The CEN ISO Standard Categorial Structure as a Top-Level Set of Constraints for Ontology Disambiguation

Jean-Marie Rodrigues[1,2,3,*], Stefan Schulz[4,5], Cédric Bousquet[1,2], and Julien Souvignet[1,2]
[1] Department of Public Health CHU University of Saint Etienne, France
[2] INSERM U 872, Eq. 20, Paris, France
[3] WHO-FIC Collaborative Centre for International Classifications in French Language, Paris, France
[4] Institute for Medical Informatics, Statistics, and Documentation, Medical University, Graz, Austria
[5] Institute for Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany

ABSTRACT

Since 1990 CEN TC 251 and ISO TC 215 have developed an approach named Categorial Structure. It is a logic-based language that aims at standardising the upper level structure of a terminological model rather than agreeing on a reference clinical terminology or on a language-independent biomedical ontology. Since 2000 methods, tools and techniques based on upper level ontology and description logic formalism have been developed in the Semantic Web and the bio-ontology communities. The objective of this paper is to analyse the relation between the two approaches in order to promote the complementary use of ontology and structured information model tools in the harmonisation between biomedical terminologies and to improve semantic interoperability.

1 INTRODUCTION

Standardisation in Health Informatics started in the U.S. with the HL7 user group. The European Standard Body CEN TC 251 WG2 (Comité Européen de Normalisation Technical Committee 251 Working Group 2) and later the International Organisation for Standardization (ISO) TC 215 WG3 elaborated and developed a standard approach for biomedical terminology named Categorial Structure (CAST) (Rodrigues, 2008), mainly based on the ontology-driven tools developed in the nineties as a result of the GALEN project (Rector,1995).

Since 2000, ontology-driven methods, tools and techniques have been developed in the Semantic Web and the bio-ontology communities. Numerous research groups have moved towards upper-level ontologies (ULO) in order to manage the organization, integration and interoperability of biomedical information in a more principled way. Not only in research circles, but also in the medical terminology business, there is an increasing belief that using formal ontology approaches helps improve the quality of terminological systems. Description logics (Baader, 2007) often using Web Ontology Language (OWL) (Horridge, 2009) has become a quasi-standard for formal ontologies, which intend to describe (as much as possible) the consensus on the nature of entities in a given scientific domain, independently of linguistic or conceptual variation.

The objective of this paper is to analyse the relation between the two approaches, viz. upper-level ontologies (ULO) vs. Categorial Structure (CAST), to ease the use of applied ontology to harmonise different biomedical terminologies. In a first part we present the definition and specifications of the CAST standard methodology (Rodrigues, 2008); whereas a second part is dedicated to upper level ontologies and the alignment of the CAST of Patient Safety (PS) Conceptual Concept Framework with the top-level ontology BFO (Smith, 2011) and the top-domain ontology BioTop (Beißwanger, 2008). In a third part we present the CAST application to the ICD 11 revision. Finally we will discuss the relation between the CAST and another ontology approach to disambiguate healthcare terminologies.

2 CEN ISO CATEGORIAL STRUCTURE APPROACH

The CEN Categorial Structure was defined, as presented in (Rodrigues, 2008), as a minimal set of constraints to represent a biomedical terminology in a given health care domain with the goal to safely communicate. It is a definition of a minimal semantic structure or ontology framework describing the main properties of the different artefacts used as terminology (controlled vocabularies, nomenclatures, reference terminologies, coding systems and classifications): a model of knowledge restricted to 1) a goal, 2) a list of semantic categories, 3) the list of semantic links or relations between semantic categories constrained by their associated semantic categories, and 4) the minimal constraints allowing the generation and the validation of well-formed terminological expressions. As a consequence any biomedical terminology artefact claiming conformance to this standard shall attach with the data sent the Categorial Structure of the terminology used. The Categorial Structure shall satisfy the four constraints, but it can add more constraints. For instance, the CAST for terminological systems of surgical procedures (Rodrigues, 2011) specifies that:

(1) The goal is to model surgical procedures,

* Corresponding Author : JM Rodrigues, CHU de St Etienne, SSPIM, Chemin de la Marandière, 42 270 Saint Priest en Jarez, France, rodrigues@univ-st-etienne.fr

(2) The main semantic categories are *Human Anatomy*, *Deed*, *Interventional Equipment* and *Lesion*;

(3) The semantic links are **has_object**, **has_site**, **has_sub_surgicaldeed**, **has_means**...

(3.1) **has_object** is authorised between *Deed* and *Human anatomy* or *Interventional Equipment* or *Lesion*;

(3.2.) **has_site** is authorised between *Interventional equipment* or *Lesion* and *Human anatomy*;

(3.3) **has_means** is authorised between *Deed* and *Human anatomy*, *Interventional equipment* or *Lesion*;

(3.4) **has_sub_surgicaldeed** is authorised between *Deed and Deed*;

(4) The minimal constraints required are:

(4.1) A *Deed* and **has_object** shall be present;

(4.2) *Human anatomy* shall always be present either with the relation **has_object** or with a **has_site**;

(4.3) Use of *Lesion* shall be restricted to macroscopic lesion and to cases where it allows differentiating the procedure from procedures using the same deed and the same human anatomy;

(4.4) When **has_sub_surgicaldeed** is used, the *Deed* on the right side of the semantic link must be conform to the rules 3.1, 3.2 and 3.3.

3 UPPER LEVEL ONTOLOGY APPROACH

3.1 BFO

Basic Formal Ontology (BFO) is a philosophically inspired top-level ontological framework (Smith, 2011), which provides general categories as a basis for domain ontologies of a given level of granularity.

BFO is divided into two main parts: *Continuants* (entities which continue to exist through time), and *Occurrents* (entities which exist in full in a single instant of time). It provides a coherent and unified understanding of basic ontological entities that are fundamental to describe a science-based reality, and it allows integrating domain terminologies especially biomedical ones. BFO v1.1 includes 39 classes. This version does not include any relations, an ontology of relationships such as OBO Relation Ontology (RO) (Smith 2005). A new release of BFO (v2) is currently under development and should incorporate RO.

3.2 BioTop

BioTop is a top-domain ontology for biomedicine, represented in OWL-DL. It provides a class and a relation hierarchy, together with description logics axioms (Baader, 2007). Furthermore, BioTop can serve as a top-level model for creating new ontologies for more specific domains or as aid for aligning or improving existing ones. BioTopLite is a lighter experimental version. BioTopLite has 49 classes and 50 object properties (including inverse properties) and presents only top level classes with a low granularity level.[1] BioTop is inspired both by BFO (Smith, 2011) and DOLCE (Gangemi, 2002), providing bridging ontologies to both.

3.3 BFO and BioTop vs. Categorial Structure

Analyzing the Categorial Structure in the light of description-logics based upper ontologies we find so many parallelisms that the hypothesis is allowed that the Categorial Structure can be interpreted, by and large, in terms of description logics axioms.

For instance, the basic categories can be declared as disjoint classes, e.g. *Human Anatomy* and *Deed* subclassOf *Nothing* (cf. (1) in section 2).

Semantic links in the model can be considered object properties with domain and range restrictions. E.g., **has_means** can be considered an OWL object property with its respective domain and range restrictions, according to (3.3):

Thing subclassOf inverse (**has_means**) only *Deed*
Thing subclassOf **has_means** only
(*Human anatomy* or *InterventionalEquipment* or *Lesion*)

Minimal constraints in the Categorial Structure model correspond to existential axioms, e.g. (3.1):

Deed subclassOf **has_object** some *Thing*

or more complex (3.2):

Deed subclassOf ((**has_object** some *HumanAnatomy*) **or**
(**has_site** some *HumanAnatomy*))

One major difference between the Categorial Structure and description logics is that OWL object properties are exclusively relations between individuals (hence they require quantifiers whenever used in axioms including classes), whereas the links in the Categorial Structure hold between classes. However, the links between classes can be defined by means of the links between individuals in a similar way as done in the context of RO (Smith 2005)[2]:

$$A \text{ rel } B =_{def} A \text{ subclassOf } \mathbf{rel} \text{ some } B$$

3.4 PS-CAST Mapping to ULO

Following the WHO Department of Patient Safety report on a conceptual basis (a list of terms and definitions of patient safety concepts[3] (Runciman, 2009) named International Classification for Patient Safety (ICPS), an ontological representation using the Categorial Structure method, named PS-CAST (Souvignet, 2011) was issued. For example, here a representation of the Fall incident :

has mereologic relation

has cause: (*Agent: Physical Environment* or *Health Services: health Intervention*)

has circumstance: (*Circumstance: Contributing Factors: Staff factors* or *Mitigating Factors: Effective protocol available*)

[1] Both versions of January 29, 2012, available at http://purl.org/biotop

[2] Relation between classes/concepts: *italics*, between individuals: **bold**

[3] http://www.who.int/patientsafety/taxonomy/en/

has consequence: (*Harm: Patient Outcomes: Fracture* or *Organizational Harm: Legal Ramifications*)

has characteristics

has type incident: (*Incident Type: Fall*)

has defining characteristics: (*Incident Characteristics: sudden, unintended, uncontrolled downward displacement of a patient body to the ground*)

has associative relation

has detection: (*Detection*)

has location: (*Care_Setting: Hospital*)

has people involved: (*Person: Health Care Profesional*)

PS-CAST aims to integrate more granular value sets for subsequent development and to test the robustness of the CAST across different PS recording systems in the world it has been decided to try to map it to upper level ontologies.

Two manual alignments were performed; first there was between classes of PS-CAST with BFO classes for BFO is the most generic ULO. The second step was to align both PS-CAST classes and relations with BioTop which is a domain ontology.

The alignment method, however, remained the same for both tasks. First, we performed an analysis of the class definitions proposed in the ontologies (BFO and BioTop), where relevant classes were kept. Then a search for connections between these classes and PS-CAST was performed. Each alignment proposed was then reviewed by a Patient Safety expert. If class definitions were unclear or inaccurate, or if a relation was questionable, new proposals were made.

The mapping was based on the OWL-DL versions of both ontologies. It used taxonomic subsumption ("is-a") (*A* subclassOf *B*), and equivalence (*A* equivalentTo *B*).

The mappings were considered complete after a few iterations when no more changes were needed. We have ensured that no violation of domains and ranges were made.

Table 1 is an excerpt of the two proposed mappings for PS-CAST with BFO and with BioTop.

	PS-CAST	**BFO (link type)**	**BioTop (link type)**
Classes	*Incident*	*Process* (subclass)	*Process* (subclass)
	Incident_Type	*Disposition* (subclass)	*Disposition* (subclass)
	Circumstance	*Process OR Independent Continuant* (subclass)	*Condition* (subclass)
	Care_Setting	*Site* (subclass)	*Material Object* (subclass)
	Detection	*Process* (subclass)	*Action* (subclass)
	Person	*Role* (equivalent class)	*Role* (equivalent class)
	Harm	*Process* (subclass)	*Condition* (subclass)
	Anatomy	*Independant Continuant* (subclass)	(*StructuredBiologicalEntity OR ImmaterialObject*) (subclass)
	…	…	…
Properties	*has_Cause*	-	**causedBy** (equivalent relation)
	has_Consequence	-	**causes** (equivalent relation)
	has_Incident_Type	-	**hasRealization** (subrelation)
	has_Means	-	**processually-RelatedTo** (subrelation)
	has_Location	-	**hasLocus** (equivalent relation)
	…	-	…

Table 1. Excerpt of Mappings between PSCAST and BFO and BioTop

As a general observation, the naming of several PS-CAST classes suggests that these classes were defined rather by functional than by ontological criteria. For instance, a *Circumstance* is difficult to describe. E.g., a certain place can be a circumstance of an adverse event just as an event or a disposition. This ambiguity could be expressed by mapping "*PSCAST:Circumstance*" to a disjunctive expression like "*bfo:Process* or *bfo:IndependentContinuant* or "*bfo:Disposition*". In BioTop, such a disjunctive class is already there, due to the need to represent ambiguous clinical terms like "allergy" or "tumor", for which different readings (processual, material, dispositional) coexist (Schulz, 2011). A difficult distinction is also the one between *Incident* and *Incident_Type* due as well to the ontology ambiguity of incident which has to be considered as a disjunctive expression. The ontological distinction between *Incident* and *Incident_Type* proved difficult. Our current hypothesis is that the intended meaning of "*Incident_Type*" is a dispositional one, i.e. an *Incident_Type* is a realizable entity which can be realized by a corresponding *Incident*.

On one hand, due to its biomedical domain coverage, BioTop provides finer alignment of some concepts such as Anatomy, which is aligned in BFO with "*BFO:Independent Continuant*" and in BioTop Lite with ("*BioTop:Material Object* > *BioTop:Polymolecular composite entity* > *BioTop:structured biological entity" OR "BioTop:Immaterial Object*"). On the other hand, for PS-CAST domain independent classes, BioTop is less suitable. Some classes do not seem to find a place in the ontological tree and/or did not match exactly the definitions while BFO was more accurate, for example, "*PS-CAST:Circumstance*", mapped with "*BFO:Process context*".

The mapping of the relations was done only for Bio Top. It also posed difficulties such as the relation "*PSCAST:***has_means**", which is part of a process, but there is no similar relation in Bio Top, the link was made on a high-class level "*BioTop:***procesually related to**".

We have not yet compared the BioTop mapping with both BFO and RO to verify our alignments and to adjust them.

4 CAST APROACH TO ICD-11 REVISION

The World Health Organization (WHO) has initiated the revision process of the International Classification of Diseases[4] (ICD-11) in 2007. Different from past revisions done by WHO-FIC collaborating centres, the ICD-11 authoring process, which involves a large community of clinical experts, is supported by ontology-driven tools (Tudorache, 2010)(Rodrigues, 2010). Another difference is the distinction between a multi-hierarchical ICD foundation component (FC) as a blend of ontology and information model to be used as a basis for target specific linearization products.

[4] www.who.int/classifications/icd/

4.1 Pre-coordinated concepts categories: the content model[5]

This blend of ontology and information model named content model (Tu, 2010) consists of 13 information items and is a mix of text, terminology management and pre-coordinated concepts :

- *ICD Entity Title*
- *Classification Properties*
- *Textual Definitions*
- *Terms*
- *Body System/Structure Description*
- *Temporal Properties*
- *Severity of Subtypes Properties*
- *Manifestation Properties*
- *Causal Properties*
- *Functioning Properties*
- *Specific Condition Properties*
- *Treatment Properties*
- *Diagnostic Criteria*

The main pre-coordinated concepts categories are *Body System/Structure* which includes *Human Anatomy and Morphology*, *Manifestation properties*, *Causal and Functioning Properties*, which can be deterministic like environmental or probabilistic as genetic.

4.2 Postcoordinated Categories :CAST approach

To try to disambiguate this blend of terminology and information model a CAST approach was developed (Rodrigues, 2009) following different use cases.

The prerequisite was that diseases, disorders, injuries, signs and symptoms (all of which matter in ICD) cannot be attributed to one ontological category either.

The Ontology for General Medical Sciences (OGMS) (Scheuermann, 2009) made the attempt to interpret diseases as dispositions and disorders at pathological body structures, both distinguished from disease courses as processes. This approach was criticized by (Schulz, 2011) who argued that this overloads the meaning of intrinsically shallow terms which are often used as synonyms. Both groups, however, agreed on the observation that there are many terms for which different meanings co-exist: “allergy” may denote an allergic disposition or an allergic manifestation, “tumor” a lump of tissue or a process, and “fracture” a broken bone or an injury and healing process.

Three different CAST were proposed for Disease, Diagnosis and Patient Findings and Problems

Disease is the most complete view. It is based on an abnormality in the *Body System OR Structure* (morphology) or in the *Functional properties* (patho-physiology), *Causal properties* and *Manifestation Properties*.

Disease/disorder furthermore has to be distinguished from *Diagnosis*, which represents the view of a clinical decision maker who shall take decisions in sometimes uncertain situations. Diagnoses are based on a set of *Manifestation properties* and *Diagnostic criteria* to be defined by the clinical domain-specific ICD-11 revision Topic Advisory Groups TAGs. Diagnoses principally include the possibility that they do not correspond to any pathological entity in the patient (suspected diagnoses or false diagnoses).

Finally *Patient Findings and Problems* (signs, symptoms, syndromes, test results, situations, etc.) are very often mentioned in health record for surveillance or other without reaching the level of a diagnosis assumption.

4.3 The Semantic Links

They are the relations **has_finding**, **has_location**, **has_abnormality**, **has_etiology** and **has_focus**.

- **has_finding** is the link authorised between *Disease or Diagnosis* and *Manifestation Properties*.
- **has_location** is the link authorised between *Disease or Diagnosis* and *Body System/structure*.
- **has_abnormality** is the link authorised between *Disease and Functioning Properties* or *Body System/structure /Morphology*.
- **has_etiology** is the link authorised between *Disease or Diagnosis* and *Causal Properties*
- **has_focus** is the link authorised between *Patient Findings and Problems* and *Manifestation Properties*, *BodySystem/structure*, *Functioning Properties*.

4.4 The Minimal Domain Constraints

Patient Findings and Problems: at least one **has_focus.**

Diagnosis: at least one **has_finding** and at least one **has_location**

Disease: at least one **has_finding**, at least one **has_location**, at least one **has_abnormality** and at least one **has_etiology**.

5 DISCUSSION

We presented the CAST approach on different healthcare sub domains (Surgical Procedures, Patient Safety, ICD 11) and the relations between this approach and ULO approaches (BFO and Bio Top).

We showed in a first step that CAST is challenging on one hand basic ontological entities and on the other the relations (Bio Top or BFO and RO).

In the following part we demonstrated that the use of the CAST approach for the ICD-11 revision is facing the ambiguity and fuzziness of the healthcare terminologies which are oversimplifying the knowledge they contain to ease their utilization. ICD 11 CAST blends ontology with information model (and therefore epistemic) aspects, as in PS CAST. Recently an ontological triad structure-named structure-disposition process (SDP) has been proposed to disambiguate complex pathological process in another healthcare terminology SNOMED CT and the creation of a BioTop disjunctive class recommended (Schulz, 2011).

[5] www.who.int/classifications/icd/revision/contentmodel/en/index.html

A preliminary analysis has shown the difficulties and possible solutions (e.g. to use dispositions), which can still be embedded into a consistent ontological framework.

Our next steps consist in the completion of DL models of both PS-CAST and ICD11-CAST, using BioTop, disjunctive entity and, in a further stage, BFO version 2.

This can contribute to the WHO/IHTSDO Joint Advisory Group (JAG) works on a common ontological basis for the Foundation Component (FC) of ICD 11 and SNOMED CT.

In another field it would be useful that ontological alignment should become a routine part of CEN/ISO CAST development, for it can be exploited to discover similarities and differences between the existing CAST suite found in CEN/ISO publications.

The empirical nature of the CAST approach influences what should be represented - if some of these requirements are epistemic then the CAST should support them. If principled ontological frameworks cannot accommodate epistemic aspects this is not a reason to marginalize or reject them from the models (and associated knowledge products).

The CAST, BFO and BioTop approaches are truly complementary. They are providing an empirical 'reality check' to quality improvement strategies that may disallow the inclusion of epistemic notions. They can support the process of ontology construction and ease ontology application when complex healthcare knowledge is concerned.

The crucial value in having computable models of these terminologies lies in the use of formal reasoning mechanisms for validation. Built into the workflow of terminology construction and maintenance, such a combined CAST ULO approach would be an important asset towards high-quality biomedical terminological systems.

ACKNOWLEDGEMENTS

PS-CAST project has been contracted by WHO (Registration #2009/33635-0, Order 200094768, Reg. File H15 APW 221) as part of ICPS development.

REFERENCES

Rodrigues JM, Kumar A, Bousquet C, Trombert B. *Standards and biomedical terminologies: the CEN TC 251 and ISO TC 215 categorial structures. A step towards increased interoperability.* Stud Health Technol Inform. 2008;136:857-62.

Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. *A Terminology Server for medical language and medical information systems.* Methods Inf Med 1995;34(1-2):147-57.

Baader F et al. (2007). *The Description Logic Handbook. 2nd edition. Theory, Implementation, and Applications*, Cambridge, U.K.: Cambridge University Press.

Horridge M, Patel-Schneider P (2009). *OWL 2 Web Ontology Language Manchester Syntax*. http://www.w3.org/TR/2009/NOTE-owl2-manchester-syntax-20091027/

Smith B, Grenon P: *Basic Formal Ontology*. 2011 [www.ifomis.org/bfo], Last accessed: Jan 30, 2012.

Beißwanger E, Schulz S, Stenzhorn H and Hahn U. *BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies*. Applied Ontology 2008;3;4:205-212.

Rodrigues JM, Casey A, Bousquet C, Kumar A, Lewalle P, Trombert Paviot B. *Comparing existing national and international classification systems of surgical procedures with the CEN/ISO 1828 ontology framework standard* Stud Health Technol Inform. 2011;169:844-8.

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. *Relations in biomedical ontologies.* Genome Biol. 2005;6(5):R46

Gangemi A, Guarino N, Masolo C, Oltramari A and Schneider L. Sweetening Ontologies with DOLCE. Lecture Notes in Computer Science, 2002, Volume 2473/2002, 223-233, DOI: 10.1007/3-540-45810-7_18

Runciman W, Hibbert P, Thomson R, Van Der Schaaf T, Sherman H, Lewalle P. *Towards an International Classification for Patient Safety: key concepts and terms*, Int J Qual Health Care. 2009;21(1):18-26.

Souvignet J, Bousquet C, Lewalle P, Trombert-Paviot B and Rodrigues JM. *Modeling patient safety incidents knowledge with the categorial structure method*. AMIA Annu Symp Proc. 2011;2011:1300-8.

Schulz S, Spackman K, James A, Cocos C and Boeker M. *Scalable representations of diseases in biomedical ontologies.* Journal of Biomedical Semantics 2011, 2(Suppl.2):S6 doi:10.1186/2041-1480-2-S2-S6

World Health Organization. *International Classification of Diseases* (ICD). www.who.int/classifications/icd/. Last accessed: Jan 2012.

Tudorache T, Falconer S, Nyulas C, Storey MA, Ustün TB, Musen MA. *Supporting the Collaborative Authoring of ICD-11 with WebProtégé*. Proceedings of the AMIA Annual Symposium 2010, 802–806.

Rodrigues JM. *ICD-11 revision: where are we now? Ontology-driven tools and the web platform.* BMC Health Services Research 2010, 10(Suppl 2):A7 doi:10.1186/1472-6963-10-S2-A7

S. W. Tu, et al. *A content model for the ICD-11 revision. Technical Report* BMIR-2010-1405, Stanford Center for Biomedical Informatics Research, 2010.

Rodrigues JM, Kumar A, Bousquet C, Trombert B. *Using the CEN/ISO standard for categorial structure to harmonise the development of WHO international terminologies*. Studies of Health Technology and Informatics 2009; 150:255–259.

Scheuermann RH, Ceusters W, Smith B: *Toward an Ontological Treatment of Disease and Diagnosis*. In Proc. of the 2009 AMIA Summit on Translational Bioinformatics. San Francisco, California. American Medical Informatics Association; 2009:116-120.