Representing 'casualties' for epidemiological data processing

Filipe Santana[1], Roberta Fernandes[1], Daniel Schober[2], Cristine Bonfim[3], Zulma Medeiros[4,5], Fred Freitas[1]*

[1]Informatics Center, Federal University of Pernambuco (CIn/UFPE), Recife, Brazil,
[2]Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center Freiburg, Freiburg, Germany,
[3]Social Research Board, Joaquim Nabuco Foundation, Recife, Brazil,
[4]Parasitology Department, Aggeu Magalhães Research Center, Oswaldo Cruz Foundation, (CPqAM/Fiocruz), Recife, Brazil,
[5]Pathology Department, Institute of Biological Sciences, University of Pernambuco, Recife, Brazil,

ABSTRACT

This paper presents an ontological approach to formally describe the events related to the passing from life to death, supporting the retrieval of mortality cases (e.g. available in mortality databases). Such representations are needed to support public decision making related to highly disabling diseases like infectious or neglected tropical diseases.

1 INTRODUCTION

Epidemiologic policies are defined as plans for actions to support preventive and control measures for diseases or others health related problems, based on knowledge about individual and collective health. One of the most important data sources to generate *information to provide policies for decision making* is mortality data (Selig *et al*., 2010).

Casualties and diseases, like tuberculosis (TB), can be considered sentinel events for health monitoring systems. (Selig *et al*., 2004). Together with HIV/AIDS and Malaria TB is one of the three most devastating worldwide diseases) (Hotez et al., 2006), resulting in nine million newly reported TB disease cases and 1.7 million casualties each year. Such diseases require new preventive strategies, like the WHO-Stop TB program (WHO, 2009), which should ideally be based on reliable information stored in data bases, providing access to reliable epidemiologic as well as individual health care data. This should ideally be embedded in their local contexts, ultimately contributing to an enhanced understanding of the dynamics of disease spread (Cohen, 2000).

Therefore, policy efforts to help fight diseases in specific countries should embrace local information and patient data in an epidemiological database management system (Selig *et al*., 2010). These should be based on ontologies to provide the rich embedded contextual data in an integrated way and as demanded above. Exploiting consensual knowledge, as formalized in ontologies, can produce new epidemiological insight and thus help in policy management and decision-making processes (Topalis et al., 2011). In order to investigate such capabilities, we created the Neglected Tropical Disease Ontology (NTDO) (Santana *et al.,* 2011).

Ontologies, from a formal point of view, intend to describe the consensus on the nature of entities in a given scientific domain, independently of linguistic variation. Accordingly, formal ontologies are expressed by means of a formal semantics, like Description Logics (DL) (Baader et al., 2007). Nowadays, most DL ontologies are , are shared in the World Wide Web Consortium (W3C) recommended exchange syntax Web Ontology Language (OWL)[1].

The aim of this study is to ontologically formalize foundational events in the life cycle of patients, i.e. events related to the passing from life to death, and representations thereof as needed to track, store and retrieve mortality cases. Ontologized mortality databases can support decision making against relevant casualty events,. Here, we used TB cases to exemplify and create this subset representation. We particularly focused on information required to support health policy management, e.g. in the case of NTD and TB using ontologic representations. We hope to show that such ontologized epidemiological data can be exploited by logics reasoners and hence render implicit data explicit and more useable by retrieval and disease monitoring tools.. Specifically we provide patterns for robust and DL-compliant ontological representations of epidemiologically important entities like Birth, Disease, and Death.

2 MATERIAL AND METHODS

NTDO (http://www.cin.ufpe.br/~ntdo) imports and re-uses the upper level ontology BioTop (http://purl.org/biotop) (Beisswanger et al., 2008). NTDO was expanded in a middle-out approach, leveraging on established ontology construction guidelines (Rector, 2003; Schober et al., 2009). The pathogen transmission pattern was based on Santana *et al.* (2011), providing the basis to describe the tuberculosis airborne transmission, its respective pathogens and affected persons.

To perform the data retrieval, the Brazilian Mortality Information System (*Sistema de Informação sobre Mortalidade* - SIM[2]) database was converted (from dBase to SQL), and views were created (using PostgreeSQL v9.0) in order to extract demographic (age, sex, among others) and epidemiological data (deceased person, place of casualty event and casualty basic cause). After, we used JENA API to generate RDF triples with individual assertions; and mappings to the NTDO (OWL2) respective classes to enable RDF-querying over the mortality data using SPARQL[3] Query Language.

[*]To whom correspondence should be addressed: fss3@cin.ufpe.br

[1] http://www.w3.org/TR/owl2-overview
[2]http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=21377
[3] http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

The mortality data gathering was approved by the Ethics Comitee of Health Sciences Center (CCS) (Federal University of Pernambuco - UFPE), as a subpart of the project "Ontologias e as Doenças Tropicais Negligenciáveis" (CAAE - 0112.0.172.000-11), in English, "Ontologies and Neglected Tropical Diseases".

3 RESULTS

3.1 Casualty representation

In this section, all ontology modeling processes, e.g.pathological processes, casualty and transmission process, and data analysis, e.g. mortality data retrieval, are scrutinized.

Our ontological representation of casualty cases, as foundationally relevant for any epistemological analysis, follows the model described in Figure 1.

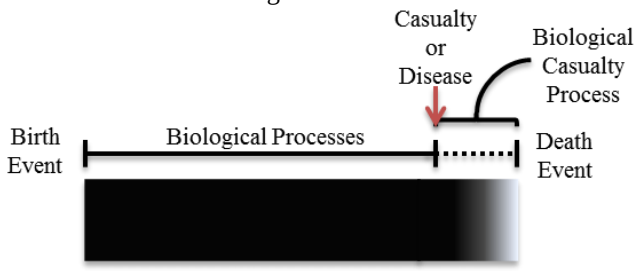

Figure 1:Organism life cycle model, including a Birth event and a Casualty Process Model.

As *birth*, *life*, *disease*, *and death* seem to be diffusely delineated concepts people tend to argue about, we introduce a semiformal notion according to Koshland (2002), who defines. "Life" as the capability of a program which has a description of the ingredients and their interaction kinetics (like the genome and the metabolome), capable of mutation and hence prone to selection . The bearer of such Life capability is an organism compartmentalized into cells and organs, which metabolize substances to generate energy for adaption, regeneration and segregation.

In addition, the notion of "birth" is based on the notion of "Living Birth", as defined by the Brazilian Geography and Statistics Institute (IBGE) as 'the expulsion or complete extraction of a product from the maternal body, after conception which after the detachment of the maternal body, breaths or gives any other life-sign, e.g. heartbeat, umbilical cord beat, or movements from voluntary muscle contraction, the umbilical cord being cut or not, and the placenta being detached or not[4]. On the other hand, 'death' means the complete extinction of any life-sign in any moment after a 'Birth"- event, i. e. cessation of the vital functions without resuscitation.

Therefore, a typical lifecycle of a living organism begins with a conception event followed by a pregnancy process which ends with an birth event (a point in time locating when it happened). Here, we are considering only the processes which happen after the conception. It is important to note that *Events* here are known to exhibit a certain behavior relative to a process (Herre *et al.*. 2007). Each organism has a lifespan and, at a later point in time, its body starts a biological death process, which can be caused by natural means, disease or casualty (as stated by the World Health Organization in "cause of death" definition[5]). It culminates, independently of how long such processes may take, with a death event.

3.2 Representational challenges

Many challenges were faced in the attempt to create a sound representation of such a structure, such as preserving identity, asserting correct cardinalities, getting support of sound background theories, and, last but not least, representing the resulting ontology in a decidable DL that could encompass the expressivity employed in the definitions. We will discuss each of these items in the following. An initial definition of a *CasualtyEvent* could b*e*

CasualtyEvent equivalentTo *Event*
and (**hasLocus** some *GeographicLocation*)
and (**hasPatient** some *DeadOrganism*)
and (**hasProcessualPart** some *BiologicalDeathProcess*)
and (**hasCasualtyInstant** some *PointInTime*)

Such a definition brings many imprecisions with regard to preserving identity and correct cardinalities. First of all, the representation purpose of this class is conveying information on the casualty of a single living organism. However the axiom does not express any cardinality constraint. Moreover, there is no guarantee that the living organism that is dying and the resulting dead organism coincide, retaining identity between these two. It makes a living organism coincides with the definition of a phased sortal, i.e., one which starts by enjoying a certain phase (in our case, a living phase) and eventually turns into a new phase (for us the phase of a dead organism).

This indeed constitutes an interesting representation problem, given that it brings about the philosophical issue of representing most (if not all) rigid classes (Guarino &Welty, 2000) as phased sortal, not to mention that it additionally provokes a discussion whether a *DeadOrganism* is still an *Organism* or not and until exactly when.

A good way to circumvent such representational problems - and probably also the common choice of the ontologists who designed all of the other biological ontologies that we found in the literature to the extent of our knowledge – is not representing a *DeadOrganism* at all.

Instead, a sound solution resides on representing *LivingOrganisms* subsumed by a *Presential*, which is a MaterialEntity sub-concept present in the General Formal Ontology, (GFO, Herre *et al.*. 2007). A *Presential* exists only at exactly one time interval (in the ontology called a *Chronoid).* As described in GFO, *Chronoids* possess two inherent and external time boundaries, *RightTimeBoundary* and LeftTimeBoundary, (Fig. 2), which can coincide in a

[4]http://www.ibge.gov.br/

[5] http://www.who.int/topics/mortality/en/

single *Chronoid*. Accordingly, *LeftTimeBoundary* and *RightTimeBoundary* represents the beginning and end, respectively, of an inner *ProcessualEntity*, realized by an object. The mereological sum of *Chronoids* represents the notion of a time region (Herre *et al.*. 2007).

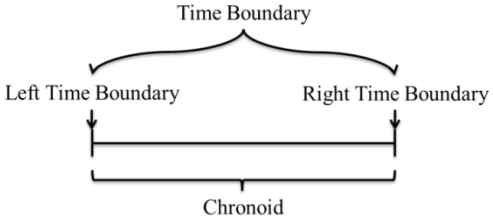

Figure 2: A *Chronoid* time interval and its boundaries along the time axis.

The axiom below should be included then:
MaterialEntity subClassOf *Presential*
departing from the premise that *LivingOrganisms* are *MaterialEntities*. Of course, we are also assuming that *MaterialEntities* are formed (*LeftTimeBoundary*) and destroyed (*RightLeftTimeBoundary*). Below are the GFO describing a *Presential* according to its *TimeBoundaries*:

Presential subClassOf MaterialEntity
and (**existsAt exactly** 1 *TimeBoundary*)

TimeBoundary subClassOf TimePoint and
(**hasRightTimeBoundary exactly** 1 *RightTimeBoundary*)
and (**hasLeftTimeBoundary exactly** 1 *LeftTimeBoundary*)

Another subtle aspect regards the range of the **hasPatient** property. For our purposes, this object property should be functional (i.e., each relation instance admits only one element in the domain and range). Nevertheless, it is not originally like that in the relation definition, which allows for more than one element from the range. In the ontology, this could mistakenly lead to the interpretation that a single casualty event can indeed represent the death of many individuals at the same time.
Even when we substitute the unknown cardinality ("some", in the axiom) by a defined cardinality (e.g. **hasPatient exactly** 1 *DeadOrganism*), the problem with preserving identity persists, and still in the case where *DeadOrganism* is replaced by a *LivingOrganism*. An alternative modeling of such a situation is depicted in figure 3, so that identity is retained by constraining the *hasCasualtyInstant* relation-value Individual to the class *LeftTimeBoundary* of the dying *LivingOrganism*, when the casualty or disease takes place. A sounder version of *CasualtyEvent*s in DL, which makes use of an object property filler-map that employs composite roles, can be seen next.

The DL agreement operator (≐, also called 'same-as') assures the coincidence between the casualty instant in which the *CasualtyEvent* occurs and the instant of the patient's casualty. For our modeling purposes, our goal was finally

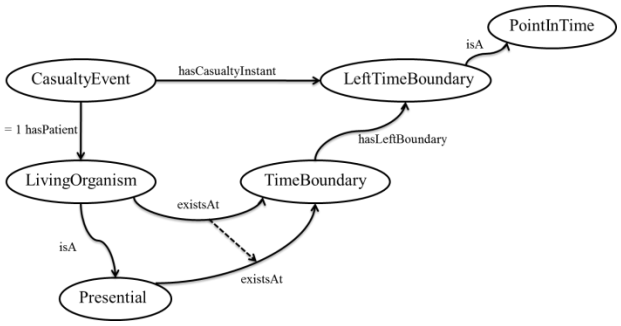

Figure 3: Graph-model of a *CasualtyEvent*, whose casualty instant coincides with the patient's casualty instant.

CasualtyEvent equivalentTo *ProcessualContext*
and (**hasLocus exactly 1** *GeographicLocation*)
and (**hasPatient exactly 1** *LivingOrganism*)
and (**hasProcessualPart** some *BiologicalDeathProcess*)
and (**hasCasualtyDate exactly 1** *PointInTime***)**
and (**hasCasualtyInstant** ≐
hasPatient o **hasPatientCasualtyInstant**)
where

hasPatientCasualtyInstant =
existsAt o **hasLeftTimeBoundary**

attained; nonetheless, for DL reasoning with the ontology, additional measures are still to be taken. If agreements are not built as role chains of functional properties in role-value-maps, then inference becomes undecidable (Schmidt-Schauss 1989). This is actually the reason why agreements are only allowed among functional roles, a constraint that is not fulfilled by our ontology yet. The properties **hasCasualtyInstant** and **hasPatientCasualtyInstant** are already functional, but **hasPatient** is not. The solution for this is creating the sub-property **hasCasualtyPatient,** which is functional and has as range the concept *LivingOrganism*. The final modeling consists of the following:

hasCasualtyPatient SubPropertyOf **hasPatient**
FunctionalProperty(**hasCasualtyPatient**)
Range *LivingOrganism*
Domain *CasualtyEvent*

Note that it is not necessary any more to qualify the property **hasCasualtyPatient,** with the class *LivingOrganism*, since its range only allows this concept.

BiologicalDeathProcess equivalentTo
BiologicalProcessualEntity
and ((**hasProcessualPart** some
(*Action* or (*BiologicalProcessualEntity*
and (not (*BiologicalDeathProcess*))))
or (**realizationOf** some Disease*Disposition*)
or (**hasLocus** some *PathologicalStructure*)))
and (**hasPatient** some *LivingOrganism*)

and (**hasTimeBoundary** some *TimeBoundary*)

BiologicalDeathProcess subClassOf

BiologicalProcessualEntity
and ((**hasProcessualPart** only
(*Action* or (*BiologicalProcessualEntity*
and (not (*BiologicalDeathProcess*))))
or (**realizationOf** only Disease*Disposition*)
or (**hasLocus** only *PathologicalStructure*)))
and (**hasPatient** only *LivingOrganism*)

It is important to note that, when a *BiologicalDeathProcess* begin (due to a casualty) in its respective *LeftTimeBoundary*, the organism is considered dead when it reaches the *BiologicalDeathProcess* respective *RightTimeBoundary*. In addition, a *BiologicalDeathProcess* is preceded by a *CasualtyEvent*, which does not have a clear process boundary, i.e. the distinction when one finishes and the other begins is sometimes fiat. But process order is known (Figure 1, dotted lines). Therefore, biologically related processes, realized dispositions or material entities (e.g. *PathologicalProcesses realizing PathologicalDispositions*, and causing *PathologicalStructures*) in some way can lead to death.

3.3 Disease and Transmission Representation

Our TB representation follows the distinction between disease and disorder (Schulz, 2010). The differences between sign and symptoms were not taken into account.

The transmission path was created based on a vector borne transmission model (Santana *et al.*, 2011), which was adapted to fit TB and its airborne transmission cycle. This model states that in a vector transmission process, the *Vector* (for TB, air or dust) is an agent and the *Pathogen* (*Mycobacterium tuberculosis*) is an additional participant and can cause a *PathologicalProcess* (TB), due to host favorable conditions.

3.4 Data Analysis

A data pre-analysis was performed in order evaluate suitability of potentially epidemiologically relevant data from differrent mortality databases. The SIM dataset contains information about most death events occurring in Brazil and was deemed suitable for populating our knowledge base.

SIM data are divided by years and has full demographical information covering the causal circumstances which led to the death of a victim. Using TB casualty data generated between 1996 and 2003, we selected cases by main cause of death, age, sex, among others. The data was added to the knowledge base as instances of classes, like *Person*, *GeographicLocation* and *DeathEvent*. An example SPARQL query for show which casualties happens by date, by location, due to which disease.

```
SELECT ?casualtyEvent ?casualityDate ?placeocurr
?placeresid ?pathologicalprocess
WHERE { ?casualtyEvent rdf:type ntdo:CasualtyEvent;
        biotop:hasLocus ?locusocurr;
        biotop:hasProcessualPart ?processualPart;
        ntdo:hasDeathPatient ?livingorganism.
        ?locusresid ntdo:name ?placeresid.
        ?locusocurr ntdo:name ?placeocurr.
        ?deathdate ntdo:casualityDate ?casualityDate.
        ?processualPart ntdo:pathologicalProcess ?patho-
logicalprocess.
FILTER( ?sex = 'Male'^^xsd:string). }
```

Using this example query, it collects 205 cases from Pernambuco State, in which 151 cases occurred in Recife. It denotes this location as a main point for health policy actions for the male population.

For our work, the choice of SPARQL, instead of SQL, lies in the graph based structure of the RDF data: SIM has two different data schemas (before 2000 and after 2001), whose differences does not affect SPARQL queries. Furthermore, this choice provides reasoning over data which cannot be found in simple SQL queries.

Acknowledgements: This work was sponsored by the German DFG grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design) and German Ministry of Education and Research (BMBF)-IB mobility project BRA 09/006.

4 REFERENCES

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

Beisswanger, E. *et al*. (2008). BIOTOP : An Upper Domain Ontology for the Life Sciences. *Appl. Ontology*, 3(4), pp.205-212.

Gruninger,M.and Fox,M. (1994). The role of competency questions in enterprise engineering. In IFIP WG 5.7, Workshop Benchmarking. Theory and Practice, Trondheim/Norway.

Guarino, N. & Welty, C. (2000). Ontological Analysis of Taxonomic Relationships. In A. Lander & V. Storey, eds. *Proceedings of ER-2000: The International Conference on Conpectual Modelling*. Springer-Verlag LNCS, pp. 1-15.

Heller, B. & Herre, H. (2004). Ontological Categories in GOL. *Axiomathes*, 14(1), pp.57-76.

Herre, H *et al*.. (2007). General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles. Research Group Ontologies in Medicine (Onto-Med), University of Leipzig.

Hotez, P.J. *et al*. (2006). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS medicine*, 3(5), p.e102.

Koshland, D.E. (2002). The seven pillars of life. *Science (New York, N.Y.)*, 295(5563), pp.2215-6.

Rector,A.L. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In Proceedings of the international conference on Knowledge capture - K-CAP'03. ACM Press, New York, USA,p. 121.

Santana, F. *et al*. (2011). Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. *Bioinformatics*, 27(13), p.i349-i356.

Selig, L. *et al*., (2004). Óbitos atribuídos à tuberculose no Estado do Rio de Janeiro. *Jornal Brasileiro de Pneumologia*, 30(4), pp.335-342.

Topalis, P. *et al*. (2011). A set of ontologies to drive tools for the control of vector-borne diseases. *Journal of biomedical informatics*, 44(1), pp.42-7.

World Health Organization (2009). *Global tuberculosis control: A short update to the 2009 report*, Geneva.