How Ontologies Can Improve Semantic Interoperability in Health Care

Stefan Schulz[*] and Catalina Martínez-Costa

Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria
{stefan.schulz,catalina.martinez}@medunigraz.at

Abstract. The main rationale of biomedical terminologies and formalized clinical information models is to provide semantic standards to improve the exchange of meaningful clinical information. Whereas terminologies should express context-independent meanings of domain terms, information models are built to represent the situational and epistemic contexts in which domain terms are used. In practice, semantic interoperability is encumbered by a plurality of different encodings of the same piece of clinical information. The same meaning can be represented by single codes in different terminologies, pre- and postcoordinated expressions in the same terminology, as well as by different combinations of (partly overlapping) terminologies and information models.

Formal ontologies can support the automatically recognition and processing of such heterogeneous but isosemantic expressions. In the SemanticHealthNet Network of Excellence a semantic framework is being built which addresses the goal of semantic interoperability by proposing a generalized methodology of transforming existing resources into “semantically enhanced” ones. The semantic enhancements consist in annotations as OWL axioms which commit to an upper-level ontology that provides categories, relations, and constraints for both domain entities and informational entities. Prospects and the challenges of this approach – particularly human and computational limitations – are discussed.

Keywords: Formal Ontology, Medical Terminologies, Health Care Standards.

1 Introduction

Semantic Interoperability had been defined in 2000 as "*…integrating resources that were developed using different vocabularies and different perspectives on the data. To achieve semantic interoperability, systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands*" [1]. Thirteen years after, the lack of semantic interoperability is, more than ever, a painful obstacle to a more rational, effective, secure, and cost-efficient data and information management in health care and biomedical research. The above citation distinguishes between

[*] Corresponding author.

D. Riaño et al. (Eds.): KR4HC 2013/ProHealth 2013, LNAI 8268, pp. 1–10, 2013.

vocabularies and perspectives, thus highlighting the deep-rooted division between ontology (“what there is”) [2] and epistemology (“what we can know”) [3]. On the level of current health informatics standards, this is mirrored by two genres of semantic resources proposed for recording health data:

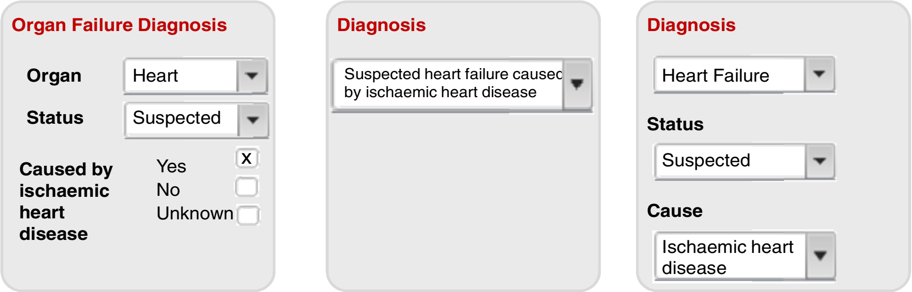
- Vocabularies, i.e. artefacts that describe and systematize meanings of terms, with the common distinctions between terminologies (which provide standardized meanings), thesauri (which introduce semantic relations between (groups of) terms), ontologies (which provide formulations of the properties and relations of domain entities [4], as denoted by domain terms), and classifications (which introduce exhaustive partitions for statistical purposes).
- Information models, which are representational artefacts that provide standardized structure (section, entry, cluster, etc.) and context (diagnosis, past history, medication order) for data acquired for a given purpose.

Typical vocabularies are the MeSH thesaurus [5], the OBO foundry ontologies [6], SNOMED CT [7], and classification systems like ICD-10 [8]. Typical information models are the ones provided by the openEHR [9] specification and the standards EN13606 [10] and HL7 [11].

To cite an example, the clinical expression "*Suspected Heart failure caused by ischemic heart disease*" would then have two components, the terminology component "*Heart failure caused by ischemic heart disease*", which is a term, which – ontologically – denotes all individual heart failure conditions caused by ischemic heart disease, and the epistemic component "*Suspected*", which expresses that the clinician has – according to the diagnostic results collected for a particular patient – a certain belief but no certainty that the patient under scrutiny hosts a medical condition of a certain type.

According to the vocabulary/information model distinction, each component should be expressed by the respective representational artefact, and the binding between both parts should be done in a uniform way. However, what we observe in practice rarely follows this paradigm. Clinical terminologies often assign a single code to complex phrases such "*Suspected Heart failure caused by ischemic heart disease*" (SNOMED CT) or "*Tuberculosis of larynx, trachea and bronchus, without mention of bacteriological or histological confirmation*" (ICD-10). In other cases a post-coordination syntax is used, such as in SNOMED CT:
84114007 | *heart failure* |: 408729009 | *finding context* |: 415684004 | *suspected* |: 42752001 | *due to* |: 414545008 | *ischaemic heart disease* |.

Fig. 1 demonstrates the different flavours of representing clinical information from the perspective of the end user. There are good reasons to tailor data acquisition forms to the users' needs and to the terminologies they are familiar with. But how can such different rendering of the same information content be reduced to an interoperable semantic representation?



Organ Failure Diagnosis

Organ: Heart

Status: Suspected

Caused by ischaemic heart disease: Yes ☒ No □ Unknown □

Diagnosis

Suspected heart failure causec by ischaemic heart disease

Diagnosis

Heart Failure

Status: Suspected

Cause: Ischaemic heart disease

Fig. 1. Heterogeneous representations of the same clinical content

This problem is addressed by the project SemanticHealthNet [12], proposing an engineering approach based on formal ontologies, using the description logics [13] language OWL-DL [14] as standardized by the Semantic Web community. Supported by easy-to-use editing tools and reasoning engines, OWL is a well-established language in biomedical ontology research and practice. SemanticHealthNet aims at developing, on a European level, a scalable and sustainable organisational and governance process for the semantic interoperability of clinical and biomedical information. The goal is to ensure that EHR systems are optimised for patient care, public health and clinical research across healthcare systems and institutions. SemanticHealthNet focuses on a cardiovascular use case, upon which the capture of the needs for evidence-based, patient-centred integrated care and for public health is based, capitalizing on existing European consensus in the management of chronic heart failure and cardiovascular prevention. Experts in EHR architectures, clinical data structures, terminologies and ontology take part of the project and tailor and pilot their best-of-breed resources in response to the needs articulated by clinicians.

2 Methods

For the purpose of interoperable descriptions within SemanticHealthNet we aim at formally describing the meaning of Health Record information entities. Each (atomic) information entity is semantically annotated, using one or more OWL DL expressions which follow predefined representational patterns. The content of these annotations addresses epistemic information (viz. whether the information is confirmed or speculative, whether it has been reported by the patient or by a caregiver, or whether it refers to the current or a past situation) and it is equally linked to clinical expressions such as the codes or combination of codes in terminologies like SNOMED CT.

Our approach to ontology follows a series of principles (cf. [15]):

- Ontologies are formal-mathematical systems that precisely describe (classes of) entities of a domain as they exist in reality [4]. This requires that the modeller always has to analyse which entities really exist. The question "what exists?" is crucial [2]. For instance, in our above example what certainly exists is the informational entity, which expresses a physician's state of knowledge related to a given type of disease. However, the existence of this disease in the reality of the patient under treatment is not guaranteed when it is referred to as "suspected". (Even the attribution "confirmed" to a diagnostic statement does not preclude a certain risk of diagnostic error).

- Each representational unit in an ontology has a strict ontological commitment supported by pairwise disjoint and exhaustive upper-level categories (process, material object, quality, information entity…); a closed set of basic relations such as '**is part of**', '**has participant**', '**is bearer of**'; constraining axioms such as that a process can only have processes as parts but not as participants.
- There is a strict bipartition between classes and individuals; what is a class and an individual is given by the domain and not at the discretion of the modeller, given that the upper level ontology provides precise elucidations of what each category means.
- Full class definitions are aimed at, as far as possible, following the Aristotelian genus / differentia principle [16].
- Naming conventions are followed, aiming at choosing self-explaining and non-ambiguous natural language identifiers [17].
- Complexity is reduced, as much as possible, by identifying reusable ontology design patterns [18].

We have chosen the upper level ontology BioTopLite [19], which provides general classes, relations (object properties), and constraints, using the description logics dialect OWL-DL [14]. Under the BioTopLite category *Information object* all (structural) information models are represented, whereas the SNOMED CT classes are placed under other BioTopLite categories like *Condition*, *Quality*, etc.

The representational challenge is two-fold. First, we have to analyse the exact meaning of the "binding" between representational artefacts and entities in a medical ontology. Second, we have to create patterns for the different distributions of content between information models and ontologies. These patterns should allow the formation of semantically equivalent (isosemantic) expressions, to be ascertained by machine reasoning. As shown in Fig. 2, information entities of clinical models will be annotated (i.e. semantically enhanced) with OWL-DL expressions conforming to certain predefined patterns, based on the proposed ontological infrastructure. These patterns are added to clinical models during their creation, in order to be filled with patient data given an appropriate software tool support. As a result, each model will have a set of annotations, which will be further processed by description logics reasoners. Finally, a query system will allow performing homogeneous queries to retrieve patient data from heterogeneously represented datasets.

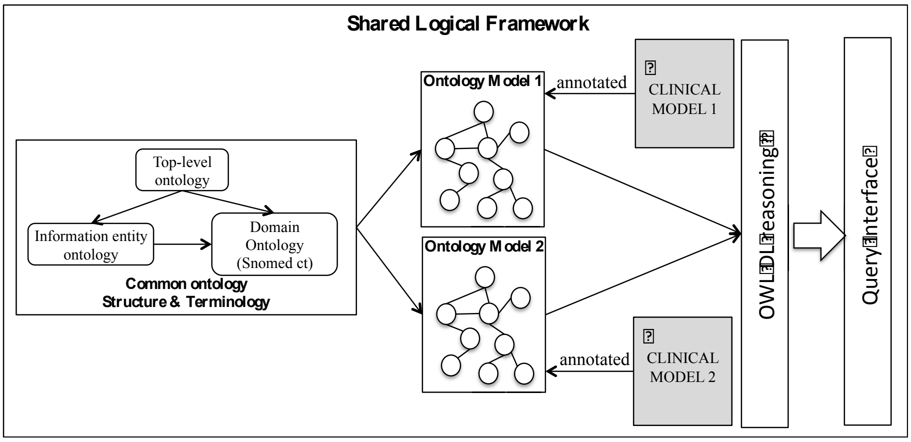


Fig. 2. Shared logical framework proposed in SemanticHealthNet

3 Results

A view of an ontology as a formal account of "what exists" reaches its boundaries when it is expected to represent statements about what does not exist, or about what only exists to a certain degree of likelihood. We reiterate that the latter abounds in medicine, where diagnoses derive from the results of clinical, lab, and imaging findings, which, synoptically, make the presence of a certain disorder very likely, but nevertheless carry a certain degree of uncertainty.

A recorded statement "Dr X suspects that Y suffers from heart failure" is, strictly spoken, a second-order statement. The subject-matter of Dr X's utterance is a sentence for which the truth-value remains undefined (the meaning of the verb "suspects", insinuates certain likelihood for a positive truth value).

As there is no way in OWL-DL to express such statements, we have suggested an approximation, the validity of which is, however, still subject to investigations. Let **d** be a diagnostic statement instance and *S* a clinical situation type referred to:

d rdf:type '*Diagnostic statement*' and
'**has information quality**' some *Suspected* and
'**is about situation**' only *S*

The use of the value restriction ('only' in Manchester Syntax) reduces the type of the clinical situation to *S*. However, it does not require that *S* is instantiated, which would be the case with the connector 'some'.

The information instance **d**, as such is characterized as a member of the class '*Information entity*', which is further specified by the attribute class *Suspected.*

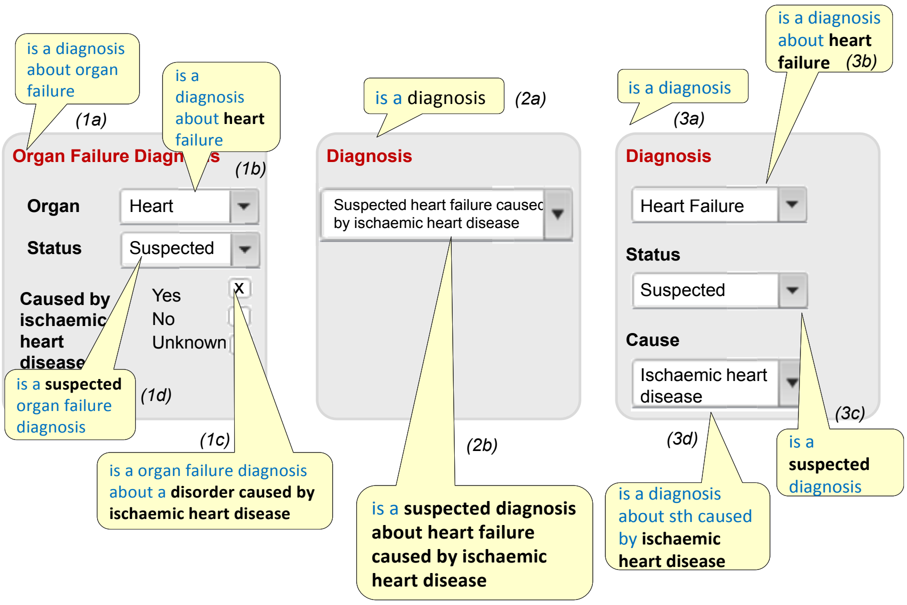

is a diagnosis about organ failure
(1a)
is a diagnosis about **heart** failure
(1b)
Organ Failure Diagnosis
Organ
Heart
Status
Suspected
Caused by ischaemic heart disease
Yes ☒
No □
Unknown □
is a **suspected** organ failure diagnosis
(1d)
(1c)
is a organ failure diagnosis about a **disorder caused by ischaemic heart disease**
is a diagnosis
(2a)
Diagnosis
Suspected heart failure caused by ischaemic heart disease
(2b)
is a **suspected diagnosis about heart failure caused by ischaemic heart disease**
is a diagnosis about **heart failure**
(3b)
is a diagnosis
(3a)
Diagnosis
Heart Failure
Status
Suspected
Cause
Ischaemic heart disease
(3c)
is a **suspected** diagnosis
(3d)
is a diagnosis about sth caused by **ischaemic heart disease**

Fig. 3. Three isosemantic representations (see Fig. 1), with free-text annotations attached, each of which with a pointer to a corresponding OWL-DL representation (1a- 3d) in the text

Fig. 3 enhances Fig. 1 by semi-formal free text annotations. Both fixed and variable elements (in bold) are referred to in the annotations, and each one can be translated into an OWL-DL axiom. Each form corresponds to one information entity instance, viz. **d1** (left), **d2** (center), and **d3** (right), and is defined by the conjunction of its annotations. Following the axioms for each of the three forms are provided.

d1 rdf:type '*Diagnostic statement*' and '**is about situation**' only '*Organ failure*' (1a)

d1 rdf:type '*Diagnostic statement*' and '**is about situation**' only '*Heart failure*' (1b)

d1 rdf:type '*Diagnostic statement*' and '**is about situation**' only ('*Disorder*' (1c)
and '**caused by**' some '*Ischaemic heart disease*')

d1 rdf:type '*Diagnostic statement*' and '**is about situation**' only '*Organ failure*' (1d)
and '**has information quality**' some '*Suspected*'

d2 rdf:type '*Diagnostic statement*' (2a)

d2 rdf:type '*Diagnostic statement*' and '**is about situation**' only ('*Heart failure*' (2b)
and '**caused by**' some '*Ischaemic heart disease*')
and '**has information quality**' some '*Suspected*'

d3 rdf:type '*Diagnostic statement*' (3a)

d3 rdf:type '*Diagnostic statement*' and '**is about situation**' only '*Heart failure*' (3b)

d3 rdf:type '*Diagnostic statement*'
and '**has information quality**' some '*Suspected*' (3c)

d3 rdf:type '*Diagnostic statement*' and '**is about situation**' only ('*Disorder*'
and '**caused by**' some '*Ischaemic heart disease*') (3d)

The above axioms have been defined by following a predefined pattern for representing a clinical diagnosis statement:

'*Diagnostic statement*' **subClassOf**
'*Information Entity*'
and '**has information quality**' some '*CertaintyAttributeC*'
and '**outcome of**' some '*DiagnosticProcedureD*'
and '**is about situation**' only '*Clinical SituationX*'

The expressions that end with a single capital letter represent variable parts, which can be filled in with specific clinical data (e.g. '*Suspected*' for '*CertaintyAttributeC*', or the expression '*Heart failure*' and '**caused by**' some '*Ischaemic heart disease*' for '*ClinicalSituationX*').

In a first study, we have identified the following seven patterns described in Table 1. They can be further specialised, as it is the case of the above one, used to represent a diagnostic statement and a specialisation of the pattern PT.1 shown in the table below.

Table 1. Initial set of patterns identified

Pattern ID	Pattern description
PT.1	Information about clinical situation with condition present
PT.2	Information about clinical situation with condition absent
PT.3	Information about quality
PT.4	Information about past clinical situation
PT.5	Information about no past clinical situation
PT.6	Plan to perform a process
PT.7	Clinical process

The use of patterns and their representation in OWL DL allows using DL reasoning to compare different distributions of content between information models and ontologies, in order to test whether they are semantically equivalent. For instance, a DL reasoner will infer that the three sample forms in Fig. 3, together with the values selected, refer to the diagnosis of heart failure caused by ischemic heart disease, independently of their representation as a pre-coordinated expression in SNOMED CT or by means of heterogeneous combination of information model structures. This means that the following DL query example will retrieve the three information entity instances **d1**, **d2** and **d3**:

'*Diagnostic statement*' and '**is about situation**' only ('*Heart failure*'
and '**caused by**' some '*Ischaemic heart disease*')
and '**has information quality**' some '*Suspected*'

4 Conclusions

Semantic interoperability of clinical information remains a largely unresolved issue. Both vocabularies and information models have been developed as semantic representations of structured clinical information. When bound together, syntactically diverse representations emerge, from which semantic equivalence can hardly be inferred. A solution proposed in the EU Network of Excellence SemanticHealthNet is based on formal ontologies using the description logics language OWL-DL. The basic idea is to add semantic annotations to atomic information entities. These annotations are OWL DL expressions which describe both informational and clinical entities, and which are rooted in the upper level ontology BioTopLite. The binding of both kinds of entities would, ideally, be achieved by a second-order expressivity, which, however is not supported by the language. This restriction is currently being circumvented by using OWL value restrictions.

The annotations follow a set of predefined patterns, which aim at explicitly stating what is represented by each representational artefact (i.e. vocabulary/information model). As they are described using a DL-based language, formal reasoning can be used to infer semantic equivalence even for quite different distributions of content between information models and ontologies, as well as different degrees of (pre/post)-coordination.

The SemanticHealthNet approach uses semantic artefacts as they exist and as they are used in practice. This descriptive approach is different from several attempts to pursue semantic interoperability by providing strict guidance how clinical information should be represented, which often turned out to be impracticable. An example of this is TermInfo [20], which attempts to solve the overlapping between the SNOMED CT and LOINC terminologies, and the HL7 information models. Other projects and initiatives that follow a more prescriptive approach are CIMI [21], CEM [22], LRA [23] or DCMs [24]. Although SemanticHealthNet values such standardization efforts, it works on the less optimistic hypothesis that current as well as future systems that represent clinical information will continue using different EHR models (proprietary or based on some standard), and that the cost of implementing a map to a “canonical” model that supports semantic interoperability will be too high. Instead, SemanticHealthNet proposes a semantic layer, which provides a consistent semantic representation of clinical information, but which is independent of a given information model standard and could even comply with information extracted from clinical narratives by advanced language technologies.

Nevertheless, we are aware of two kinds of significant bottlenecks, which will require special attention. First, human factors matter insofar as the semantic annotations of information model components and values require considerable intellectual effort.

We postulate that the use of semantic patterns can mitigate this effort, as it presents a simplified representation for the user from which OWL expressions are automatically generated. The task would therefore boil down to the selection of the patterns to be used by clinical models. Ideally this could be supported by appropriate clinical model editing tools.

Second, reasoning performance has to be controlled. OWL DL entailment is known to be complete for NEXPTIME, which limits its scalability. However, preliminary tests in SemanticHealthNet, using OWL-DL design patterns for the SNOMED CT context model [25] on medium-size ontology modules have shown favourable runtime performance. Besides, DL reasoning does not perform well with a big amount of instance data, which would make SPARQL [26] the query language of choice. However, SPARQL does not support most of OWL entailments, which means that it lacks reasoning capabilities which are essential in our approach. Other alternatives and query languages, e.g. combinations of SPARQL with OWL have to be further investigated.

Summing up, SemanticHealthNet is breaking new ground by consequently using Semantic Web techniques, above all description logics, as an intermediate representation for both ontological and epistemic aspects of the electronic health record. The main goal is to address the needs for an improved interoperability of clinical data, which is an important prerequisite for clinical data management and research support in the beginning era of data-intensive personalized medicine. Still in an experimental phase, SemanticHealthNet is currently focussing on the representation of a heart failure summary. Although it could be shown that semantic interoperability can be supported, the technological uptake of this approach will require a series of challenges (human, computational) to be met, as well as a consensus process within a heterogeneous group of stakeholders.

Acknowledgements. This work has been funded by the SemanticHealthNet Network of Excellence within the EU 7th Framework Program, Call: FP7-ICT- 2011-7, agreement 288408. http://www.semantichealthnet.eu/

References

1. Heflin, J., Hendler, J.: Semantic Interoperability on the Web (2000), http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf (last accessed July 17, 2013)
2. Quine, W.V.: On what there is. In: Gibson, R. (ed.) Quintessence-Basic Readings from the Philosophy of W. V. Quine. Belknap Press, Cambridge (2004)
3. Bodenreider, O., Smith, B., Burgun, A.: The Ontology ‐ Epistemology Divide: A Case Study in Medical Terminology. In: Proceedings of FOIS 2004, pp. 185–195. IOS Press, Amsterdam (2004)
4. Hofweber, T.: Logic and Ontology. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, 2013th edn. (Spring 2013), http://plato.stanford.edu/archives/spr2013/entries/logic-ontology/ (last accessed July 17, 2013)
5. United States National Library of Medicine (NLM). Medical Subject Headings, MeSH (2013), http://www.nlm.nih.gov/mesh (last accessed July 17, 2013)

6. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.: OBI Consortium. In: Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S. (eds.) The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology, vol. 25(11), pp. 1251–1255 (November 2007)
7. Systematized Nomenclature of Medicine - Clinical Terms, SNOMED CT (2008), `http://www.ihtsdo.org/snomed-ct` (last accessed July 17, 2013)
8. World Health Organization (WHO). International Classification of Diseases (ICD) (2013), `http://www.who.int/classifications/icd` (last accessed July 17, 2013)
9. OpenEHR. An open domain-driven platform for developing flexible e-health systems, `http://www.openehr.org` (last accessed July 17, 2013)
10. En13606 Association, `http://www.en13606.org/` (last accessed July 17, 2013)
11. Health Level Seven International, `http://www.hl7.org/` (last accessed July 17, 2013)
12. SemanticHealthNet Network of Excellence, `http://www.semantichealthnet.eu/` (last accessed July 17, 2013)
13. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook. Theory, Implementation, and Applications, 2nd edn. Cambridge University Press, Cambridge (2007)
14. W3C OWL working group. OWL 2 Web Ontology Language, Document Overview. W3C Recommendation (December 11, 2012), `http://www.w3.org/TR/owl2-overview/` (last accessed July 17, 2013)
15. Schulz, S., Jansen, L.: Formal ontologies in biomedical knowledge representation. Yearbook of Medical Informatics (2013)
16. Cohen, S.M.: Aristotle's metaphysics. Stanford Encyclopedia of Philosophy (2012), `http://plato.stanford.edu/entries/aristotle-metaphysics/` (last accessed July 17, 2013)
17. Schober, D., Smith, B., Lewis, S.E., Kusnierczyk, W., Lomax, J., Mungall, C., Taylor, C.F., Rocca-Serra, P., Sansone, S.A.: Survey-based naming conventions for use in OBO Foundry ontology development. BMC Bioinformatics 10, 125 (2009)
18. Seddig-Raufie, D., Jansen, L., Schober, D., Boeker, M., Grewe, N., Schulz, S.: Proposed actions are no actions: re-modeling an ontology design pattern with a realist top-level ontology. J. Biomed. Semantics 3(suppl. 2), S2 (2012)
19. Schulz, S., Boeker, M.: BioTopLite: An Upper Level Ontology for the Life Sciences. In: Evolution, Design and Application. Workshop on Ontologies and Data in Life Sciences, Koblenz, Germany, September 19-20 (2013)
20. TermInfo Project, `http://www.hl7.org/special/committees/terminfo/` (last accessed July 17, 2013)
21. Clinical Information Modeling Initiative (CIMI), `http://informatics.mayo.edu/CIMI/` (last accessed July 17, 2013)
22. Clinical Element Model (CEM), `http://informatics.mayo.edu/sharp/` (last accessed July 17, 2013)
23. Logical Record Architecture (LRA), `http://www.connectingforhealth.nhs.uk/systemsandservices/data/lra` (last accessed July 17, 2013)
24. Detailed Clinical Models (DCMs), `http://www.detailedclinicalmodels.nl/` (last accessed July 17, 2013)
25. Martínez Costa, C., Schulz, S.: Ontology-based reinterpretation of the SNOMED CT context model. In: Fourth International Conference in Biomedical Ontologies (ICBO 2013), Montreal, Canada, July 6-9 (2013)
26. SPARQL Query Language For RDF. W3C Recommendation (January 15, 2008), `http://www.w3.org/TR/rdf-sparql-query/` (last accessed July 17, 2013)