no pancreatic head mass or biliary dilation. Moderate atherosclerotic changes are noted in the abdominal aorta. Maximum AP diameter is 2.4 cm. Several cysts arise from each kidney. Largest right renal cyst measures 4.7 x 6.6 cm. Largest left renal cyst measures 2.3 x 2.3 cm. By report, these have enlarged since 2005. No hydronephrosis is seen. Renal parenchymal echotexture is mildly increased. Transabdominal images reveal mild splenomegaly (3.6 x 5.0 x 5.7 cm).
impression	1. Hyperechoic liver consistent with fatty infiltration or hepatocellular disease. 2. Interval enlargement of renal cysts bilaterally. Mildly echogenic renal parenchyma consistent with history of decreasing renal function. 3. Prostatomegaly. No evidence of bladder outlet obstruction.

A.3. Notation

Naming of Entities

The IRI of an entity has two parts: the namespace and the local identifier. E.g. in the IRI `http://www.w3.org/2002/07/owl#Class`, has the namespace `http://www.w3.org/2002/07/owl#` and the local identifier `Class`. Within one document the namespace might be associated by a shorter prefix. For instance, the namespace IRI `http://www.w3.org/2002/07/owl#` is commonly associated with the prefix `owl:` and one can write `owl:Class` instead of the full IRI when the prefix is defined accordingly in the corresponding document. Within the biomedical domain the local identifier is often an alphanumeric ID without any semantics [Sch+09] - e.g. `OGMS_0000073`. To enhance readability, the (preferred) label from the ontology is used for the corresponding entity. That is instead of writing `obo:OGMS_0000073` the entity is denoted as `obo:diagnosis`. If the label contains spaces, the entity may be surround the by hyphens, e.g. `'obo:value specification'`, to unambiguously distinguish subsequent entities. If the ontology namespace is clear by the context the prefix is sometimes omitted (e.g. `value specification`) when referring to the corresponding entity. The OBO namespace is used by many ontologies; it is however often the case that one wants to make transparent by which ontology exactly, a certain entity is defined. Then an entity is denoted as `ogms:diagnosis` instead of `obo:diagnosis` to make clear that this entity is defined by OGMS.

Size of Reused Models

To give the reader an impression of the size of described ontologies the number of classes and (if relevant) also the number of properties are listed. Since one ontology often imports other ontologies and the official release often contains deprecated classes which are not of relevance for us - and which make the ontology appear bigger than it actually is - a clear definition for these numbers is necessary: For instance OBI contains in total 2800 classes, 61 of them are deprecated. From of the remaining 2739 classes, 2216 classes are defined by OBI (that is they have an ID starting with `OBI_`) while the rest (523 classed) is imported from other ontologies. Thus, it is written, that OBI *defines* 2216 classes and *contains* 2739 classes *in total*. The same approach is applied when the corresponding numbers for properties are counted. Note that the numbers might differ from those displayed at BioPortal [Whe+11], where the number of classes in OBI is given as 2799, i.e. they include deprecated classes as well as imports in their calculation.

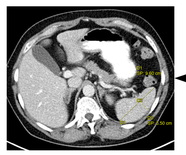
A.4. The Biological Spatial Ontology

The Biological Spatial Ontology (BSPO) [14q] is a small ontology, specialized for *cross species* representation of spatial anatomical entities. BSPO contains 146 classes, defining material and immaterial anatomical such as anatomical boundaries, planes and axes. In general the defined classes and relations are very useful. However, some issues of textual definitions and logical axioms do not fulfil our expectations regarding the *human* anatomy: Firstly, in BSPO the orthogonal_to relation has domain `anatomical axis` and range `anatomical axis`. In general orthogonality is a symmetric property and thus should have the same class as domain and range. In defining range and domain orthogonal_to relation in BSPO is indeed a relation between orthogonal complements within the three dimensional space. Thus, the orthogonal relation defined by BSPO cannot be used to represent orthogonality between lines. Secondly, the main human body planes and axes are the following: the sagittal plane which is orthogonal to the left-right axis, the coronal plane which is orthogonal to the anterior-posterior axis and the transverse plane which is orthogonal to the craniocaudal axis. The coronal plane and transverse plane are not defined in this way in BSPO. In BSPO anterior posterior axis has synonym craniocaudal axis which is not correct for the human anatomy. In this work the FMA classes for the main body planes were reused instead and *orthogonal to* relations as well as the main body axis were defined in MCI itself.

Clinical Data

Model for Clinical Information

Ontologies



ogms:healthcare encounter
mci:start date time '2010-10-12'^^xsd:dateTime
mci:end date time '2010-11-02'^^xsd:dateTime
mci:has unique ID 123456789
obi:has participant
ro:has part
mci:patient
mci:date of birth '2010-10-12'^^xsd:dateTime
mci:age at admission '2010-11-02'^^xsd:dateTime
mci:has unique ID 12002889
obi:hospital
ro:has part
mci:hospital department
mci:imaging device
obi:has participant
mci:has previous examination
mci:radiological examination
obi:has participant
obi:has participant
mci:date time '2010-11-02'^^xsd:dateTime
mci:has examination modality
mci:has examination modality
mci:patient position
obi:has specified output
mci:imaging modality
mci:dicom study
mci:has unique ID 12002889
ro:has part
mci:dicom image series
mci:has unique ID 100000123
ro:has part
iao:image
mci:has unique ID 053
obi:specified input of
iao:documenting
obi:has specified output
iao:report
mci:creation date time '2010-11-02T20:14:03'^^xsd:dateTime
ro:has part
iao:document title
mci:textual representation 'Thorax-Abdomen CT ohne KM'
oa:hasSource
mci:findings section
mci:textual representation 'Als Voraufnahme liegt ein...'
ro:has part
mci:finding subsection
iao:is about
mci:anatomical structure
mci:textual representation 'Thorax: mediastinal keine...'
mci:finding subsection
iao:is about
mci:anatomical structure
mci:textual representation 'Abdomen: Im Bereich des...'
ro:has part
mci:sentence
mci:assessment section
mci:textual representation 'Weiterhin keine Auffälligkeiten im ...'
mci:documented in
mci:has preceding finding
ogms:clinical finding
mci:size finding
iao:is about
mci:anatomical structure
ro:has part
obo:measurement datum
obi:has value specification
obi:scalar value specification
iao:has measurement value '3.5'^^xsd:float
iao:has measurement unit label
uo:unit
ro:has part
obi:specified input of
obi:specified input of
obi:specified input of
mci:is about symptom
ogms:diagnostic process
dc:date '2010-10-18'^^xsd:date
bfo:has participant
mci:hospital department
obi:has specified output
mci:main diagnosis
obi:has specified output
mci:secondary diagnosis

Annotations

iao:spatial coordinate datum
iao:has x coordinate value '144.4'^^xsd:float
iao:has y coordinate value '98.2'^^xsd:float
iao:has z coordinate value '204.4'^^xsd:float
oa:hasTarget
mci:geometrical shape
oa:hasSelector
mci:region of interest
oa:hasTarget
oa:Annotation
oa:hasBody
oa:Composite
oa:hasTarget
oa:Annotation
oa:hasBody
oa:hasSelector
oa:item
nif:Word
oa:Selector
oa:start '130'^^xsd:integer
oa:end '172'^^xsd:integer
oa:end 'Lymphknoten … bei colon'^^xsd:string
oa:hasSelector
oa:TextPositionSelector
oa:start '130'^^xsd:integer
oa:end '141'^^xsd:integer
oa:end 'Lymphknoten'^^xsd:string

Knowledge Models

bfo:quality
bfo:inheres in
mci:anatomical structure
mci:is range specification of quality
mci:range specification
obi:has value specification
obi:scalar value specification
iao:has measurement value '1.0'^^xsd:float
iao:has measurement unit label
uo:unit
bfo:inheres in
obi:population
mci:disease symptom relation
mci:is about disease
do:disease
mci:disease leading symptom relation
mci:is about disease

orgerl:hospital department
radlex:imaging device
radlex:clinical finding
rdfs:subClassOf
radlex:pathophysiologic finding
rdfs:subClassOf
radlex:cyst
radlex:colic lymph node
radlex:anatomical structure
rdfs:subClassOf
radlex:thorax
radlex:abdomen
radlex:colic lymph node
rdfs:subclass of
radlex:abdominal lymph node
rdfs:subclass of
radlex:lymph node
rdfs:subclass of
radlex:anatomical entity
rdfs:subclass of
rdfs:subclass of
icd10:malignant neoplasm of colon
icd10:malignant neoplasm of gallbladder