Ontology-based knowledge base model construction-OntoKBCF

Xia Jing, Stephen Kay, Nicholas Hardiker, Tom Marley

SHIRE, IHSCR, University of Salford, UK

Abstract

Semantic web technologies are used in the construction of a bio-health knowledge base model, which, when coupled with an Electronic Health Record (EHR), is to be used by clinicians. Specifically, this ontology provides the basis for a domain knowledge resource that attempts to bridge biological and clinical information. The prototype is focused on a Cystic Fibrosis exemplar, and the content of the model includes: Cochrane reviews; a time-oriented description; gene therapy; and the most common cystic fibrosis gene mutations. The facts within the model range from nucleobase mutation and amino acid change to clinical phenotypic. The knowledge is represented by layers from the micro level to the macro level. Here, emphasis is placed upon the details between levels (i.e., the ‘vertical’ axis) and these are made available to bridge the knowledge from different levels. The description of gender, age, mutation and clinical manifestations are clues for matching points within an EHR system. OWL is the ontology representation language used and the output from Protégé-OWL is a XML-based file format, which facilitates further application and communication.

Keywords:
Knowledge representation; OWL; Cystic fibrosis; OntoKBCF

Introduction

The ultimate goal is to provide relevant bio-health information to doctors in a clinical environment. To move towards this goal, we construct an ontology-based knowledge base (KB) model (called OntoKBCF) to provide the required domain knowledge and then consider how to make it available to an electronic health record (EHR) system as a plug-in. Figure 1 provides a schematic overview of the layered knowledge base exploiting the metaphor of an EHR as a container drawing from clinical and bio-information.

The content produced for the model, depicted by the horizontal levels in figure 1, is merely sufficient to organize the relationships between the vertical levels for this research. This paper introduces the OntoKBCF construction and the results obtained thus far.

Semantic web technology is the major technology used in the OntoKBCF construction, recognising that this is the next evolution of the current World Wide Web. Simply put, the semantic web provides more “labels” for data so that machines can process the data more precisely and in more meaningful ways. The semantic web requires common formats and language; the former are useful for data sharing, reuse and exchange; the latter is used for expressing data meaningfully and relating it with real objects[1].

Figure 1 - The EHR as container of diverse and complex data

An ontology is an important part of the infrastructure for semantic web applications and shares similar characteristics with it. However, it is particularly useful for providing a “common understanding of a domain” and provides opportunity to improve knowledge management[2][p5]. Considering these characteristics i.e., universal format (URI, uniform resource identifiers), expression capability (to represent data related with real objects), extensibility and standardization, we select the ontology-based KB to construct the application’s knowledge source. Cystic Fibrosis is chosen as an exemplar.

Procedures, technology and tool

General procedures

The general procedures for OntoKBCF construction are outlined: (A): To understand doctors’ information needs; (B): To define OntoKBCF’s border, axes and granularity; (C): To dissect knowledge resources into basic concepts; (D): To list relevant vocabularies; (E): To extract target concepts from UMLS (Unified Medical Language System)[3] or GO(Gene Ontology)[4]; (F): To organize and arrange concepts; (G): To
create terms necessary for OntoKBCF; (H): To discuss with colleagues and to iterate all the procedures until the concepts are stable. The OntoKBCF is constructed bottom-up. Most of the hierarchy of the basic concepts in OntoKBCF follow UMLS and GO when appropriate, and so do the semantic types in UMLS. However, we also add, adjust relationships and vocabularies if there are no appropriate choices. For example, we use “Gender_group” to connect “Population_group” and “Female”; but “Gender-group” is not an independent concept in UMLS. Likewise we create “Patient_CF” and all its subclasses and adjust the concept “CFTR gene” (cystic fibrosis transmembrane regulator) in UMLS to “Human CFTR gene”. We justify these changes with respect to the OntoKBCF border, which is discussed in more detail in the “model construction” section of this paper.

Technologies and Tool: OWL and Protégé-OWL

OWL (web ontology language) is used to explicitly represent vocabularies and their relationships in an ontology and this representation is directly usable by machine[5]. OWL has more semantic representation ability than XML (extensible markup language), RDF, and RDMS (resource description framework schema)[5] and OWL facilitate greater machine interoperability. OWL ontology can be encoded in different syntactic forms including RDF/XML, which is defined by XML syntax for RDF[6].

OWL Lite is an ontology language supporting a subset of the OWL language, specially designed to develop OWL tools; OWL DL (description logic) and the complete OWL language (OWL Full) support the same set of OWL language, however the former has desirable functions for reasoning system with more restrict constraints. We choose OWL Full as the ontology language in the OntoKBCF construction, as it provides the maximal RDF compatibility. OWL Full ontologies contain all the RDF content in a consistent manner, and OWL Full also assigns more meaning for certain RDF triples[7]. The OWL ontology uses the Protégé-OWL 3.2 platform and is encoded in XML. OWL statements make the OntoKBCF consistent, extensible and computer-readable.

OntoKBCF Construction

OntoKBCF's border, main axes and granularity

The first thing in order to construct a KB model is to define its scope or border[8]. The research assumes that the OntoKBCF will be a plug-in to an EHR system, with the target end-users being clinicians (more specifically doctors). Initial scoping must therefore include the doctors’ information need.

There are several ways to obtain details of information need: e.g., survey, questionnaire, interview, observation, system log files analysis and so on. We utilized the published literature to set the initial scope. According to references [9-11], the most common clinical question is about treatment. Since we are trying to construct a knowledge base prototype to show proof of concept rather than produce a full commercial product, pragmatically we limited our search to a few reputable sources. In particular, we chose the Cochrane review topics as our starting clinical requirements.

The content of OntoKBCF comprises four parts: cystic fibrosis related Cochrane reviews[12]; time related cystic fibrosis description[13]; gene therapy of cystic fibrosis[14, 15]; the most common CFTR mutations and their characteristics[16, 17]. All the detailed content is available on request.

The two main axes developed in the OntoKBCF are time and problem orientation. With respect to Time the interest is in how cystic fibrosis may announce itself at different ages; with respect to problem-orientation the emphasis is from the Cochrane review on treatment, gene therapy and the most common CFTR mutations and their characteristics. We did not include the entire Cochrane review's conclusions about cystic fibrosis, only some definite positive conclusions. Negative and uncertainty conclusions were excluded.

We only represent the definite knowledge facts about cystic fibrosis in the OntoKBCF-content i.e., the “what”, without disease mechanism and reasons explanation i.e., “how” and “why”. In the OntoKBCF the granularity of genetic level starts from nucleobase, which is the important component for the elemental unit (nucleotide) of RNA and DNA; the most of representation granularity in phenotype level starts from those concepts, such as diarrhea, nausea and coughing; other part of representation granularity in phenotype level depends on the border of the OntoKBCF; if the subclasses of the concepts are out of the border, we will not split the concepts any more. In OntoKBCF construction, the border and granularity are also the major criteria in considering an upper class concept's inclusion or exclusion.

Construction

We will explain construction of the OntoKBCF in detail by using the example knowledge fact (1) and biological concepts organization (2). The example given is from the time-oriented cystic fibrosis description. We organize the biological concepts since we have been unable to find an agreed reference hierarchy, and therefore we explicitly introduce what we include and how we have done it. Space determines we only give examples rather than provide every statement.

The following font conventions are used in this section: the class name is in bold, and the class defined by us is in bold and underline; the property we defined is in italic and underlined; description from “Asserted Conditions” panel in “OWL Classes” tab on the Protégé-OWL interface is in italic. We connect the sub class with super class by symbol “<” and super class with sub class by symbol “>“. The general procedures, mentioned earlier, are given here (e.g., C, E etc).

1. In female adolescent cystic fibrosis patient: infertile with scanty cervical mucus, bronchiectasis
   - C) To analyze and dissect knowledge fact into basic concepts

In female adolescent cystic fibrosis patient group, the clinical manifestations include infertile, scanty cervical mucus and bronchiectasis. Basic concepts include: gender: female; different age groups: adolescent; diseases: cystic fibrosis and infertility; human being group: patient; quantitative concept: scanty; body substances: cervical mucus; respiratory tract diseases: bronchiectasis.
   - E) To select proper terms and upper class terms used
in UMLS according to their definitions

We choose the corresponding terms from UMLS according to the concepts acquired from step (C) and their super classes. Examples include in figure 2.

<table>
<thead>
<tr>
<th>Conceptual_entity</th>
<th>Phenomenon_or_process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Pathologic_function</td>
</tr>
<tr>
<td>Population_group</td>
<td>Disease_or_syndrome</td>
</tr>
<tr>
<td>Female</td>
<td>Diseases</td>
</tr>
<tr>
<td>Age_group</td>
<td>Cystic_fibrosis</td>
</tr>
<tr>
<td>Adolescent_age_group</td>
<td>Infertility</td>
</tr>
</tbody>
</table>

*Figure 2- Example terms & hierarchy from UMLS*

Patient—there is no proper terms in UMLS for patients in different diseases categories.

- F) To organize, create and arrange the terms with original identities and alternative vocabularies

In this case the original identity is a concept unique identifier (CUI) from UMLS, which can be used to track the concept over time and also can be used like a “primary key” in further application or communication. Alternative vocabularies are synonyms. Both of CUI and synonyms are recorded in the “Annotations” panel in the “OWL Classes” tab

The exact terms found in UMLS and their upper class terms can be organized hierarchically from bottom up. We create the new term “Patient CF” as a subclass of “Human being” to represent patient diagnosed with cystic fibrosis. Because the cystic fibrosis patients have different properties, such as different age groups, different mutations and different therapies, and they are the subjects for representation as knowledge facts, we split the concept “Patient CF” further. An example hierarchy for this concept is:

Patient CF > Patient CF with age group > Adolescent CF >
Adolescent female CF

We define the “Patient CF” as intersection of (1) Human being; (2) has_diagnosis some Cystic_fibrosis.

“has_diagnosis” is a property used to connect human being with diseases or syndrome.

“Patient CF with age group” : this class is only an abstract and interim concept used for keeping its subclasses tidy.

“Adolescent female CF” is an intersection of (1) Adolescent CF; (2) occur_in_gender_group some Female.

“occur_in_gender_group” is a sub property of “occur” to describe human being in male or female group.

- G) To represent the knowledge fact by combination of the terms and properties with logic symbols

The final representation of this fact is under the “Adolescent female CF” concept as a necessary condition: has_manifestation some (Infertility and Scanty_cervix_mucus). There are also some inherited descriptions from upper classes. The final representation of this knowledge fact showed in figure 3.

“has_manifestation” is a property to describe human being’s manifestation specifically related with diseases or syndromes.

We define “Scanty_cervix_mucus” as in intersection of (1) Cervix_mucus; (2) has_quantitative_property some Scanty.

*Figure 3- Representation of example 1*

2. Introduce the hierarchy of the bio concepts in the model

We explain the construction idea top down. There are two major parts: “Nucleobase mutation”, “Translation change”, both are subclasses of “Phenomenon_or_process” for mutation related concepts; “Amino_acid_peptide_or_protein”, “Nucleic_acid_nucleoside_or_nucleotide”, both are subclasses of “Chemical”. We will introduce them separately.

- Hierarchy of “Nucleobase mutation” and “Translation change” (Figure 4, 5)

*Figure 4- Hierarchy of Nucleobase mutation and Translation change*

<table>
<thead>
<tr>
<th>Nucleobase mutation</th>
<th>Translation change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleobase deletion</td>
<td>Amino acid deletion</td>
</tr>
<tr>
<td>Nucleobase insertion</td>
<td>Amino acid insertion</td>
</tr>
<tr>
<td>Nucleobase transition</td>
<td>Amino acid substitution</td>
</tr>
<tr>
<td>Nucleobase transversion</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleobase transversion</th>
<th>Nucleobase deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_transversion C</td>
<td>Del A</td>
</tr>
<tr>
<td>A_transversion T</td>
<td>Del C</td>
</tr>
<tr>
<td>C_transversion A</td>
<td>Del G</td>
</tr>
<tr>
<td>C_transversion G</td>
<td>Del T</td>
</tr>
<tr>
<td>G_transversion C</td>
<td>Del 394 TT</td>
</tr>
<tr>
<td>G_transversion T</td>
<td>Del U</td>
</tr>
<tr>
<td>T_transversion A</td>
<td></td>
</tr>
<tr>
<td>T_transversion G</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 5- Hierarchy of Nucleobase transversion and Nucleobase deletion*

“Nucleobase insertion” is similar with “Nucleobase deletion”, and Ins_A, C, G, T, U are its subclasses.

There are four possible nucleobase transitions, “A_transversion”, “C_transversion”, “G_transversion”, and “T_transversion”. Hierarchy of “Translation change” is shown in figure 6.

- Hierarchy of “Amino_acid_peptide_or_protein” and “Nucleic_acid_nucleoside_or_nucleotide”

The subclasses of “Amino_acid_peptide_or_protein” are: “Amino_acids” and “Human_CFTR_protein”. Under
“Amino_acids” we list all the three letters abbreviation names. Hierarchy of “Nucleic_acid_nucleoside_or_nucleotide” is shown in figure 7.

We choose “Nucleobase” as subclass of “Nucleic_acid_nucleoside_or_nucleotide” and super class of all the nitrogenous bases’ abbreviation name. We also classified the nucleobases into “Purines_and_derivatives” or “Pyrimidines_and_derivatives” correspondingly.

![Image](image.jpg)

Figure 6- Hierarchy of Translation_change

Discussion

The complete OntoKBCF is a subset of cystic fibrosis’ clinical and biological knowledge facts. These are appropriate to our model border so as to support our application’s aim. We have provided some comments about the construction, its limitations, and the upper ontology.

Explanation of the construction approach

The construction idea starts from the top-down analysis of the knowledge facts and for the dissection from general to specific; construction work starts from basic concepts, from bottom to top and from specific to general-to modify step by step until the basic concept is turned into complex one with more meaning or restrictions. For example we may need to represent patient with the G542X mutation, and need several steps to achieve it. Thus: we define “G542X” as an amino acid location in the human CFTR protein; we define “G542X” as an amino acid substitution in human CFTR protein; finally we define “Patient_CF with G542X” as a cystic fibrosis patient group with amino acid change G542X.

Nucleotides are units of nucleic acid[18] (p47). According to its chemical structure, nucleotides can be broken into nucleosides and phosphate. The nucleosides can be broken down further into nitrogenous bases and ribose (RNA) or deoxyribose (DNA). It is the nitrogenous base (A, T, C, G, or U) decides the nucleotide type. “The one letter abbreviation can be used for either the bases alone or for the nucleotides containing them”[18] (p185). So when mutation occurs in DNA or RNA chain, it is mutation in nitrogenous base, also named nucleobase. We choose “Nucleobase” in the classes naming for description of bases and their mutations.

There are 22 types of amino acids, so there are many more possibilities for amino acid changes; we only list those involved in our border or scope. There are 5 types of nucleobases, so there is less possibility for nucleobase mutations; we list the entire possible point mutation types.

The amino acid three letters names are listed and the single letter names and full names are recorded in the “Annotation” panel as reference. The nomenclature follows the IUPAC recommendation. For description of amino acid changes we use single letter abbreviation name (such as “G551D”) and three letters abbreviation name (“Gly 551 Asp”) is recorded in the “Annotation” panel as well. The single abbreviation letter name also is used for amino acid change location, such as: “G542”, which is shorter and simpler.

Single letter abbreviation of amino acid can be confused with nucleobase: A, C, G, T have different meanings. In the OntoKBCF amino acids change with three letter abbreviation names and full names in the “Annotation” panel. Those nucleobase mutation location name with “minus” or “plus”, or start from “Ins” or “Del”, all are “labels” for nucleobase mutation location (such as “G621 plus 1” or “Ins3905”); it is not easy to distinguish the location between nucleobase mutation and amino acid change if they don’t carry the “labels”, although we did not encounter an example in construction.

There is a dilemma: according to IUPAC recommendation “X” can refer to any amino acid; according to recommendation of protein sequence variants[19] “X” is used to designate a translation termination codon. In the OntoKBCF we use “X” to represent stop codon. Because we list all the 22 amino acids, we can use the specific name if required.

Protégé-OWL can not accept the class name start from number, and also can not accept “>” or “<” symbols in the class name, which is the necessary symbols for nomenclature of mutation[19], so we adjust the way it represented. We use “Del_A” and “Ins_A” to represent nucleobase deletion or insertion. Specific name like “Ins3905_T” means insertion in nucleotides 3905 (position) with thymine. For nucleobase substitution’s name, we use “minus” or “plus” to replace the symbol and start from the replaced nucleobase and end with the new nucleobase (avoid to use “>”): such as “G1717 minus 1 A”, which means at nucleotides 1717-1 (means the end of the intron, in this case it is intron 10, 1717 is the first nucleotide for exon 11) guanine change to adenine. To name amino acid substitution is in the similar way, such as “Gly 551 Asp”, amino acid 551- glycine changes to aspartate . Although we didn’t follow strictly the nomenclature recommendation[19] about mutation, we believe it is acceptable considering our aim for the model, which attempts to explain explicitly the meaning of mutation, instead of a reference dictionary to map all the possible mutations. We hold to the principle of the “recommendation” and adjust it according to the naming rules of Protégé-OWL.

In OntoKBCF and its further application within an EHR system, we only consider the human being as the subject of care, not other species. So we narrow down the classes into human being type, such as: under “Gene_or_genome”, we list “Human_CFTR_gene”; under “Exon” we list “Human_CFTR_gene_exon”. At this stage all the basic
concepts or combined concepts are treated as class. We only consider individual if it belongs to a specific person, but this can be revised later if required.

Scope: border, granularity and axes

It is important to decide the border and granularity of the OntoKBCF as knowledge itself is endless; here we only construct the OntoKBCF within the defined scope to complete our goal. Indeed, construction of the OntoKBCF is only part of our research and we need use it in the next stage. We prefer to extend the properties and constraints in future according to the application scenario rather than complete all properties, some of which will never be used. Without establishing clear granularity, it would also be an endless process to split the knowledge into subunits. The main axes are helpful in deciding the position of terms and will provide clarity and keep the “knowledge unit” organization consistent.

Upper ontology

There are some higher level formal ontologies which can be followed: Medical Ontology, SnapBFO (snapshot ontologies, indexed by times) and SpanBFO (videoscopic ontology) [20]. However at this stage it is difficult to say if the upper ontology view from SnapBFO and SpanBFO can make sense in this application. The OntoKBCF tests with the EHR system may provide more practical evidence for upper ontology mapping.

Representation limitation

An attempt was made to represent all the selected content fully in the OntoKBCF, but there are some precise details that cannot be represented. For example, “nasal polyps, especially if recurrent”, is a highly suspected evidence for cystic fibrosis for older children. But it is difficult to represent “especially” within OWL. And it is also difficult to represent the difference between “very common” and “sometimes”, both of which are quite common in bio and clinical knowledge description.

Conclusion

The OntoKBCF includes knowledge facts from nucleobase mutation, amino acid change to clinical phenotype. Although the horizontal layers of knowledge in the model are far from complete, the vertical relationships and structures partially succeed in bridging the knowledge from different levels. The link to a record system as modelled by openEHR will be the next stage of our work.

Acknowledgements

Sincere thanks to Matthew Horridge and Hai Wang from Manchester University for their kind help in Protégé-OWL.

References


Conclusion

The OntoKBCF includes knowledge facts from nucleobase mutation, amino acid change to clinical phenotype. Although the horizontal layers of knowledge in the model are far from complete, the vertical relationships and structures partially succeed in bridging the knowledge from different levels. The link to a record system as modelled by openEHR will be the next stage of our work.

Acknowledgements

Sincere thanks to Matthew Horridge and Hai Wang from Manchester University for their kind help in Protégé-OWL.

References


Conclusion

The OntoKBCF includes knowledge facts from nucleobase mutation, amino acid change to clinical phenotype. Although the horizontal layers of knowledge in the model are far from complete, the vertical relationships and structures partially succeed in bridging the knowledge from different levels. The link to a record system as modelled by openEHR will be the next stage of our work.

Acknowledgements

Sincere thanks to Matthew Horridge and Hai Wang from Manchester University for their kind help in Protégé-OWL.

References


Conclusion

The OntoKBCF includes knowledge facts from nucleobase mutation, amino acid change to clinical phenotype. Although the horizontal layers of knowledge in the model are far from complete, the vertical relationships and structures partially succeed in bridging the knowledge from different levels. The link to a record system as modelled by openEHR will be the next stage of our work.

Acknowledgements

Sincere thanks to Matthew Horridge and Hai Wang from Manchester University for their kind help in Protégé-OWL.

References


Address for correspondence

Prof. Stephen Kay
Room PO33, SHIRE, Brian Blatchford Building
Frederick Road Campus, University of Salford, M6 6PU, UK
S.Kay@salford.ac.uk