Disease Compass – A Navigation System for Disease Knowledge based on Ontology and Linked Data Techniques

Kouji Kozaki[1,*], Yuki Yamagata[1], Riichiro Mizoguchi[2,*], Takeshi Imai[3] and Kazuhiko Ohe[3]

[1] ISIR, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, Japan

[2] Research Center for Service Science School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

[3] Department of Medical Informatics, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

ABSTRACT

This paper discusses a navigation system for disease knowledge named *Disease Compass*. It navigates the users through a disease ontology defined based on River Flow Model of diseases which captures a disease as causal chains of abnormal states. The disease ontology is published using linked data techniques so that medical information systems can use it as knowledge infrastructure about disease with other related knowledge sources. Because the disease ontology has been developed under a tight collaboration between ontology engineers and medical experts, it could be a valuable knowledge base for advanced medical information systems. Furthermore, linked data techniques enable us to obtain related information from other linked data or web services. Based on these techniques, the users of *Disease Compass* can browse causal chains of a disease and obtain related information about the selected disease and abnormal states from the following two web services. One is general information from linked data such as DBpedia, and the other is a 3D image of anatomies. Such a functionality was enabled thanks to the disease ontology which is successfully combined with other web resources. As a result, *Disease Compass* can support the users to explore disease knowledge with related information from various point of views.

1 INTRODUCTION

Recently, medical information resources storing considerable amount of data are available. Semantic technologies are expected to contribute to the effective use of such information resources and many medical ontologies such as SNOMED-CT[1], OGMS (Scheuermann 2009) have been developed for realizing sophisticated medical information systems. Although medical ontologies consist of various domains such as diseases, anatomy, drug, clinical information etc., disease is an important concept because its complicated mechanisms are deeply related to other concepts across many of these medical domains.

This is why we focus on developing disease ontology. Some disease ontologies such as DOID (Osborne 2009), and IDO (Cowell 2010) have been developed. They mainly focus on the ontological definition of a disease with related properties. On the other hand, we proposed a definition of a disease that captures it as a causal chain of abnormal states and a computational model called the River Flow Model of a Disease (Mizoguchi 2011). Our disease ontology consists of rich information about causal chains related to each disease. The causal chains provide domain-specific knowledge about diseases, answering questions such as "What disorder/abnormal state causes a disease?" and "How might the disease advance, and what symptoms may appear?" We believe it could be a valuable knowledge base for advanced medical information systems.

This paper discusses a navigation system for disease knowledge named *Disease Compass* based on the disease ontology which we developed. The system has two special features. Firstly, its user can browse disease knowledge according to causal chains of diseases which are defined in the disease ontology. Secondly, the user can obtain related information about the selected disease based on linked data techniques. These functionalities of the *Disease Compass* can support the users to understand disease knowledge from various points of view.

This paper is organized as follows. In Section 2, we introduce our disease ontology. In Section 3, we outline how to publish the disease ontology as linked data. In Section 4, we discuss a navigation system for disease knowledge named *Disease Compass*. Finally, in Section 5, we present concluding remarks and a give an outline of future work.

2 A DISEASE ONTOLOGY

2.1 Definition of a Disease

A typical disease, as a dependent continuant, enacts extending, branching, and fading processes before it disappears. As a result of these processes, a disease can be identified as a continuant that is an enactor of those processes. Such an entity (a disease) can change according to its phase while maintaining its identity. On the basis of this observation, we defined a disease and related concepts as follows (Mizoguchi 2011).

Definition 1: A disease is a dependent continuant constituted of one or more causal chains of clinical disorders (abnormal state) appearing in a human body and initiated by at least one disorder.

When we collect individual causal chains belonging to a particular disease type (class), we can find a common causal

* To whom correspondence should be addressed: kozaki@ei.sanken.osaka-u.ac.jp and mizo@jaist.ac.jp

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

chain (partial chain) that appears in all instance chains. By generalizing such a partial chain, we obtain the notion of a core causal chain of a disease as follows.

Definition 2: A core causal chain of a disease is a sub-chain of the causal chain of a disease, whose instances are included in all the individual chains of all instances of a particular disease type. It corresponds to the essential property of a disease type.

Definition 2 provides a necessary and sufficient condition for determining the disease type to which a given causal chain of clinical disorders belongs. That is, when an individual causal chain of clinical disorders includes instances of the core causal chain of a particular disease type, it belongs to that disease type. We can thus define such a disease type, which includes all possible variations of physical chains of clinical disorders observed for patients who contract the disease. According to a standard definition of subsumption, we can introduce an *is-a* relationship between diseases using the chain-inclusion relationship between causal chains.

Definition 3: *Is-a* relationship between diseases. Disease A is a supertype of disease B if the core causal chain of disease A is included in that of disease B. The inclusion of nodes (clinical disorders) is judged by taking an *is-a* relationship between the nodes, as well as sameness of the nodes, into account.

Definition 3 helps us systematically capture the necessary and sufficient conditions of a particular disease, which roughly corresponds to the so-called "main pathological conditions." Assume, for example, that (non-latent) diabetes and type-I diabetes are, respectively, defined as <*deficiency of insulin* → *elevated level of glucose in the blood*> and <*destruction of pancreatic beta cells* → *lack of insulin I in the blood* → *deficiency of insulin* → *elevated level of glucose in the blood*>. Then, we get <*type-I diabetes is-a (non-latent) diabetes*> according to Definition 3.

2.2 Types of causal chains in disease definitions

In this paper, we call causal chains that appear in the disease definition disease chains. In theory, we can consider three types of causal chains that appear in the disease definition, when we define a disease:

General Disease Chains are all possible causal chains of (abnormal) states in a human body. They can be referred to by any disease definition.

Core Causal Chain of a disease is a causal chain that appears in all patients of the disease.

Derived Causal Chains of a disease are causal chains obtained by tracing general disease chains upstream or downstream from the core causal chain. The up-stream chains imply possible causes of the disease, and the downstream ones imply possible symptoms in a patient suffering from the disease.

Fig.1 shows the main types of diabetes constituted by the corresponding types of causal chains. The figure shows that

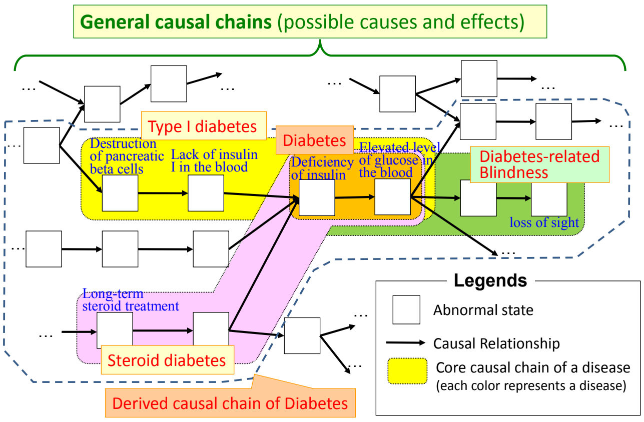

Fig. 1 Types of diabetes constituted of causal chains.

subtypes of diabetes are defined by extending its core causal chain according to its derived causal chains upstream or downstream.

Note here that it is obviously difficult to define all general causal chains in advance, because it is impossible to know all possible states in the human body and the causal relationships among them. In order to avoid this difficulty, we define the general disease chains by generalizing core/derived causal chains of every disease defined by clinicians in bottom-up approach. That is, we ask clinicians to define only core causal chains and typical derived causal chains of each disease, according to their knowledge and the textbooks on the disease. And then define the general disease chains by generalizing them.

Our disease ontology has been developed by clinicians in the 13 special fields. As of 11 May 2013, it has about 6,302 disease concepts and about 21,669 disorder (abnormal state) concepts with causal relationships among them.

3 DISEASE ONTOLOGY AS LINKED DATA

3.1 Basic policy to publish the disease ontologies as linked data

There are several approaches for system development based on ontologies. One of typical approaches is to use some APIs for ontology processing. Because our disease ontology is built using Hozo[2], we can develop application systems using APIs for Hozo ontologies. We can also use API for OWL since Hozo has OWL exporting function. On the other hand, linked data techniques are very efficient to develop applications across several datasets published on the Web. Because disease knowledge is related to various knowledge in other medical domains, we take an approach to publish the disease ontology as linked data to develop application system based on it.

[2] http://www.hozo.jp/

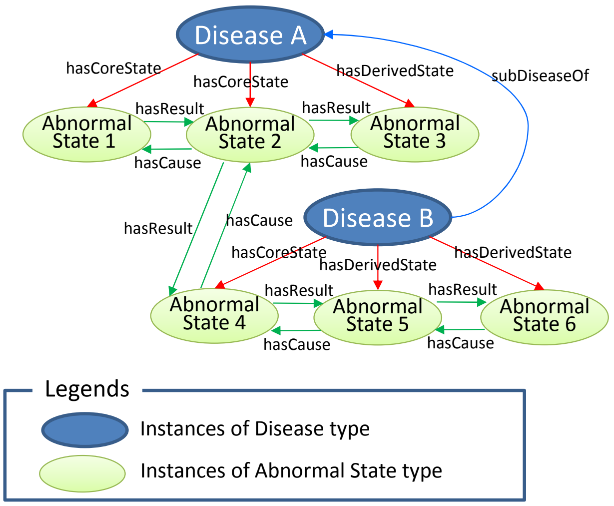

Fig.2 An RDF representation of diseaes.

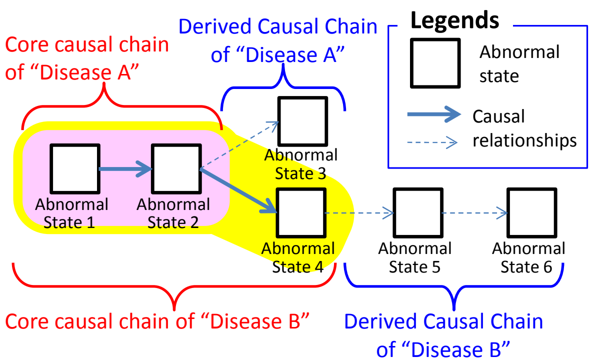

Fig.3 Causal chains of diseases shown in Fig.2.

Because the standard format for linked data is RDF, it may be regarded an easy task to publish ontologies in RDF formats using OWL or RDF(S) as linked data. However, an ontology language such as OWL is designed for mainly class descriptions, and the assumption is that the language will be used for reasoning based on logic; yet finding and tracing connections between instances are main tasks in linked data. Therefore, OWL/RDF and RDF(S) are not always convenient or efficient for linked data because of their complicated graph structures. This is problematic, especially when we want to use an ontology's conceptual structures as a knowledge base with rich semantics.

Consequently, we consider to design an RDF data model for publishing our disease ontology as linked data (Kozaki 2013). We outline the RDF model in the next section.

3.2 RDF Model for causal chains of diseases

After we constructed the disease ontology, we extracted information about causal chains of diseases from it and converted them into RDF formats as a linked data. We call the dataset *Disease Chain-LD*. It consists of diseases, abnormal states, and the relationships among them. Abnormal states

(1) Get all abnormal states.
```
select ?abn
where {?abn rdf:type dont:Abnormal_State}
```
(2) Get all cause of abnormal state <abn_id>.
```
select ?o
where {<abn_id> dont:hasCause* ?o }
```
(3) Get all causal chains which appear in definitions of disease <dis_id> as a list of abnormal state.
```
select DISTINCT ?o
where { <dis_id> dont:subDiseaseOf* ?dis .
 {?dis dont:hasCoreState ?o }
 UNION {?dis dont:hasDerivedState ?o }}
```

Fig.4 Example queries. In this figure "dont:" represents a prefix of the *Disease Chain-LD* and <abn_id> and <dis_id> represents id of a selected abnormal state and disease respectively.

are represented by instances of *Abnormal_State* type. Causal relationships between them are represented by describing *hasCause* and *hasResult* which are inverse properties. Abnormal states connected by these properties are a possible cause/result. Therefore, general disease chains can be obtained by collecting all abnormal states according to these connections.

Diseases are represented by instances of *Disease* type. Abnormal states that constitute a core causal chain and a derived causal chain of a disease are represented by *hasCoreState* and *hasDereivedState* properties, respectively. *Is-a (sub-class-of)* relationships between diseases and abnormal states are represented by *subDiseaseOf/subStateOf* properties instead of *rdfs:subClassOf* because diseases and abnormal states are represented as RDF resources, while *rdfs:subClassOf* is a property between *rdfs:Classes.*

Fig.2 shows an example of RDF representation of diseases. It represents *disease A* and its sub-disease *disease B*, whose causal chains are shown in Fig.3. Note that causal chains consist of abnormal states and causal relationships between them. Therefore, when we obtain a disease's core causal chain or derived causal chain, we have to obtain not only abnormal states connected to the disease by *hasCoreState/hasDereivedState* properties but also causal relationships between them. Although causal relationships are described without determining whether they are included in the causal chains of certain diseases, we can identify the difference by whether abnormal states at both ends of *hasCause/hasResult* properties are connected to the same disease by *hasCoreState/hasDereivedState* properties. Furthermore, when we obtain the causal chain of a disease that has a super disease, such as disease B in Fig.2, we have to obtain causal chains of its super disease in addition to the causal chain directly connected with it, and aggregate them.

The processing is not complicated; it just requires simple procedural reasoning. In summary, we can obtain a disease's causal chains, which define the disease through several SPARQL queries to the dataset. Fig.4 shows some

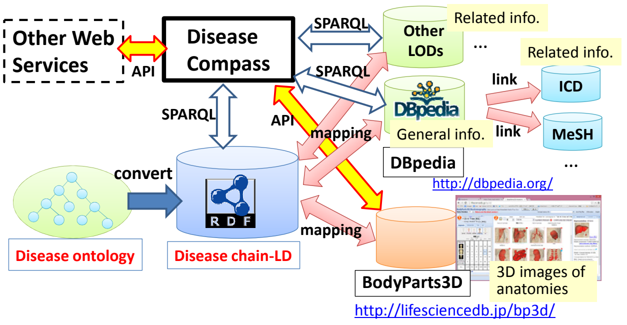

Fig.5 System architecture of Disease Compass.

example queries to obtain disease chains. We confirmed that we can obtain every information about disease chains using SPARQL queries (Kozaki 2013).

We published the disease ontology as linked data based on our RDF model. It includes definitions of 2,103 diseases and 13,910 abnormal states in six major clinical areas extracted from the disease ontology on May 11, 2013. The dataset contained 71,573 triples[3].

4 DEVELOPMENT OF DISEASE COMPASS

4.1 Disease Compass

It is not easy to use SPARQL for medical experts while every piece of information about disease chains can be obtained using SPARQL queries. Therefore, we developed a navigation system for disease knowledge named *Disease Compass*[4]. We designed the system so that the users can easily explore disease knowledge with related information even if they do not know about ontology or linked data.

4.2 System architecture

Fig.5 shows the system architecture of *Disease Compass*. The system obtains disease knowledge from Disease Chain-LD which is converted from the disease ontology. It has mapping information with other LODs (Linked Open Data) and web services. The system can obtain related information through these mappings. Though the system currently has mappings only to DBpedia and BodyPart3D, we can extend mappings to other LODs using existing approach to generate such linkages such as Song 2013).

Technically, the system uses two ways to access these datasets. One is SPARQL queries for linked data and the other is API for web services. If related resources (ontologies and other datasets) are published as LOD, the system is easily extended to link such related information using SPARQL. It is a large benefit to use linked data techniques. Please note that many linked data includes links to others. For example, DBpedia includes links to major medical codes such as ICD10 and MeSH. It means that the system can follows these links through mappings between Disease Chain-LD and DBpedia.

Disease Compass is developed as a web service that supports not only PCs but also tablets or smartphones. It is implemented using Virtuoso for its RDF database and HTML 5 for visualizations of disease chains and other information. All modules of the systems provides APIs for other web services. It enable others to use all functions of Disease Compass so that they work with related services.

4.3 User interfaces for navigation

Fig.6 shows user interface of *Disease Compass*. The users select a disease according to *is-a* hierarchy of diseases or search a disease chain by disease name or abnormal state which is included in it. The system visualizes disease chains of selected disease in a user friendly representation on the center of the window.

At the same time, the system obtains and shows related information about the selected disease and abnormal state from the following two web services. One is general information from linked data such as DBpedia, and the other is a 3D image of anatomies.

DBpedia[5] is a linked open dataset extracted from Wikipedia. It provides general information about diseases while its content is not authorized by medical experts. We suppose its contents is valuable enough to get an overview of diseases. In fact, it also gives links to major medical terminology and codes such as ICD10 and Mesh. The users can follow these links when they want to know more special information about the disease. This technology to obtain related information from other web resources (ontologies, medical codes, datasets etc.) through mappings is easy to apply to other linked data. We plan to extend the target linked data in the near future.

On the other hand, a 3D image of anatomies are generated using a web service named BodyPart3D/Anatomography (Mitsuhasi 2009). The target area of the image is decided by *Disease Compass* to combine all target of abnormal states appearing in the definition (causal chains) of the selected disease chain. Then, the system highlights a part of 3D image which is target of the selected abnormal state in the disease chains.

Such a functionality was enabled thanks to the disease ontology which is successfully combined with other web resources based on linked data technologies. As a result, *Disease Compass* can support the users to explore disease knowledge with related information through various web resources towards integrations of disease knowledge.

[3] Although the disease ontology includes definitions diseases in 13 clinical areas, we published parts of them that were well reviewed by clinicians. We will publish the latest version on the end of March, 2014. We provides a SPARQL endpoint to access the disease ontology at http://lodc.med-ontology.jp/ .

[4] Disease Compass is available at http://lodc.med-ontology.jp/ .

[5] We use DBpedia English (http://dbpedia.org) and Japanese (http://ja.dbpedia.org).

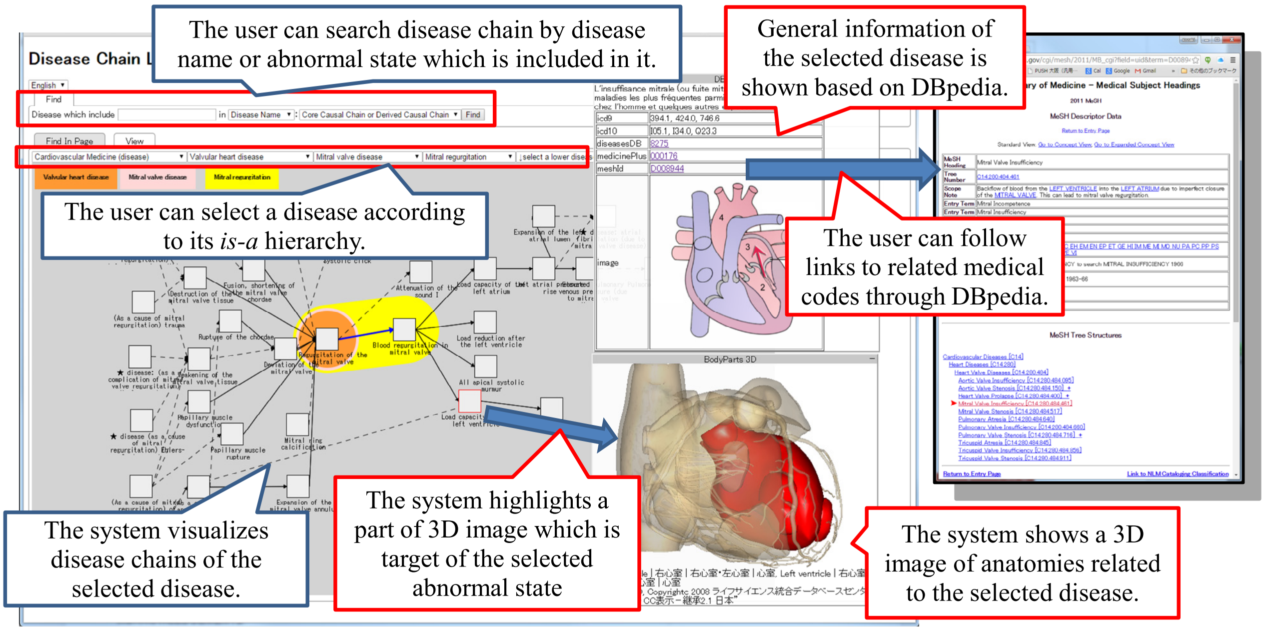

Fig.6 User interface of Disease Compass for navigation of disease knowledge.

We tried to extend the system towards integration of disease knowledge through ontologies and linked data technologies. As its first step, we investigated the differences in the hierarchical structure of biomedical resources and made a trial integration of our abnormality ontology and related resources such as PATO, HPO and MeSH based on ontological theory (Yamagata 2014).

As a result, we developed a prototype of the abnormality ontology as linked data with a browsing system. Thanks to mapping information with other resources, users can access disease knowledge through not only our abnormality ontology but also other open resources. We plan to extend this integration to our disease ontology and *Disease Compass*.

5 CONCLUDING REMARKS

This paper discusses a navigation system for disease knowledge based on the disease ontology and linked data technologies. Our disease ontology defines diseases based on causal chains of abnormal state (disorder) and a browsing system for it. It allows users to browse definitions of diseases with related information obtained from other linked data. We suppose that it can help them to understand about diseases from various point of views according to the users' interests and intention. The system was demonstrated for medical experts in some meetings and workshops and got positive comments while an evaluation by users is a future work.

Future work includes extension of related resources using linked data and developments of more practical applications using the Disease Chain LD. We also continue to improve the system including bug fixes and developments of new functions. The latest version of Disease Compass is available at the URL http://lodc.med-ontology.jp/ .

ACKNOWLEDGEMENTS

A part of this research is supported by the Japan Society for the Promotion of Science (JSPS) through its "FIRST Program"" and the Ministry of Health, Labour and Welfare, Japan. The authors are deeply grateful to Drs. Natsuko Ohtomo, Aki Hayashi, Takayoshi Matsumura, Ryota Sakurai, Satomi Terada, Kayo Waki, and other, at The University of Tokyo Hospital for describing disease ontology and assisting us with their broad clinical knowledge. We also would like to thank other team members, Drs. Yoshimasa Kawazoe, Masayuki Kajino, and Emiko Shinohara from The University of Tokyo for useful discussions related to biomedicine.

REFERENCES

Scheuermann, R. H., Ceusters, W., and Smith, B. (2009): Toward an Ontological Treatment of Disease and Diagnosis. *Proc. of the 2009 AMIA Summit on Translational Bioinformatics*, San Francisco, CA, 116-120.

Mizoguchi, R., Kozakil, K. and et al. (2011): River Flow Model of Diseases, *Proc. of ICBO2011*, Buffalo, USA, 63-70.

Osborne, J. D., et al. (2009): Annotating the human genome with Disease Ontology. *BMC Genomics* **10**(1):S6.

Cowell, L. G. and Smith, B (2010): Infectious Disease Ontology. *Infectious Disease Informatics*, Chapter 19, Sintchenko V., 373-395.

Dezhao Song and Jeff Heflin. (2013): Domain-Independent Entity Coreference for Linking Ontology Instances. *ACM Journal of Data and Information Quality* **4**(2), Article 7.

Mitsuhashi, N. et al. (2009): BodyParts3D: 3D structure database for anatomical concepts, *Nucleic Acids Research* **37**, D782-D785.

Kozakil, K., and et al.(2013): Publishing a Disease Ontologies as Linked Data, *Proc. of JIST2013*, Seoul, Korea, 110-128.

Yamagata Y, Kozaki K, Mizoguchi R, Imai T and Ohe K.(2014):Towards the Integration of Abnormality in Diseases, Proc. of ICBO 2014, 7-12.

