eHealth2015 – Health Informatics Meets eHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-524-1-65

Semi-Automated Evaluation of Biomedical Ontologies for the Biobanking Domain Based on Competency Questions

Philipp HOFER[a,*], Sabrina NEURURER[a,*], Helga HAUFFE[a,b], Thomas INSAM[a,c], Anette ZEILNER[a,d] and Georg GÖBEL[a,1]
[a]*Department of Medical Informatics, Statistics and Health Economics, Medical University of Innsbruck, Innsbruck, Austria*
[b]*Department of Urology, Medical University of Innsbruck, Innsbruck, Austria*
[c]*Department of Gynecology and Obstetrics, Medical University of Innsbruck, Innsbruck, Austria*
[d]*Division of Human Genetics, Medical University of Innsbruck, Innsbruck, Austria*

Abstract. Background: Biosample collections and biobank information systems have become a key enabler for medical research. Therefore it is important to identify potentially relevant ontologies to semantically enrich information related to the biobanking domain. Objectives: We present a three-stage semi-automated evaluation approach which allows identifying relevant ontologies for the biobanking domain based on competency questions. Methods: After identifying candidate biobanking ontologies (Stage 1) and competency questions (Stage 2), a six-step lexical evaluation approach, which assesses the coverage of concepts, properties or instances defined by competency questions is suggested and described (Stage 3). Results: We were able to perform a proof-of-concept evaluation of the OMIABIS ontology using our proposed three-stage approach together with a sample competency question. Conclusion: Our evaluation approach allows a swift evaluation of candidate ontology entities based on a search for higher hierarchy key terms that exist in comprehensive medical vocabularies in order to state the usability of specific ontologies for the biobanking domain.

Keywords. Biological Specimen Bank, Biomedical Ontologies, Evaluation, Natural Language Processing.

1. Introduction

Biobanks store biological samples along with clinical data, informed consent declarations, processing and storage conditions. Biological samples in biobanks are widely used, from in-house clinical trials to collaborative research projects. Biobank information systems play an integral role in the organization of data amounts produced by sample collections and medical research projects over a longer period of time. The European Biobank Research Organization (BBMRI) currently works on building a common infrastructure and standardized description terminology to share biobank contents [1]. A recent study explored the regional, semantic and ontological variations

[1] Corresponding Author: Georg Göbel, Department of Medical Informatics, Statistics and Health Economics, Medical University of Innsbruck, Schöpfstraße 41/1, 6020 Innsbruck, Austria, E-Mail: georg.goebel@i-med.ac.at
* Philipp Hofer and Sabrina Neururer contributed equally to the study

of definitions of biobank terms preferably used across Europe [2] and highlights the importance to define biobank terms covering both, the clinical application and research context. Moreover, the study leads to the conclusion that the development of a global ontology for the biobanking domain would be beneficial.

Ontologies are defined as an "explicit specification of a shared conceptualization" [3]. In other terms, ontologies represent classes of entities of the real world and focus on the principled definition of concepts and relations between them [4]. Biomedical ontologies gain growing attention as they can provide highly structured knowledge representation models of different bio-/medical research fields. As of today, competency questions are mainly used to define the scope of ontologies within the ontology engineering process [5]. Medical Subject Headings (MeSH) [6] based query expansion approaches are used to improve retrieval of (bio-) medical literature [7]. The aim of this paper is to present a systematic semi-automated three-stage evaluation approach which, in a first step, allows identifying applicable ontologies for solid materials within the biobanking domain .These stages assess the lexical coverage of concepts, object properties and instances in existing biomedical ontologies in order to evaluate the usefulness of specific ontologies for the biobanking domain based on competency questions. Furthermore, we want to evaluate whether the definition of the ontology terms correspond with those used by the medical experts. Our evaluation method can be re-used and applied to other biomedical ontologies and competency question types.

The remainder of this paper is organized as follows: section 2 includes a detailed description of the methods involved in the proposed methodological approach. In section 3, the results of each correspondent step are described, while in section 4 the obtained results are discussed and finally our conclusions are stated.

2. Methods

Herein, we propose a systematic evaluation approach to assess the usability of biomedical ontologies for the biobanking domain which consist of three stages: In Stage 1, we identify candidate ontologies which might be of use for the biobanking domain. Stage 2 includes the identification of relevant competency questions that need to be covered by the ontology, which is of interest for biobanking applications. Stage 3 includes a six-step evaluation approach which assesses, whether concepts, properties or instances of the ontology identified in Stage 1 meet the requirements defined by the competency questions in Stage 2. In the following, the three stages of our proposed evaluation approach are explained in detail.

2.1. Stage 1: Identification of candidate ontologies

For the identification of candidate biobanking ontologies developed in Web Ontology Language (OWL), we suggest the screening of different ontology portals, such as the National Center for Biomedical Ontology [8] or the Open Biological Ontologies Foundry [9], a screening of biobanking projects as well as a systematic literature review. In order to provide a proof of concept of the proposed evaluation approach, we decided to use OMABIS [10] as the reference ontology. OMIABIS is based on the MIABIS [11] data model, which represents an international standard representation of biobank contents. As MIABIS neither covers all medical areas nor includes information on sample level [11], it does not reflect the user requirements for a comprehensive biobanking ontology.

Therefore it cannot be used as a starting point for an ontology evaluation approach, whereas competency questions describe use cases and therefore reflect the user demands on an ontology. According to our knowledge, OMIABIS is still under development and not used by any biobank or medical research organization yet.

2.2. Stage 2: Identification of competency questions

One approach to define ontology evaluation criteria is based on the formulation of competency questions and the expected answers that should be provided by the ontology [12]. In other words, competency questions might be used to determine whether typical information requests within a specific work environment can be satisfied by re-using existing ontologies according to the current state of the art. Therefore, a set of competency questions was created together with members of the Department of Urology and the Division of Human Genetics at the Medical University of Innsbruck in Stage 2. The competency questions were identified in collaboration with two BBMRI medical experts, each working at one of the two departments. Both members deal with regular incoming requests for tissue and body fluid samples and related clinical data from medical researchers. In our evaluation, we only focus on the coverage of concepts and relationships. Therefore, it is not necessary to consider different types [13] of competency questions.

2.3. Stage 3: Competency evaluation approach

Based on the set of competency questions, our goal is to prove whether the elements in the competency questions are covered by the concepts and object properties of selected biomedical ontologies that were identified from different ontology portals. As we want to present an international approach and most of the ontologies are provided in English, we assume that a limitation to the English language is justified. The steps including stop word removal, lemmatization and matching of the ontology entities are processed automatically. The proposed algorithm takes a competency question *CQ* as an input. As an output it states, whether a competency question is covered by an ontology *O*. This is true if each relevant token of the *CQ* can be mapped to a concept, object property or instance of the ontology *O*. The algorithm consists of the following six steps:

(i) Split the *CQ* into a *sequence S* of tokens.
(ii) Remove all stop words from *S*.
(iii) Apply a lemmatization or stemming algorithm to the remaining tokens.
(iv) For each token t_i from *S*, match t_i against all concepts c_i from the *ontology O*.
   a. If there exists a concept c_j where $t_i=c_j$ then add the concept c_j to the results, remove token t_i from the sequence *S* and match the next token t_{i+1}.
   b. Otherwise try to identify all synonyms for t_i. If t_i exists in the WordNet database, a list of synonyms for t_i might be obtained.
   c. It might be the case that no synonyms exist for t_i. However, there might be a more general enclosing parent term p_i of t_i with a corresponding concept in the ontology. Therefore, we also try to match potential higher terms that might be obtained from MeSH.
   d. Again, repeat Step *(a)* for each synonym and parent term p_i of token t_i.

(v) Repeat Step (i)-(iv) for all object properties P (set of concepts C is replaced by the set of properties P)
(vi) Repeat Step (i)-(iv) for all instances I of O.

In Step *(i)*, each competency question is split into a sequence of single tokens. In Step *(ii)*, all stop words are removed from the input sequence. A list of English stop words was obtained from the MYSQL sources. After removing all stop words, the Python based NLTK WordNet Lemmatizer [14] [15] library was used to gain the word stem of each of the remaining tokens *(Step iii)*. In Step *(iv)(a)*, each of the remaining tokens is being matched against concepts, object properties and instances of the ontology. The matching between the competency questions terms and ontology entities was performed with a SPARQL query based on a regular expression matching any occurrence of the given term.

```
SELECT ?label ?concept
{
?concept a owl:Class .
?concept rdfs:label ?label .
FILTER(REGEX(STR(?label), „<CQ term>“))
}
```

Figure 1. SPARQL query for matching concepts in the ontology with a CQ term as input.

We also want to incorporate possible synonyms of the competency question terms as potential matching candidates *(iv)(b)*. MeSH was the vocabulary thesaurus used to identify broader terms in the hierarchy related to the competency question terms *(iv)(c)*. MeSH provides a very high coverage of clinical concepts and relationships among medical terminologies. We used descriptors and entry terms from the MeSH vocabulary. Entry terms of MeSH represent synonyms and related terms of descriptors.

The identification of synonyms of competency questions terms was performed manually using the Python based NTLK WordNet interface [14] via console. All MeSH terms were obtained automatically with an implementation of a java based tree structure of MeSH terms extracted from the source files. The matching steps described above are repeated for matching potential object property (v) and individual (vi) candidates. True and false positive entity matches of were determined manually by comparing the desired information content defined by the domain experts and the entity definitions from Ontobee [16].

3. Results

The following competency questions were elaborated with medical experts from the Department of Urology at the Medical University of Innsbruck:

1) Which radical prostatectomies with biopsies, recurrence as well as subsequent radiation and medical therapies existed in the year 2010?
2) Which radical prostatectomies together with histology samples as well as the corresponding PSA curves do exist between the year 2000 and now?

3) Which registered biopsies with DG-GLS6 and PSA value between 2 and 10 as well as with subsequent radical prostatectomies exist between 2000 and 2011?
4) Which PSA curves and corresponding radical prostatectomies with subsequent treatments exist?
5) How many patients having biopsies died from neoplasms of the prostate between 1990 and now?

Applying stop word removal and the lemmatization algorithm (Step *(i)-(iii)*) on competency question 1) yields the following set of relevant terms: *S={biopsy, radiation, medical, therapy, 2010, existed, prostatectomy, radical, year, subsequent, recurrence}*

Compared to other stemming algorithms, such as Porter and Lancaster [17], the highest number of corrects word stems was obtained with the NLTK WordNet Lemmatizer. Two OMIABIS concepts were found for the keyword "medical" after performing SPARQL queries for concepts with the remaining stem words (see Table 1). Table 1 also includes a presentation of entities matched to the keyword along with their entity type and definition.

Table 1. Two entity matches were obtained from Step (iv)(a) after stop word removal and lemmatization. Each token in the remaining sequence is matched against entities of the OMIABIS ontology OWL file obtained from http://purl.obolibrary.org./obo/omiabis.owl.

Keyword	Matched entity	Entity type	Definition
medical	medical record	concept	"A document that contains information representing health-relevant qualities of a patient written in a chronological manner and is primarily used for patient care in a clinical setting"
medical	sample medical record	concept	"A medical record of specimen donor"

We used the NLTK WordNet Interface to detect possible synonyms of the keywords that could be used for matching concepts. The results are shown in Table 2. Only one false concept match was obtained when entering the keyword "group". Entering the split term "free" in the synonym "free radical" resulted in seven false concept matches. Both of these synonyms were returned by the WordNet database for the competency question keyword "radical".

Table 2. All synonyms that were returned by the NLTK WordNet interface from the CQ keywords of the sequence.

Keyword	WordNet synonym
existed	exist
radiation	Radiation Sickness radiotherapy
radical	group free radical root extremist
medical	checkup aesculapian
year	class

Table 3. Two correct object property matches were obtained in Step *(v)* from the synonym "exist" in Table 2 that were extracted from the WordNet database in Step *(iv)(b)*.

Keyword	Matched entity	Entity type	Definition
exist	exists at	object property	"b exists_at t means: b is an entity which exists at some temporal region t"
exist	during which exists at	object property	"b exists_at t means: b is an entity which exists at some temporal region t"

The two correct object property matches in Table 3 were returned for the synonym "exist" which was derived from the original term "existed". No matches were obtained for the other synonyms detected. There were no synonyms identified for the other keywords in the competency question.

In Step *(iv)(c)* a list of MeSH terms was searched with the keywords which had no match in Step *(iv)(a)*: *S={biopsy, radiation, therapy, 2010, prostatectomy, radical, year, subsequent, recurrence}*. A list of terms that were obtained for the keyword "prostatectomy" is shown in Figure 2.

E04:Surgical Procedures, Operative
E04.950:Urogenital Surgical Procedures
E04.950.774:Urologic Surgical Procedures
E04.950.774.860:Urologic Surgical Procedures, Male
E04.950.774.860.625:Prostatectomy

Figure 2. MeSH tree hierarchy for the CQ keyword "prostatectomy". The upper terms were integrated into the matching process.

The extracted MeSH terms were matched again (Step *(iv)(a)*) against all ontology entities. The results for the keyword "prostatectomy" are shown in Table 4.

Table 5 depicts the total number of concepts and object properties identified in the OMIABIS ontology, based on the competency question keywords and related synonyms and MeSH terms.

Table 4. Concept matches that were obtained in Step *(iv(a))* from the MeSH terms identified for the keyword "prostatectomy"

Keyword	Matched entity	Entity type	Definition
surgical, procedures	surgical procedure	concept	"A planned process that uses operative manual and instrumental techniques on a patient to investigate and/or treat a pathological condition such as disease or injury, or to help improve bodily function or appearance"
surgical	time of surgical removal of specimen	concept	"A data item that reports the time when a specimen was collected by a surgical procedure"
surgical	specimen surgical removal	concept	"A collecting specimen from organism process that involves removing the specimen from an individual through a surgical procedure"
male	male	concept	"A biological sex quality inhering in an individual or a population whose sex organs contain only male gametes"
male	female	concept	"A biological sex quality inhering in an individual or a population that only produces gametes that can be fertilized by male gametes"

Table 5. Total number of matches obtained for each single keyword (superscript digits highlight related keywords and OMIABIS matches).

Keywords	True positive Matches	False positive matches	Total matches	Positive OMIABIS matches
biopsy[1]	3	63	66	medical record[2]
radiation	0	1	1	sample medical record[2]
medical[2]	2	0	2	diagnostic process[1]
therapy	0	0	0	surgical procedure[1,4]
2010	0	0	0	specimen surgical removal[1,4]
existed[3]	2	0	2	diseased state specimen data[5]
prostatectomy[4]	2	3	5	specimen disease state data[5]
radical	0	8	8	exists at[3]
year	0	1	1	during which exists[3]
subsequent	0	0	0	
recurrence[5]	2	3	5	

4. Discussion

We developed a semi-automated approach to assist the identification of relevant entities in existing biomedical ontologies based on the input of competency questions. We contribute to the current state of the art as we provide an approach that combines competency evaluation and query expansion in order to assess the usefulness of existing ontologies for the biobanking domain. Nevertheless, this evaluation method is not restricted to the biobanking domain and therefore can be re-used for other domains as well. The proposed method allows matching with higher terms from a hierarchical medical terminology which can be useful for the detection of potential higher ontology classes including a given search object. Moreover, relevant concepts might exist as a synonym of a given keyword. Nevertheless, a clear predefinition of the meaning of the competency question keywords was necessary for a correct identification of semantically equivalent ontology concepts. At a current state, there is no possibility to automatically proof whether identified ontology concepts are semantically equivalent to the competency question terms. Therefore, a manual plausibility check of the results needs to be performed. Since we are providing a proof of concept evaluation, the competency questions used here were only focusing on solid biomaterials. It remains to be evaluated if the proposed approach can be used for body fluids and biomolecules by using an extended set of competency question. A major limitation of our evaluation approach is the matching of single tokens, which leads to a higher number of false positive matches as words in competency questions often comprise two or three tokens, e.g. “radical prostatectomy”. In order to reduce the false positive matches, we plan to extend and refine the term matching process towards integrating composite terms.

For next steps, we plan to conduct a broader study on several different biomedical ontologies using an extended set of competency questions that is not limited to solid material.

We conclude that the proposed evaluation approach allows a swift detection of candidate ontology entities based on an expanded search for higher hierarchy key terms that exist in comprehensive medical vocabularies.

References

[1] G.J.B. van Ommen et al, BBMRI-ERIC as a resource for pharmaceutical and life science industries: the development of biobank-based Expert Centres, *European Journal of Human Genetics* (2014), 1–8.

[2] M.N. Frannson, E.R. Rial Sebbag, M. Brochhausen and J.E. Litton, Toward a common language for biobanking, *European Journal of Human Genetics* **23** (2005), 22–28.

[3] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, *Data & Knowledge Engineering* **25**(1-2) (1998), 161-197

[4] A. Burgun, Desiderata for domain reference ontologies in biomedicine, Journal of Biomedical Informatics **29** (3) (2006), 307-313

[5] C. Bezerra, F.Freitas and F. Santana, Evaluating Ontologies with Competency Questions, *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences*, vol.3, 284-285, 2013.

[6] Medical Subject Headings, Introduction to MeSH 2015, http://www.nlm.nih.gov/mesh/introduction.html, last access: 17.12.2014.

[7] L. Zhiyong, W. Kim, W.J. Wilbur, Evaluation of Query Expansion Using MeSH in PubMed, *Inf Retr Boston* **12**(1) (2009), 69–80

[8] The Board of Trustees of Leland Stanford Junior University, NCBO BioPortal, http://bioportal.bioontology.org/ontologies, last access: 24.1.2015.

[9] Berkeley Bioinformatics Open Source Project, The Open Biological and Biomedical Ontologies Foundry, http://www.obofoundry.org, last access: 22.1.2015.

[10] M. Brochhausen, M.N Frannson, N.V. Kanaskar, M. Eriksson, R.Merino-Martinez, R.A. Hall, L. Norlin, S. Kjellqvist, M. Hortlund, U. Topaloglu, W.R. Hogan and J.E. Litton, Developing a semantically rich ontology for the biobank-administration domain, *Journal of Biomedical Semantics* **4**(1) (2013), 856–890.

[11] L. Norlin, M.N. Fransson, M.Eriksson, R. Merino-Martinez, M. Anderberg, S. Kurtovic and J.E. Litton. A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS, *Biopreservation and Biobanking* **10** (4) (2014), 966–974.

[12] A.D. Ramos, M. Boeker, L.Jansen, S.Schulz, M. Iniesta, J. Tomás and F. Breis, Evaluating the Good Ontology Design Guideline (GoodOD) with the Ontology Quality Requirements and Evaluation Method and Metrics (OQuaRE), *PLoS ONE* **9**(8) (2014).

[13] C. Bezerra, F.Freitas and F. Santana, CQChecker: A tool to check the Satisfaction of Description Logic Competency Questions on Ontologies, 2013 X National Conference on Artificial and Computational Intelligence (ENIAC).

[14] Bird, Steven, E. Loper and E. Klein, Natural Language Processing with Python. (2009) O'Reilly Media Inc.

[15] Bird, Steven, E. Loper and E. Klein, Natural Language Toolkit, http://www.nltk.org/, last access: 20.1.2015.

[16] University of Michigan Medical School, Ontobee, http:// http://www.ontobee.org/, last access: 22.1.2015.

[17] B. Steven, E. Loper and E. Klein, Stemming and Lemmatization with Python NLTK, http://text-processing.com/demo/stem/, last access: 22.1.2015.