Oncology Ontology in the NCI Thesaurus

Anand Kumar[1] and Barry Smith[1,2]

[1] IFOMIS, University of Saarland, Saarbruecken, Germany
[2] Department of Philosophy, SUNY at Buffalo, NY, USA
akumar@ifomis.uni-saarland.de, phismith@buffalo.edu

Abstract. The National Cancer Institute's Thesaurus (NCIT) has been created with the goal of providing a controlled vocabulary which can be used by specialists in the various sub-domains of oncology. It is intended to be used for purposes of annotation in ways designed to ensure the integration of data and information deriving from these various sub-domains, and thus to support more powerful cross-domain inferences. In order to evaluate its suitability for this purpose, we examined the NCIT's treatment of the kinds of entities which are fundamental to an ontology of colon carcinoma. We here describe the problems we uncovered concerning classification, synonymy, relations and definitions, and we draw conclusions for the work needed to establish the NCIT as a reference ontology for the cancer domain in the future.

Keywords: Ontology, Oncology, NCI Thesaurus, Clinical Bioinformatics.

1 Introduction

The NCI Thesaurus (NCIT)[1] is a public domain Description Logic-based terminology produced by the National Cancer Institute's Center for Bioinformatics as a component of its caCORE distribution.[2] NCIT was initially conceived as a terminology server to be used within the various NCI departments. However, it has slowly gained acceptance also outside the NCI as a source for carcinoma terminology. The Thesaurus spans clinical and biological domains. It is one of the earliest terminologies to operationally federate with another ontology system (the MGE Ontology) and to embrace the goal of harmonizing with external ontology modeling practices. The NCIT is thus to be welcomed because it covers an unusually large domain with limited resources. This means also however that it is marked by certain inaccuracies, which are currently managed in an ad hoc way via updates on the basis of criticisms received. The NCIT's impressively broad coverage makes it a good dictionary or lexicon. As argued in [Ceusters et al 2004], however, if the NCIT is to conform to the best practices of ontology building, then it needs to do more than resolve reported inaccuracies in an ad hoc fashion. It should rather be rebuilt on the basis of a sound ontology and its constituent terms should be evaluated in light of this ontology. Such an exercise will guarantee that it has the sort of robust organization that can support

[1] http://nciterms.nci.nih.gov/NCIBrowser/Startup.do
[2] http://ncicb.nci.nih.gov/core

S. Miksch et al. (Eds.): AIME 2005, LNAI 3581, pp. 213 – 220, 2005.

the drawing of inferences involving the different sub-domains of oncology in a maximally efficient and reliable way.

In a series of earlier papers we have drawn attention to certain characteristic families of errors in biomedical terminologies and ontologies such as SNOMED-CT,[3] the UMLS Semantic Network[4] and the Gene Ontology[5] [Kumar & Smith 2004a, Kumar & Smith 2004b, Ceusters et al 2004, Kumar et al 2004, Kumar & Smith 2003], pointing especially to problems with is_a and part_of relations and to logical errors in the formulation of definitions. [Ceusters et al 2004] continues this work in relation to the NCIT. Here we concentrate on one particular aspect of the Thesaurus: its representation of the entities involved in colon carcinoma. To this end we draw on our previous work in collaboration with the Swiss Institute of Bioinformatics [Kumar et al 2005], which uses SNOMED-CT and Gene Ontology annotations, the Swissprot mutant protein database and the Foundational Model of Anatomy to construct an onco-ontology within the Protégé 2000 ontology editing environment, supplementing these digital information resources with deVita's text book [deVita 2004 CD].

Our strategy was to assess the degree to which the NCIT lives up to its goal of serving as a reference ontology for the cancer domain, by attempting to represent within it the entities belonging to our onco-ontology for colon carcinoma. As concerns the salient anatomical entities we also compared the NCIT with the Foundational Model of Anatomy[6] (FMA).

2 The Colon and Adjacent Anatomical Structures

We begin with the normal anatomy and physiology of the colon. The NCIT incorporates the UMLS Semantic Network (SN), meaning that it provides for each class the SN semantic type provided within the UMLS Metathesaurus. Problems arise because SN sometimes conflicts with the subsumption relations provided by the Thesaurus itself and also with the assertions derivable from its incorporation of relational expressions (such as 'Anatomic_Structure_is_Physical_Part_of') in which the types of the relata are explicitly stated.

The ontology of *colon carcinoma* revolves around the *colon*, the anatomical structure which bears the carcinoma and upon whose existence the carcinoma depends. NCIT has:

colon is_a *gastrointestinal system part*

gastrointestinal system part is_a *body part, organ or organ component*

The NCIT does not provide definitions of 'is_a' or of its other relational expressions. For present purposes, however, we can define 'is_a' (meaning: is a subkind of / is a subtype of) as follows:

$$A \text{ is\_a } B =_{\text{def}} \forall x\ (Ax \rightarrow Bx)$$

[3] http://www.snomed.org/

[4] http://www.nlm.nih.gov/research/umls/

[5] http://www.geneontology.org/

[6] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

Here 'Ax' abbreviates 'individual x is an instance of kind A.' '$\forall$' is the standard universal quantifier of first-order logic (signifying 'for all values of') and '$\rightarrow$' is the logical connective 'if … then'. In a full version of the ontology we would need to take account of time for continuant entities and write 'Axt' for 'x is an instance of A at t'. We could then assert for example an axiom to the effect that

$$A \text{ is\_a } B \rightarrow \exists xt(Axt \,\&\, Bxt),$$

which means that is_a statements imply also the existence of corresponding instances. (Here '$\exists$' is the standard existential quantifier, signifying 'for some value of'.)

The NCIT has:

colon is_a *body part*, *organ or organ component*

Here the use of the disjunctive class name '*body part, organ or organ component*' does not reflect good ontology authoring practices [Ceusters et al 2005]. Moreover its use means that, unlike the FMA, NCIT does not define what an *organ* is and makes no statement from which we can infer that *colon* is in fact an *organ*. *Organs* have specific properties and belong to a specific level of granularity, which is different from that of both *organ systems* and *organ parts*.

3 Large Intestine

NCIT also has:

colon Anatomic_Structure_is_Physical_Part_of *large intestine*

We presume that the oddly named relation: Anatomic_Structure_is_Physical_Part_of comes close to the part_of relation for anatomical structures as defined within the FMA. We could then reformulate the above as:

colon part_of *large intestine*

where A part_of B is defined as: for every instance x of A there is some instance y of B which is such that x **part_of** y. In symbols:

$$A \text{ part\_of } B \text{ =def } \forall x\ (Ax \rightarrow \exists \mathrm{y}\ (\mathrm{B}y \,\&\, x \ \textbf{part\_of}\ y))$$

(Here **part_of** is the instance-level parthood relation obtaining between individuals, illustrated for example by Mary's_colon **part_of** Mary.)

Unfortunately NCIT also has:

colon synonym *large intestine*

This is problematic first of all because synonymy relations (unlike part relations) hold not between classes or kinds, but rather between the corresponding *names*. Thus we should more properly write:

'colon' synonym 'large intestine'

though even this (as any dictionary will verify) is an error.

Such mistakes have adverse implications when for example we attempt to use a resource like Swissprot to derive information pertaining to those mutant proteins which are involved in colon carcinoma specifically and not in carcinomas of both the colon and the rectum, or to draw on the information in Swissprot pertaining to the markers present within rectum carcinomas of the *squamous* and of the *adenocarcinoma* type.

4 Colorectal Carcinoma

‘Colorectal carcinoma’ is a term used in some contexts to represent a carcinoma which affects both the colon and the rectum, and in others to represent a kind of carcinoma which is sometimes present in the colon, sometimes in the rectum, and sometimes in both. Unfortunately NCIT, which should have provided some regimentation in the use of this term, has not only:

colon carcinoma is_a *colorectal carcinoma*

but also assertions to the effect that the colorectal carcinoma is located both within the colon and the rectum and within the small intestines.

5 Colon Epithelium

Colon epithelium in the FMA is an *organ part* which is asserted to stand also in a parthood relation to *colon*. The NCIT, in contrast, has

epithelium is_a *tissue*

tissue is_a *other anatomic concept*

other anatomic concept is_a *anatomic structure, system or substance*

The classification of *tissue* as *other anatomic concept* reflects a characteristic confusion, found still today in many terminologies, between concepts and entities in the world. This represents a departure from the principles of good ontology not least because it blocks inferences on the basis of the physical characteristics of the entities at issue [Smith et al 2005b]. Thus it blocks such inferences regarding *adenomatous polyposis coli*, one of the prime predisposing factors for *colon carcinoma*, for which we have:

adinomatous polyposis coli Disease_Has_Normal_Tissue_Origin *epithelium*

We find analogous mistakes with respect to specific organ parts, e.g. in:

large intestinal muscularis mucosa is_a *large intestinal mucosa*

which, because the muscularis mucosa is not a type of but rather a part of the mucosa, confuses mereology with subsumption.

Consider also:

large intestinal muscularis mucosa is_a *large intestinal wall tissue*

large intestinal wall tissue is_a *normal tissue*

normal tissue is_a *microanatomy*

which involves a confusion between a kind of entity and a branch of science.

6 Is_a Overloading

The *is_a* relation can reflect different types of partition of reality. For example in a partition on the basis of pathology we have:

colon carcinoma is_a *disease of colon*

The former is a specification of the latter reflecting the added factor of (carcinomatous) *pathology*. In a partition on the basis of location, in contrast, we have

colon carcinoma is_a *carcinoma*

the specification here deriving from the factor: *location within the colon*. The specification is in each case a child (type) of the relevant partitioning entity. Thus, carcinomatous pathology is a type of pathology and colon location is a type of location. Where distinct specification factors are combined within a single tree errors often result. Thus in the NCIT we find:

mutagen is_a *chemical modifier*

chemical modifier semantic_type *chemical viewed functionally*

drugs and chemicals, functional classification semantic_type *classification*

drugs and chemicals, functional classification is_a *drugs and chemicals*

so that *functional classification* is classified as a subtype of *drugs and chemicals.*

DNA damage, which plays an important role within the pathogenesis of colon carcinomas, is asserted to belong to the *semantic type: Cellular or Molecular Dysfunction*. In:

DNA damage Biological_Process_Has_Associated_Location *chromosome structure*

DNA damage Biological_Process_Has_Initiator_Chemical_or_Drug *mutagen*

however, NCIT asserts also that *DNA damage* is a *process*. The problem here is that functions and processes belong to ontologically distinct top-level categories. The former are continuants, the latter occurrents.[7] Functions are powers or potentials which can be realized in corresponding processes. That function and process cannot be identified follows also from the fact that many functions are never realized.

Even if *DNA damage* were a biological process (which we doubt), then one would still not need to represent this fact twice, once by means of an explicit assertion of an is_a relation, and again by making it explicit within the assertion of relationships in which *DNA damage* enters as a term. The NCIT's complex relational expressions bring not only redundancy (and occasional contradiction) but also serve as an obstacle to the goal of integration with other ontologies, for which it is important that relations be both clearly defined and maximally general in scope. [Smith et al 2005]

[7] http://ontology.buffalo.edu/bfo/BFO.htm

7 Predisposing Factors

Alcohol consumption is one of the predisposing factors for colon carcinoma. There are both physical and behavioral aspects of such consumption, and NCIT mixes the two together by taking over definitions of 'alcohol consumption' from two distinct sources, defining it both as: 'consumption of liquids containing ethanol; includes the behavior of drinking the alcohol (CSP2003)' and as: 'behaviors associated with the ingesting of alcoholic beverages, including social drinking (MeSH2001)'. The Thesaurus thus asserts that *alcohol consumption* is:

a. *physical consumption* (from the definitions)

b. *individual behavior* (from SN)

c. (on some occasions) *social behavior* (from MeSH)

This is a good example of the mistakes which result when term-to-term matching is used to create an "ontology of ontologies" on the basis of component parts which are of varying quality and without careful consideration of the meanings associated with the terms in the different sources.

The Thesaurus assigns to *physical activity* the semantic type: *Daily or Recreational Activity* (!). At the same time it asserts:

physical activity is_a *health behavior*

Each case of physical activity, then, is a case of health behavior (!).

The confusion between physical activity and behavior is extended in the case of *obesity* which is on the one hand assigned the semantic type *Sign or Symptom* and is on the other hand classified as follows:

obesity is_a *symptom*

symptom is_a *other finding*

Here the physical parameters are not considered at all. Rather *obesity* is classified as a *finding*, an extremely broad class (analogous to *concept*), which is applied on the basis of how the corresponding knowledge is gained by healthcare professionals. Such jumps from one partition to another leave gaps which cannot be spanned by inference.

For *old population* we have:

old population is_a *population group*

population group is_a *social concept*

A population group, then, is a certain kind of concept.

Confusions also arise with relations which are not well defined. For example, *APC gene* is asserted to stand in the following relations

APC gene Gene_is_Element_in_Pathway *TGF beta signaling pathway*

APC gene Gene_is_Element_in_Pathway *WNT signaling pathway*

APC gene Gene_Plays_Role_in_Process *cell adhesion*

APC gene Gene_Plays_Role_in_Process *cytoskeletal modeling*

Pathways are built out of multiple subprocesses, here called ‘Elements’. Yet the processes of *cell adhesion* and *cytoskeletal modeling* mentioned in the above relations are also composed of many subprocesses and thus do not differ in their ontology from those classes here called ‘Pathways’. It therefore makes little sense to represent a gene as an Element of the one and as a Role Player in the other. We do not really understand what the Thesaurus means by ‘Element’ but we surmise that it is meant to represent the fact that there are other subprocesses in which the particular gene at issue is not involved. This is also true, however, for cell adhesion and for cytoskeletal modeling, and thus the relations involved should be identical in all of the four cases.

8 Clinical Manifestations

The inaccuracies related to the representation of the clinical management of colon carcinoma are very similar to those already mentioned above. For example, *obstruction* and *perforation* are assigned the semantic type: *Finding*. *Rectal hemorrhage* is classified as *other finding*.

There are situations where the Thesaurus puts *neoplasm*, a continuant and *neoplastic process*, an occurrent, together under a single heading. This *mucinous neoplasm*, one of the most aggressive kinds of colon carcinoma, is classified as follows:

mucinous neoplasm is_a *neoplasm by morphology*

neoplasm by morphology is_a *neoplasm*

mucinous neoplasm semantic type *neoplastic process*

This is rather as if one were to identify fracturing process with the fracture itself, the result of such a process. On the therapeutic side, there are many cases where an agent used within a therapy is classified together with the therapy itself. *BCG therapy*, which involves the use of an immunomodulator, is assigned three distinct semantic types – *bacterium*, *immunologic factor* and *pharmacologic substance*, none of which represent it as a therapy.

On the other hand there are cases where a drug combination itself is represented as a therapy even if it mentions only the names of the involved drugs. Thus *Capecitabine/DJ-027* is assigned the semantic type *Therapeutic or Preventive Procedure* and is classified as follows:

Capecitabine/DJ-027 is_a *chemotherapeutic regimen*

9 Conclusion

In adhering to its legacy in the UMLS Semantic Network, the NCI Thesaurus has increased the number of its inaccuracies. But there are also mistakes which are the responsibility of the Thesaurus itself. If the Thesaurus were to be used for representing entities involved in the location, pathogenesis or management of carcinomas, then it needs to be thoroughly restructured, and this is all the more the case if Electronic Health Records are to use the Thesaurus as the source of terms for the entities involved in carcinomas.

The NCIT does provide a rich terminology for carcinomas, which makes it a good starting point for ontology work in the cancer domain. Moreover, the problems which are present within the Thesaurus are, as we have seen, not new. One of the reasons why current Electronic Health Record standards restrict the use of standard terminologies and ontologies as code providers is to ensure that employing a particular code would convey a single corresponding term. And when one attempts to go further than that on the basis of current approaches, in order to use the structures of terminologies and ontologies in a way that supports the drawing of inferences, then experience has shown that one is confronted by formidable obstacles. If the necessary integration is to be accomplished, then the structure found within those terminologies and ontologies must be aligned with each other and with those found within Electronic Health Records on the basis of robust formal principles. We established in a controlled comparison that FMA, though its representation is restricted to the domain of (non-pathological) anatomy, does far better in this respect. We therefore recommend a thorough audit of the NCI Thesaurus on the basis of the principles followed by the FMA.

Acknowledgements. Work on this paper was carried out under the auspices of the Wolfgang Paul Program of the Humboldt Foundation and also of the EU Network of Excellence in Semantic Datamining and the project “Forms of Life” sponsored by the Volkswagen Foundation.

References

1. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine.(2005). In press
2. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-Based Error Detection in SNOMED-CT® Medinfo. 2004 (2004) 482-6.
3. Devita VT, Hellman S, Rosenberg SA. Principles and Practices of Oncology. Chapter 33. Cancers of the Gastrointestinal Tract. 33.7 Cancer of the Colon. 6th edition. CD
4. Kumar A, Yip L, Smith B, Grenon P. Bridging the Gap between Medical and Bioinformatics Using Formal Ontological Principles. Computers in Biology and Medicine. 2005. In press
5. Kumar A, Smith B. On Controlled Vocabularies in Bioinformatics: A Case Study in Gene Ontology. Drug Discovery Today: BIOSILICO, 2, (2004) 246-252 [2004a]
6. Kumar A, Smith B. Towards a Proteomics Metaclassification. IEEE Fourth Symposium on Bioinformatics and Bioengineering, Taichung, Taiwan. IEEE Press. (2004) 419-427 [2004b]
7. Kumar A, Smith B, Borgelt C. Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations. CompuTerm Aug 29, 2004: 3rd International Workshop on Computational Terminology: 31-38.
8. Kumar A, Smith B. The Universal Medical Language System and the Gene Ontology: Some Critical Reflections. Lecture Notes in Computer Science. (2003) Sep; 2821/2003: 135 – 148.
9. Smith B, Ceusters W, Koehler J, Klagges B, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies Genome Biology. 2005;6:R46. 2005a
10. Smith B, Ceusters W, Temmerman R. Wuesteria. MIE 2005. (2005) in press. 2005b