

Grading lung tumors using OWL-DL based reasoning

Olivier Dameron¹, Élodie Roques¹, Daniel Rubin², Gwenaëlle Marquet¹ and Anita Burgun¹

¹UPRES-EA 3888, Université de Rennes1, France

²Stanford Medical Informatics, Stanford University School of Medicine,
Stanford CA 94305, USA

`olivier.dameron@univ-rennes1.fr`

1 Introduction

The treatment and prognosis for a tumor depend to a large extent on its stage. Assessing the grade of a tumor consists in finding that the description of the location and features of the tumor and of its possible metastasis matches the definition of a grade. However, the grade definitions and the tumor descriptions generally do not have the same granularity. The former have to encompass all the corresponding situations and tend to rely on general criteria, such as “the cancer has spread to the heart” or “metastasis in the brain”. The latter are typically as precise as possible and may refer to kinds or parts of the anatomical entities mentioned in the grade definitions.

Some general knowledge of the domain is necessary in order to fill the gap between the tumor descriptions and the grade definitions. This is done naturally when a human user interprets the data. However, having formal representations of the grade definitions is not enough for an application to be able to provide some automatic assistance. The domain knowledge has to be explicitly and formally represented. The typical representation format for symbolic knowledge is formal ontologies.

Assessing the grade of a tumor is a classification process consisting in finding the grade which criteria are met by the tumor’s features. Description logics is a family of formal representation languages for ontologies and are designed for classification-based reasoning. Therefore, representing knowledge about oncology in description logics looks like a desirable contribution for performing automatic grading of tumors.

The goal of this article is to analyze to what extent tumor grading can be performed automatically using the OWL-DL description logic language. We focused on the grading of lung tumors. Section 2 is a review of the authoritative cancer ontology, in which we conclude that the NCIT has to be extended in order to perform automatic tumor grading. Section 3 is a description of the TNM classification used for defining the grade of a tumor and the anatomical

concepts that we used. Section 4 compares how Description Logic’s class-based classification and instance-based classification support tumor grading.

2 Studying the NCI thesaurus

The NCI Thesaurus¹ (NCIT) is a controlled vocabulary designed to meet the needs of the cancer research community [1, 2]. Its goal is to provide definitions for basic and clinical concepts used in cancer research. These definitions, the taxonomic structure of the thesaurus and the presence of explicit relationships between classes make the NCIT more than a controlled vocabulary.

Several works analyzed the NCIT as a terminology and as an ontology. They identified limitations concerning the term-formation principles, the missing or inappropriately assigned verbal and formal definitions [3], as well as classification, synonymy, relations and definitions [4].

The NCIT was originally designed using a proprietary description-logic-based languages and has been converted into OWL-Lite [2]. The current version is composed of 41,717 named classes, all of which are primitive classes, i.e. they can have constraints, but do not have any necessary and sufficient definition that can be used to infer that an individual is an instance of this class.

The published limitations (although some of them have been fixed, and other are on the process of being fixed) and the lack of defined classes, specially for the T, N and M criteria make the NCIT useless for our grading purpose. As we focus on the OWL-DL language capabilities, rather than on the model, it was easier to devise our own ontology, drawing inspiration from the NCIT whenever possible.

3 Methods: OWL-DL Ontology for tumor grading

The TNM Classification of Malignant Tumours[5] (TNM) is the cancer staging system developed and maintained by the International Union Against Cancer (UICC) and the American Joint Committee on Cancer (AJCC) to maintain consensus on one globally recognized standard for categorising cancer².

The TNM criteria for staging tumors involve (among others) the anatomical location of the primary tumor and of its possible metastasis as well as the involved lymph nodes. Therefore, different types of tumors (for example lung tumors or colon tumors) have different sets of TNM criteria. Section 3.1 describes the criteria used for grading lung tumors. Section 3.2 presents the anatomical entities used for describing tumors.

3.1 The T, N and M axis and the grades

Staging a tumor is a two steps process.

The first step consists in giving a score along three axis describing the tumor, the spreading into lymphatic nodes and the possible metastasis:

¹<ftp://ftp1.nci.nih.gov/pub/cacore/EVS/>

²<http://en.wikipedia.org/wiki/TNM>

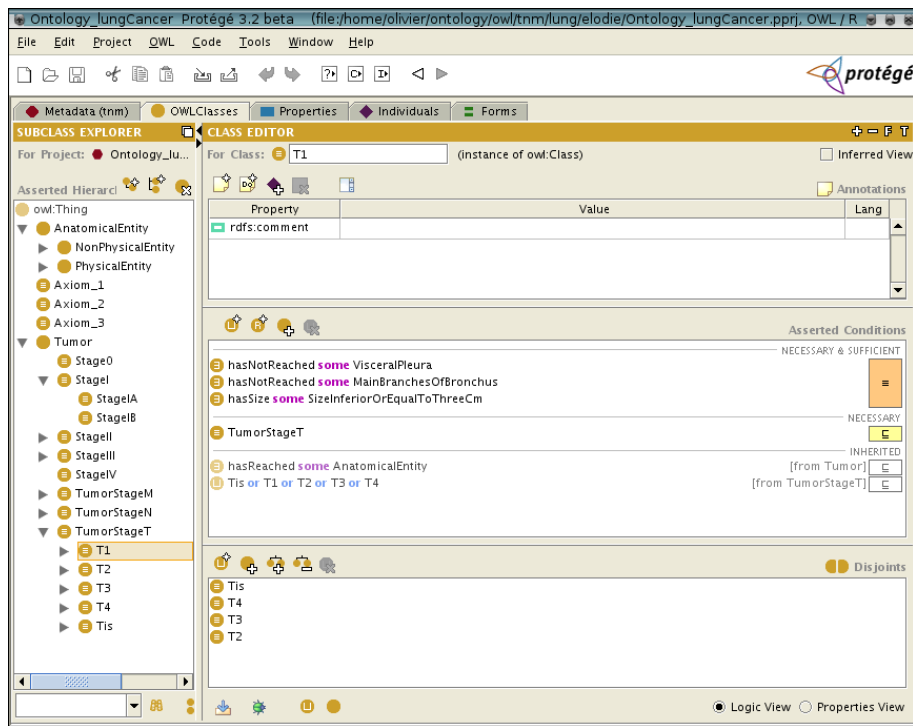


Figure 1: Definition of the T1 criterion

- T (a, is, (0), 1-4): size of the tumor Relates to Tumour size and local spread.
- N (0-3): spreading into lymphatic nodes
- M (0/1): spreading into other organs (metastasis)

The second step consists in determining the stage according to the previous scores. Each combination of a score on the T, N and M axis define the unique stage of the tumor, from 0 to IV. Note that several combinations of T, N and M scores can lead to the same stage.

For lung tumors, we used the lung carcinoid tumors staging guide³. Table 1 presents the definitions for the T, N and M criteria. Table 2 presents the definitions for the five stages of tumors.

We used the previous descriptions to provide necessary and sufficient definitions for the various T, N and M criteria, and for the various stages. Figure 1 illustrates the definition of the T₁ criterion.

³http://www.cancer.org/docroot/CRI/content/CRI_2.4.3X.How_is_lung_carcinoid_tumor_staged.56.asp?sitearea=

Tis	The cancer is found only in the layer of cells lining the air passages.
T1	The cancer is no larger than 3 cm, has not spread to the membranes that surround the lungs (visceral pleura), and does not affect the main branches of the bronchi.
T2	The cancer has 1 or more of the following features: (1) It is larger than 3 cm. (2) It involves a main bronchus but is not closer than 2 cm to the point where the windpipe (trachea) branches into the left and right main bronchi. (3) It has spread to the visceral pleura. (4) The cancer may partially clog the airways, but this has not caused the entire lung to collapse or develop pneumonia.
T3	The cancer has 1 or more of the following features: (1) It has spread to the chest wall, the breathing muscle that separates the chest from the abdomen (diaphragm), the membranes surrounding the space between the 2 lungs (mediastinal pleura), or membranes of the sac surrounding the heart (parietal pericardium). (2) It involves a main bronchus and is closer than 2 cm to the point where the windpipe (trachea) branches into the left and right main bronchi, but does not involve this area. (3) It has grown into the airways enough to cause 1 lung to entirely collapse or to cause pneumonia of the entire lung.
T4	The cancer has 1 or more of the following features: (1) It has spread to the space behind the chest bone and in front of the heart (mediastinum), the heart, the windpipe, the tube connecting the throat to the stomach (esophagus), the backbone, or the point where the windpipe branches into the left and right main bronchi (carina). (2) Two or more separate tumor nodules are present in the same lobe. (3) There is a fluid containing cancer cells in the space surrounding the lung.
N0	The cancer has not spread to lymph nodes.
N1	The cancer has spread to lymph nodes within the lung, hilar lymph nodes (located around the area where the bronchus enters the lung). The cancer has metastasized only to lymph nodes on the same side as the cancerous lung.
N2	The cancer has spread to lymph nodes around the point where the windpipe branches into the left and right bronchi or to lymph nodes in the mediastinum (space behind the chest bone and in front of the heart). The lymph nodes on the same side of the cancerous lung are affected.
N3	The cancer has spread to lymph nodes near the collarbone on either side, to hilar or mediastinal lymph nodes on the side opposite the cancerous lung.
M0	The cancer has not spread to distant sites.
M1	The cancer has spread to distant sites such as other lobes of the lungs, lymph nodes farther than those mentioned in N stages, and other organs or tissues such as the liver, bones, or brain.

Table 1: The T, N and M criteria for lung tumors

Stage 0	(Tis, N0, M0)
Stage IA	(T1, N0, M0)
Stage IB	(T2, N0, M0)
Stage IIA	(T1, N1, M0)
Stage IIB	(T2, N1, M0 or T3, N0, M0)
Stage IIIA	(T1, N2, M0) or (T2, N2, M0) or (T3, N1, M0) or (T3, N2, M0)
Stage IIIB	(T1, N3, M0) or (T2, N3, M0) or (T3, N3, M0) or (T4, N0, M0) or (T4, N1, M0) or (T4, N2, M0) or (T4, N3, M0)
Stage IV	(Any T, Any N, M1)

Table 2: Definitions for the different stages of tumors

3.2 Anatomical entities

Some of the T, N and M criteria refer to anatomical entities as location landmarks. In order to address the difference of granularity between clinical descriptions and the definitions of the staging criteria, it is important that our model contains not only the anatomical entities mentioned in the staging criteria, but also the various direct and indirect parts of these entities.

In order to demonstrate this principle, we selected some anatomical entities of interest such as Heart and Lung, for which we added direct and indirect parts. For this anatomical decomposition, we drew inspiration from the Foundational Model of Anatomy (FMA) [6], which is a canonical model of healthy human anatomy. We were not trying to be exhaustive in the description of anatomy, as we expect that it would require to deal with other problems such as the automatic extraction of the relevant portion of the FMA (including pruning classes with a too small level of granularity such as cells), or such as dealing with the great number of additional classes.

The staging criteria definitions were adapted in order to support such part-hood descriptions.

4 Results: Mechanisms for grading tumors

The generated OWL-DL ontology was composed of 123 classes, among which 66 were defined. There were 57 classes representing anatomical entities.

This ontology was imported by an ontology in which simulated patient conditions were represented using classes, and by another ontology in which the same conditions were represented using individuals.

In both cases, the classification time ranged from a few seconds to a few minutes.

The files for the ontology and the test cases are available online⁴.

4.1 Class-based reasoning

For performing class-based reasoning, we represented clinical situations by creating specific subclasses for each entity. Depending on the goal, the reasoning

⁴<http://www.med.univ-rennes1.fr/~dameron/ontology/tnmClassification/>

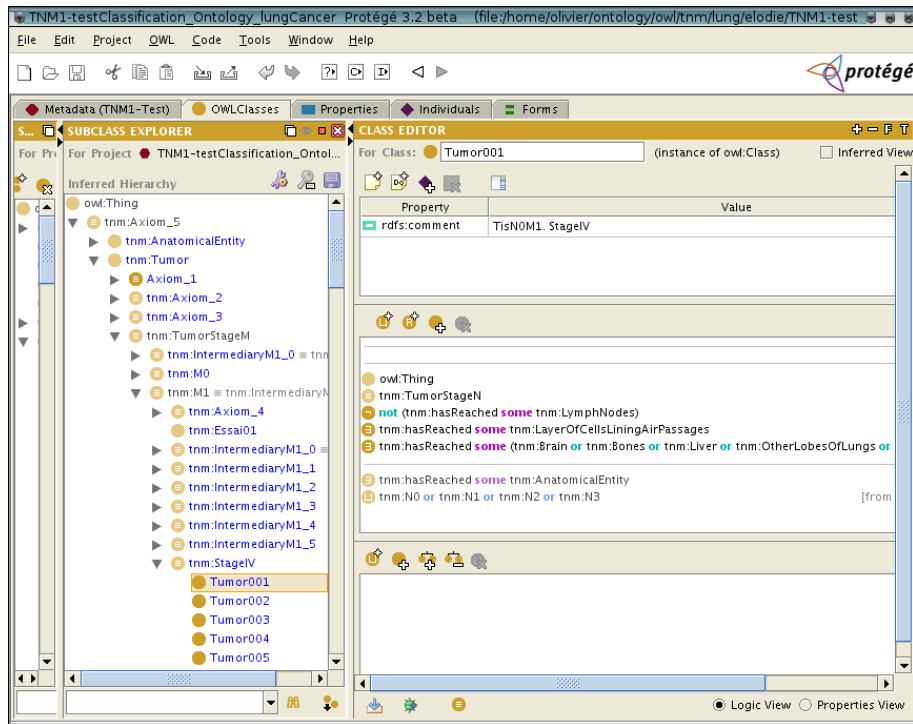


Figure 2: Class-based reasoning: the class representing a tumor is correctly inferred to be a subclass of the Stage IV tumor class

process consisted either in retrieving the inferred subclasses of a particular stage, or in retrieving the inferred superclass of a tumor that is also a stage.

We devised 240 tests, all of which were correctly classified. Figure 2 illustrates the result of classification for a class.

4.2 Instance-based reasoning

For performing instance-based reasoning, we represented clinical situations by creating instances of the corresponding classes. Depending on the goal, the reasoning process consisted either in retrieving the inferred instances of the class representing a particular stage, or in retrieving the inferred type of the tumor individual that is also a stage.

We devised 150 tests, all of which were correctly classified. Figure 3 illustrates the result of classification for an instance.

5 Discussion

We demonstrated that automatic staging of lung tumors can be performed using OWL-DL classification. First, this result was obtained by providing logical definitions to a limited number of classes. Second, this result shows that applications can be expected to reuse leverage the symbolic knowledge represented in

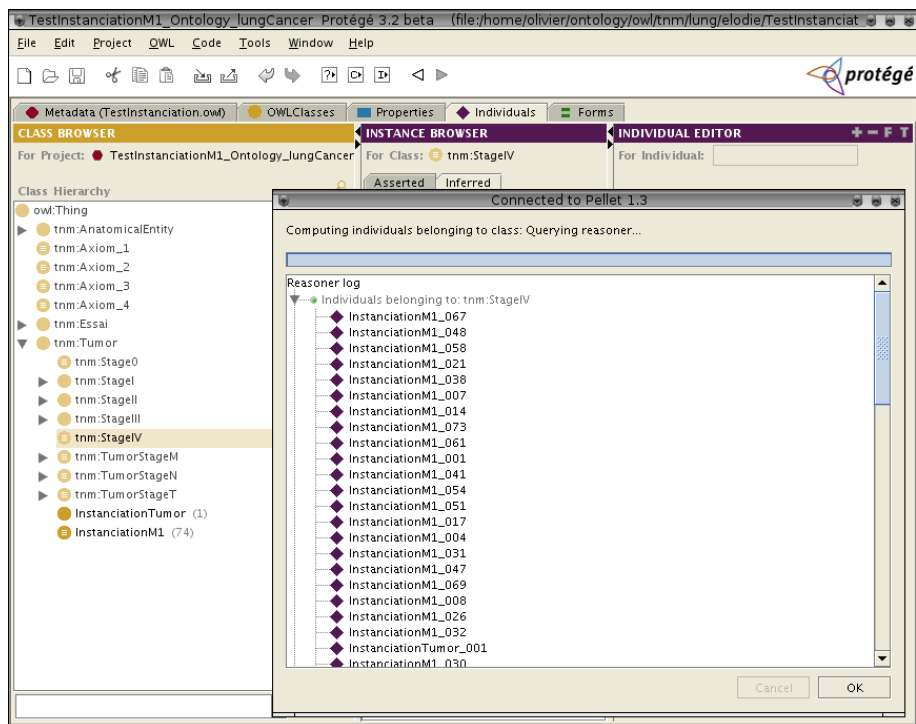


Figure 3: Instance-based reasoning: the individual representing a tumor is correctly inferred to be an instance of the Stage IV tumor class

separate ontologies. The logical and computational feasibility of reusing (portions of) existing ontologies such as the NCI Thesaurus, the Foundational Model of Anatomy and possibly bioinformatics ontologies [7] remains to be studied.

We also highlighted some of the limitations of description logics-based classification for this task of tumor staging.

- Particularly, the open-world assumption makes it clumsy to perform instance classification (though theoretically possible). In practice, when the patient’s record mentions that the primary tumor is located in the lower lobe of the right lung and involved the right hilar lymph nodes, it implicitly indicates that no other lymph nodes were involved and that there were no detectable distant metastasis. If we do not represent this assumption, the open world assumption will make it impossible for the classifier to infer that this is a N_1 situation, as there might be unspecified involved lymph nodes somewhere else, or some unspecified distant metastasis. It would be possible to make the individual representing the tumor an instance of the class of tumors involving ipsilateral lymph nodes and having no distant metastasis, but such classes and the corresponding closure would have to be generated on the fly. The problem is that in this context, we try to perform some closed-world based reasoning.
- The previous limitation can easily be solved by modeling the patient condition using classes, for which we can add closure constraints. Admittedly, this is tweaking the OWL-DL possibility for achieving a computational goal and is not very elegant. However, we can also consider that this method adequately represents that fact that the class of all the tumors located in the lower lobe of the right lung and involving only the right hilar lymph nodes are N_1 tumors.
- Another limitation of OWL-DL is the impossibility to use numeric range constraints for representing “the cancer is larger than 3cm” or “the tumor is closer than 2 cm to the point where the trachea branches into the left and right main bronchi”. This makes impossible the automation of tumor staging using only description logics reasoning. An additional functionality has to be added to the system that makes the individual or the class representing a patient’s tumor an instance or a subclass of `TumorEqualOrBiggerThan3cm` or `TumorSmallerThan3cm` depending not the actual value of its size.
- For our model to be complete, we should have included classes for all the possible anatomical entities (for example by generating them from the FMA). This would have lead to a major decrease of the classification performance in terms of computing time and of memory footprint. Moreover, the FMA is a model of healthy anatomy, so we should even have doubled the classes for considering all the possibly abnormal entities.

We noticed that the way we modeled things mattered. For example, it was easier to define N_3 and to reuse its definition for N_2 and N_1 , rather than to start with the definition of N_1 and to have to handle complex closures for N_2 and N_3 .

Although we have not actually done it, we assume that our approach could be applied to other kinds of tumors and that similar conclusions would have been reached. The definitions we provided for $T_0... T_4$; $N_0... N_3$ and M_0, M_1

are specific to lung tumors. They should be renamed T₀-Lung... in order to avoid confusion between the name of these classes and their logical definitions. Other definitions could be defined for other kinds of tumors, and they should be named accordingly T₀-Colon...

Acknowledgments

This work was partially funded by the SMI, Stanford University, while Olivier Dameron was a postdoctoral fellow.

References

- [1] Jennifer Golbeck, Gilberto Fragoso, Franck Hartel, Jim Hendler, Jim Oberthaler, and Bijan Parsia. The national cancer institute's thesaurus and ontology. *Journal of Web Semantics*, 1(1):75–80, 2003.
- [2] Franck W. Hartel, Sherri de Coronado, Robert Dionne, Gilberto Fragoso, and Jennifer Golbeck. modeling a description logic vocabulary for cancer research. *Journal of Biomedical Informatics*, 38:114–129, 2005.
- [3] Werner Ceusters, Barry Smith, and Louis Goldberg. A terminological and ontological analysis of the NCI thesaurus. *Methods of Information in Medicine*, 44:498–507, 2005.
- [4] Anand Kumar and Barry Smith. Oncology ontology in the nci thesaurus. In *Proceedings of the Artificial Intelligence in Medicine Europe conference AIME 2005*, pages 213–220, 2005.
- [5] Vincent T. DeVita, Samuel Hellman, and Steven A. Rosenberg. *Cancer: Principles and Practice of Oncology*. Lippincott Williams and Wilkins, 7th edition, 2005.
- [6] C. Rosse and J.L.V Mejino. A reference ontology for bioinformatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003.
- [7] Anand Kumar, Yum Lina Yip, Barry Smith, and Pierre Grenon. Bridging the gap between medical and bioinformatics: An ontological case study in colon carcinoma. *Computers in Biology and Medicine*, page In press, 2005.