A Semantic Web-Based Approach for Harvesting Multilingual Textual Definitions from Wikipedia to Support ICD-11 Revision

Guoqian Jiang[1,*] Harold R. Solbrig[1] and Christopher G. Chute[1]

[1] Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN

ABSTRACT

In the beta phase of the 11th revision of International Classification of Diseases (ICD-11), the World Health Organization (WHO) intends to accept public input through a distributed model of authoring, in which creating textual definitions for ICD categories is a core use case. In a previous study, Wikipedia has been demonstrated as a useful source for textual definitions of diseases. The objective of the study is to develop and evaluate a semantic web-based approach for harvesting multilingual textual definitions from Wikipedia to support ICD-11 revision and its public review. In a prototype implementation, we developed a semantic web service application known as LexReview that automates the harvesting process in a dynamic way through invoking and integrating three online web services: 1) WHO ICD-11 content services; 2) NCBO BioPortal annotation services; and 3) DBpedia SPARQL endpoint query services. The Simple Knowledge Organization System (SKOS) lexical and mapping properties are used to represent the harvested definitions. The LexReview service application could be extended to integrate the textual definitions from other resources and subsequently consumed by a review application to support ICD-11 revision.

1 INTRODUCTION

The 11th revision of International Classification of Diseases (ICD-11) was officially launched by the World Health Organization (WHO) in March 2007 (1). The beta phase of the ICD-11 revision started in May 2012, and WHO intends to accept public input through a distributed model of authoring. An ICD-11 Beta Browser application has been developed and released by WHO (2). The browser provides simple commenting functionality to allow the domain professionals to make comments on existing contents, and it intends to introduce more social computing capabilities.

Lexical properties of ICD categories including titles, synonyms, and textual definitions should be reviewed following a standard and homogeneous terminological approach. The provision of textual definitions has been regarded as one of important criteria for measuring the quality of a terminology/ontology (3). In our previous study (4), we demonstrated that the textual definitions from the Unified Medical Language System (UMLS) (5), the formal definitions of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (6) and the linked open data (LOD) from DBpedia (7) are potentially useful resources for supporting ICD-11 textual definitions authoring. We argued that the ICD-11 project might potentially take advantage of the crowd-souring model of Wikipedia (8). Using this model, each ICD-11 category would be seeded as a Wikipedia page for public input and the definitions of ICD categories would be harvested using the DBpedia.

The objective of the study is to develop and evaluate a semantic web-based approach for harvesting multilingual textual definitions from Wikipedia to support ICD-11 revision and its public review. In a prototype implementation, we developed a semantic web service application known as LexReview that automates the harvesting process in a dynamic way through invoking and integrating a number of online web services: 1) WHO ICD-11 content services; 2) NCBO BioPortal annotation services; and 3) DBPedia SPARQL endpoint query services. The Simple Knowledge Organization System (SKOS) lexical and mapping properties are used to represent the harvested definitions.

2 BACKGROUND

2.1 WHO ICD-11 Content Model and Services

An ICD-11 content model has been developed by WHO to present the knowledge that underlies the definitions of an ICD entity. The content model is composed of three layers: a foundation component, a linearization component and an ontological component (9). The foundation component stores the full range of knowledge of all classification units in ICD. The linearization component corresponds to the classical print versions of ICD. The ontological component provides references to formal definition of terms and relationships. Currently, there are 13 defined main parameters in the content model to describe a category in ICD, in which "Textual Definitions" is one of main parameters for describing an ICD category.

Recently, an ICD URI scheme is proposed for naming and supporting web services by WHO. A base URI of http://id.who.int has been proposed, with http://id.who.int/icd/schema as the prefix for the vocabulary terms that related to ICD classification efforts maintained by WHO, http://id.who.int/icd/entity for the fundamental foundation entities related to ICD concepts.

2.2 BioPortal Annotation Services

The National Center for Biomedical Ontology Annotator is an ontology-based web service for annotating the textual biomedical data with biomedical ontology concepts (10, 11).

[*] To whom correspondence should be addressed: jiang.guoqian@mayo.edu

The biomedical community can use the Annotator service to tag datasets automatically with concepts from more than 300 ontologies coming from the two most important biomedical ontology & terminology repositories: the Unified Medical Language System (UMLS) Metathesaurus and NCBO BioPortal. Such annotations contribute to create a biomedical semantic web that facilitates translational scientific discoveries by integrating annotated data. In this study, the Medical Subject Headings (MeSH) (12) was configured to annotate the preferred labels of ICD-11 categories.

2.3 DBpedia SPARQL Endpoint

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web (7). DBpedia adopts Semantic Web Linked Open Data technology and its datasets are rendered in RDF format and can be accessed online via a public SPARQL query endpoint at http://dbpedia.org/sparql. The endpoint is provided using OpenLink Virtuoso as the back-end RDF database engine.

DBpedia also defines an ontology to organize its datasets. The ontology is a shallow, cross-domain ontology and covers 359 classes that form a subsumption hierarchy and are described by 1,775 different properties. In this study, we used one of the classes http://dbpedia.org/ontology/Disease and extracted all instances of the class for obtaining textual definitions.

2.4 Semantic Web Technologies

The World Wide Web consortium (W3C) is the main standards body for the World Wide Web (13). The goal of the W3C is to develop interoperable technologies and tools as well as specifications and guidelines to lead the web to its full potential. The resource description framework (RDF), web ontology language (OWL), and SPARQL (a recursive acronym for **S**PARQL **P**rotocol and **R**DF **Q**uery **L**anguage) specifications have all achieved the level of W3C recommendations, and are becoming generally accepted and widely used.

The SKOS data model views a knowledge organization system as a concept scheme comprising a set of concepts (14).The vocabulary used in the SKOS data model is a set of URIs that specifies the notion of SKOS concepts, concept schemes, lexical labels, notations, documentation properties and semantic relations. SKOS data are expressed as RDF triples. An increasing number of SKOS datasets in RDF are publicly available.

3 SYSTEM ARCHITECTURE

Figure 1 shows the system architecture of our approach. The LexReview service appplication invoked and integrated mainly three web services: 1) WHO ICD-11 content services for retrieving preferred label and definition for a target ICD entity; 2) NCBO BioPortal annotation services for retrieving the MeSH term annotation and its ID; and 3) DBpedia SPARQL endpoint query services for retrieving textual definitions by MeSH ID.

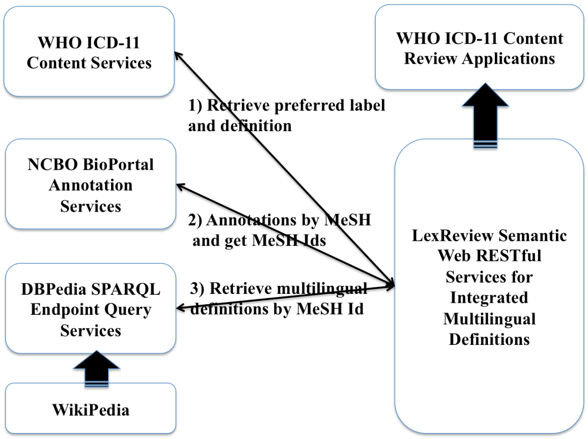

WHO ICD-11 Content Services
NCBO BioPortal Annotation Services
DBPedia SPARQL Endpoint Query Services
WikiPedia
WHO ICD-11 Content Review Applications
LexReview Semantic Web RESTful Services for Integrated Multilingual Definitions
1) Retrieve preferred label and definition
2) Annotations by MeSH and get MeSH Ids
3) Retrieve multilingual definitions by MeSH Id

Figure 1. System architecture of the LexReview service application.

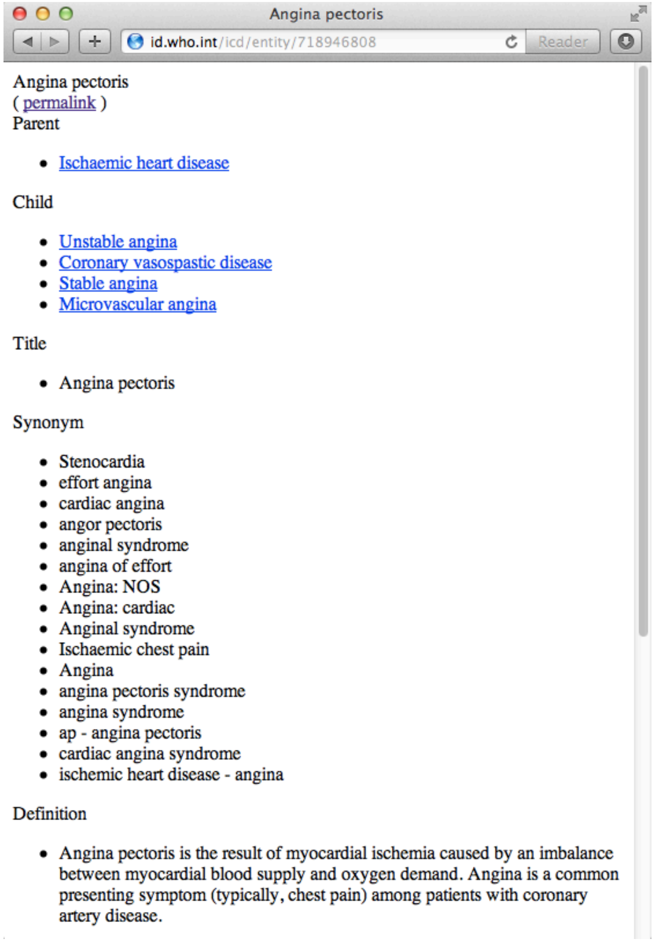
4 PROTOTYPE IMPLEMENTATION

The LexReview service application was implemented using a Java-based RESTful web services JAX-RS API known as Jersey (15) and a Jena ARQ API (16) that is a Java-based query engine for Jena that supports SPARQL RDF query language.

The service application accepts a standard URI of a single ICD entity as input. For example, the URI - http://id.who.int/icd/entity/718946808 represents an ICD entity Angina pectoris. Figure 2 shows the HTML rendering of the ICD entity Angina pectoris dispalyed through a web browser.

The content of an ICD-11 entity can be accessed through Content Negotiation that is a mechanism of RESTful services that makes it possible to serve different representation of a resource at the same URI. The WHO ICD content services provide the content representation in the formats of HTML, RDF and JSON. First, the system retrieved the title and definition of a target ICD entity based on its RDF rendering, in which the SKOS lexical properties skos:prefLabel and skos:definition are used to represent the values.

Second, the system invoked NCBO BioPortal annotation services using the title of a target ICD entity as the input. The annotation services were configured to use the ontology MeSH only and the semantic types within the semantic group Disorders (17)(see Table 1). The annotation services provide a score for each annotation that is the weight based on the annotation context. In this prototype implmentation, we harvested those annotation with the score=10, meaning that a direct annotation is matched with a concept preferred name. We then retrieved the MeSH ID, preferred name and URI of each annotation.



Angina pectoris
(permalink)
Parent

- Ischaemic heart disease

Child

- Unstable angina
- Coronary vasospastic disease
- Stable angina
- Microvascular angina

Title

- Angina pectoris

Synonym

- Stenocardia
- effort angina
- cardiac angina
- angor pectoris
- anginal syndrome
- angina of effort
- Angina: NOS
- Angina: cardiac
- Anginal syndrome
- Ischaemic chest pain
- Angina
- angina pectoris syndrome
- angina syndrome
- ap - angina pectoris
- cardiac angina syndrome
- ischemic heart disease - angina

Definition

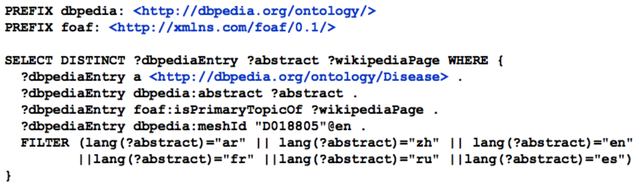
- Angina pectoris is the result of myocardial ischemia caused by an imbalance between myocardial blood supply and oxygen demand. Angina is a common presenting symptom (typically, chest pain) among patients with coronary artery disease.

Figure 2. The HTML rendering of the ICD entity Angina pectoris.

Table 1. A list of semantic types within the semantic group Disorders

DISO|Disorders|T020|Acquired Abnormality
DISO|Disorders|T190|Anatomical Abnormality
DISO|Disorders|T049|Cell or Molecular Dysfunction
DISO|Disorders|T019|Congenital Abnormality
DISO|Disorders|T047|Disease or Syndrome
DISO|Disorders|T050|Experimental Model of Disease
DISO|Disorders|T033|Finding
DISO|Disorders|T037|Injury or Poisoning
DISO|Disorders|T048|Mental or Behavioral Dysfunction
DISO|Disorders|T191|Neoplastic Process
DISO|Disorders|T046|Pathologic Function
DISO|Disorders|T184|Sign or Symptom

Third, when the system had a MeSH term annotated for a target ICD entity, the system invoked the DBpedia SPARQL enpoint to retrieve the textual definitions of a DBpedia entry coded in a MeSH ID. Figure 3 shows the SPARQL query used to retrieve mulilingual textual definitions from the instance entries of a DBpedia class "Disease" (i.e., http://dbpedia.org/ontology/Disease).



```
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?dbpediaEntry ?abstract ?wikipediaPage WHERE {
  ?dbpediaEntry a <http://dbpedia.org/ontology/Disease> .
  ?dbpediaEntry dbpedia:abstract ?abstract .
  ?dbpediaEntry foaf:isPrimaryTopicOf ?wikipediaPage .
  ?dbpediaEntry dbpedia:meshId "D018805"@en .
  FILTER (lang(?abstract)="ar" || lang(?abstract)="zh" || lang(?abstract)="en"
         ||lang(?abstract)="fr" ||lang(?abstract)="ru" ||lang(?abstract)="es")
}
```

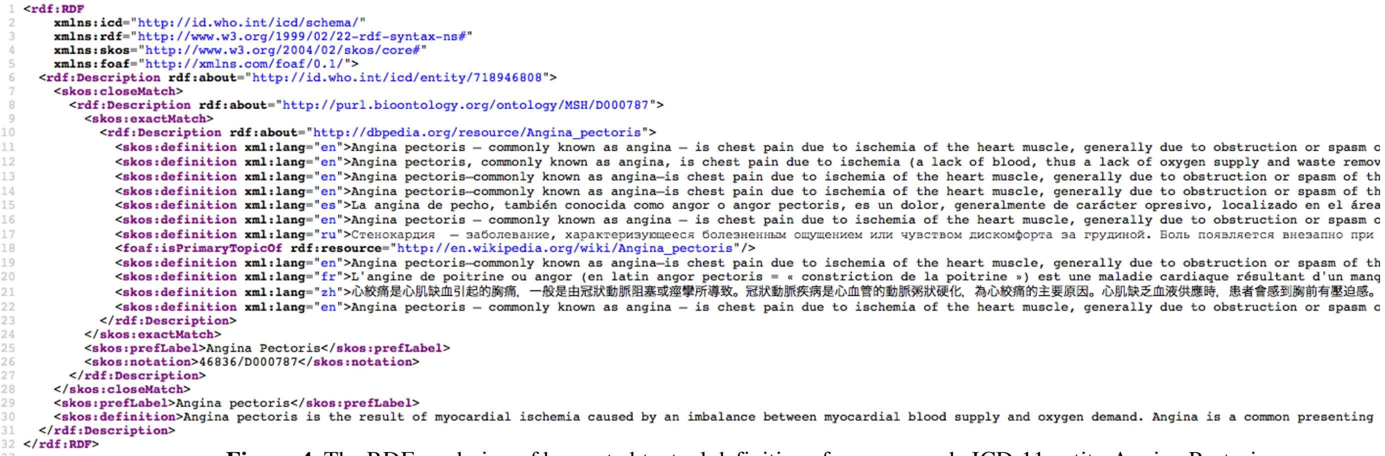
Figure 3. The SPARQL query used to retrieve multilingual textual definitions from DBpedia for a MeSH ID (e.g., D018805 for the MeSH term Sepsis)

Here, we asserted that the values of the predicate dbpedia:abstract are candidates for textual definitions. We used the language tags as a filter to retrieve those textual definitions in six official languages adopted by the WHO (18), i.e. "ar" standing for Arabic, "zh" for Chinese, "en" for English, "fr" for French, "ru" for Russian, and "es" for Spanish.

Finally, we represented the MeSH mapping based on BioPortal annotation services and the multilingual textual definitions retrieved for a target ICD category in RDF format, in which the SKOS lexical and mapping properties (skos:prefLabel, skos:definition, skos:closeMatch, skos:exactMatch) are used. We then exposed the RDF rendering through a RESTful service API. Figure 4 shows an example of RDF rendering of multilingual textual definitions for a target ICD-11 entity Angina pectoris. As illustrated in the figure, we used the predicate skos:closeMatch to represent the relationship between the target ICD entity and its MeSH annotation http://purl.bioontology.org/ontology/MSH/D000787. We used the predicate skos:exactMatch to represent the relationship between the MeSH annotation with the DBpedia entry http://dbpedia.org/resource/Angina_pectoris because they share the same MeSH ID. There are 11 definition entries in 5 languages available for the DBpedia entry and the predicate skos:definition is used to represent them. In addition, we also put the original title and definition of the target ICD entity in the RDF rendering using the predicates skos:prefLabel and skos:definition.

The prototype implementation will be accessible soon through http://informatics.mayo.edu/rest/project/icd11/lexreview/definition?uri=http://id.who.int/icd/entity/718946808, in which the uri parameter can be replaced by any other ICD entity URIs.

Table 2 shows a list of ICD-11 entity examples (n=10) that have Wikipedia definition matches. The first column in Table 2 shows the ICD-11 entity URI and its preferred label; the second column shows the corresponding Wikipedia URI for each ICD-11 entity matched by the system, and the codes for available languages; the third column shows the MeSH ID being an anchor between an ICD-11 entity and an Wikipedia entry. For each ICD-11 entity in Table 2, the Wikipedia definition entries are available at least in two language codes (range from 2-5 codes). The first author of



```
<rdf:RDF
    xmlns:icd="http://id.who.int/icd/schema/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdf:Description rdf:about="http://id.who.int/icd/entity/718946808">
    <skos:closeMatch>
      <rdf:Description rdf:about="http://purl.bioontology.org/ontology/MSH/D000787">
        <skos:exactMatch>
          <rdf:Description rdf:about="http://dbpedia.org/resource/Angina_pectoris">
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
            <skos:definition xml:lang="en">Angina pectoris, commonly known as angina, is chest pain due to ischemia (a lack of blood, thus a lack of oxygen supply and waste remov
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="es">La angina de pecho, también conocida como angor o angor pectoris, es un dolor, generalmente de carácter opresivo, localizado en el área
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
            <skos:definition xml:lang="ru">Стенокардия  — заболевание, характеризующееся болезненным ощущением или чувством дискомфорта за грудиной. Боль появляется внезапно при
            <foaf:isPrimaryTopicOf rdf:resource="http://en.wikipedia.org/wiki/Angina_pectoris"/>
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="fr">L'angine de poitrine ou angor (en latin angor pectoris = « constriction de la poitrine ») est une maladie cardiaque résultant d'un manq
            <skos:definition xml:lang="zh">心絞痛是心肌缺血引起的胸痛，一般是由冠狀動脈阻塞或痙攣所導致。冠狀動脈疾病是心血管的動脈粥狀硬化，為心絞痛的主要原因。心肌缺乏血液供應時，患者會感到胸前有壓迫感。
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
          </rdf:Description>
        </skos:exactMatch>
        <skos:prefLabel>Angina Pectoris</skos:prefLabel>
        <skos:notation>46836/D000787</skos:notation>
      </rdf:Description>
    </skos:closeMatch>
    <skos:prefLabel>Angina pectoris</skos:prefLabel>
    <skos:definition>Angina pectoris is the result of myocardial ischemia caused by an imbalance between myocardial blood supply and oxygen demand. Angina is a common presenting
  </rdf:Description>
</rdf:RDF>
```

Figure 4. The RDF rendering of harvested textual definitions for an example ICD-11 entity Angina Pectoris

the paper (GJ) reviewed all definition entries in Chinese (n=5) available from the 10 ICD-11 entity examples, and concluded that the quality of the definitions in Chinese are reasonably good and could be useful for supporting ICD-11 multilingual definition authoring.

5 DISCUSSION

In this study, we developed a semantic web service application that provides a dynamic way to harvest textual disease definitions of Wikipedia to support the ICD-11 textual definitions authoring and its public review. The "Dynamic" means that the service application would always retrieve the most current textual definitions stored in the DBpedia dataset. We found that MeSH IDs (i.e., dbpedia:meshId) are used to code the DBpedia entries under the class "Disease", which provide a good anchor to access the textual definitions of a DBpedia entry. As of April 14, 2013, there are 5,126 entries under the class "Disease", of which 2809 (54.8%) entries have MeSH IDs annotated (covering 2505 unique IDs). In total, 19,696 (71.5%) of 27, 540 textual definitions are available for those DBpedia disease entries with MeSH IDs. In future, we will build an approach to match those DBpedia disease entries that do not have MeSH IDs coded.

To get a MeSH term mapping to a target ICD entity, we invoked the BioPortal annotation services. We used a heuristic configuration by restricting the ontology to the MeSH only and setting up the semantic types within the semantic group Disorders. In our previous study, we used the UMLS CUIs to convert the ICD-10 codes to MeSH IDs. Considering that the ICD-11 covers many new terms other than ICD-10 terms, our approach in this prototype implementation may potentially provide a better coverage though a rigorous evaluation would be needed in the future.

In addition, we used SKOS lexical and mapping properties to represent the annotations and harvested textual definitions. The main reason is that the SKOS model provides a set of semantic web friendly signatures with well-defined semantics as we demonstrated in our previous study (19).

In summary, we developed a prototype of semantic web RESTful services that automates harvesting multilingual textual definitions of Wikipedia to support ICD-11 textual definition authoring and its public review. The LexReview service application could be extended to integrate the textual definitions from other resources and subsequently consumed by a review application to support ICD-11 revision. In the future, we plan to evaluate the quality and usefulness of the harvested multilingual definitions in collaboration with WHO ICD-11 revision community.

ACKNOWLEDGEMENTS

This work was supported in part by the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

Table 2. A list of ICD-11 entity examples that have Wikipedia definition matches.

ICD-11 Entity URI (Preferred Label)	Wikipedia URI (Avalilable Language Codes)	MeSH ID
http://id.who.int/icd/entity/1719064637 (Blind Loop Syndrome)	http://en.wikipedia.org/wiki/Blind_loop_syndrome (ru, en)	D005734
http://id.who.int/icd/entity/162683166 (Acute and subacute endocarditis)	http://en.wikipedia.org/wiki/Endocarditis (ru, fr, es, en)	D004696
http://id.who.int/icd/entity/761947693 (Essential (primary) hypertension)	http://en.wikipedia.org/wiki/Hypertension (zh, ru, fr, es, en)	D006973
http://id.who.int/icd/entity/925320484 (Deep vein thrombosis)	http://en.wikipedia.org/wiki/Thrombosis (es, en)	D013927
http://id.who.int/icd/entity/884453307 (Sinoatrial block)	http://en.wikipedia.org/wiki/Sinoatrial_block (fr, en)	D012848
http://id.who.int/icd/entity/1034471684 (Atrial flutter)	http://en.wikipedia.org/wiki/Atrial_flutter (fr, es, en)	D001282
http://id.who.int/icd/entity/1208831985 (Long QT syndrome)	http://en.wikipedia.org/wiki/Long_QT_syndrome (zh, fr, es, en)	D008133
http://id.who.int/icd/entity/1250136584 (Brugada syndrome)	http://en.wikipedia.org/wiki/Brugada_syndrome (zh, fr, es, en)	D053840
http://id.who.int/icd/entity/1026224967 (Lactose intolerance)	http://en.wikipedia.org/wiki/Lactose_intolerance (zh, ru, fr, es, en)	D007787
http://id.who.int/icd/entity/587895568 (Intussusception of small intestine)	http://en.wikipedia.org/wiki/Intussusception_(medical_disorder) (zh, ru, fr, es, en)	D007443

REFERENCES

1 WHO. Revision of the International Classification of Diseases (ICD). . [cited April 14, 2013]; Available from: http://www.who.int/classifications/icd/ICDRevision/en/index.html

2 WHO. ICD-11 Beta Browser. [cited April 14, 2013]; Available from: http://apps.who.int/classifications/icd11/browse/f/en

3 Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007 Nov;**25**(11):1251-5.

4 Jiang G, Solbrig HR, Chute CG. Using semantic web technology to support ICD-11 textual definitions authoring. ACM International Conference Proceeding Series; 2011; 2011. p. 38-44.

5 UMLS. [cited April 14, 2013]; Available from: http://www.nlm.nih.gov/research/umls/

6 SNOMED CT. [cited April 14, 2013]; Available from: http://www.ihtsdo.org/snomed-ct/

7 DBpedia. [cited April 14, 2013]; Available from: http://dbpedia.org/About

8 Wikipedia. [cited April 14, 2013]; Available from: http://wikipedia.org/

9 ICD-11 Information Model. [cited April 14, 2013]; Available from: http://informatics.mayo.edu/icd11model

10 Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on translational bioinformatics. 2009;**2009**:56-60.

11 NCBO Annotator. [cited April 14, 2013]; Available from: http://www.bioontology.org/annotator-service

12 MeSH. [cited April 14, 2013]; Available from: http://www.nlm.nih.gov/mesh/

13 The World Wide Web Consortium (W3C). [cited November 26, 2012]; Available from: http://www.w3.org/

14 SKOS. [cited April 14, 2013]; Available from: http://www.w3.org/TR/skos-primer/

15 Jersey API. [cited April 14, 2013]; Available from: http://jersey.java.net/

16 Jena ARQ API. [cited April 14, 2013]; Available from: http://jena.apache.org/documentation/query/

17 The UMLS Semantic Groups. [cited April 14, 2013]; Available from: http://semanticnetwork.nlm.nih.gov/SemGroups/

18 WHO Multilingualism. [cited Apirl 14, 2013]; Available from: http://www.who.int/about/multilingualism/en/

19 Jiang G, Solbrig HR, Chute CG. Building Standardized Semantic Web RESTful Services to Support ICD-11 Revision. ACM International Conference Proceeding Series 2012; 2012.