Universidade do Minho
Escola de Engenharia

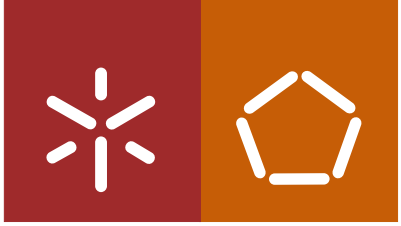
Fábio Humberto Pinto França

Ontological Representation of Tumor-Node-Metastasis Classification and an Ontology-Driven Classifier: A Study on Colorectal Cancer

outubro de 2015

Fábio Humberto Pinto França **Ontological Representation of Tumor-Node-Metastasis Classification and an Ontology-Driven Classifier: A Study on Colorectal Cancer**

UMinho|2015



Universidade do Minho
Escola de Engenharia

Fábio Humberto Pinto França

Ontological Representation of Tumor-Node-Metastasis Classification and an Ontology-Driven Classifier: A Study on Colorectal Cancer

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Ramo de Informática Médica

Trabalho efetuado sob a orientação de:
Martin Boeker
e
Paulo Jorge Freitas de Oliveira Novais

outubro de 2015

Declaração

Nome: Fábio Humberto Pinto França

Endereço eletrónico: fabiofranca92@gmail.com

Cartão de Cidadão: 14194724

Título da Dissertação: Ontological Representation of Tumor-Node-Metastasis Classification and an Ontology-Driven Classifier: A Study on Colorectal Cancer

Orientadores: Martin Boeker e Paulo Jorge Freitas de Oliveira Novais

Ano de conclusão: 2015

Designação do Mestrado: Mestrado Integrado em Engenharia Biomédica
Ramo: Informática Médica

Universidade do Minho, __ / __ / __

Assinatura: _____

Acknowledgements

Germany

In the first place, I'd like to thank my supervisor Dr. Martin Boeker for the opportunity, the time spent and the confidence placed in me and my work. I want to express my gratitude for everything that I learned this year and for the opportunities that you gave me to expand my experience. It was a great experience and a pleasure to work with such a good professional and person. Still within my experience in Germany, I'd like to thank Prof. Dr. Stefan Schulz for also helping me in my thesis work with very good insights and advise. I'd also like to acknowledge Dr. Peter Bronsert for opening the doors of the Pathology Center of the Freiburg University Hospital, where he provided material essential for my thesis. A special thank you to all the people that work in the IMBI institute, that opened me the doors of their offices any time I needed some help and for providing me a pleasant and professional work environment. A very big thank you to all my brothers in arms in my Erasmus adventure. We shared a lot of adventures, stories, we supported each other and finally said goodbye. However, I'll never forget anyone and I'll cherish each moment that we passed in my memory and my heart. I'd like to leave a personal note to Adrian Fernandez that was there any time I needed both in good or bad moments and to Maria João Sousa, my portuguese Erasmus mate, for all the help and support given during this year.

Portugal

Em primeiro lugar queria agradecer ao meu orientador Professor Doutor Paulo Novais por acompanhar de muito perto o meu projeto mesmo estando a milhares quilometros de distância. Queria agradecer por toda ajuda e tempo que disponibilizou para mim sempre com muita sinceridade, profissionalismo e também boa disposição. Queria agradecer também à minha família, especialmente aos meus pais que amo muito, que trabalham e se esforçam todos os dias para que eu chegasse a este ponto. Estiveram sempre ao meu lado e nunca me impediram de nada.

Acreditaram em mim e fizeram que eu desse sempre o meu melhor . Quero que vocês vejam este trabalho e o meu futuro como a maior prova do meu agradecimento. Aos meus avós que me sempre apoiaram e sempre quiseram estar a par da minha vida académica, um pelas ajudas extras que tanto fazem falta e o outro por me querer ver vestido de negro por fora mas também por dentro. E as minhas irmãs, uma por ser um exemplo que sempre seguirei e que me faz cada vez lutar mais e a outra por ser sempre a minha companheira número um quando estou em casa. Gostaria também agradecer a todos os meus amigos de curso que partilharam as mesmas lutas que eu até chegar a este día. Dentro destes gostava de dar um obrigado muito especial a Manel Zamith, João Macedo e Filipe Fernandes, que os considero como meus irmãos que sempre estiveram ao meu lado como sempre estarão! Queria deixar uma palavra a aqueles que patilharam casa comigo, João Prates e Cláudia Rodrigues (é como se lá vivesses) que juntos com o Macedo foram a minha família na Universidade do Minho. Nunca esquecer da minha malta de espinho - João Martins, Diogo Remoaldo, João Vitorino, Ricardo Pacheco, Guilherme Mendes, Afonso Couto, que já estão comigo durante muitos anos e foram o meu ponto seguro quando ia a casa. Foram sempre magníficos e nunca se esqueceram de mim. Um grande obrigado ao Jorge França, que além de família é também um amigo especial. Fizeste-me abrir os olhos para a vida e tornaste-me uma melhor pessoa! Por fim , o obrigado mais especial para a Rita Roxo por tudo que ela passou durante este ultimo ano que também acreditou sempre em mim. Nada disto seria possível se não fosse o teu apoio e carinho incondicional.

Resumo

O sistema para classificação de tumores malignos mais aceite globalmente é o Tumor-Nódulos-Metásteses Classificação de Tumores Malignos (TNM). O procedimento de classificação compreende diversos parametros patológicos baseados na Extenção da Doença (EOD).

Os objetivos deste trabalho consistem na apresentação da ontologia *TNM-O*, uma ferramenta utilizada na representação do sistema de classificação TNM; na implementação da ontologia *Colon and RectumTNM-CR*, uma ontologia modular que representa as regras de classificação TNM referentes aos cancros no cólon e no recto, no desenvolvimento de uma aplicação, cuja base de conhecimento é a ontologia TNM-O e no teste de viabilidade desta abordagem com dados reais.

A ontologia TNM representa todas as definições e regras presentes na classificação TNM. Esta ontologia é o ponto central de um sistema desenvolvido com base numa arquitetura modular. Cada módulo consiste numa ontologia que representa as regras de classificação respetivas aos diferentes tumores. Estas ontologias podem ser importadas para a ontologia central, sendo que todas utilizam o *Foundational Model of Anatomy* (FMA) para representar os conceitos anatómicos e o *BioTopLite 2* como ontologia de domínio. A aplicação desenvolvida para a classificação de ontologias tem como base de conhecimeto a ontologia TNM. Esta foi programada em JAVA utilizando a OWL-API como ponte entre a aplicação e a base de conhecimento.

Neste estudo foram avaliados dois dataset com dados reais. O primeiro continha 382 registos que foram classificados pelos nódulos regionais. Comparando classificação automática com a manual obteve-se uma precisão de 55%. No entanto, a aplicação apontou inconsistências e erros feitos na documentação do tumor que causou este resultado. O segundo dataset consistia em 292 registos produzidos e classificados manualmente por um patologista através de documentos em texto. A classificação automática revelou resultados ótimos para todos os tipos de classificação

Este estudo mostrou que a aplicação desenvolvida melhora a consistência e eficiência dos dados na documentação de tumores assim como providencia classificação automática exata durante o processo de diagnóstico do tumor.

Abstract

The most important staging system for cancer is the TNM Classification of Malignant Tumors (TNM) classification. The staging procedure compiles several clinical and pathological parameters based on the Extent of Disease (EOD).

The objectives of this work are to present the Tumor-Nodes-Metastasis Ontology (TNM-O), a framework for the representation of the TNM classification of malignant tumors (TNM) system; to implement the TNM Colon and Rectum ontology, a modular ontology that represents the TNM classification for the colorectal tumors based on this framework; to develop an ontologically driven classifier application with the TNM-O as it's knowledge base and to show the feasibility of this approach on real data.

TNM Ontology (TNM-O) and TNM Colon and Rectum Ontology (TNMCR-O) use the Foundational Model of Anatomy (FMA) for representing anatomical entities and BioTopLite2 (BTL2) as a domain top-level ontology. The classification rules of the TNM classification for colorectal tumors were represented as described in the literature. The automatic classifier for pathological data uses these ontologies as knowledge base. It was developed with JAVA using the Ontology Web Language (OWL)-application programming interface (API) to make the bridge between the application level and knowledge base.

In this study, two datasets with real data where evaluated. The first dataset contained 382 entries that was classified by the regional lymph nodes. This study compared automatic classification with the expert one and obtained an accuracy of 55%. However, the classifier flagged inconsistencies and errors made during the manual tumor documentation that caused the misclassification. The second dataset contained 292 records carefully classified by a pathologist. In this dataset, automatic classification was optimal to all types of assessment.

Therefore, this study proved that an ontology-driven automatic classifier enhances the consistency in tumor documentation and provides accurate instance classification during pathological assessment of tumors.

Contents

List of Figures

List of Tables

Acronymes

AJCC American Joint Committee on Cancer. 2
API application programming interface. vii, 41, 42, 58

BFO Basic Formal Ontology. 9, 11–13, 17, 20
BTL2 BioTopLite2. vii, 17, 39, 40, 44, 45

CS Collaborative Stage Data Collection System. 2, 67

DL Description Logic. 2, 6, 13, 17, 20, 28–30, 40
DOLCE Descriptive Ontology for Linguistic and Cognitive Engineering. 9, 12–14, 20

EOD Extent of Disease. vii, 31, 39

FMA Foundational Model of Anatomy. vii, 2, 39, 40, 44, 66
FME Foundational Model Explorer. 39, 48

GO Gene Ontology. 15
GoodOD Good Ontology Designed. 20
GSS Government Statistical Service. 2
GUI Graphical User Interface. 41, 56, 58–60

IARC International Agency for Research on Cancer. 1
ICD-10 International Classification of Diseases - 10th Revision. 4
ICD-O International Classification of Diseases for Oncology. 32
IMBI Institute of Medical Biometry and Medical Informatics. 2, 17
ISI Information Sciences Institute. 26

MeSH Medical Subject Headings. 4, 15

ODE Ontology Design Environment. 27
OKBC Open Knowledge Base Connectivity. 28
OWL Ontology Web Language. vii, x, 5–9, 12, 13, 17, 20, 28, 30, 40–42, 56

RDF Resource Description Framework. x, 6, 7, 9, 28

SEER The Surveillance, Epidemiology, and End Results. 2
SNOMED CT Systematized Nomenclature of Medicine - Clinical Terms. 4

TNM TNM Classification of Malignant Tumors. vii, 1–3, 30–32, 36–41, 44–46, 50, 54, 56, 58–60, 64–67
TNM-O TNM Ontology. vii, xi, 3, 15, 38–41, 44, 48, 50, 56, 61, 66, 67
TNMCR-O TNM Colon and Rectum Ontology. vii, 3, 38, 40, 41, 44, 47, 48, 50, 66, 67

UDO Upper Domain Ontologies. 15
UICC Union for International Cancer Control. 1, 31, 38, 39
ULO Upper Level Ontologies. 9, 13
UPM Technical University of Madrid. 26

W3C World Wide Web Consortium. 6, 7
WHO World Health Organization. 1

Chapter 1

Introduction

The last estimations made by the International Agency for Research on Cancer (IARC), a specialized cancer agency of the World Health Organization (WHO), verified that in the year 2012 there were 14.1 million new cancer cases, 8.2 million cancer deaths and 32.6 million people living with cancer. In these estimations, colorectal cancer was the third most common type of tumor in men and the second in women [1]. Thus, research in new methods for diagnosing and treatment of cancer is the main goal of the WHO cancer programs [2].

One of the most globally accepted staging system for cancer is the TNM Classification of Malignant Tumors (TNM) [3] published by the Union for International Cancer Control (UICC). This system compiles various pathological and clinical parameters for three types of assessment: primary tumor (T), regional lymph nodes (N) and distant metastasis (M). It also provides a distinct and specialized classification for each tumor site. The primary tumor classification generally evaluates the infiltration and size of the carcinoma; the regional lymph nodes assessment concerns the number of metastatic lymph nodes in the regional area of the primary tumor and the presence and absence of distant metastasis.

TNM Classification has been used for more than fifty years, being under an developmental process for updating and revising its documentation. This process made this classification one of the most complete and precise tumor classifications of today, requiring a high level of knowledge and expertise in the domain. However, it is very difficult to this system to keep up with the overwhelming changes and updates in this field [4, 5]. Despite the importance of the TNM classification, no formal logic-based representation has been developed.

Ontologies are information artefacts that formally represent knowledge from a certain domain in order to be machine processable. In the biomedical domain, they are used to describe the structure of their complex domains and to relate their data to shared representations of biomedical knowledge. They provide reference encyclopaedic knowledge and enable computer reasoning of biomedical data [6].

Examples of biomedical ontologies in the literature are: the Foundational Model of Anatomy (FMA) ontology [7] for representation of concepts and definitions about the human anatomy, the HL7 Reference Information Model (HL7-RIM) ontology [8] to represent the messaging standard HL7 and the Gene Ontology (GO) [9] that seeks to provide a set of vocabularies for biological domains that can be used to describe gene products in any organism. Therefore, an ontological representation of the TNM Classification system would be a solution to fill the gap of a missing formal representation of this system.

This project was developed within the Institute of Medical Biometry and Medical Informatics (IMBI) in Freiburg - Germany, which has the goal to provide a full ontological representation of the TNM classification system. As prior work, an ontology for representation of the TNM classification rules for breast cancer was already developed by Rita Faria [10, 11]. Despite representing a different type of tumor, this work provided some useful definitions and concepts to the development of the ontology for colorectal classification.

Today, tumor registries collect data on the diagnosis and staging of cancer to generate reports for the physicians and hospital cancer registries. Maintaining a consistent and updated cancer registry positively influences the quality of prognosis and treatment protocols. A project conducted by the American Joint Committee on Cancer (AJCC), the Collaborative Stage Data Collection System (CS), consists in a software equipped with algorithms capable to translate TNM staging information in order to be used across cancer statistical databases such as the The Surveillance, Epidemiology, and End Results (SEER) and Government Statistical Service (GSS) [12]. Other examples of software related to the TNM system are: an ontology-driven classifier that processes physicians annotations in images to reason the TNM classification [13] and a semi-automatic tool that classifies tumor documentation in the ESTHER system also based in the TNM classification system [14]. However, none of these studies present a tool for classification based on a formal representation of the TNM classification system. Additionally, no study was found were they provided a feasibility test of these systems.

One advantage of pursuing an ontology-based approach is that ontology maintenance and updating is done in a quicker and more consistent way. Thus, all the modifications made on TNM can be centrally done in the ontology, with only little changes needed on the application level. Using Description Logic (DL) semantics in the ontology, adds the advantage of detecting logical inconsistencies and coding problems that can happen due to the system's complexity.

Having a formal representation of the TNM system provides uniformity of knowledge that enhances interoperability and robustness between distinct systems. Maintaining independence between knowledge base and application allows the developer to maintain and update the knowledge base without making substantial

changes in the application level.

With this work we propose to close the gap of a missing formal representation by presenting the TNM Ontology (TNM-O) and the TNM Colon and Rectum Ontology (TNMCR-O). The first aims to represent the main structure of the TNM classification in order to provide support to the TNMCR-O, that represents the concepts and classification rules for the colorectal tumors. Additionally, we also present an ontology-driven automatic classifier that uses these ontologies as knowledge base to provide instance classification and consistency evaluation on the pathological data registry.

Therefore, the objectives of this project are to present the TNM-O, an ontological framework for the TNM classification system; implement the TNMCR-O, a modular ontology that represents the TNM classification for the colon and rectum tumors based on this framework; develop an ontology driven classifier application with the TNM-O as its knowledge base and show the feasibility of this approach on real data.

Chapter 2

Backgrounds

2.1 Ontologies

In computer science, ontologies are information artefacts that formally standardize, describe and order concepts and definitions in a certain domain. Its relevance has been increasing over the years, however there is still no consensus about what criteria an information artefact has to meet in order to be an ontology [15].

Many definitions of ontologies exist in literature. In 1991 a definition by R. Neches et al. stated that "An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary." [16]. A few years later, the definition was adapted by Gruber et al. to "An ontology is a formal, explicit specification of a shared conceptualization" [17]. In this definition, an ontology is defined as a *conceptualization*, which means that an ontology is an abstract model that identifies relevant concepts and their relations within a certain domain. It also characterizes ontology as *explicit*, since all the concepts and their constraints are explicitly defined in order to be machine-processable. Finally, another aspect is shareability, as long as an ontology captures a consensual knowledge it will be shared in the community [18, 19].

Many definitions of ontologies complement each other. The one presented above is probably one that might reflect consensus in the ontological community.

Ontologies are used to represent shared knowledge of a certain domain in order to be handled by a machine. The interest in building ontologies has grown as researchers had problems to keep track of all scientific publications published in the medical domain. To render all this knowledge, documentation specialist have developed large terminologies, such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), International Classification of Diseases - 10th Revision (ICD-10) and Medical Subject Headings (MeSH). However, these termi-

nologies cannot cope with problems that are inherent to human language, where, for example, one term can have multiple meanings. A solution is the use of ontologies in which the terms of a domain are used to build logical relations between them.

Another motivation came from the inevitable incompatibilities of the current relational databases, in which, for example, labels can have the same name but different meaning, and vice versa. Despite its similar functions, there are some key characteristics that distinguishes an ontology from a relational database. First, ontologies are syntactically and semantically far richer than common databases. Second, knowledge is described in a formal language instead of the tabular information tuples. Finally, an ontology provides a consensual theory for a domain and not only the structure of a data container [15].

An ontology can been seen as a form of semantic network. One example is displayed in Figure 2.1 that presents knowledge as classes of individuals connected with *is-a* relations. These relations create a hierarchy based on super/subclass relationships. For example, *Student* **is-a** *Person* means that an entity that belongs to the class *Student* is also an entity of *Person*, therefore *Student* is a subclass of *Person*. Additionally, in the same figure it is also possible to identify relations between different types of concepts, such as *Student* **studiesAt** *University*. *Student* and *University* are two distinct types of entities, however, they can be used to represent the class that contains all the students that study at an university.

Even so, not every ontology can be represented as in Figure 2.1. For more detailed and complex axioms and restrictions there is no appropriate representation other than using one of the available ontology languages like the Ontology Web Language (OWL) [20] (on section 2.1.1). Besides providing the syntax for ontology representation, OWL is also used for serialization and transport of the ontologies [18].

Biomedical ontologies generally present taxonomic representations of concepts with a broad and sometimes non-consensual theoretical support, so the need for ontology alignment and integration is broadly accepted by the community. This can be done with mappings between terminology based ontologies. Although, most of these ontologies are of poor quality and the mappings between them do not constitute significant improvement. One of the most accepted methods nowadays is a vertical integration of ontologies with different scopes. So, ontologies can be distinguished in the following types [15, 21]:

- **Top-level ontology**
- **Upper-domain ontology**
- **Domain-ontology**

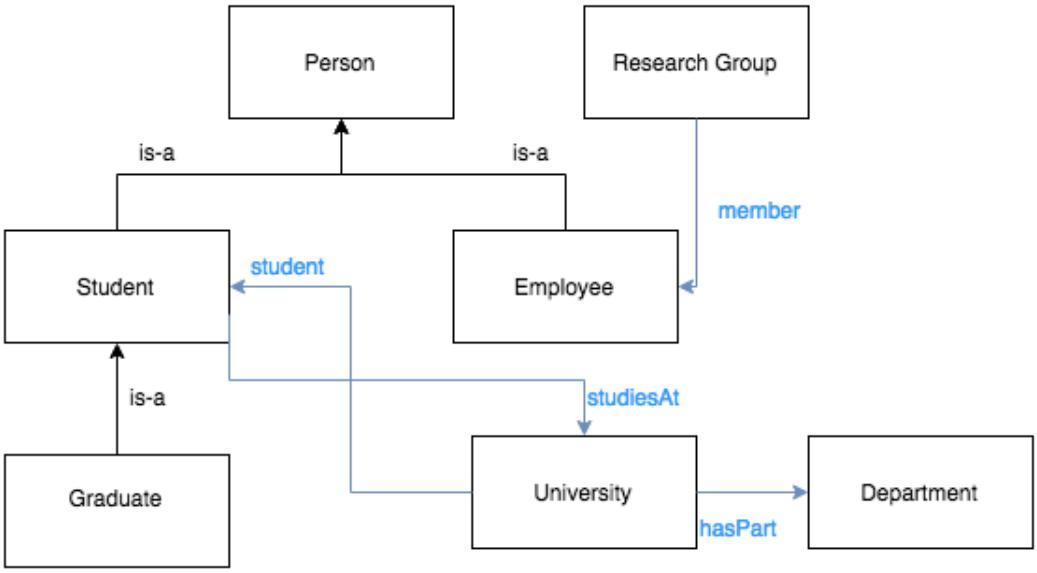


Figure 2.1: An example of a semantic network [18].

Ontologies have the potential to deeply change how intelligent systems are built. Building good knowledge bases and sharing them publicly will make libraries of ontologies available to every developer. This should decrease the time spent developing new software since the knowledge is already represented in a commonly accepted way [22]. Besides the biomedical domain, ontologies are used in domains such as agriculture, aviation, chemistry, civil engineering, business and others.

2.1.1 Web Ontology Language - OWL

OWL [20] is one of the most recent ontology language developed by the World Wide Web Consortium (W3C) in the Web Ontology Working Group. Its first goal was to represent information about categories of objects and their logical connections, what we now can call an ontology. Furthermore, OWL can also represent data about the object themselves [20, 23].

However, the development of the OWL was not made from scratch. At its core there is DL (section 2.2) that provides formalization of the semantics, language constructors, data types and data values. Moreover, OWL can be viewed as expressive DL, where an ontology in OWL is equivalent to a DL knowledge base. [15, 23].

Another major influence in the OWL design is the Resource Description Frame-

work (RDF) [24]. This is a general-purpose language for representing information in the semantic web [25]. It provides meta-data for descriptors of resources on the web and it works as a framework capable of representing data. Compatibility between web languages is the main reason of the big influence of RDF on OWL. A basic way to do this was to provide OWL the same syntax as RDF. However, the W3C found out that for some cases, ontologies would require much more expressiveness than the one provided by RDF. [26, 27]

So, extending RDF brings a trade-of between expressiveness and reasoning efficiency [23]. This lead to the development of three types of OWL that could meet the needs of each developer [26, 27]:

- **OWL Full** - The entire OWL language is called OWL Full. The main advantage of using this is full compatibility with RDF both semantically and syntactically. So, any valid RDF document is a valid OWL Full document and vice-versa. However, this language becomes to heavy that limits drastically the usage of reasoning support;
- **OWL Description Logic (OWL DL)** - this sub-language of the OWL Full is used when better computational efficiency is wanted. This is made by reducing the amount of OWL constructors to ensure that the language corresponds to a well defined description logic. Although, this will decrease the compatibility with RDF since an RDF document will need some extension to be a valid OWL DL document. On the other hand, an OWL DL document is still a valid RDF document;
- **OWL Lite** - This is a subset of OWL DL where more restrictions to its limits were applied. The main goal of this language is to simplify usage and implementation in new frameworks;

Syntactically, as said before, OWL is an extension of RDF syntax. In Figure 2.2 there is a graphical representation of an example of how the concepts are arranged. Without entering in a lot of detail, and not forgetting that an OWL document is generally a RDF document, the constructors of the OWL syntax are [20, 27]:

- **Header** - despite of being the root of an OWL ontology, this is a RDF element `rdf:RDF`. This is the element where all the namespaces used are specified. Then, any OWL ontology should start with a set of assertions for purposes like: version control, comments and the inclusion of other ontologies with the element `owl:Ontology`;
- **Class elements** - classes in ontologies are defined with the `owl:Class` element. In order to represent the relation of subsumption between class its

used the element `rdfs:subClassOf`. Besides this , it is also possible to create disjoint classes with `owl:disjointWith` and equivalent classes using `owl:equivalentClass`. Finally, there are two predefined classes: `owl:Thing` that corresponds to the root class and the `owl:Nothing` which is an empty one;

- **Property elements** - In OWL there are two kinds of properties: object properties that make logical connections between objects (**owl:ObjectProperty**) and data type properties that relates objects to values (**owl:DatatypeProperty**). Since OWL doesn't have any defined type of values, XML Schema data types are used to define them.

- **Property restrictions** - In OWL is also possible to specify that all the instances inside certain class satisfy one or more conditions. This made by creating a subclass, which can be anonymous, where all these instances belong to. This new subclass should contain the restriction that specifies which instances belongs there using the element `owl:Restriction`. For this restriction it is necessary to identify two things: first the property with the `owl:onProperty` and the type of restriction. Within the types of restrictions there are:

  - `owl:allValuesFrom` that is used to restrict the range of the property to the instances of a specific class;
  - `owl:hasValue` that identifies the exact value that the property must have to satisfy the restriction;
  - `owl:someValuesFrom` that says that an instance should at least satisfy this restriction to be a instantiated;
  - `owl:maxCardinality` and `minCardinality` for specifying the max and minimum of a specific number of properties are needed to satisfy the condition respectively;

Besides this, the OWL syntax provides much more elements that reinforce its great expressiveness and justify its great use for ontology development. Although, extensions for this syntax are in the making with the goal to provide further logical features. This syntax was the one used for the development of this project. For that reason, further discussion about existing ontologies will be done within the OWL syntax. Therefore, *Italic* font will be used for classes and **bold** for the relations.

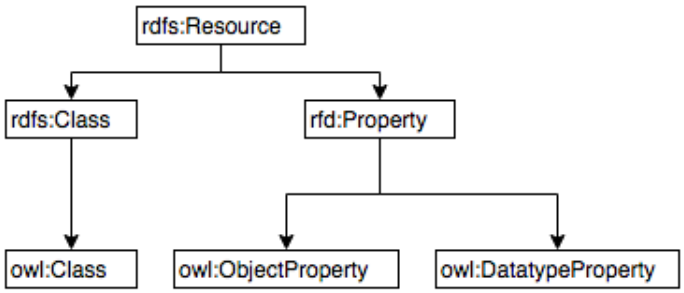


Figure 2.2: Subclass relationships between OWL and RDF [27]

2.1.2 Upper Level Ontologies

The biomedical domain is highly complex and some overlapping between ontologies occurs. Therefore, it is fundamental to consistently incorporate multiple ontologies. One approach is to build well designed and documented ontologies as high level structures with general concepts and relations on which domain ontologies can be used.

These ontologies are the Upper Level Ontologies (ULO).They aim to provide reusable and reliable definitions of concepts and their relations for independent domains, facilitating the integration and development of new domain ontologies. They are differentiated by the entities they include, the theory of space and time as well as the relation between individuals to these theories. They contain rich definitions and axioms that are applicable across multiple domains.

Nowadays there is no agreement on what makes a good top level ontology. However , there are some candidates already published such as Basic Formal Ontology (BFO) and Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [28–31]

Basic Formal Ontology - BFO

BFO, is an ULO still in development by Ontology Research Group (ORG) leaded by Barry Smith at the Department of Philosophy in the University of Buffalo. Formal representations on the biomedical domain are focused in static and dynamic entities of biological reality, while BFO tries to combine these two perspectives in order to address the issue of representing both in a consistent way. Thus, this ontology starts to provide formal distinctions between:

- Universal and Particular

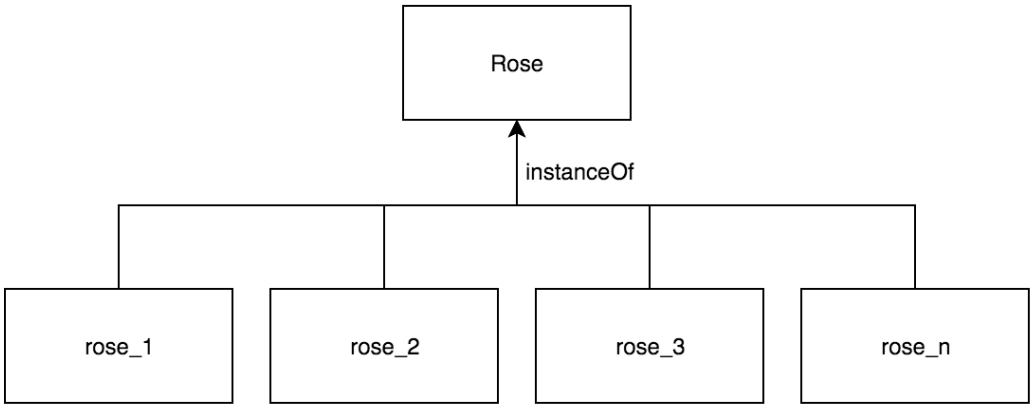


Figure 2.3: Example of the relation of **instanceOf** between an *Universal* and *Particular*

- Continuant and Occurrent
- Dependent and Independent
- Formal and Material

Universal and Particular

The *Universals* are the real invariants in the domain. On the other hand a *Particular* is an entity or individual that is the instantiation of a certain *Universal* (Figure 2.3). Another important relation is the subsumption between two *Universals*. For example, *Rose* **subClassOf** *Plant* indicates that any instance of *Rose* is also an instance of *Plant*.

Continuant and Occurrent

Continuants are entities that exist through time. They persist with the same identity even when undergoing through changes. A *Continuant* is bound to the space which it occupies but not in time. A *Continuant* is segmented in terms of space but will endure during time. A instance *Continuant* is our body, it can be divided in terms of space although it will keep its identity through time.

On the other hand, there are the *Occurrent*, *Event* or *Process*, that instead of existing in full at a single moment in time they unfold themselves in phases. Thus, in contradiction to *Continuant* they are bound with respect to time, that is, the *Occurrent* is segmented in terms of time. With this in mind, an example of

this is the process of embryological development which processes in a succession of stages.

Dependent and Independent

This is the distinction between entities which have the ability to exist without the support of others and the ones who are dependent. For example, a *Quality* is dependent on the *Thing* which it bears.

Both *Continuant/Occurrent* and *Dependent/Independent* distinctions are applied to both *Universals* and *Particulars*. An example of this can be the functioning of my kidney in this moment, which is a *Particular/Occurrent* that depends on my kidney and its function (both *Particular Continuant*). This can also reflect the same dependence between the corresponding *Universals*.

Formal and Material

Biomedical terms reference mainly material objects like organs, cells, organisms, etc. However, ontologies have to deal with vast formal relations in which these entities are related together. So *Material* is a class that is confined to its domain while *Formal* relations are used across multiple ones. Examples of formal relations can be dependence, instantiation, subsumption, etc .

SNAP and SPAN Ontologies

BFO ontology is divided in two different types: SNAP ontologies for representation of *Continuants* and SPAN ontologies for the *Occurrents*.

The main classes in SNAP ontologies (Figure 2.4) are:

- *Independent Continuant* with the subclasses *Object*, *Object Aggregate*, *Site*, *Boundary* and *Part of Object*;
- *Dependent Continuant* with the subclasses *Quality* and *Realizable*. These also have as subclasses *Function*, *Role* and *Disposition*;
- *Spatial Region* with subclasses *Volume*, *Surface*, *Line* and *Point*.

And the main classes in the SPAN ontologies (Figure 2.5) are:

- *Processual Entity* with subclasses *Process*, *Process Aggregate*, *Process Part*, *Processual Context* and *Boundary of Process*;

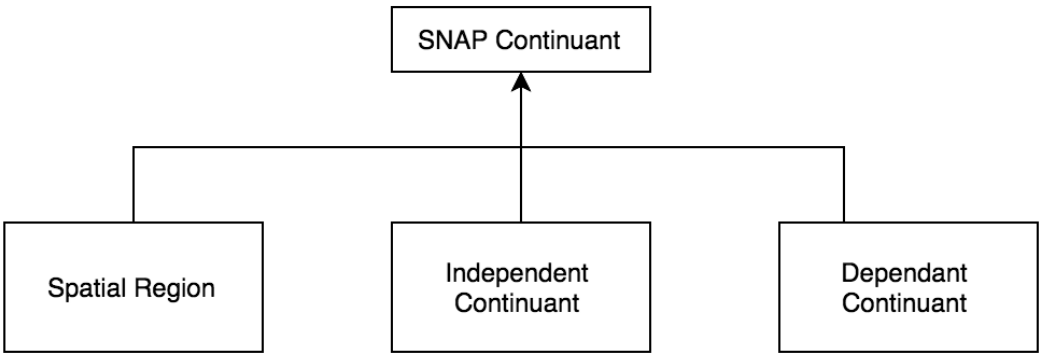


Figure 2.4: Main classes in BFO SNAP ontologies [15]

- *Spatiotemporal Region* with subclasses *Scattered Spatiotemporal Region* and *Connected Spatiotemporal Region*. The last one has the sublasses *Spationtemporal Interval* and *Spatiotemporal Instant*;

- *Temporal Region* with subclasses *Scattered Temporal Region* and *Connected Temporal Region*, which has the subclasses of the last one are *Temporal Interval* and *Temporal Instant*

BFO was implemented in OWL and is freely available to the community. It contains the top class *Entity*, seventeen SPAN classes and eighteen SNAP classes. Its application is mainly biomedical and is applied e.g. for ontology development in the domain of trials on cancer [15, 32–34].

Descriptive Ontology for Linguistic and Cognitive Engineering - DOLCE

DOLCE is the first module of the Library of Foundational Ontologies being developed as part of the WonderWeb project headed by Nicola Guarino at the Laboratory for Applied Ontology in Trento. Differently to other ontologies that follow a minimal taxonomic structure satisfying the needs of a specific domain, DOLCE aims to establish consensus in the multi-area community where artificial intelligence meets humans.

Contrarily to BFO that represents the world as it is, this ontology inclines to the cognitive point of view introducing ontological categories as cognitive artefacts depending on human perception, cultural background and social conventions.

Even with different perspectives to the world, BFO and DOLCE share many similarities (compare Figures 2.4 and 2.5 with Figure 2.6). For example an *En-*

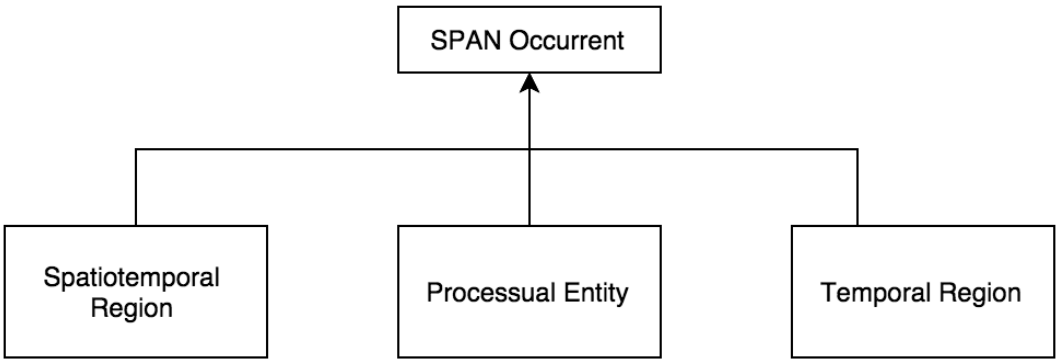


Figure 2.5: Main classes in BFO SPAN ontologies [15]

durant in DOLCE corresponds to a *Continuant* in a SNAP ontology; an *Occurrence* in DOLCE to a *Occurent* in BFO. In addition to this there are categories that almost share the same name like *Temporal Region* and *Spatial Region*.

Besides all the similarities there are some essential differences between these two ontologies worth noticing. In the BFO ontology there are no classes for abstract entities like cognitive and social objects. Also, there are no quality regions or values, instead, these are subclasses of the correspondent quality. Other differences are:

- DOLCE *Processes* can have qualities, while in BFO not;
- in DOLCE both *SpatialRegion* and *TemporalRegion* are *Abstract* entities;
- BFO does not contain subclasses for processes.

DOLCE was implemented in First Order Logic with OWL. It contains around 100 terms and the same number of axioms. Many projects uses the DOLCE ontology, including the LOIS Project, an international research project on information retrieval from legal databases; SmartWeb, another prestigious research project on artificial intelligence technologies and their application on web based systems and AsIsKnown which is a semantic-based knowledge system form home textile industries. DOLCE Lite is the OWL DL representation of DOLCE [15, 32, 35, 36].

2.1.3 Upper Domain Ontologies

Developing ontologies using the same ULO, which means, reusing the same classes, relations and even high-level restrictions does not guarantee interoperability. Therefore, an additional common terminological framework is necessary to

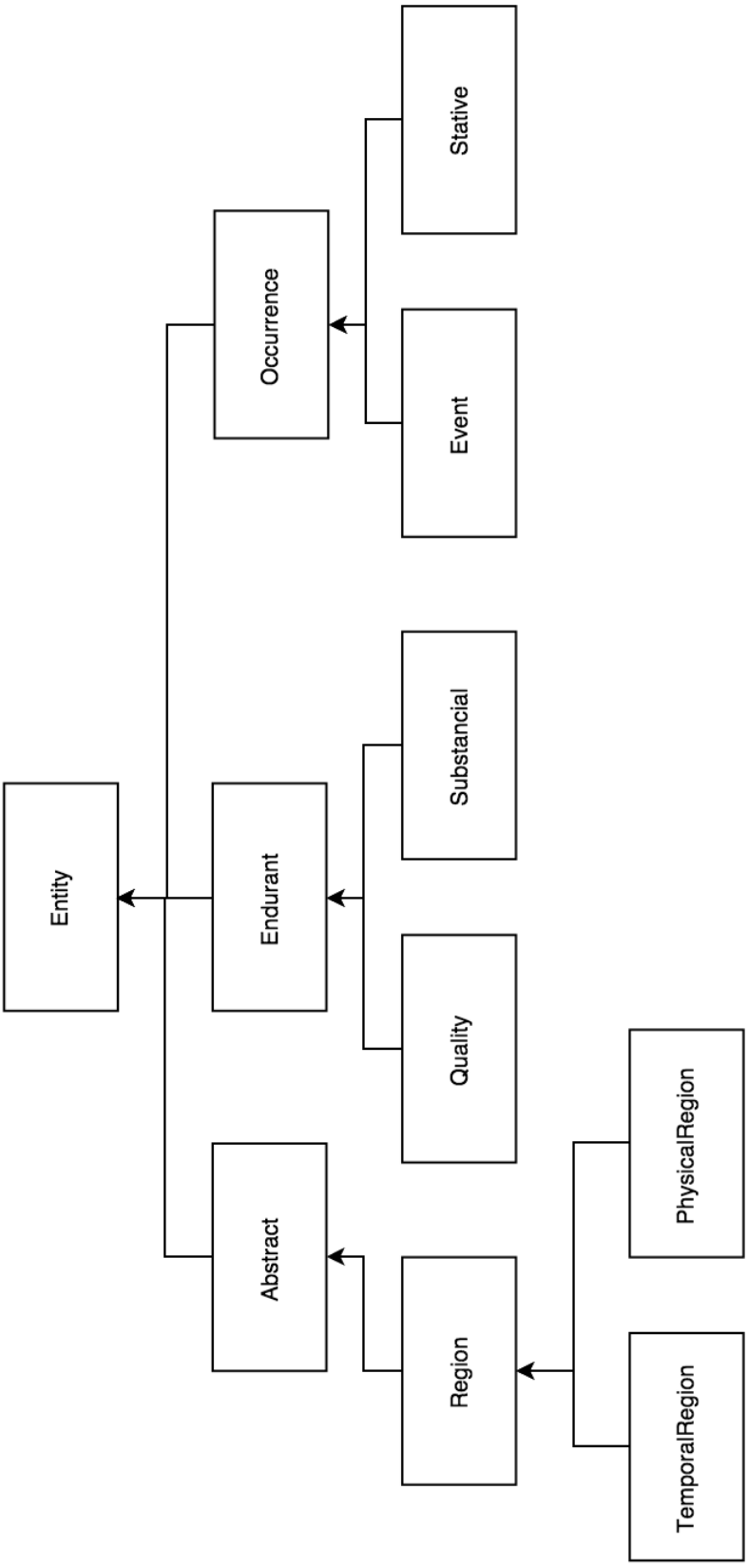


Figure 2.6: Structure of the DOLCE ontology [15]

obtain a soft, seamless transition between the most generic classes like *Continuant* and *Occurrent* and more granular ones.

This intermediary is an Upper Domain Ontologies (UDO) and it defines the types and relations essential to represent a specific domain. Examples of UDO are Biotop [29], that is used as UDO for the TNM-O and GENIA [37] which originally motivated the development of BioTop [21, 29, 38].

GENIA

The GENIA ontology was designed to provide a semantic annotation to the GENIA *corpus*. The latter is an aggregate of extracted articles from the MEDLINE database. Its purpose is to provide high quality materials for Natural Language processing and be used as a high performance standard for the evaluation of text mining systems.

Nowadays the ontology is in its second version and it was divided in two ontologies:

- **Term Ontology** - This ontology is designed to support the GENIA corpus term annotation. It represents and classifies the most significant biological terms found in literature. It defines biological, anatomic and organism entities, most of which are mapped to the MeSH repository. This ontology is also subdivided in three sub-categories:
  - **GENIA Chemicals** - which is intended to define any chemical substance;
  - **GENIA Anatomy** - it corresponds to the MeSH Anatomy category;
  - **GENIA Organisms** - this corresponds to the MeSH Organism category;
- **Event Ontology** - this ontology is designed to provide a semantic platform for the GENIA *corpus* event annotation. Its main purpose is to match Natural Language expressions within biological processes and molecular functions. It was designed to be interconnected with the Gene Ontology (GO) to improve its utility.

GENIA is free to the public and was implemented with XML and the DAML+OIL ontology language [39, 40]

BioTop

BioTop is an UDO developed with the aim to aid engineers with an ontological framework for the life sciences. It provides a layer for connecting and integrating

various domain ontologies with the biomedical domain. It is intended to integrate with more extensive domain ontologies to enhance the capabilities of current applications in areas like information retrieval and text mining.

The initial aim for developing BioTop was to redesign and expand the GENIA ontology adding fundamental principles of formal explicitness and precision of ontological axioms.

The top class of BioTop is *Particular* which has ten subclasses (Figure 2.7) :

- *Material Object* - is a *Continuant* entity that has one mass and one volume at a certain point of time. It is used to represent everything that is material in the domain. Instances of *Material Object* can be related to each other in terms of location and constitution with relations like *spatiallyRelatedTo* and its subrelations (in particular *hasLocus*, *locusOf* and *physicallyConnectedTo*) [15] ;

- *Immaterial Object* - is a subclass of *Continuant* with n-spatial dimensions like points, lines or planes. Instances of *Immaterial Object* are related to other physical entities regarding their location, connected by the same relations as the instances of *MaterialObject*. Subclasses of *ImmaterialObject* are *Wave* and *Cavities* [15];

- *Information Object* - represents information. An instance of *InformationObject* is dependent on a physical carrier that is *bearerOf* or *inheresIn*, but independent of a carrier with regard to its encoded content. For example, a treatment plan exists independently of the planned procedure but the planned procedure is dependent on the plan for its realization [15];

- *Disposition* - A disposition is a realizable entity that inheres in something and can bring itself to existence in a process. It depends on the physical make-up of the agent that participates. Although, even if a disposition exists it does not mean that its manifestation exists. Humans have the disposition for reproduction even if they never do [15].

- *Role* - In opposition to *Disposition*, a *Role* is brought into existence by its participation in a certain process. In this case, a human can have the role of a customer and salesman depending on his participation in a trading procedure;

- *Process* - Is an *Occurrent* that has temporal parts which are not always simultaneously present. It can have *Material Objects* and *Immaterial Objects* as participants;

- *Time* - represents a point or interval in the time axis;

- *Quality* - represents a feature of some other entity and cannot exist independently of it;

- *ValueRegion* - is a temporal, abstract or spatial region in which qualities are located, it corresponds to the values qualities can have;

- *Condition* - is the result of the union between *Material Object*, *Process* and *Disposition* with the aim to represent the ambiguous nature of a condition in the medical domain. Some terms can have different meanings such as tumor that can be a pathological process and also an abnormal growth of malignant tissue. This class provides a common class where these terms can be added without having to resolve this ambiguity [15].

BioTop ontology was aligned with the BFO upper level ontology and implemented in OWL-DL language. Today it is composed by 175 classes interconnected with 171 axioms. It has been developed at the IMBI at the University Medical Center Freiburg, Germany, the Department of Computer Linguistics at the University of Jena, Germany and and in the Institute of Medical Informatics, Statistics and Documentation at the Medical University Graz. It is still under development and maintenance in the IMBI [15, 29, 41].

BioTopLite

BioTopLite is a smaller, simpler and computationally more efficient version of the BioTop ontology. Both share the same objective: to provide an upper domain ontological framework for ontology developers in the biomedical sciences. Provides a core of 53 classes with 240 logical axioms using a set of 37 ontological relations (Figure 2.8).

BioTopLite2 (BTL2) is the current version and like its predecessor, it was implemented in OWL-DL. The main changes comparing to its previous version are:

- **Additional Classes** - The use of biomedical terminologies motivated the creation of a class *Life* which represents the process of an organism during its lifetime. In medical diagnoses, time references are made in segments of the *Life* of the living organism;

- **Simplified Relation Hierarchy** - Relations were distinguished between processes and objects which turned out to complicate the use of this ontology. After the abolition of this distinction the number of relations was reduced to 37.

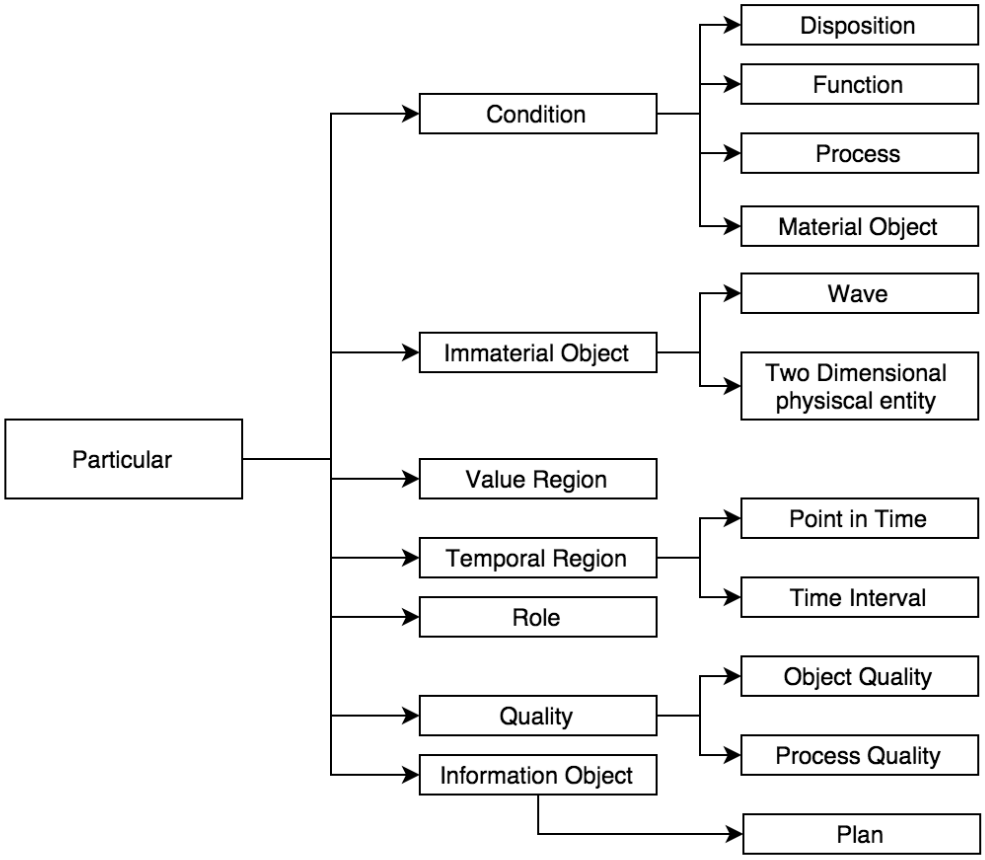


Figure 2.7: Fragment of the main BioTop class hierarchy [15]

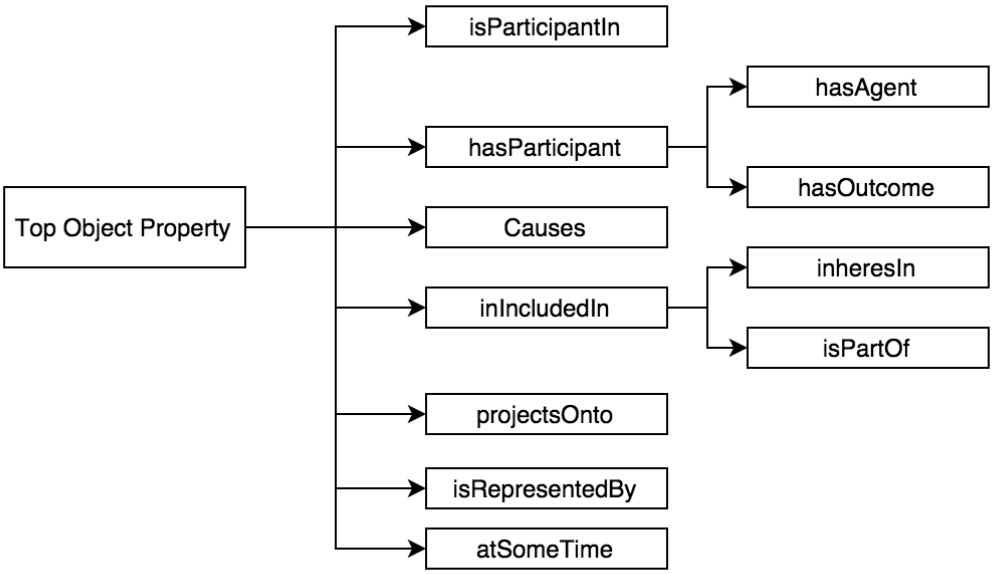


Figure 2.8: Fragment of the main BioTopLite relation hierarchy [15]

- **Substitutions of Domain/Range axioms** - The simplification of the relations hierarchy resulted in the fusion of several relations. Consequently, necessary constraints could not be fully expressed. Therefore, new class axioms were added increasing the number of axioms to 572.

- **More Intuitive Labels** - Relation labels were modified for better comprehension on the linguistic level. For example *has locus/locus of* was changed to *is included in/includes*. Class labels remained the same;

- **Representation of Time Relevant Entities** - BioTop, BFO and DOLCE share the same lack of representation between continuants since OWL-DL does not provide a genuine representation of time. This becomes an issue when the same individual is an instance of disjoint classes in different points in time. This can lead to ontological inconsistencies due to the impossibility to express time as an *occurrent*. In order to surpass this problem, the class *Entity at some time* was introduced. Using this with the relations **is referred to a time**/**at some time** turned out to be a possible solution.

BioTopLite2 was used as a upper lever ontology in the project Good Ontology Designed (GoodOD), which provided an extensive guideline for good practices in ontology design in biomedical domain. It was also used as upper level ontology in the SemanticHealthNet project that ontologically integrates diverse semantic resources in order to increase interoperability between electronic health records and data [42].

2.1.4 Methodologies for Building Ontologies

Each development team follows its own criteria for the development of a ontology. However, the absence of methods and guidelines decreases the ontology shareability.

The common practice of switching directly from knowledge acquisition to implementation poses some problems: commitment and design criteria are implicit; domain experts and end users have more difficulties in understanding the formal ontology; direct coding of the knowledge acquisition is too abrupt and ontology developers may have more difficulties to extend or reuse such ontologies.

The ontology development process identifies tasks and activities that the developer should carry out, when building an ontology. These activities are presented in the IEEE 1074-1995 standard [43] that describes how the software development process should be structured. Since ontologies are software artefacts, they should also be developed according to the same standard, but slightly adapted to the ontology environment. This standard applied on the ontology development process comprises to the following activities:

- **Software Life Cycle** - the life cycle of a software should specify in which order the activities and tasks defined below should be performed. A methodology should specify at least one life cycle;

- **Project Management** - all the processes within this activity, recommended in the standard, should be applied also to ontology development. They are activities related to the project initiation, monitoring and ontology quality management;

- **Development** - concerns the production, installation, operation, maintenance and retirement from its use. These processes are divided in three stages:

  - **Pre-Development** - involves the study of the environment in which the ontology will be used, the possibilities of integration in other systems and a feasibility study;
  - **Development** - this includes the requirements, design and implementation process;
  - **Post-Development** - is related to the installation, operation, support and maintenance of an ontology.

- **Integral Processes** - these can include the training of the personnel responsible to the usage and maintenance of the ontology;

Depending on the size or purpose of the ontology some steps can be skipped. On the other side, if correctness and completeness of an ontology must be assured these activities should be performed during the whole process of ontology development. In the next sections are presented some methods which are already applied in the ontology development process. In Appendix A it is possible to identify the similarities or small deviances between methodologies to the IEEE standard. [44, 45]

Cyc KB Project method

Since the beginning, the main goal of the Cyc project [46] was to build a large knowledge base that contained a vast formal knowledge background that could be suitable for a variety of domains. In the last twenty years it has been building a knowledge base capable to represent a vast selection of common-sense knowledge in order to support unforeseen future knowledge representation and reasoning tasks.

The method behind this project is divided in three phases:

1. **Phase 1** - the codification of articles and pieces of knowledge where implicit common-sense knowledge is extracted manually;
2. **Phase 2** - the extraction of common-sense knowledge is aided by tools, however still mainly performed by humans;
3. **Phase 3** - similar to phase two, although the acquisition of knowledge is mainly performed by tools.

At this moment, the Cyc KB contains more than 2.2 million assertions used to describe more than 250.00 terms with around 15.000 predicates. The Cyc project is already available online providing tools like the OpenCyc that is a subset of the knowledge base Cyc KB and the Knowledge Server that includes a reasoning tool and others for acessing, utilizing and extending the knowledge base. [47, 48]

Uschold and King's method

The Uschold and King's method [44, 48–50] consists of four activities :

- **Identifying the purpose and level the formality** - this activity is focused on clarifying why the ontology is wanted and used for. This stage is important to know if the ontology should be built or not. If the developer can't find its purpose, he shouldn't proceed. After clarifying the purpose follows the decision about the level of formality. This level increases with the degree of automation in the tasks that the ontology will support. For example, if it is intended to support reusing and sharing of knowledge bases, then a more formal representation is needed.
- **Building the ontology** - this step concerns the development of the ontology. For this, Uschold and King's method gives 4 different approaches (Figure 2.9):
    1. The first approach is ignoring all the stages above and start the development by defining terms and axioms in an ontology editor. This is the best approach when only a prototype is intended.
    2. The second approach is more adequate for more simple and small ontologies. This approach already needs to have a proper identification of purpose and scope.
    3. The third approach starts by producing a prototypical ontology mainly structured in natural language with the terms and definitions of the domain. This process is mainly driven by hypothetical scenarios and competency questions. If this approach is taken, this informal document should be revised and evaluated before developing the formal ontology;

4. The last approach starts by identifying the formal within the informal set of terms using these to convert the informal competency questions into formal ones. Then specify the axioms and definitions that comprise the ontology.

- **Evaluation and Revision** - The evaluation and revision of an ontology can follow a more general or more specific criteria:
  - the general criteria for evaluation are the clarity, consistency and reusability of an ontology. However, this method is limited since there is no proper way to do this. Although, automated support to evaluate the ontologies by the the criteria is available.
  - the specific criteria involves techniques like manually checking the ontology against the identified purpose. These criteria is more appropriate for evaluating informal ontologies

Gruninger and Fox

The Gruninger and Fox method [51] is manly targeted to the development of ontologies in the enterprise domain. It is inspired by the problems that can be found with particular enterprises using them to define a motivation for an ontology. This motivation often have the form of problems that could not be addressed by existing ontologies. Intuitively, knowing what the problem is, possible solutions comes to mind. These solutions provide the first glance of an informal semantics for terminology included in the ontology.

Defining the motivation and possible solutions, requirements come next. These requirements are transformed in competency questions that an ontology must answer. These questions are a set of natural language competency questions that are used to determine the scope of the ontology or it's competency. This also provides an initial evaluation of the ontology that determines whether develop it or reuse existing ontologies.

The next step is to define the terminology. It will consist of concepts and definitions represented as axioms that should provide the necessary depth to restate the informal competency questions. If designing a new ontology, for every competency question there must be terms, relations and definitions on the ontology that should be able to intuitively answer the question.

After defining the terminology, the informal competency questions should be formally represented using the axioms of the ontology. These new formal questions will work as constraints on which axioms will be included. All the terms stated in these new formal competency questions should also be added to the terminology.

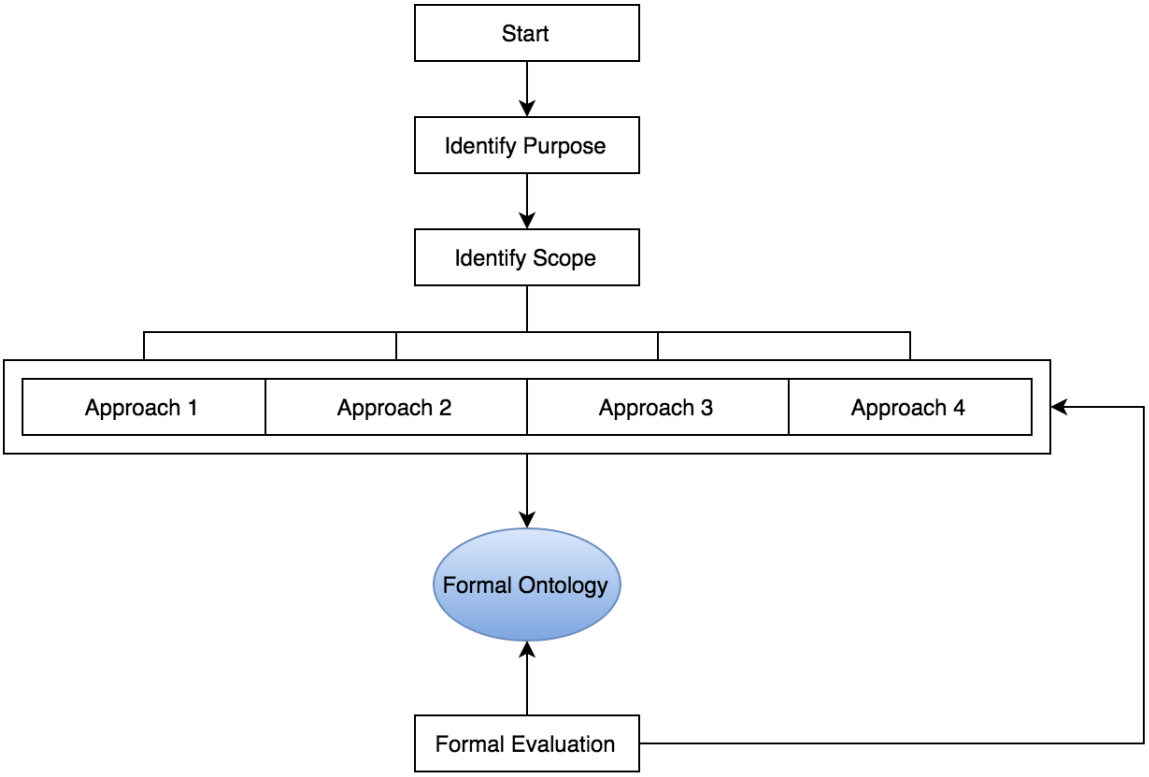


Figure 2.9: Flowchart of the methodology for building ontologies from the Uschold and King's methodology [50]

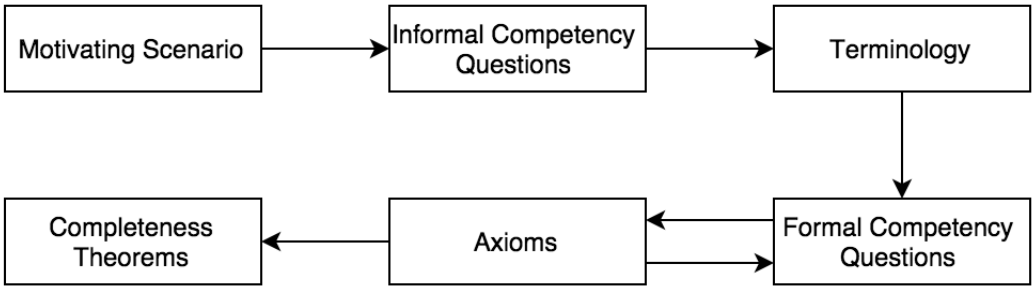


Figure 2.10: Gruninger and Fox procedure for ontology design and evaluation [51]

All the components are formally expressed in first order logic inhering it's intrinsic robustness. This model is also used as guide to convert informal scenarios into computable models [44, 48, 51].

KACTUS

The main objective of the KACTUS [52] project is to investigate the feasibility of reusing knowledge bases in complex technical problems and the role of ontologies in supporting it. This approach is conditioned by the number of applications being developed (bottom-up strategy). This means that when more applications are built, more general the ontology becomes. It all starts with building a knowledge base to a specific domain, further knowledge bases will be developed in order to be included in the existing ones. Therefore, when a new application is developed, the following steps are needed:

- **Specification of the application** - the first insight on what the ontology must represent;
- **Preliminary design based on relevant top-level ontological categories** - this process involves looking at previous ontologies developed that are possible candidates to be extended to this new application;
- **Ontology refinement and structuring** - this is made to assure that all the modules are not very dependent on each other and the most coherent as possible.

In summary, this new ontology can be built by reusing others and possibly integrated into ontologies of future applications. Applying this method along the

time, the ontology will evolve to represent the consensual knowledge for all the applications [45, 48].

Sensus

The ontology SENSUS [53] was developed in the Information Sciences Institute (ISI) and is used in natural language processing in order to provide a a conceptual structure for developing automatic translators. This ontology has more than 50.000 terms organized in a hierarchy according to their level of abstraction. The method behind the development of SENSUS is a top-down approach where domain ontologies are derivations of more broad ones. For this, the following steps were taken:

1. Identification of a set of *seed* terms that are relevant to the domain;
2. Then, this *seed* is linked by hand to a broader ontology;
3. All the concepts in the path between the *seed* terms and the upper ontology are included;
4. The terms that are relevant for the domain that are not yet included in the ontology are then added manually (this step is repeated until all terms necessary are represented);
5. Finally, for the nodes that have a large number of paths between them, the entire sub-tree under the node is added. This step is mostly done by hand since it requires a deep knowledge of the domain.

Using this method, knowledge-based applications for air campaign have been developed in a conjunct work with the ISI, ARPA Rome Planning and DARPA Joint Forces Air Component Commander. These include the Strategy Development Assistant, a tool that supports intelligent guided plan development. The method of using the same base ontology to develop ontologies in particular domains provides a high level of shareability [45, 48].

METHONTOLOGY

METHONTOLOGY [54] is a methodology developed in the Artificial Intelligence Lab from the Technical University of Madrid (UPM) for building ontologies either starting from zero or reusing other ontologies. This method enables the construction of ontologies at the knowledge level, that includes the identification of the ontology development process, a life cycle on evolving ontologies (Figure 2.11) and the techniques to carry out all the process.

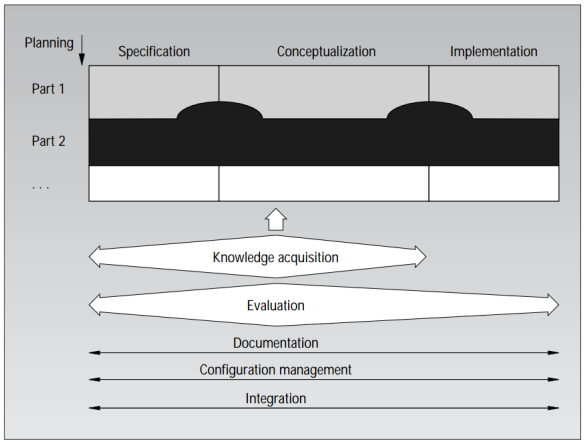


Figure 2.11: Methontology ontology development life cycle [54]

The development process includes a set of tasks that should be done during the ontology building process. These tasks are schedule in the life cycle of the ontology and they are:

- **Specification** - the goal of this task is to develop a prototypical document with the ontology's primary goal, purpose, granularity level and scope;
- **Conceptualization** - after most of the knowledge acquisition is done, the ontology developer must organize all this unstructured data.
- **Knowledge Acquisition** - the level of knowledge acquisition decreases with the increase of familiarity of domain and with the progression of the ontology development. The acquisition follows three stages:
  1. Meetings with experts to give an overview knowledge about the domain;
  2. Studying the documentation about the domain;
  3. After having a good insight on the domain, knowledge is acquired by looking from general knowledge to more particular one.
- **Integration** - During the development, some terms can be included in other ontologies. The target ontologies must be checked if they have been validated and verified. Since there is no automatic tool for this, the guidelines given by Asunción Gómez-Pérez [55] were followed.
- **Implementation** - Tools like Ontology Design Environment (ODE) [56] and the WebODE [57] provide support to the METHONTOLOGY.

The most important ontologies built based on this method were the CHEMICALS within the domain of chemical elements and crystalline structures; Environment pollutants ontologies with methods of detecting and evaluating different pollutants in various systems (soil, water, etc...); The Reference Ontology that works as yellow pages of ontologies that gathers, describe and links existing ontologies and others. This method follows a top down strategy where the most important terms are defined first and through specialization the most domain specific terms are obtained [45, 48, 54].

2.1.5 Protégé

The Protégé system is a framework being developed by the Stanford Medical Informatics group in the last two decades. It is an environment for knowledge-based systems development. Nowadays Protégé is the leading ontology editor, used by a world-wide community of about 50 000 users, who themselves are contributing to its evolution.

It was originally developed for representing frame-based ontologies within the Open Knowledge Base Connectivity (OKBC) protocol and its original goal was to minimize de role of the engineer in the ontology design process and consequently reduce the knowledge acquisition bottleneck. Currently, and in collaboration with University of Manchester, it evolved to represent ontologies based on DL in a variety of ontology languages.

Besides this new features, recent updates let Protégé export ontologies in a big variety of formats like RDF, OWL , XML and others. It has an open architecture which can be extended through plug-in components created by other developers [58, 59].

2.2 Description Logics

Description Logics are a family of languages for representing knowledge in a formal and structural way. It allows the representation of a model for a certain domain using a syntax constituted by classes, individuals, relations and the logical connections between them [**?**].

In the 70's , knowledge representation started to gain popularity and the approach was divided into two types [60] :

- **Logic Based** - where new facts can be intuitively deducted by *predicate calculus* - an axiomatized form of predicate logic;
- **Non-Logic Based** - this representation unfolds more cognitive notions mainly derived from human memory and execution of tasks. Network struc-

tures and rule-base representations are examples of this type of representation.

Between these two types, Non-Logic Based representations became more appealing from a practical viewpoint. However, they are designed for a specific problem or task where the knowledge is represented as structured data sources and reasoning is made by manipulation of them bringing some application limitations. On the other hand, on a logic based approach, the representational language uses a set of relational descriptions and variables to build predicates in which consistency and knowledge can be inferred by means of reasoning.

DL is the latest name on the knowledge representation family which is equipped with a formal, logic based semantic. It is called description logics because the important notions within the domain are described by concept *descriptions*. These can be represented by atomic concepts (unary predicates) and atomic roles (binary predicates) where the concept and role constructors are given by the particular DL [61].

Knowledge representation with description logics starts by first identifying and defining the most relevant concepts of the domain, this means its terminology, and uses this concepts to build the descriptions that specifies the properties of the individuals or objects in the domain. Unlike the other languages, DL are equipped with a logic-based semantics and reasoning. The latter allows to infer implicitly new knowledge from the already explicit knowledge representation. This new knowledge can be used by humans to structure and better understand the domain that is being represented with two types of classification [62]:

- **Concept Classification** - this type of classification is based on the principle of subsumption. This means that the classification is done by determining sub/superconcept relationships between concepts providing a terminology structured as an hierarchy. This provides useful information about the connection between concepts and also increases the performance of inference services;

- **Individual Classification** - this classification tries to determine if a certain individual is in fact an instance of a class. Knowing this, The properties of the individual are easily extracted.. Also, this may flag some inconsistencies on the knowledge base that forces the engineer to add or modify the knowledge base, thus contributing for the improvement of its own efficiency and robustness.

Inside of a knowledge base, it is possible to see a distinction between what is the general knowledge about the domain and the knowledge specific to the problem. Thus, a DL knowledge base is also divided in two components: a *TBox* and a

ABox. The first one contains the general knowledge, or the *intentional knowledge*, in form of a terminology, and is built by declarations that characterizes the general properties of concepts. The *TBox* is the definition of a new concept by using other previously defined concepts. For example:

$$\mathbf{Man} \equiv \mathbf{Person} \sqcap \mathbf{Male}$$

This is a logical equivalence that specifies both sufficient and necessary conditions to classify an individual as a **Man**. This type of definition is usually considered as a feature of DLs knowledge bases. Classifying in *TBox* basically means placing a new concept in the proper place. This is accomplished by checking the subsumption relation between all the hierarchy of concepts and the new concept.

In the other side there is the *ABox* that contains the assertions made about the individuals. These are also called as *membership assertions* since they refer to an individual being an instance, or member, of a certain concept. For example:

1: Man ⊓Person(MICHAEL)
2: hasFather(MICHAEL,CHARLES)

The first assertion states that Michael is a **Man**. Concerning also the assertion made before, it is possible to state that Michael is an instance of **Male**. This type of assertions are called *concept assertions*. The second assertion describes that Michael has a father called Charles. These kind of assertions are denominated as *role assertions*. The reasoning task in *ABox* is to check if a given individual is an instance of a specific concept [15, 60, 62].

DLs have demonstrated their practical usage by being implemented in many systems in various domains. Software Engineering was one of the first target domain. One example took place in the *ATT* that developed the *Classic* system that helped the software developer in finding out information about a large software system. Another domain is for configuration tasks. DLs are useful to support the design of complex systems by combining multiple components. On the biomedical domain, DLs proved to be very useful in the development of decision support systems besides the complexity of the medical domain [60]. In the ontological domain, DL is the core of the OWL 2 ontology language.

2.3 Medical Scope

2.3.1 The TNM Classification

The staging of malignant tumors is essential to the diagnosis, prognostic and management of cancer. The TNM was developed between 1943 and 1952 by Pierre

Table 2.1: Objectives of the TNM Classification [3, 5]

To aid the clinician in planning treatment
To give some indication of prognosis
To assist in evaluating the results of treatment
To contribute to continuing investigations of human malignancies
To facilitate the exchange of information between treatment centres

Denoix and was first published by the UICC in 1968 . This system is used for more than 50 years and with time and different editions, it has been evolving to meet the explosive growth in medical research, knowledge and information.

Today the TNM classification is in its 7th edition and is considered a worldwide tool for reporting the Extent of Disease (EOD) and prognosis of the outcome of patients with cancer evaluating the anatomic EOD. This system is the base of decision-making systems and clinical practice guidelines making it immeasurably useful.

The UICC established a set of objectives, presented in Table 2.1, in which they believe will maintain their prime motivation to have a broad and unified system where a common language is used and understood by clinicians in all specialities.

This system evaluates the attributes of the tumor including local growth and extension (T), spread to regional lymph nodes (N) and distant metastasis (M). T and N usually provide different levels with increasing severity, however for the distant metastasis, generally there is only a binary combination: 0 (no evidence) and 1 (evidence). Besides this complex classification, a series of different symbols exits to complement the classification increasing substantially its complexity. For example, each one of these levels can also have a suffix as a sub-classification (ex. T1a , N2b etc..) that can add specific information. This can become problematic because this varies in each tumor location. We can also have "X" when we face a clinical and pathological situation with incomplete or inaccurate information and "is" is needed for classifying a *carcinoma in situ*. The staging of the tumor corresponds to the combination of the three types of assessment.

There are two types of classification differing in the way the evidence was obtained:

- **Clinical Classification** - this consists as the pre-treatment clinical classification, which means that is based on evidence gathered before treatment and physical examination, and is designated as c. This is essential in the process of choosing and evaluating the proper therapy. This classification requires the use of the prefix "c" e.g cT1 , cN2;
- **Pathological Classification** - designated as pTNM , this classification is used to guide through further therapy and provides new data to the progno-

sis estimation. This is based on evidence acquired before treatment supplemented by new informations acquired from surgery and pathological examination. This type of classification must be identified by the prefix "p" in the TNM e.g pT2 , pN1.

The TNM system brings three additional advantages over other staging systems. This system is data orientated and has continuous improvement based on ongoing expert review of existing data. It is constituted by a comprehensive set of rules and definitions that guarantees the uniformity of use. Last , it is multidisciplinary and is suitable to all modern techniques of staging.

With all the different tumor sites and different classifications each one with its specific suffixes and prefixes makes the coding and interpretation very difficult and complex for the medical community. Because of that, the need of efficient and accurate information systems based on this system has been increasing [3–5,63–65].

2.3.2 TNM Classification for Colon and Rectum Tumors

The TNM classification for colon and rectum tumors (International Classification of Diseases for Oncology (ICD-O) C18-20) provides more detail than any other staging systems. The Colon and Rectum Staging is based on the depth of the tumor invasion on the wall of the intestine (T), the number of regional lymph nodes involved (N) and the presence and absence of distant metastasis (M). It is applied both types of classification however, to this particular cancer site, the pathological and clinical classification are based in the same rules. [66]

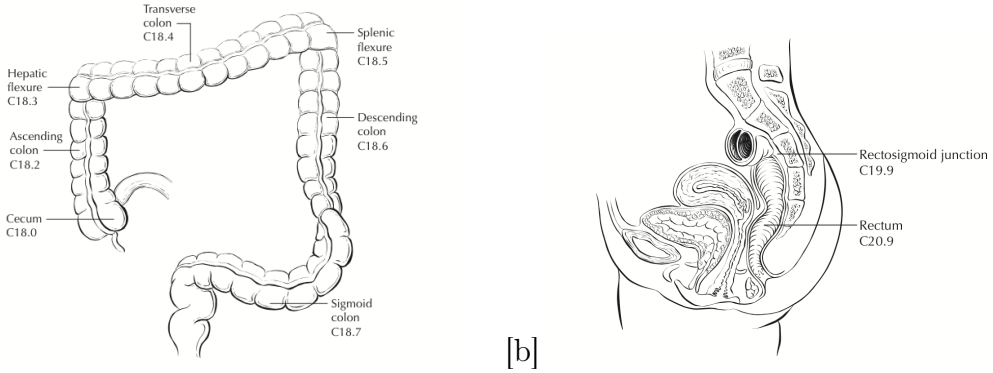
The colon and rectum classification is subdivided in some anatomical sites and subsites, each one with their respective ICD-O coding (see Figure 2.12). As the principal anatomic components we have the Colon (C18) , Rectosigmoid Junction (C19) and Rectum (C20). As subdivisions of Colon there are : Caecum (C18.0), Ascending Colon (C18.2), Hepatic Flexure (C18.3), Transverse Colon (C18.4), Splenic Flexure (C18.5), Descending Colon (C18.6) and Sigmoid Colon (C18.7). [3]

The regional lymph nodes are located near the major vessels that supply the colon and rectum, along the vascular arcades of the marginal artery and adjacent to the colon. They can be seen in the Figure 2.13. For the pN classification the only information needed is the amount of metastatic regional lymph nodes. Any non-regional metastatic lymph node is recorded as a distant metastasis

Definitions for Colon and Rectum TNM

The same classification is applied to both pathological and clinical classification [3, 66].

T - Primary Tumor



[a] [b]

Figure 2.12: Anatomical sites and subsites of colon [a] and rectum [b] [66]

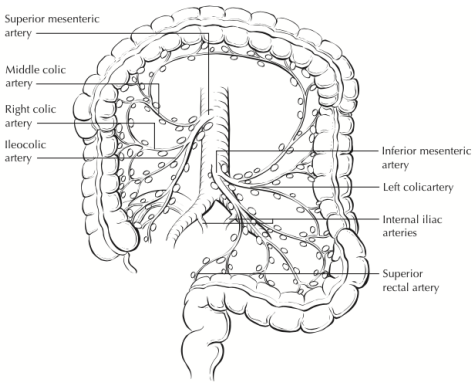


Figure 2.13: Identification and location of the regional lymph nodes [66]

The primary tumor classification, as mentioned before, is focused mainly in the confinement or invasion of the carcinoma inside the gastrointestinal wall. From the inner layer there are the mucosa, lamina propria, submucosa, muscularis propria, subserosa, pericolic and perirectal tissue and serosa (see Figure 2.14). For assessing the primary tumor the clinician gets his evidence from physical examination, imaging, endoscopy and/or surgical exploration. [3, 67, 68]

- **TX** - Primary tumor cannot be assessed
- **T0** - No evidence of primary tumor
- **Tis** - Carcinoma in Situ : intraepithelial or invasion of lamina propria
- **T1** - Tumor invades submucosa
- **T2** - Tumor invades muscularis propria
- **T3** - Tumor invades subserosa or non-peritonealized pericolic or perirectal tissues
- **T4** - Tumor directly invades other organs or structures and/or perforates visceral peritoneum
  - **T4a** - Tumor perforates visceral peritoneum
  - **T4b** - Tumor directly invades other organs or structures

N - Regional Lymph Nodes

The N classification concerns the number of metastatic regional lymph nodes and the presence or absence of tumor deposits. To determine this, the clinician proceeds with physical examination, imaging and/or surgical exploration.

- **NX** - Regional Lymph Nodes cannot be assessed
- **N0** - No regional lymph node metastasis
- **N1** - Metastasis in 1-3 regional lymph nodes
  - **N1a** - Metastasis in 1 regional lymph node
  - **N1b** - Metastasis in 2-3 regional lymph nodes
  - **N1c** - Tumor deposit(s), i.e. satellites, in the subserosa or in non-peritonealized pericolic and perirectal soft tissue withou regional lymph node metastasis

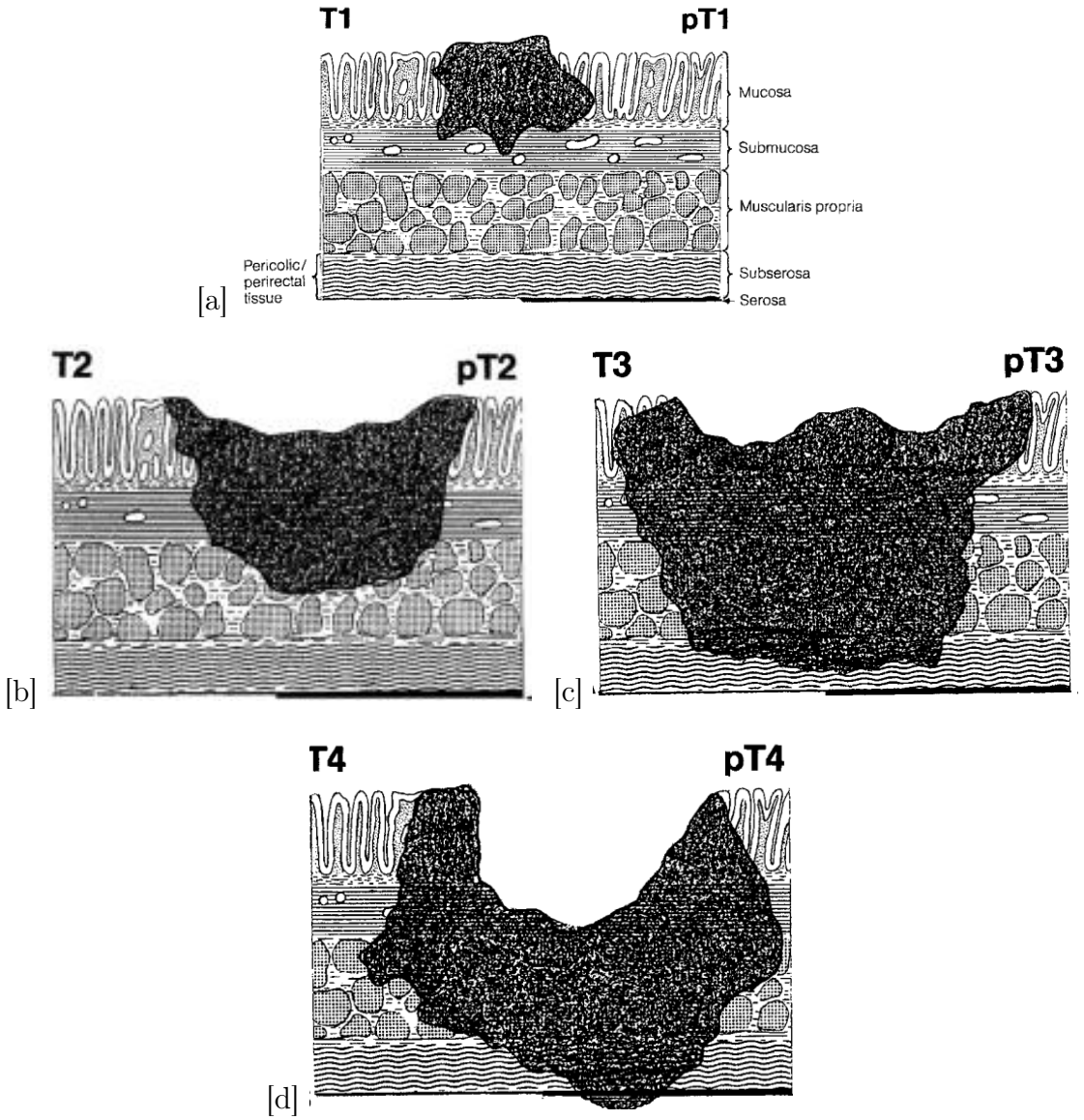


Figure 2.14: Representations of the primary tumor classification for colon and rectum tumor [67]

- **N2** - Metastasis in 4 or more regional lymph nodes
  - **N2a** - Metastasis in 4-6 regional lymph nodes
  - **N2b** - Metastasis in 7 or more regional lymph nodes

To a pathological classification recommends a minimum of 12 regional lymph nodes must be examined. However, if these nodes are negative and the number of examined lymph nodes is not met the classification maintains as pN0.

M - Distant Metastasis

Although metastasis of the colon and rectum tumor can occur in any organ, the lung and liver are the most common sites. Metastatic non-regional lymph nodes are considered as distant metastasis.

- **M0** - No distant metastasis
- **M1** - Distant metastasis
  - **M1a** - Metastasis confined to one organ (liver, lung, ovary, non-regional lymph node(s))
  - **M1b** - Metastasis in more than one organ or the peritoneum

Staging

The staging of the tumor is done after knowing the full TNM code . Each stage is associated to a different combination of classifications (Table 2.2). The higher staging value correspond to increasingly worse scenarios.

Table 2.2: Correspondence between the TNM classification and Staging [66]

Stage	T Classification	N Classificatio	M Classification
Stage 0	Tis	N0	M0
Stage I	T1, T2	N0	M0
Stage IIA	T3	N0	M0
Stage IIB	T4a	N0	M0
Stage IIC	T4b	N0	M0
Stage IIIA	T1, T2	N1	M0
	T1	N2a	M0
Stage IIIB	T3, T4a	N1	M0
	T2,T3	N2a	M0
	T1, T2	N2b	M0
Stage IIIc	T4a	N2a	M0
	T3, T4a	N2b	M0
	T4b	N1, N2	M0
Stage IVA	Any T	Any N	M1a
Stage IVB	Any T	Any N	M1b

Chapter 3

Methods

3.1 Ontology Development

During the development of this project, some methodologies for building ontologies that were proven to be very effective in the past. So, for this project it was decided to follow an adaptation and combination of the all methodologies studied. The first step is to determine the domain, scope and purpose to the ontology. Then the extraction of all the concepts in order to provide a terminology. After the extraction of the terminology necessary, follows the development of the knowledge base.

3.1.1 Specification

The domain of the ontology is the TNM Classification published by the UICC. The scope represented was restricted to the classification of colon and rectum tumors.

For every cancer site TNM-O provides a set of ontologies that can be imported to it. So, TNM-O works as a connecting hub containing a set of classes that will be common between all the other ontologies. This modular architecture forces the main TNM-O represent the most general definitions where the other modules can connect to. Therefore, this ontology should contain the concepts that are transversal to all cancer sites.

For this project, TNMCR-O was also developed, as modular ontology, that represents the classification rules of the colon and rectum tumors. This ontology should contain all the concepts and definitions for the classification of colorectal tumors as described in the Section 2.3.2. For that, it was necessary to provide a representation to:

- all the anatomic concepts related to the classification of colorectal tumors;

- qualities and possible values for the tumor;
- and the EOD;

All these representations were done in consideration with the upper level classes of the TNM-O. As a domain top-level ontology, BTL2 was used [69]. This ontology provides some predictability to the development of the TNM-O and its modules. This is intended since it facilitates the development and implementation of new modular ontologies to the main TNM-O. Besides a formal representation of the TNM classification of malignant tumors, this ontology should be capable of correctly classifying instance data.

3.1.2 Terminology

As reference, the seventh edition of the "TNM Classification of Malignant Tumours" published by the UICC and edited by L. Sobin et al. was used, where all the classification rules for all cancer sites are described very extensively with natural language.

Extraction of Anatomical Structures

The representation of anatomical structures was based on the FMA ontology [7]. Besides the ontology, it also provides a web framework Foundational Model Explorer (FME), this allows the user to search for any concept about the human anatomy and the relations between them.

Each modular ontology should import its own anatomical concepts and hierarchy. Some ontologies can share some anatomical categories or even entities in it's classification. Because of this, is necessary to design the base structure to the TNM-O to provide better categorization during the import of anatomical entities for each modular ontology.

Additionally, to provide a correct implementation, this process of building a general hierarchy must be iterative. It means that, each time a new module is developed, the anatomical tree in the TNM-O should be updated, which is not a problem since ontologies are very easily updated and maintained. By now, the anatomical representations in the TNM-O manly concern the colorectal tumor since this was the module also developed in this project.

In order to chose the best categories to add to the main ontology, it was necessary to take anatomical concepts from the colon and rectum classification and search them in the FME. In the Figure 3.1 there is a screenshot with the result given when searched the submucosa, which is a component in the wall of the colon and essential to the classification. Although, representing all the FMA ontology would increase the computational resources needed to use the TNM-O.

Therefore, it was necessary to do some pruning to the hierarchy tree in the FMA to better suit our needs. This process was mainly done by hand and consisted in removing some intermediary concepts between the top concept and the one needed to represent. Considering that this is part of an iterative process, some concepts that today are present in the TNM-O can in the future be removed and vice versa.

Qualities and Value Regions

The *Quality* and *ValueRegion* classes are defined by the BTL2 (section 2.8). Each *Quality* and *ValueRegion* for a specific type of tumor must be imported with the respective modular ontology. These classes correspond to certain features of the colorectal cancer that are relevant in the TNM classification rules, such as:

- *Confinement* - this quality concerns the confinement of the primary tumor within the wall of the colon and rectum. The respective values are *Confined* and *Invasive*;
- *Cardinality* - this quality represents a quantity. In this ontology it is used, for example, to represent the number of metastatic regional lymph nodes found;
- *AssessmentQuality* - this quality represents cases where the assessment was not done (*NoAssessment*) or no evidence was found (*NoEvidence*).

3.1.3 Classification

The goal of the TNM Classification is to properly classify malignant tumors. Analogue to this, the TNM-O plus the TNMCR-O should also be able to perform such classification. For this, it is necessary to attach each classification rule to the respective TNM code. Ontologically, each code corresponds to a *RepresentationalUnit*. *RepresentationalUnit*s were defined as a subclass of *InformationObject* which is provided by the BTL2.

3.1.4 Implementation

Both TNM-O and the TNMCR-O were implemented in the Semantic Web standard OWL-DL. This standard is a sub-language of the OWL strictly based on Description Logics and currently adopted by the ontology editor Protégé. For reasoning purposes was used the HermiT reasoner.

3.2 Software Development

One of the proposed objectives of this investigation project was to provide an ontology-driven automatic classifier that uses the TNMCR-O as knowledge base to classify colorectal tumors based on the TNM classification.

3.2.1 Requirements

Before coding the TNM Classifier it is needed to determine the target of the application and what is the purpose of it. There are two types of classification: the Clinical and the Pathological. The first one is done by a team of doctors in the hospital while the second is done by pathologists in pathology centres. During it's the development a visit to the Univirsitats Klinikum Pathology Centre in Freiburg was possible. While visiting the facilities, some problems where spotted on their registry. The data was inserted without any help of a framework designed for that purpose which could lead to some understandable inconsistencies. Therefore, this served as motivation to develop this tool for pathological classification.

Based on what was verified, the classifier should be able to:

- To assist the pathologist in the classification of malignant tumors based on TNM definitions;
- To help detect data inconsistencies in the clinical information systems and different sources.

In order to attend to both objectives the classifier provides a friendly-user Graphical User Interface (GUI) to guide the pathologist through the assessment of the tumor. In addition to this, the classifier allows classification of both instance and tabular data. Instance data input is made through GUI and for the tabular data a Comma Separated Values (*csv*) file is needed.

Like the TNM-O, the underlying ontology, this system was developed as a modular system. This way, it enables future extension of classification to other tumor sites. In this work the classifier was developed to provide the automatic classification of colon and rectum tumors since it was also the modular ontology developed.

3.2.2 Application Development

For the application development, JAVA programming language and following libraries were used: OWL [70] for parsing, rendering and manipulation of ontologies; HermiT application programming interface (API) [71] for consistency evaluation and classification of instance data and Opencsv JAVA library for manipulation of tabular data in *.csv* format.

OWL API

The OWL API is implemented in JAVA and has been available since 2003 [70]. It's main purpose is to allow engineers to make the bridge between the OWL ontologies and the domain applications.

It contains a set of classes and interfaces providing the developer the necessary tools to render, parse, reason, structure and manipulate ontologies. The main interface OWLOntology works as access point to the axioms in an ontology. As instance of this interface, the OWLOntologyManager allows actions like creating, loading, changing and saving ontologies.

This API was used as a bridge between the two main parts of the classifier, the JAVA application code and the ontology guaranteeing the independence of both. This way the ontology is responsible for providing all the knowledge necessary while the application just has to manipulate or query the knowledge base. Additionally, the ontology allows easy knowledge maintenance without making changes in the source code of the application.

The OWL API is implemented in JAVA and is available as open source under an LGPL licence [70].

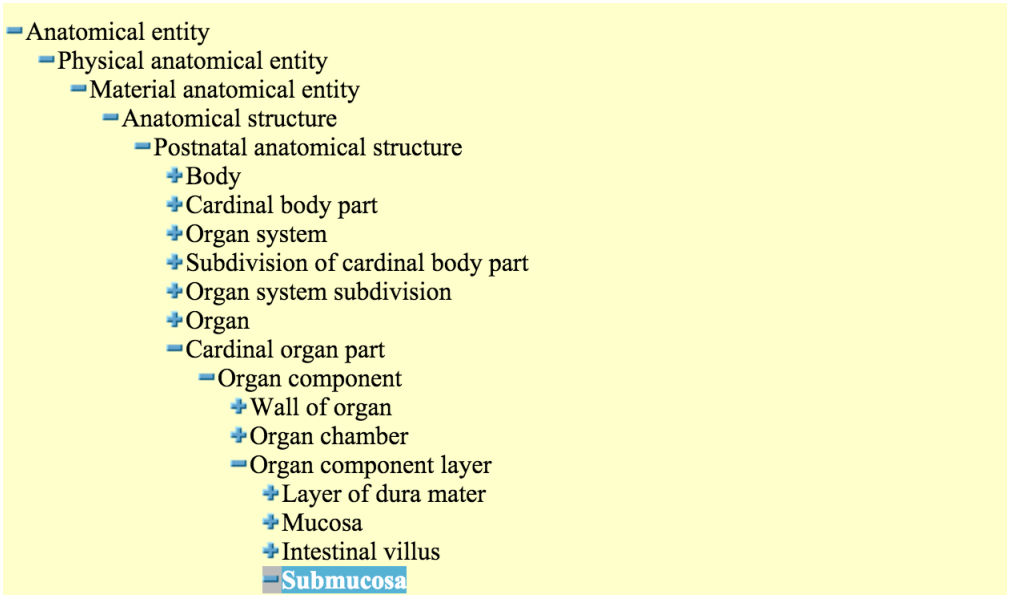


Figure 3.1: Screenshot of the FMA Explorer when searching for the concept *Submucosa*

Chapter 4

Results

4.1 TNM Ontology

Two versions of the TNMCR-O were already developed concerning the version 6 and 7 of the TNM classification. Version 7 contains the same rules as the previous version plus some new ones. Thus, this work will focus on the Version 7 of the ontology (*TNM-O colorectal 7.owl*) mainly the representations of the classification rules and the necessary qualities and respective values.

The whole ontological system is composed by 231 classes and 993 axioms where 489 of them are logical axioms including, but not only, 385 *SubClassOf*, 27 *EquivalentClasses*, 34 *DisjointClasses* axioms. The consistency and performance of the ontology were checked by the HermiT reasoner which revealed the need of good computational resources due to the high complexity of the system. The TNM ontologies complete set is available at [72].

4.1.1 TNM Structure

A medical expert, when performing the diagnosis of a cancer patient, only uses the TNM classification to a particular cancer site. Therefore, this justifies the choice for having a modular architecture for this ontology. Without it, classifying instances or querying the ontology would require a lot of time and computational resources. Thus, this modular architecture of ontologies was created with the aim to increase the efficiency and performance of the entire system.

In Figure 4.1 there is an extract of the main structure of the TNM-O. Until the *StructuralBiologicalEntity* class there are the hierarchy of concepts from the BTL2 upper domain ontology.

AnatomicalStructure is one of the top class of the FMA ontology and represents every *StructuralBiologicalEntity* but restricted to the human body. The ontology is divided in two main groups: the *BodyPortion* plus the *BodyPart* for representing

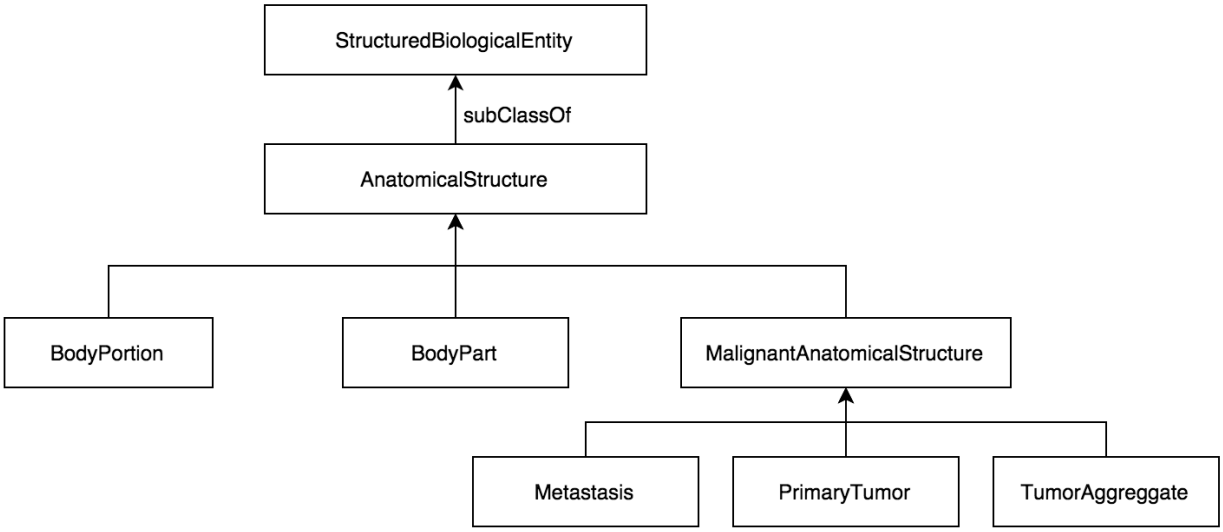


Figure 4.1: Main structure of the TNM-O

the anatomical structures and the *MalignantAnatomicalStructure* that will contain all the concepts related to the tumor. The latter is also divided in three other categories :

- *Metastasis* - this class represents any type of metastasis such as a distant metastasis, metastatic regional lymph node and others;
- *PrimaryTumor* - this class will contain both clinical and pathological classification rules definitions related to the primary tumor;
- *TumorAggregate* - this class contains all the definitions related to the Regional Lymph Nodes and Distant Metastasis classification, also for both clinical and pathological. It is called the aggregate because it contains all the structures that are a consequence of the primary tumor existence.

Another crucial point to the classification is the TNM coding. Each tumor site has its specific coding, so it is important link the central and modular ontology. Figure 4.2 shows the classes used for this purpose. *InformationObject* is a BTL2 class that defines pieces of information that, in this domain, are the TNM codes used for classification. Each code is represented as a *RepresentationalUnit* and they are imported with the respective modular ontology.

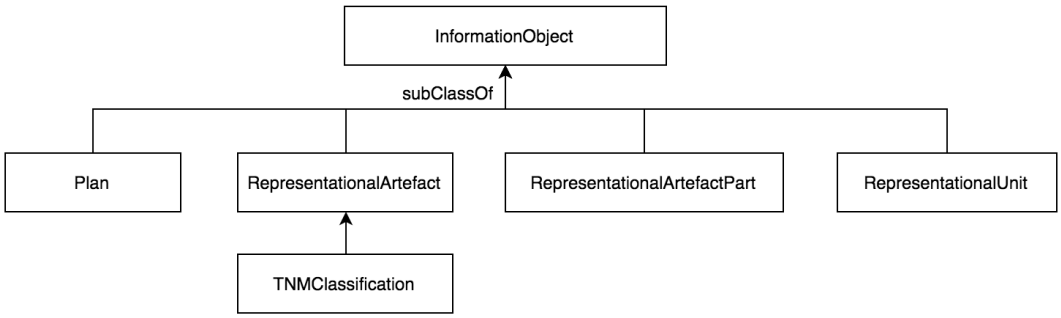


Figure 4.2: Hierarchies of classes for including the *RepresentationalUnits* of each modular ontology

4.1.2 Representational Units

The representation of the TNM classification can be break down in three major units each for a different type of assessment: T for the Primary Tumor, N for the Regional Lymph Nodes and M for the Distant Metastasis. Every TNM rule is represented by a separate class and for each one there is a *SubClassOf* axiom which is a **btl2:isRepresentedBy** relation with a *RepresentationalUnit* that codifies the respective TNM code. Ontologies representing the cancer sites will have their own *RepresentationalUnits* which will be imported with it.

InvasiveTumorOfMuscularLayerOfColonAndRectum subClassOf
ColonAndRectumTumor and
btl2:isRepresentedBy some (*ColonRectumTNM_T2*
or *ColonRectumTNM_pT2*) and
btl2:isRepresentedBy only (*ColonRectumTNM_T2*
or *ColonRectumTNM_pT2*)

Classification of *Individuals* is a promising use of this ontology. The reasoning task determines to which class a certain *Individual* belongs to. If it is an instance of any class that defines a classification rule, the *RepresentationalUnit* attached is the classification. In Figure 4.3 there are all the RepresentationalUnits needed to classify clinically the colon and rectum tumor. Exceptionally, in the colorectal cancer, both pathological and clinical classification share the same rules. The pathological *RepresentationalUnits* are distinct from the clinical by the prefix "p".

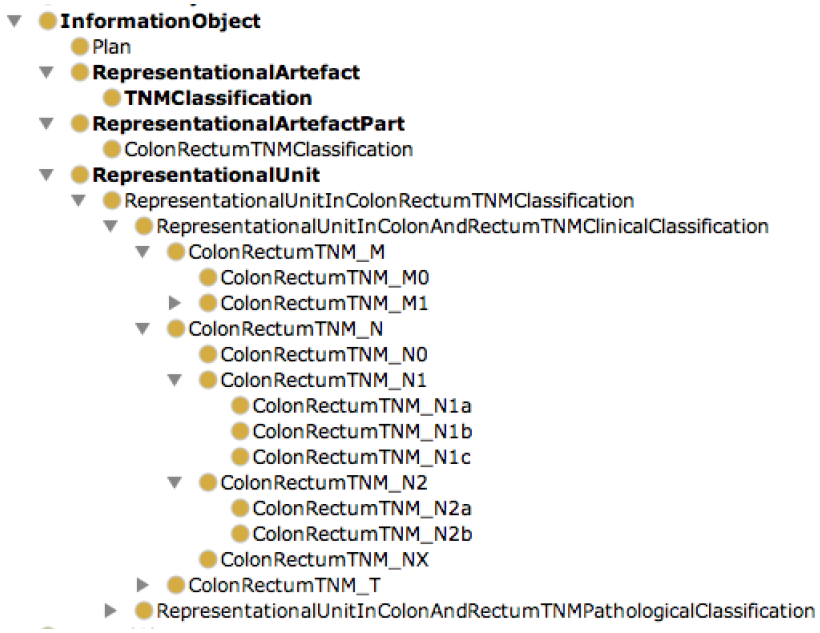


Figure 4.3: Example of hierarchy and classes of all *RepresentationalUnits* imported by the TNM Colon and Rectum ontology

4.1.3 Tumor Aggregate

To provide a classification to the regional lymph nodes and the distant metastasis it is necessary to know the *PrimaryTumor*. Therefore, *TumorAggregate* was created in order to represent the aggregate between the primary tumor and all the affected structures, regional lymph nodes or metastasis :

TumorAggreagate subClassOf
MalignantAnatomicalStructure and
btl2:hasPart some *PrimaryTumor*

4.1.4 Quality and ValueRegion

Figure 4.4 shows the qualities and respective value regions represented in the TNMCR-O.

In order to assign certain quality and value region to a certain entity, we need two *Object Properties*: **btl2:isBearerOf** and **btl2:projectsOnto**. The first one connects the class to the quality while the second refers to its value. For example:

btl2:isBearerOf some (*Confinement* and
(**btl2:projectsOnto** only (*Invasive*))

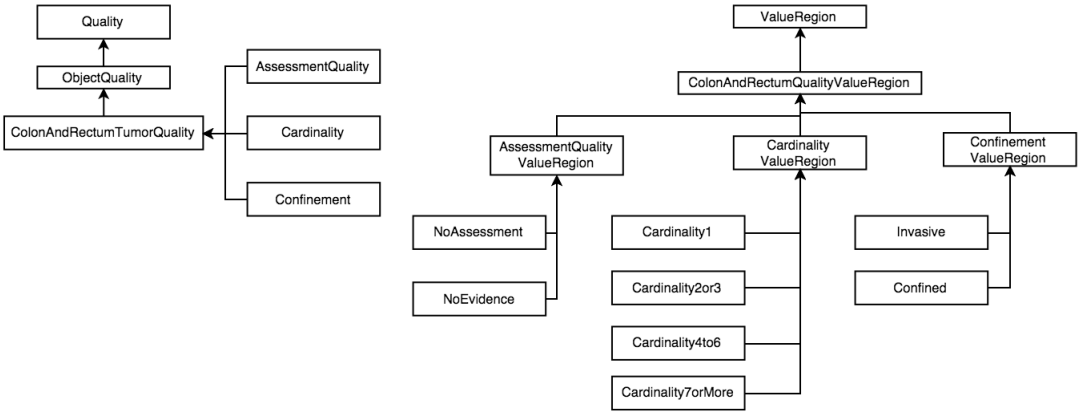


Figure 4.4: Quality and ValueRegions Classes of the TNMCR-O

All qualities and the correspondent values are subclasses of *ColonAndRectumQuality* and *ColonAndRectumQualityValueRegion* respectively. These two classes are the ones that make the bridge between the TNMCR-O and the TNM-O.

4.1.5 Representation of Anatomical Structures

Figure 4.5 depicts the current state of the anatomical hierarchy of concepts in the TNM-O.

For the TNMCR-O, every anatomical concept in the terminology was searched in the FME. Although not every term has the same label as the correspondent in the FME. However, the ontology allows the developer to add metadata to each concept where is possible to attach comments with all the changes and correspondences made, so that every user understands what was done. In Figure 4.6 there is the same tree of concepts as in Figure 4.5 but now with the TNMCR-O imported. Although, the hierarchies of concepts presented by the FME are too extent and contain a very large of anatomical concepts that are not necessary to represent between the top concept and the pretended concept. Therefore, a pruning of the hierarchies was done that resulted in the hierarchy presented in Figures 4.6 and 4.5.

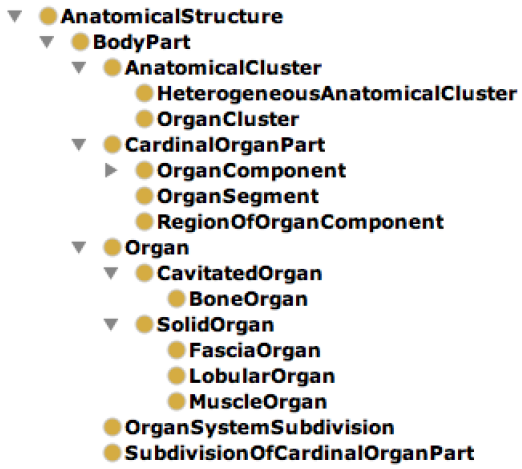


Figure 4.5: TNM-O current hierarchy of anatomic related classes

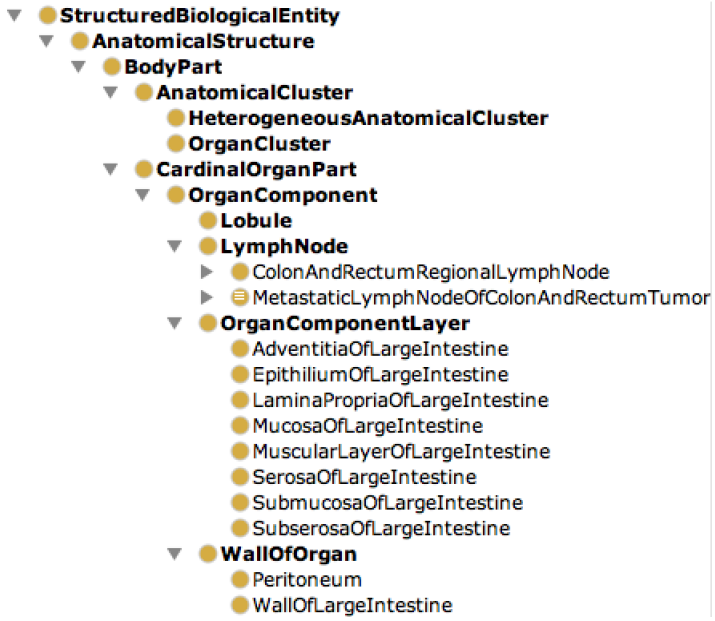


Figure 4.6: Hierarchy of anatomical classes of TNM-O when TNM Colon and Rectum Ontology imported

4.1.6 Representation of the Primary Tumor

The *PrimaryTumor* is a subclass of *MalignantAnatomicalStructure* in the TNM-O. For the TNMCR-O the *ColonAndRectumTumor* was created that makes the logical connection between both ontologies. This class represents all the colorectal classification rules regarding the primary tumor. Each classification rule contains axioms that represent the tumor confinement with respect to neighboring organs and/or the quality of the assessment (no assessment, no evidence or carcinoma in situ). For example:

InvasiveTumorOfSubserosaOfColonAndRectum EquivalentTo
ColonAndRectumTumor and
(**btl2:isBearerOf** some (*Confinement* and
(**btl2:projectsOnto** some *Invasive*))) and
btl2:isIncludedin some
(*AdventitiaOfLargeIntestine* or *SubserosaOfLargeIntestine*)

Qualities and respective *ValueRegions* are added to the axioms as showed in section 4.1.4. Besides that, it is necessary to specify which layer of the gut wall is invaded by the tumor. The previous axiom is a full representation of a invasive primary tumor on the subserosa layer of the gut wall defined by the class *InvasiveTumorOfSubserosaOfColonAndRectum*. Nevertheless, all the other classes that describe the primary tumor follow the same pattern, with the exception when there is no evidence or assessment.

When describing a classification rule it is also required to connect them to the correspondent *RepresentationalUnit*. This will create the bridge between the rule and the TNM code, using the **btl2:isRepresentedBy** relation:

InvasiveTumorOfSubserosaOfColonAndRectum subClassOf
btl2:isRepresentedBy some (*ColonRectumTNM_T3*
or *ColonRectumTNM_pT3*)

In Figure 4.7 there is a graph with all the classes and relations necessary to represent a tumor that invaded the muscular layer of the gut wall, the *InvasiveTumorOfMuscularLayerOfColonAndRectum* class. It's possible to identify the major branches in the hierarchy necessary to the classification: *BodyPart* to define the extension of the invasion of the tumor, *MalignantAnatomicalStructure* that contains all representations of the classification rules and the *Quality* with respective *ValueRegion* to identify the tumor as invasive or confined.

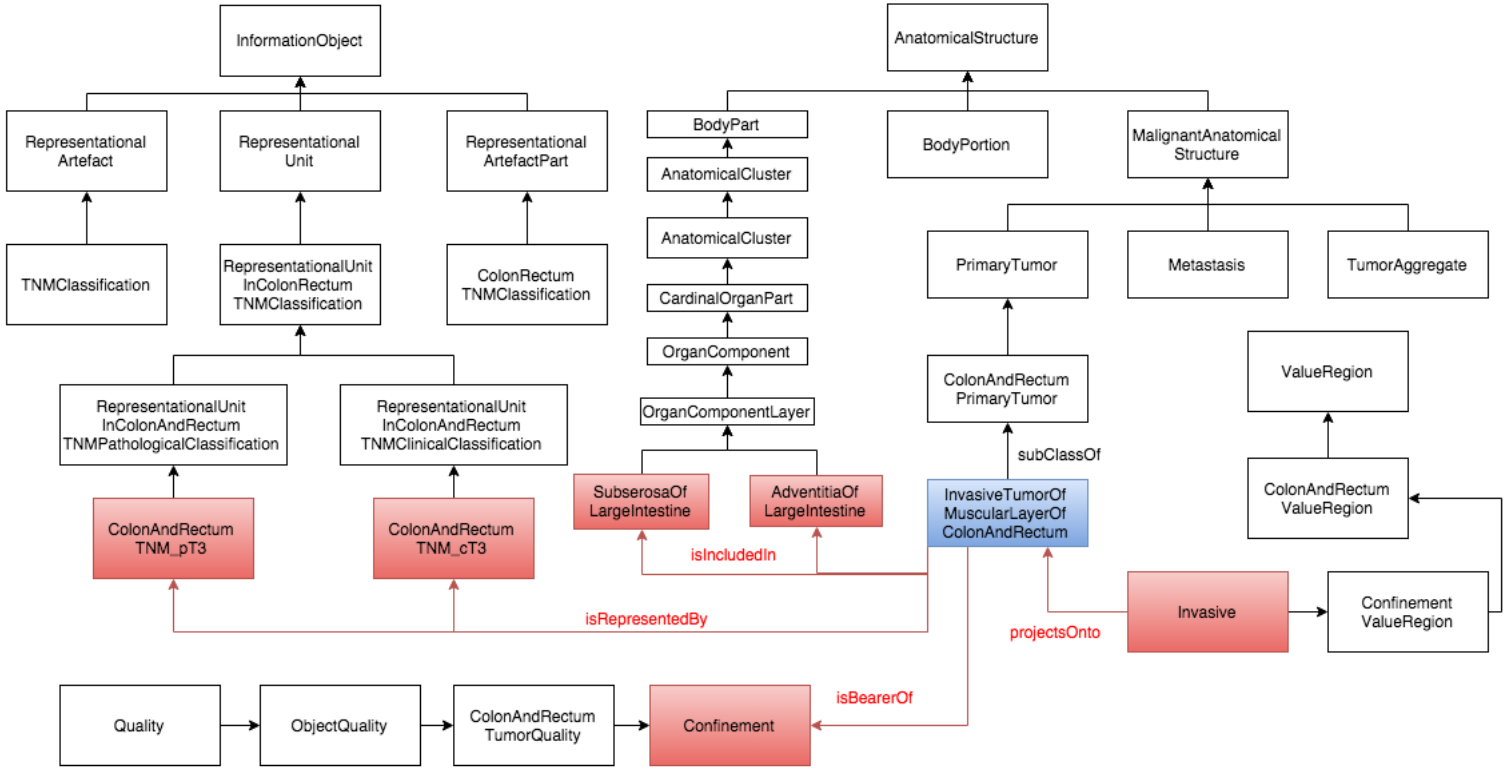


Figure 4.7: Graph of the patho-anatomical structures represented by a T3/pT3 representational unit of the TNMCR-O

4.1.7 Representation of Regional Lymph Nodes

The representational unit N of the classification is relative to the number of the metastatic regional lymph nodes. However, if none was found during the evaluation, the physicist can determine the presence of tumor deposits in the subserosa, non-peritonealized pericolic or perirectal soft tissue.

The anatomy of the metastatic lymph node was represented as it follows:

MetastaticLymphNodeOfColonAndRectumTumor EquivalentTo
LymphNode and
btl2:hasPart some *MetastasisOfColonAndRectumTumor*)

Additionally, this assessment only evaluates the metastatic lymph nodes which are located in the regional areas of the colon and rectum. They were represented as:

MetastaticRegionalLymphNodeOfColonAndRectumTumor EquivalentTo
ColonAndRectumRegionalLymphNode and
MetastaticLymphNodeOfColonAndRectumTumor

The aggregate of the infiltrated regional lymph nodes and the primary tumor are represented as one entity (*TumorAggregate*). The number of metastatic regional lymph nodes is described with the *Cardinality* quality and respective *ValueRegion* :

TumorOfColonAndRectumWith4to6MetastaticRegionalLymphNodes EquivalentTo
TumorOfColonAndRectumWith4orMoreMetastaticRegionalLymphNodes and
(**btl2:isBearerOf** some (*Cardinality* and
(**btl2:projectsOnto** some (Cardinality4to6))))

For classification purposes it is also necessary to add the respective *RepresentationalUnit* with the code:

TumorOfColonAndRectumWith4to6MetastaticRegionalLymphNodes subClassOf
btl2:isRepresentedBy some (*ColonRectumTNM_N2a*
or *ColonRectumTNM_pN2a*)

The Figure 4.8 show the graph that represents the classes and the relations between them in order to represent the *TumorOfColonAndRectumWith4to6Metastatic RegionalLymphNodes*.

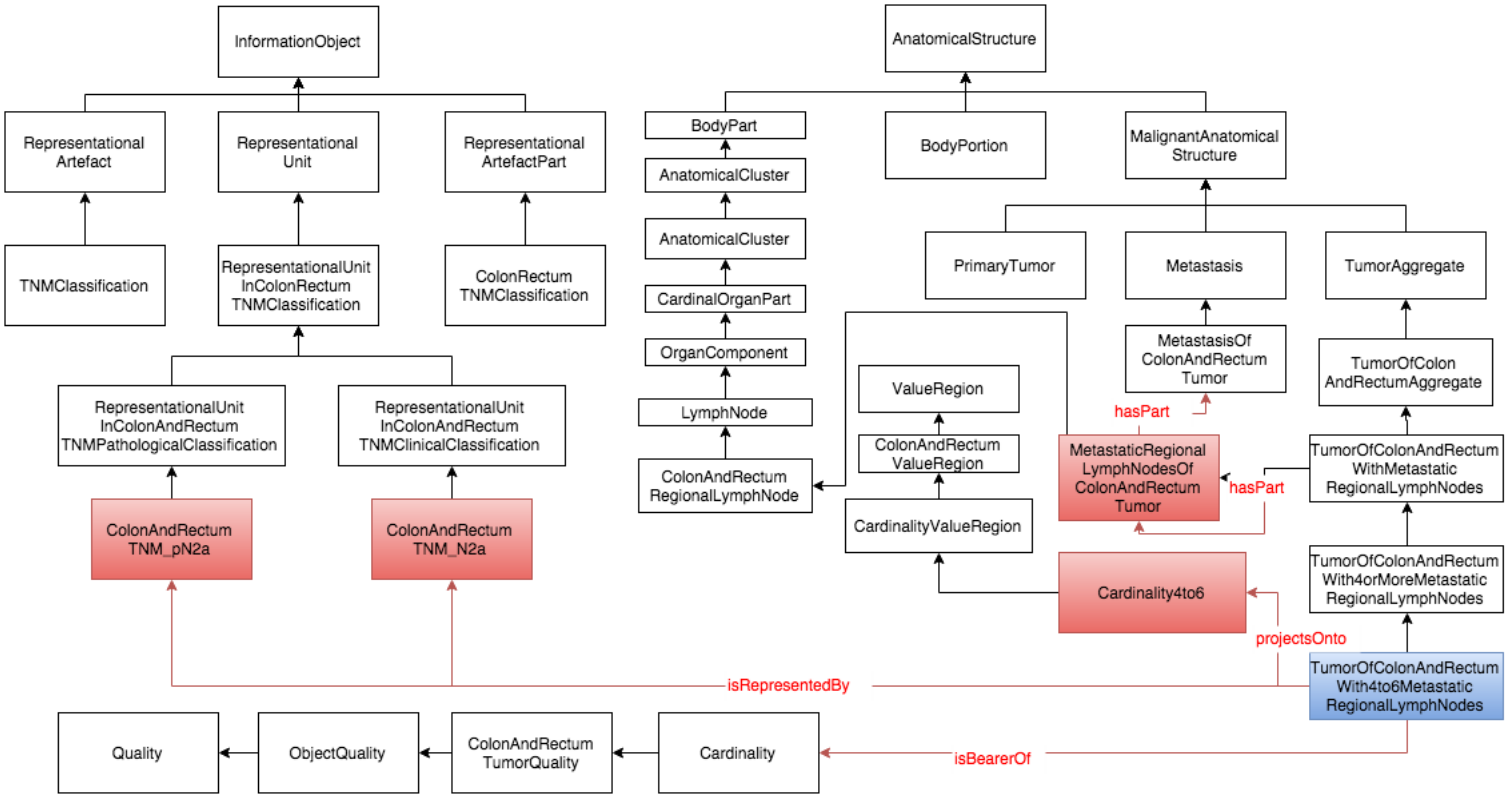


Figure 4.8: Graph of the patho-anatomical structures represented by a N2a/pN2a representational unit of the TNMCR-O

4.1.8 Representation of Distant Metastasis

The assessment for the representational unit M of the TNM classification system evaluates the presence or the absence of distant metastasis. If metastasis are found it is also required the number of organs that contain metastasis and whether or not the peritoneum is affected. A non-regional metastatic regional lymph node is a distant metastasis.

The definition of distant metastasis is :

DistantMetastasisOfColonAndRectumTumor EquivalentTo
MetastasisOfColonAndRectumTumor and
(not (**btl2:isIncludedIn** some
ColonAndRectumRegionalLymphNode)) and
btl2:isIncludedIn some *BodyPart*

is the combination between the primary tumor and respective distant metastasis. So a tumor with distant metastasis is represented as:

TumorOfColonAndRectumWithDistantMetastasis EquivalentTo
TumorOfColonAndRectumAggregate and
(**btl2:hasPart** some *DistantMetastasisOfColonAndRectum*)

Like any other representation of a classification rule from the TNM system, it is required to attach the respective *RepresentationalUnit* to the description :

TumorOfColonAndRectumWithDistantMetastasis subClassOf
btl2:isRepresentedBy some (*ColonRectumTNM_M1*
or *ColonRectumTNM_M1*)

The Figure 4.8 show the graph that represents the classes and the relations between them in order to represent a *TumorOfColonAndRectumWithDistantMetastasis*.

4.1.9 Staging

To determine the stage of the tumor it is needed the complete TNM classification. Table 2.2 in the section 2.3.2 contains the definitions for each stage.

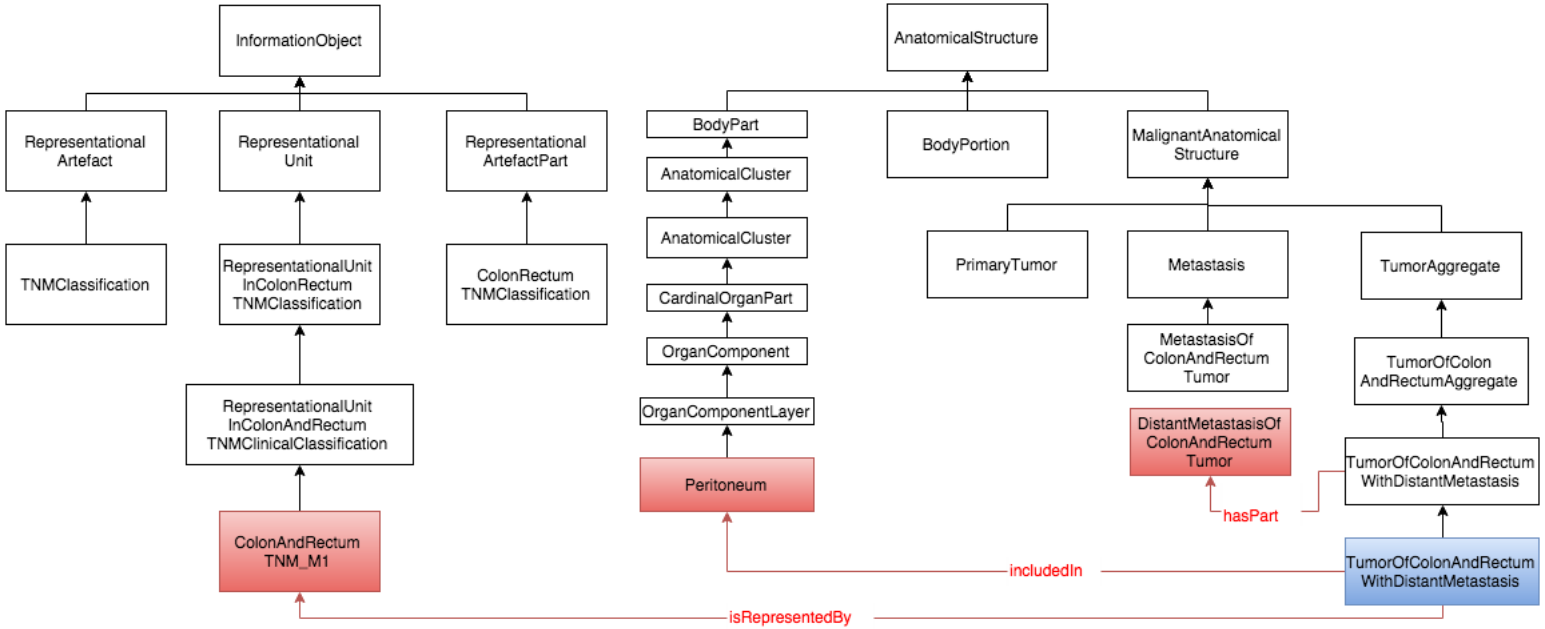


Figure 4.9: Graph of the patho-anatomical structures represented by a M1 representational unit of the TNMCR-O

Ontologically, each stage is represented as a *RepresentationalUnit* and contain the axiom that correctly represent the TNM code combination correspondent. For example:

ColonRectumStage_I EquivalentTo
RepresentationalUnitInColonOrRectumTNMStaging and
btl2:isRepresentedBy some (*ColonRectumTNM_M0*
and (*ColonRectumTNM_N0* or ColonRectumTNM_pN0)
and ((ColonRectumTNM_T1 or ColonRectumTNM_pT1)
or (ColonRectumTNM_T2 or ColonRectumTNM_pT2))))

The example above is the representation of the Stage I of the colon and rectum tumors that corresponds to a (p)T1 or (p)T2 plus (p)N0 and M0 classification.

4.2 TNMO-Classifier

Architecture

This section describes the architecture of the classification system as shown in Figure 4.10. The main classes are:

- Ontology_Handler - Analogue to the TNM-O, the classifier was developed in a modular way. Thus, this class works as the connecting point to all the modules. It contains methods that allows loading ontologies, read classes, adding and erasing individuals, setting up the reasoner and starting the classification process.

- ColonAndRectum - When the colorectal ontology is loaded, it automatically instantiate this class. This is responsible for preparing and processing the data that is inserted by the pathologist. It provides the necessary methods to translate the information in the respective axioms. These axioms will be connected to *Individuals* in the ontology for further classification;

- AnaliseCsv - This class is responsible for reading the *.csv*. After reading the data all the processing is performed by the class above.

The GUI guides the pathologist through the process of the classification. The first step is to choose the ontology correspondent to the tumor site that is being assessed. Doing this, the ontology becomes the knowledge base and the GUI for this specific tumor becomes available to the input of data by the pathologist.

After the input, the data is processed by the classifier with the classes and methods explained above. All the ontology management is made by the OWL

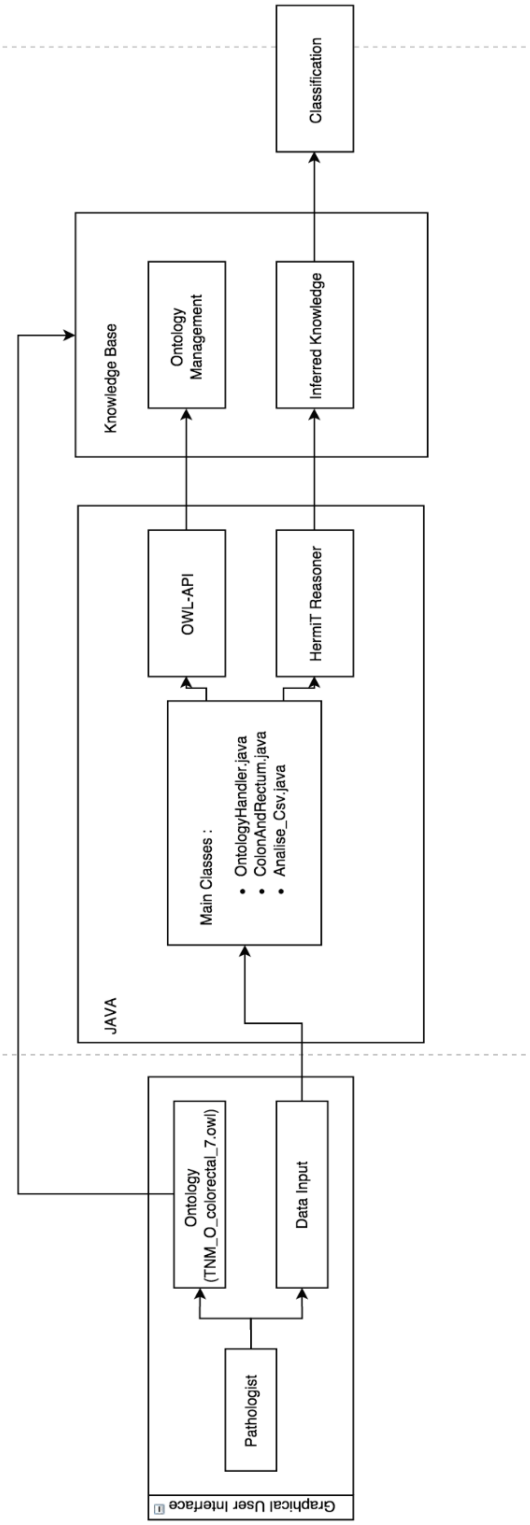


Figure 4.10: Technical architecture of the classifier application

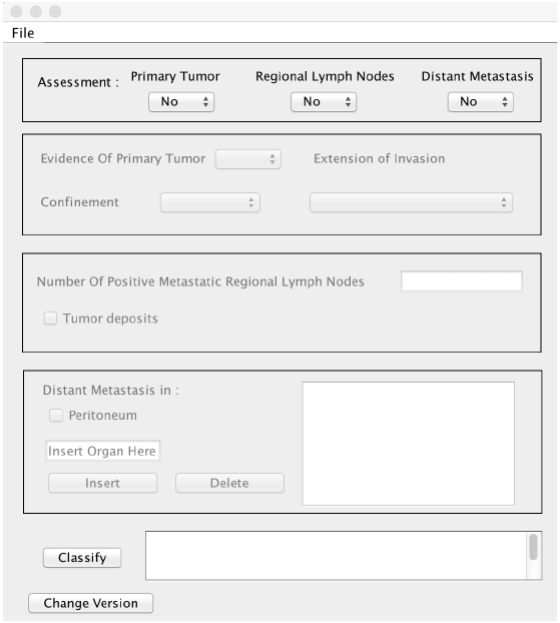


Figure 4.11: Graphical User Interface of the TNM-O Classifier

API while the HermiT reasoner evaluates the consistency of it proceeds with the classification.

Graphic User Interface - GUI

This classifier provides a GUI for manual data input (see Figure 4.11). Due to it's modular architecture, the layout is dependent on the tumor site selected since each site has it's own set of rules. In this implementation, the GUI presents the user the exact fields for data input in coherence with the rules of the colon and rectum TNM classification.

The GUI was divided in 5 regions:

- Assessment - When the GUI appears, all the components are disabled except the top three ones. These ones are where the pathologist is going to inform if the correspondent assessment was made or not. When the user states that there was a certain assessment, the correspondent components for adding information becomes available;

- Primary Tumor - This region corresponds to the T (or Primary Tumor) classification. The first step is to specify whether or not the tumor is invasive. In case of invasiveness, then it's necessary to indicate the extension of the tumor.

- Regional Lymph Nodes - Regarding the N classification the number of the metastatic regional lymph nodes is asked. Besides this, it is important to know if any tumor deposits exists in the subserosa or in non-peritonealized pericolic or perirectal tissues;

- Distant Metastasis - The M classification is based on the existence or absence of distant metastasis in the organism. In case they exist, it is necessary to know in which organs. Specially in the peritoneum since this has a different classification;

- Classification - This region has a button to start the classification based on the information given to the system. As a result, the calculated TNM code is displayed to the pathologist.

The main goal of this GUI is to guide the pathologist through all the process of assessing the tumor. As a result, it prevents mistakes and inconsistencies that can be made during the data input.

4.2.1 Automatic Classification

This system provides classification for two types of data input:

- Manual data input - the pathologist uses the graphical user interface;

- Tabular data - the data is on a .csv file format where each row corresponds to one assessment.

Although, in the low level, the classification process follows the same pattern for both types. There are two different classes in charge of doing a preparation of the data in order to be classified. The process can be seen in the Figure 4.12.

The classification starts by reading the data and determining which type of data is, tabular or instance. This data is processed by the class responsible to convert it into logical axioms. Building these axioms converts this information to be processable by the ontology in a logical way. The axioms are connected to the correspondent *Individual* that represents each assessment:

- *PrimaryTumor* - T classification

- *RegionalLymphNodes* - N classification

- *DistantMetastasis* - M classification

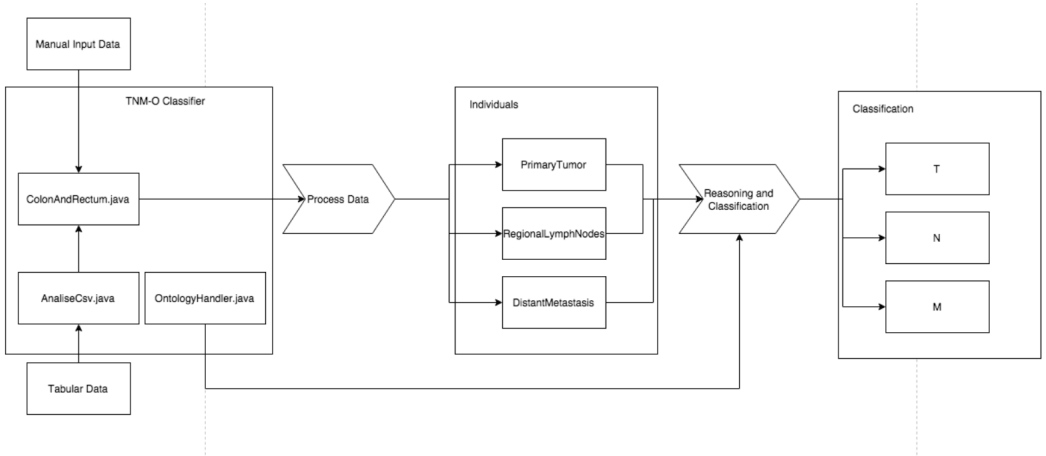


Figure 4.12: Classification process

After connecting each *Individual* to the respective axioms, they are added to the ontology. Then, the reasoner is started and the new knowledge is inferred. These *Individuals* will become members of the class that corresponds to a classification rule (Figure 4.13).

The TNM code for each assessment is the *RepresentationalUnit* obtained from the class of the defining rule. The final classification is the combination of the three *RepresentationalUnits*.

Even though the automatic classification is very similar between the two types of input data, it is important to show the distinctions between them.

Classification from GUI

During the data input to the GUI, the pathologist has all the options needed to perform a correct assessment. These options are imposed, which is intended because it provides a high level of uniformity of concepts. This uniformity allows less processing of data, higher efficiency and prevents inconsistencies that can be made by the user.

In Figure 4.14 it is possible to see how the axioms are built during this the classification process.

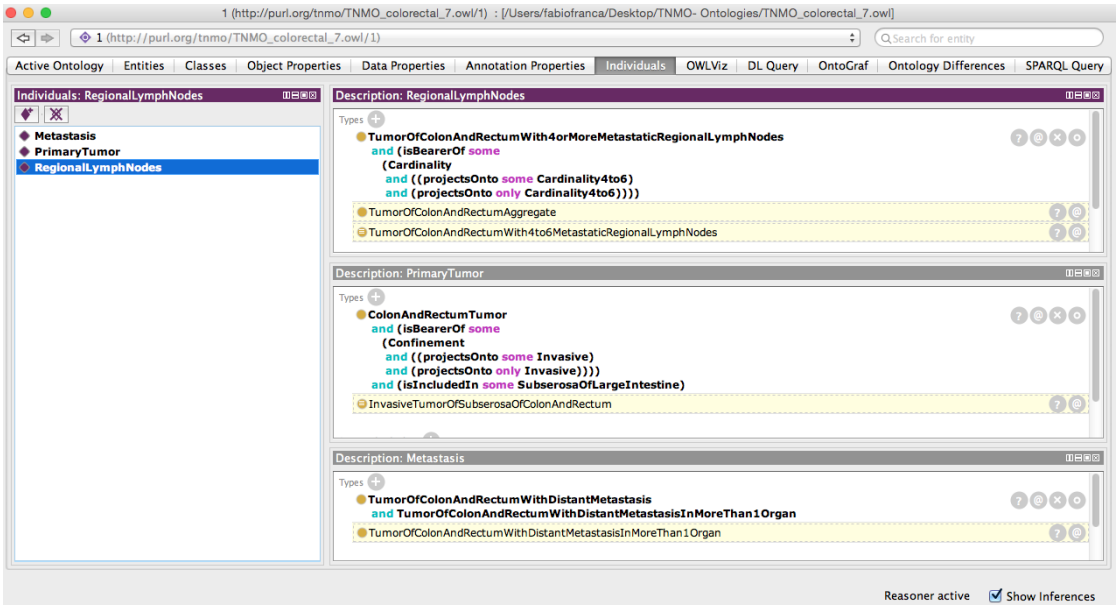


Figure 4.13: Screenshot of the ontology editor *Protege* with the TNM-O loaded during the classification process

Table 4.1: Each assessment criteria present on the *.csv* file for tabular classification

	Criteria			
Primary Tumor	Assessment	Evidence	Extension of Invasion	
Regional Lymph Nodes	Assessment	Evidence	Nr of Regional Lymph Nodes Examined	Nr of Positive Metastatic Regional Lymph Nodes
Distant Metastasis	Evidence	Nr of Metastasis	Metastasis in Peritoneum	

Classification from Tabular Data

As said before, tabular data must be given in a *.csv* file format. But in order for this to work a template must be given to the pathologists where they can write all the data recovered by their assessment necessary for a proper classification by the application. The criteria used can be seen in Table 4.1.

The classifier extracts the information from each row and builds the axioms for each assessment for further classification.

4.3 Evaluation

Real data was anonymously provided by the by the Institute of Clinical Pathology in Freiburg, Germany in two datasets. The first dataset had 382 records documented by the pathologist during the assessment of the tumor. The second ,

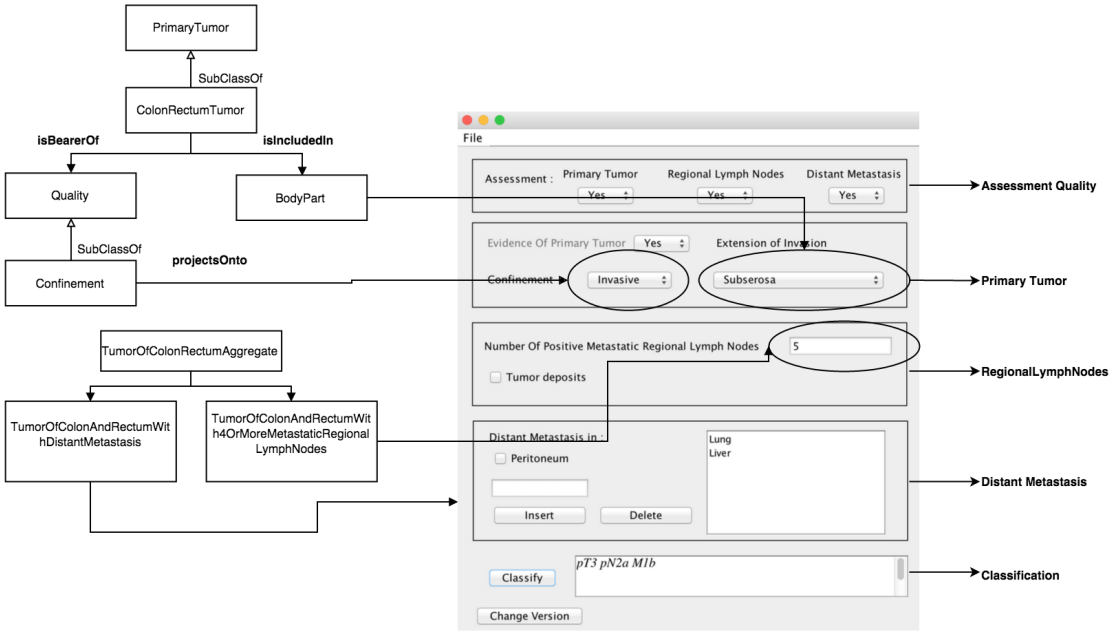


Figure 4.14: Example of classification from manual input data with respective diagram of the involved classes from TNM-O

Table 4.2: Examples of correct classifications when comparing the number of metastatic regional lymph nodes and automatic classification

Metastatic Lymph Nodes	Automatic Classification
	pNX
	pNX
6	pN2a
0	pN0
	pNX
2	pN1b
3	pN1b
6	pN2a

information of 292 patients was extracted from records in text format and classified by an expert pathologist, in order to compare to the automatic classification.

4.3.1 Classification of Metastatic Regional Lymph Nodes

This dataset consisted in 382 entries which were classified by the N classification rules - the metastatic regional lymph nodes. The test was made in two stages. The first one was a comparison between the number of metastatic lymph nodes and automatic classification. In the Table 4.2 there is an extract from the results obtained. The first column is the number of metastatic lymph nodes found and the second the classification made by the classifier. For example, when there is not any information about the metastatic regional lymph nodes the correct classification is *pNX* which can be seen in the Section 2.3.2.

Concerning the TNM classification rules, all data was classified correctly by the automatic classifier tool. This step proved that the system is capable of correctly classify instance data for the N classification.

The second phase of testing was the comparison between the expert classification and the automatic one. In this experiment the system only revealed an efficiency around the 55%. Although, this results revealed inconsistencies made in the data during the input by the specialist.

In Table 4.3 both classifier and pathologist correctly classified all the instances. Although, in the Table 4.4 shows some examples were the expert classification was made incorrectly. In the same table, for example, the pathologist gave a classification without specifying the number of metastatic regional lymph nodes. In this case the classification is pNX (*TumorOfColonAndRectumWithNoAssessmentOfRegionalLymphNodes*).

Table 4.3: Examples of correct classifications both from pathologist and classifier

Metastatic Lymph Nodes	Expert Classification	Automatic Classification
0	pN0	pN0
0	pN0	pN0
7	pN2b	pN2b
4	pN2a	pN2a

Table 4.4: Examples of inconsistencies found when comparing classifications between classifier and pathologist

Metastatic Lymph Nodes	Expert Classification	Automatic Classification
0	pN1	pN0
2	pN1	pN1b
		pNX
	pN2b	pN2a

After carefully analysing the results, the reasons for this inconsistencies were determined:

- Incomplete information - Information gaps due to human error were made because of the absence of a automatic tool for assessment;
- Lack of identification of TNM version used - the TNM classification is now on the version 7 and previous versions have different codings, old registries can be identified as inconsistent if there is no version attached;
- Incorrect classification;

This test reinforced the necessity of having support of automatic classifiers in the classification task. These systems will help detect inconsistencies while documenting tumor assessments. Besides, they reveal helpful on guiding the expert during the classification process. Therefore, preventing documentation errors during data curation.

4.3.2 Classification of all Assessments

In this evaluation a dataset was used and classified by an expert pathologist while reading the text records from 292 patients. The dataset followed the criteria showed in Table 4.1, that was designed for latest implementation of tabular classification on the system.

Table 4.5: Results obtained in percentage by automatic classification for the TNM version 6 of the colon and rectum tumors

Assessment	Accuracy
T Classification	100.00 (292/292)
N Classification	99.31 (290/292)
M Classification	98.97 (289/292)

Table 4.6: Results obtained in percentage by automatic classification for the TNM version 7 of the colon and rectum tumors

Assessment	Accuracy
T Classification	100.00 (292/292)
N Classification	99.31 (290/292)
M Classification	98.63 (288/292)

The classification was divided in the two versions of the TNM classification which are supported by the classifier. The results are on the Tables 4.5 and 4.6 for version 6 and 7 respectively.

The results showed that the automatic classification accuracy was very near 100%. Although, looking at the results, it is possible to identify some misclassified data. Therefore, this test revealed that the classifier was able to correctly classify data for all types of assessment in the colorectal classification.

Chapter 5

Discussion

The TNM-O and the respective TNMCR-O proved to be able to represent all TNM classification concepts and definitions. The modular architecture allows better reasoning performance reducing the time cost of the system. The ontology driven classification system provided accurate classification of pathological data and detected inconsistencies made by experts during tumor documentation. Therefore, this study proved that automatic classification, based on the TNM classification system, improves data validity and consistency.

Prior work has documented the impact and importance of biomedical ontologies in managing the increase of knowledge and the massive amounts of information in this domain [6]. Some of the ontologies with bigger impact in this domain are the FMA ontology [7] for representation of concepts and definitions about the human anatomy, the HL7 Reference Information Model (HL7-RIM) ontology [8] that represents the messaging standard HL7 and the Gene Ontology (GO) [9] that seeks to provide a set of vocabularies for biological domains that can be used to describe gene products in any organism. Although, a complete representation of the TNM classification of malignant tumors is still missing.

This project was developed in the Institute of Medical Biometry and Medical Informatics in Freiburg - Germany, which goal is to provide a formal representation of the TNM classification. As prior work, it was already developed an ontology that represented the TNM classification rules for the breast tumor [10, 11]. Although, the results obtained with the colorectal ontology met the specifications imposed to this project better than the previous ontology. Therefore, some re-designing of the breast tumor ontology was done in order to be integrated in the current state of the TNM ontological representation.

This work presents a formal representation of the classification system and the TNM rules for colon and rectum tumors, represented as in the literature [3], by developing the TNM-O and the TNMCR-O. The first one contains the most general concepts and definitions of the TNM classification, where each modular

ontology e.g. TNMCR-O can connect to. The latter, contains all the concepts and definitions specific to the classification of colorectal tumors. This architecture provides better efficiency during the classification task. So far , it was developed the TNMCR-O that represents the classification rules of the colorectal tumors for both 6 and 7 editions of the TNM Classification.

These ontologies close the gap of the lack of formal representation of such a complex and extent system as the TNM classification. Having such ontology will decrease the time spent on building the knowledge base for a new applications in this domain. Besides, it also increases the uniformity of concepts and relations between applications and consequently the interoperability between intelligent systems.

The TNM classification has become an important and dynamic system to describe the anatomical extent of malignant tumors and is a major prognostic factor in predicting the outcome of patients with cancer [5]. Examples of automatic tools for the TNM classification system are the CS, a software equipped with algorithms capable to translate TNM staging information in order to be used across cancer statistical databases [73], an ontology-driven classifier that processes physicians annotations in images to reason the TNM classification [13] and a semi-automatic tool that classifies tumor documentation in the ESTHER system also based in the TNM classification system [14]. However, none of these studies presents a tool for classification based on a formal representation of the TNM classification system.

For this purpose it was developed the an ontology driven classification system that uses the TNM-O, and its modules, for its knowledge base. This automatic classifier provides a correct and efficient classification of tumor as well as detects inconsistencies made on the actual handmade datasets. In opposite to previous studies, this work presents a feasibility test based on real data that proves the accuracy of such system and showed that it is capable to improve data validity and consistency during tumor documentation.

5.1 Limitations and Future Work

Although, this project is limited to the number of tumors already ontologically represented. So far, only ontologies to represent colorectal tumors was developed. Besides this, some reasoning issues were detected with classes that were restricted with cardinality axioms. Another limitation is the lack of bigger datasets with more information to test. Since the classifier uses the ontologies developed for its knowledge base, only colorectal tumors can be automatically classified by it.

Future work will therefore include the expansion of the classifier to other tumor sites and to add more useful functionalities to aid the pathologist in the classification process. Also, the presented prototypical TNM classifier showed its potential

in the integration to the existent health information system for classification and documentation of tumors in cancer registries.

Chapter 6

Conclusion

This work presents the first version of Tumor-Node-Metastasis Ontology (TNM-O) which represents the Tumor-Node-Metastasis (TNM) classification of malignant tumors, one of the most important tools in clinical oncology. This ontology works as connecting hub to other modular ontologies, each containing the classification rules for the corresponding cancer type. As modular ontology the TNM Colon and Rectum Ontology was developed representing the classification rules of colorectal tumors. This ontology provides a formal representation of this system, providing a coherent knowledge base that can be used in future applications.

A classifier application was developed to provided automatic classification of tumors with the TNM ontology as knowledge base. This application was developed to guide the pathologist during the assessment of the tumor, provided a efficient classification of tabular and instance data and detect inconsistencies made on datasets by expert manual data input.

The whole system proved to be a real asset to this domain where few software development has been made and tested. This work provides a foundation of what tumor classification and documentation can be.

Bibliography

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[2] "International agency for research on cancer, globocan 2012 : Estimated cancer incidence, mortality and prevalence worldwide in 2012globocan 2012 : Estimated cancer incidence, mortality and prevalence worldwide in 2012."

[3] L. Sobin, M. Gospodarowicz, and C. Wittekind, *TNM Classification of Malignant Tumours*, 7th ed. Chichester, West Sussex, UK ; Hoboken, NJ: Wiley-Blackwell, 2009.

[4] M. K. Gospodarowicz, D. Miller, P. A. Groome, F. L. Greene, P. A. Logan, and L. H. Sobin, "The process for continuous improvement of the tnm classification," *Cancer*, vol. 100, no. 1, pp. 1–5, 2004. [Online]. Available: http://dx.doi.org/10.1002/cncr.11898

[5] C. Webber, M. Gospodarowicz, L. H. Sobin, C. Wittekind, F. L. Greene, M. D. Mason, C. Compton, J. Brierley, and P. A. Groome, "Improving the tnm classification: Findings from a 10-year continuous literature review," *International Journal of Cancer*, vol. 135, no. 2, pp. 371–378, 2014. [Online]. Available: http://dx.doi.org/10.1002/ijc.28683

[6] D. L. Rubin, N. H. Shah, and N. F. Noy, "Biomedical ontologies: a functional perspective," *Briefings in bioinformatics*, vol. 9, no. 1, pp. 75–90, 2008.

[7] C. Rosse and J. L. M. Jr., "A reference ontology for biomedical informatics: the foundational model of anatomy," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 478 – 500, 2003, unified Medical Language System. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046403001278

[8] A. Iqbal, "An owl-dl ontology for the hl7 reference information model," in *Toward Useful Services for Elderly and People with Disabilities*, ser. Lecture Notes in Computer Science, B. Abdulrazak, S. Giroux, B. Bouchard, H. Pigot, and M. Mokhtari, Eds. Springer Berlin Heidelberg, 2011, vol. 6719, pp. 168–175.

[9] G. O. Consortium *et al.*, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D258–D261, 2004.

[10] R. Faria, "Ontology based knowledge representation for tumors," Master's thesis, Universidade do Minho, Escola de Engenharia, Departamento de Informática, 2014.

[11] M. Boeker, R. Faria, and S. Schulz, "A proposal for an ontology for the tumor-node-metastasis classification of malignant tumors: a study on breast tumors." ODLS 2014, 2014.

[12] I. D. Fleming, "Ajcc/tnm cancer staging, present and future," *Journal of surgical oncology*, vol. 77, no. 4, pp. 233–236, 2001.

[13] E. Luque, D. Rubin, and D. Moreira, "Automatic classification of cancer tumors using image annotations and ontologies," in *Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on*, 2015.

[14] R. Mösges, A. Heinen, and J. Höfener, "[semi-automatic tnm classification of malignant tumors with the ester system exemplified by the larynx]," *HNO*, vol. 39, no. 10, pp. 396–400, 1991.

[15] S. Schulz, D. Seddig-Raufie, N. Grewe, J. Röhl, D. Schober, M. Boeker, and L. Jansen, "Guideline on developing good ontologies in the biomedical domain with description logics," 2012.

[16] R. Neches, R. E. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout, "Enabling technology for knowledge sharing," *AI magazine*, vol. 12, no. 3, p. 36, 1991.

[17] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[18] M. M. Gaber, *Scientific data mining and knowledge discovery*. Springer, 2009.

[19] D. Fensel, "Ontologies: A silver bullet for knowledge management and electronic commerce. secaucus," 2003.

[20] S. Bechhofer, "Owl: Web ontology language," in *Encyclopedia of Database Systems*. Springer, 2009, pp. 2008–2009.

[21] C. Rosse, A. Kumar, J. L. Mejino Jr, D. L. Cook, L. T. Detwiler, and B. Smith, "A strategy for improving and integrating biomedical ontologies," in *AMIA Annual Symposium proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 639.

[22] W. Swartout and A. Tate, "Guest editors' introduction: Ontologies," *IEEE Intelligent Systems*, no. 1, pp. 18–19, 1999.

[23] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, "The semantic web: The roles of xml and rdf," *Internet Computing, IEEE*, vol. 4, no. 5, pp. 63–73, 2000.

[24] O. Lassila and R. R. Swick, "Resource description framework (rdf) model and syntax specification," 1999.

[25] D. Brickley and R. V. Guha, "Rdf vocabulary description language 1.0: Rdf schema," 2004.

[26] I. Horrocks, P. F. Patel-Schneider, and F. Van Harmelen, "From shiq and rdf to owl: The making of a web ontology language," *Web semantics: science, services and agents on the World Wide Web*, vol. 1, no. 1, pp. 7–26, 2003.

[27] G. Antoniou and F. Van Harmelen, "Web ontology language: Owl," in *Handbook on ontologies*. Springer, 2004, pp. 67–92.

[28] L. Magee, "Upper-level ontologies," *Towards A Semantic Web: Connecting Knowledge in Academic Research*, p. 235, 2011.

[29] E. Beisswanger, S. Schulz, H. Stenzhorn, and U. Hahn, "Biotop: an upper domain ontology for the life sciences," *Applied Ontology*, vol. 3, no. 4, pp. 205–212, 2008.

[30] R. Hoehndorf, "What is an upper level ontology?" *Ontogenesis*, 2010.

[31] F. Gibson, "Upper level ontologies," *Ontogenesis*, 2010.

[32] V. Mascardi, V. Cordì, and P. Rosso, "A comparison of upper ontologies." in *WOA*, 2007, pp. 55–64.

[33] B. Smith, A. Kumar, and T. Bittner, "Basic formal ontology for bioinformatics," *Journal of Information Systems*, pp. 1–16, 2005.

[34] P. Grenon, B. Smith, and L. Goldberg, "Biodynamic ontology: applying bfo in the biomedical domain," *Studies in health technology and informatics*, pp. 20–38, 2004.

[35] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with dolce," in *Knowledge engineering and knowledge management: Ontologies and the semantic Web*. Springer, 2002, pp. 166–181.

[36] V. R. Benjamins, *Knowledge engineering and knowledge management: ontologies and the semantic web*. Springer, 2003, vol. 2473.

[37] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.

[38] S. Schulz, E. Beisswanger, L. van den Hoek, O. Bodenreider, and E. M. van Mulligen, "Alignment of the umls semantic network with biotop: methodology and assessment," *Bioinformatics*, vol. 25, no. 12, pp. i69–i76, 2009.

[39] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.

[40] J.-D. Kim, T. Ohta, Y. Teteisi, and J. Tsujii, "Genia ontology," *Tech Rep TR-NLP-UT-2006-2, Tsujii Laboratory, University of Tokyo*, 2006.

[41] S. Schulz, H. Stenzhorn, and M. Boeker, "The ontology of biological taxa," *Bioinformatics*, vol. 24, no. 13, pp. i313–i321, 2008.

[42] S. Schulz and M. Boeker, "Biotoplite: An upper level ontology for the life sciencesevolution, design and application." in *GI-Jahrestagung*, 2013, pp. 1889–1899.

[43] D. J. Schultz *et al.*, "Ieee standard for developing software life cycle processes," *IEEE Std*, pp. 1074–1997, 1997.

[44] A. Gómez-Pérez, "Ontological engineering: A state of the art," *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence*, vol. 2, no. 3, pp. 33–43, 1999.

[45] M. Fernández-López, "Overview of methodologies for building ontologies," 1999.

[46] D. B. Lenat and R. V. Guha, *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[47] C. Matuszek, J. Cabral, M. J. Witbrock, and J. DeOliveira, "An introduction to the syntax and content of cyc." in *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. Citeseer, 2006, pp. 44–49.

[48] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, "Methodologies, tools and languages for building ontologies. where is their meeting point?" *Data & knowledge engineering*, vol. 46, no. 1, pp. 41–64, 2003.

[49] M. Uschold and M. King, *Towards a methodology for building ontologies.* Citeseer, 1995.

[50] M. Uschold, "Building ontologies: Towards a unified methodology," *Technical Report-University of Edinburg Artificial Intelligence Applications Institute AIAI TR*, 1996.

[51] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," 1995.

[52] Th, "The KACTUS Booklet version 1.0. Esprit Project 8145. September, 1996," Tech. Rep., 1996. [Online]. Available: http://www.swi.psy.uva.nl/prjects/NewKACTUS/Reports.html

[53] B. Swartout, R. Patil, K. Knight, and T. Russ, "Toward distributed use of large-scale ontologies," in *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, 1996, pp. 138–148.

[54] M. F. López, A. Gómez-Pérez, J. P. Sierra, and A. P. Sierra, "Building a chemical ontology using methontology and the ontology design environment," *IEEE intelligent Systems*, no. 1, pp. 37–46, 1999.

[55] A. Gómez-Pérez, "Towards a framework to verify knowledge sharing technology," *Expert Systems with Applications*, vol. 11, no. 4, pp. 519–529, 1996.

[56] J. Blázquez, M. Fernández, J. M. García-Pinar, and A. Gómez-Pérez, "Building ontologies at the knowledge level using the ontology design environment," 1998.

[57] J. C. Arpírez, O. Corcho, M. Fernández-López, and A. Gómez-Pérez, "Webode: a scalable workbench for ontological engineering," in *Proceedings of the 1st international conference on Knowledge capture.* ACM, 2001, pp. 6–13.

[58] O. Bodenreider and R. Stevens, "Bio-ontologies: current trends and future directions," *Briefings in bioinformatics*, vol. 7, no. 3, pp. 256–274, 2006.

[59] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of protégé: an environment for knowledge-based systems development," *International Journal of Human-computer studies*, vol. 58, no. 1, pp. 89–123, 2003.

[60] F. Baader, *The description logic handbook: theory, implementation, and applications.* Cambridge university press, 2003.

[61] F. Baader, I. Horrocks, and U. Sattler, "Description logics as ontology languages for the semantic web," in *Mechanizing Mathematical Reasoning.* Springer, 2005, pp. 228–248.

[62] F. Baader and W. Nutt, "Basic description logics." in *Description logic handbook*, 2003, pp. 43–95.

[63] F. L. Greene and L. H. Sobin, "The tnm system: Our language for cancer care," *Journal of Surgical Oncology*, vol. 80, no. 3, pp. 119–120, 2002. [Online]. Available: http://dx.doi.org/10.1002/jso.10114

[64] L. Greene and L. Sobin, "A worldwide approach to the tnm staging system: Collaborative efforts of the ajcc and uicc," *Journal of Surgical Oncology*, vol. 99, no. 5, pp. 269–272, 2009. [Online]. Available: http://dx.doi.org/10.1002/jso.21237

[65] C. C. Compton and F. L. Greene, "The staging of colorectal cancer: 2004 and beyond," *CA: A Cancer Journal for Clinicians*, vol. 54, no. 6, pp. 295–308, 2004. [Online]. Available: http://dx.doi.org/10.3322/canjclin.54.6.295

[66] A. J. C. on Cancer, *AJCC Cancer Staging Atlas*, 2nd ed. Springer-Verlag New York, 2012.

[67] F. L. Greene, A. K. Stewart, and H. J. Norton, "A new tnm staging strategy for node-positive (stage iii) colon cancer: An analysis of 50,042 patients," *Annals of Surgery*, vol. 236, no. 4, pp. 416–421, 10 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422595/

[68] F. L. Greene, C. M. Balch, D. G. Haller, and M. Morrow, *AJCC Cancer Staging Manual (6th Edition)*, 6th ed. Springer, May 2002. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387952713

[69] S. Schulz and M. Boeker, "Biotoplite: An upper level ontology for the life sciencesevolution, design and application." *GI-Jahrestagung*, pp. 1889–1899, 2013.

[70] M. Horridge and S. Bechhofer, "The owl api: A java api for owl ontologies." *Semantic Web*, vol. 2, no. 1, pp. 11–21, 2011.

[71] (2015, December). [Online]. Available: http://hermit-reasoner.com/

[72] (2015, 11) Tnm ontologies repository. [Online]. Available: http://purl.org/tnmo

[73] S. B. Edge and C. C. Compton, "The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm," *Annals of surgical oncology*, vol. 17, no. 6, pp. 1471–1474, 2010.

Appendix A

Comparison between ontology development methodologies and the IEEE 1074-1995 standard

Table A.1: Comparison between ontology development methodologies and the IEEE 1074-1995 standard [45]

Feature			Cyc	Usdhold and King	Gruninger and Fox	KACTUS	Methodontology	Sensus
Project Managment Processes	Project initiation		Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
	Project monitoring		Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
	Quality Management		Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
Development Processes	Pre-Development	Concept exploration	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
		System allocation	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
	Development	Requirements	Not proposed	Proposed	Proposed	Proposed	Described in detail	Proposed
		Design	Not proposed	Not proposed	Described	Described	Described in detail	Not proposed
		Implementation	Proposed	Proposed	Described	Proposed	Described in detail	Described
	Post-Development	Installation	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
		Operation	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
		Support	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
		Maintenance	Not proposed	Not proposed	Not proposed	Not proposed	Purposed	Not proposed
		Retirement	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	Not proposed
Integral Processes	Knowledge Acquisition		Proposed	Proposed	Proposed	Not proposed	Described in detail	Not proposed
	Verification and Validation		Not proposed	Proposed	Proposed	Not proposed	Described in detail	Not proposed
	Configuration		Not proposed	Not proposed	Not proposed	Not proposed	Described in detail	Not proposed
	Documentation		Proposed	Proposed	Proposed	Not proposed	Described in detail	Not proposed
	Training		Not proposed	Not proposed	Not proposed	Not proposed	Not proposed	

Appendix B

Published Papers

B.1 TNM-O an Ontology for the Tumor-Node-Metastasis Classification of Malignant Tumors: a Study on Colorectal Cancer

Authors

Martin Boeker, Fábio França, Peter Bronsert and Stefan Schulz

Conference

International Conference on Biomedical Ontology 2015

Abstract

Objectives: To (1) present an ontological framework for the TNM classification system, (2) implement an ontology of the TNM classification system of the tumors of the colon and rectum based on this framework, and (3) evaluate this ontology with a classifier for pathology data.

Methods: The TNM ontology uses the Foundational Model of Anatomy for anatomical entities and BioTopLite 2 as a domain top-level ontology. The general rules for the TNM system and the specific TNM classification for colorectal tumors (ICD-O C19-C23) were represented as described in the literature. Additional information was collected from daily practice in tumor documentation in the university level Comprehensive Cancer Center. Based on the ontology, an automatic classifier for pathology data was developed.

Results: TNM was represented as an information artifact which consists of single representational units. Corresponding to every representational unit, tumors and tumor aggregates were defined. Tumor aggregates consist of the primary tumor and (if existent) of infiltrated regional lymph nodes and distant metastases. TNM codes depend on the location and certain qualities of the primary tumor (T), the infiltrated regional lymph nodes (N) and the existence of distant metastases (M). Tumor data from clinical and pathological documentation were successfully classified with the ontology.

Conclusion: This work presents a first version of the TNM Ontology which represents the TNM system for the description of the anatomical extent of malignant tumors which is one of the most important tools in clinical oncol- ogy. The presented work is already sufficient to show its representational correctness and completeness as well as its applicability for classification of instance data. This work provides a foundation for a TNM Ontology.

Relation to this work

This paper is related to the first two main goals for this project : to present the TNM Ontology and the TNM Colon and Rectum Ontology. This was important to evaluate the impact of this project in the ontological community.

B.2 Feasibility of an Ontology Driven Tumor-Node-Metastasis Classifier Application: a Study on Colorectal Cancer

Authors

Fábio França, Stefan Schulz, Peter Bronsert, Paulo Novais, Martin Boeker

Conference

2015 International Symposium on INnovations in Intelligent SysTems and Applications

Abstract

The objectives of this work are (1) to develop a classifier application for tumor staging based on a formal representation of the Tumor-Node-Metastasis classification system (TNM), and (2) to show the feasibility of this approach on real data. This paper presents a classifier application for colorectal tumors based on the TNM-O ontology. It was developed in the JAVA using the OWL-API. The TNM-O uses the Foundational Model of Anatomy for representing anatomical entities and BioTopLite2 as a domain-top-level ontology. The classifier application processes input data via a user interface or tabular data. The classification starts with the creation of RDF Individuals for each pathological information item formally described in the ontology. These Individuals are then classified by the HermiT Description Logics reasoner by A-Box classification. A dataset with 382 entries was provided by the pathology department of a university hospital. It was automatically classified with regard to metastatic regional lymph nodes. Results or expert classification by pathologists and automatic classification were compared. The automatic process helped to detect and explain inconsistencies between expert and automatic classifications. This work, we demonstrate the use of semantic technologies in a TNM classifier application separating underlying medical knowledge represented in OWL from process logics. The pre- sented prototypical TNM classifier application shows the potential to be integrated in larger software systems.

Relation to this work

This paper is related to the two last objectives of this work: to present an automatic classifier application with the TNM-O plus TNM Colon and Rectum ontology as knowledge base and to test the feasibility of this approach.