Semi-Automated Ontology Evaluation Based On Competency Questions and Query Expansion

Philipp HOFER\textsuperscript{1a}, Sabrina NEURURER\textsuperscript{a}, Helga HAUFFE\textsuperscript{b}, Thomas INSAM\textsuperscript{c}, Anette ZEILNER\textsuperscript{d} and Georg GÖBEL\textsuperscript{a}

\textsuperscript{1}philipp.hofer@i-med.ac.at

\textsuperscript{a}Department for Medical Statistics, Informatics and Health Economics, Medical University of Innsbruck

\textsuperscript{b}Department of Urology, Medical University of Innsbruck

\textsuperscript{c}Department of Obstetrics and Gynecology, Medical University of Innsbruck

\textsuperscript{d}Division of Human Genetics, Medical University of Innsbruck
Overview

Introduction

Which Information is Stored in a Biobank?
Relevance of Biomedical Ontologies for Biobanking

Objectives

Semi-Automated Ontology Evaluation Approach

Methods

Query Expansion, Medical Subject Headings & WordNet
Query-Expansion Approach

Results

Competency Questions and Relevant Ontology Concepts
Evaluation Results

Discussion & Outlook
Which Information is Stored in a Biobank?

- **Sample related information**
  - Collection process, sample processing and storage
- **Donor related information**
  - Medical history, current disease, medication, family history, therapies
- **Information about access control on samples**
Relevance of Biomedical Ontologies for Biobanking

Introduction

• Searching for samples across different biobanks needs a common semantic foundation\(^1\)

• Harmonization of biobank contents via biomedical ontologies

• It remains unclear which biomedical ontologies are relevant for the biobanking domain

• E.g. anatomical structures and body fluids are represented by several ontologies (FMA, SNOMED, MeSH, etc.)

We want to propose a semi-automated competency evaluation approach to identify candidate ontologies for the biobanking domain.

Competency questions (CQs)

- Representation of typical user requests to an information system
- Used to define the scope of ontologies within the ontology engineering process
Query Expansion, Medical Subject Headings (MeSH) & WordNet

- **Query Expansion**
  - Extension of the original query by additional search terms to improve retrieval performance
  - Identification of semantic relationships between competency question terms and biobank ontologies by using lexical-semantic relationships from other hierarchical terminologies

- **Medical Subject Headings (MeSH)**
  - Hierarchical terminology of medical terms and synonyms

- **WordNet**
  - Database containing words and related synonyms for the English language

- **Evaluation of the OMIABIS biobanking ontology**
Query-Expansion Approach

Input:
Competency question CQ
Ontology O

Which radical prostatectomies with biopsies, recurrence as well as subsequent radiation and medical therapies existed in the year 2010?

Lexical Processing
Stemming and stopword removal

{biopsy, radiation, medical, therapy, 2010, existed, prostatectomy, radical, year, recurrence, subsequent}

Lexical matching
SPARQL

SELECT ?label ?concept
{
?concept a owl:Class .
FILTER(REGEX(STR(?label), "<CQ\ term=""))}

Output:
Matched entities

prostatectomy → OMIABIS_0001041: surgical procedure,
medical → OMIABIS_0001026: medical record,
...

Query expansion

MeSH
WordNet

no direct match
{prostatectomy, biopsy, ...}
direct match

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

Legend:
- Direct Match
- No Direct Match

Lexical Processing
Stemming and stopword removal
Competency Questions and Relevant Ontology Concepts

- Competency questions were defined in collaboration with medical experts from different departments at the Medical University of Innsbruck

- Manual definition of relevant ontology entities for the evaluation process

<table>
<thead>
<tr>
<th>CQ Terms</th>
<th>Example: Relevant entity relationships (OMIABIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>biopsy</td>
<td>primary diagnosis, diagnosis, sample handling, diagnostic process, surgical procedure, specimen surgical removal</td>
</tr>
<tr>
<td>radiation</td>
<td>treatment</td>
</tr>
<tr>
<td>medical therapy</td>
<td>medical record, sample medical record</td>
</tr>
<tr>
<td>2010 existed</td>
<td>value specification</td>
</tr>
<tr>
<td>prostatectomy</td>
<td>exists at, during which exists</td>
</tr>
<tr>
<td>radical</td>
<td>surgical procedure, specimen surgical removal</td>
</tr>
<tr>
<td>year recurrence</td>
<td>surgical procedure, specimen surgical removal</td>
</tr>
<tr>
<td>subsequent</td>
<td>date of data entry, time unit</td>
</tr>
<tr>
<td></td>
<td>specimen disease state data, disease state specimen data</td>
</tr>
</tbody>
</table>
Evaluation Results

• “Which radical prostatectomies with biopsies, recurrence as well as subsequent radiation and medical therapies existed in the year 2010?”

Evaluation of OMIABIS

535 concepts, 77 object properties (including imports from higher ontologies) (22.2.2015)

<table>
<thead>
<tr>
<th>Keywords</th>
<th>True positive</th>
<th>False positive</th>
<th>Total matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>biopsy</td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>radiation</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>medical therapy</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>therapy</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>existed</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>prostatectomy</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>radical</td>
<td>0</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>year</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>subsequent</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>recurrence</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Discussion & Outlook

• Pilot-approach for systematical evaluation of candidate ontology entities based on a query expansion to state the usability of specific ontologies for the biobanking domain

• However there are still limitations to be considered
  – Compound nouns, e.g. “radical prostatectomy”, “medical therapy”
  – We plan to extend and refine the term matching process towards integrating composite terms to reduce false positive matches.

• For next steps we plan to conduct broader study on several different biomedical ontologies using an extended set of competency questions
Thank you for your attention!

Philipp Hofer
Department of Medical Statistics, Informatics and Health Economy
Medical University of Innsbruck

Sabrina Neururer
Department of Medical Statistics, Informatics and Health Economy
Medical University of Innsbruck

Georg Goebel
Department of Medical Statistics, Informatics and Health Economy
Medical University of Innsbruck

Helga Hauffe
Department of Urology
Medical University of Innsbruck

Anette Zeilner
Division of Human Genetics
Medical University of Innsbruck

Thomas Insam
Department of Obstetrics and Gynecology
Medical University of Innsbruck

Funding
BBMRI.AT