Data Integration Framework of Pharmacology Databases Using Ontology

Phimphan Thipphayasaeng [1], Poonpong Boonbrahm [1], Marut Buranarach[2] and Anunchai Assawamakin[3]

[1] School of Informatics, Walailak University, NakornsiThammarat, Thailand
[2] National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand
[3]Faculty of Pharmacy, Mahidol University, Bangkok, Thailand

Abstract. This paper presents linked data of pharmacology domain generated with ontology as central schema. To link data from several formats, data are transformed into database format, and they are mapped to ontology. The ontology is developed with concepts provided in the dataset. Mainly, the developed ontology contains a concept of drug, disease, genetic and drug-gene interaction with their details. The ontology is used as central schema to link concepts together; thus, linked data are created. Within the linked data, we found three types of links, i.e. addition of instances, addition of attributes and changing of variable data field to a link to another table. In this paper, actual scenarios of the found links with exemplified data are explained.

Keywords: Linked data; Ontology; Pharmacological data; Data integration

1 Introduction

In pharmacology field, data of drugs, their usage, study, and description have been digitalized and provided on many sites. Those data give different characteristics of drugs; hence, a schema of the databases was designed specifically for their purpose. Additionally, these data are live data that have regularly been updated for experts to reference. These data sources are open to use and are important for experts in the field to consult for a case result and lengthen their researches.

These databases apparently contain a large amount of data, and their schema is complex. Users are needed to understand database schema and have pharmacological background to read through the data. In fact, provided data have been gathered based on a ground from where they are designed. Every database has its own strength and specified to their locality. In usage, experts commonly require searching through many databases to assure correctness and coverage of data, and they need to be aware of different appearance terms referring to a same concept or instance (synonymy) or a same term with many definitions (polysemy).

To support users of the data, this work aims to link those open pharmaco-genetic data together using Resource Description Framework (RDF) standard. RDF standard [1] is often used as the metadata interchange format since its expression was designed

to represent as a model of information using a variety of syntax notations and data serialization formats. This work proposes a method for interoperability between the datasets from different formats and schema using Linked Data framework. Moreover, a method for integration of data is designed to recognize an overlapping of data and extend range of data with other data sources. A complete data integration solution should provide data influent to be trusted by within crosschecking from a variety of sources since a volume of data will be increased and coverage of scope will be extended.

2 Background

Pharmacology is the study of drug action and effects of the drug on biological systems. Nowadays, available linked data were created and distributed as accessible data for experts in the field. In this work, we review some of the well-known linked data in pharmacology and summarized them as follows.

ChEMBL[2] provides chemical entities of biological activities against drug targets. That could be used as a reference for drug researchers. DrugBank [3] gathers information on drugs and their targets that include in drug target discovery, drug design, drug screening or docking, interaction prediction, metabolism prediction and pharmaceutical education. Diseasome [4] provides a resource for biomedical researchers that include disease-gene associations using network maps and understanding of the genetic origins of disease. DisGeNET [5] representation knowledge in the molecular mechanisms that combines detailed gene data with disease and calculate a score in order to rank of these associations for support research in the biomedical science. The Linking Open Drug Data project focuses on linking various sources of drug data to answer scientific questions [6].

3 Methodology

This work aims to combine provided pharmacological data from several sources into linked information. Ontology is chosen as an intermediate schema to relate data into uniform concepts. The method involves with four processes as summarized in Fig. 1.

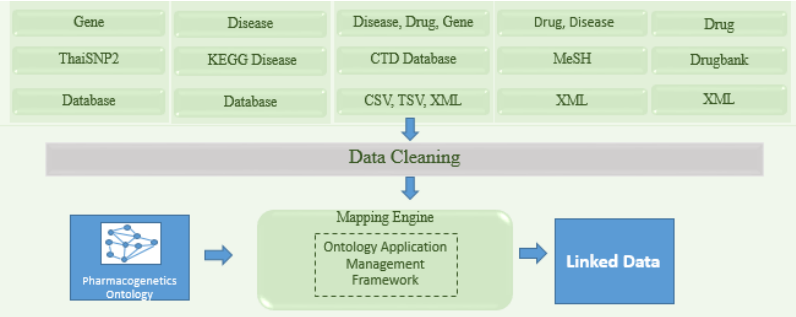

Fig. 1. An overview of Data Integration of Pharmacology Databases Using Ontology

3.1 Data preparation

In this paper, five pharmacological datasets, i.e. KEGG Disease [7, 8, 9], ThaiSNP [10], Comparative Toxicogenomics Database [11], Drugbank [3] and MeSH [12], are chosen as input data for integration. The details of the dataset are given in Table 1.

Table 1. Details of the chosen pharmacological datasets

Dataset name	About	Format	Concepts
KEGG Disease	disease entry knowledge on genetic and other relevant	relational database	Drug, Disease
ThaiSNPs	Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs) in genetic of Thai population	relational database	Gene, SNPs
CTD	chemical–gene/protein interactions, chemical–disease and gene–disease relationships	CSV, TSV, XML	Gene, Disease
Drugbank	bioinformatics and cheminformatics resource including detailed drug data with their drug target	XML	Drug, Pharmacology
MeSH	mapping drug-disease relationships in research	XML	Drug, Disease

The commonness of data in these dataset in Table 1 is about drug, gene, and disease. All datasets represent in several data formats such as database and XML. Please be noted that these datasets contain some data tags involving in data management and referring, such as sorting key and ID for their related application, in which we will ignore in data integration since they are not semantically important. To combine these data together, we need to uniform data format to database format. These databases will be mapped to ontology in later process.

3.2 Ontology design

Since the core of the aforementioned data is about drug, we decide to initiate with drug concept as a main class and expand relations from this concept. Our ontology was designed on and created by Hozo ontology editor [13] following the development guideline by Mizucuchi [14].

Firstly, terms in these datasets were gathered from the table heads and fields. With the gathered terms, concepts of terms were recognized and relations to link concepts were decided. Relations of concepts were decided based on following criteria:

- is-a relation : forming superclass-subclass relation in which the concepts must be the same kind, and all properties of superclass inherit to its subclass
- object property : forming belonging relation of two concepts in which representing of a part in another concept

- data property : forming concept-data type relation to signify a concept containing a value such as number or string
- instance of : providing a relation to link a concept to real data or instance, this relation is to link ontology class to real data in database

Regarding to data in dataset, our ontology is designed to gather all the concepts. The major concepts are Drug, Gene, Disease, interaction and SNP, and their properties are the fields of their dataset. Some parts of the ontology are demonstrated in Fig. 2.

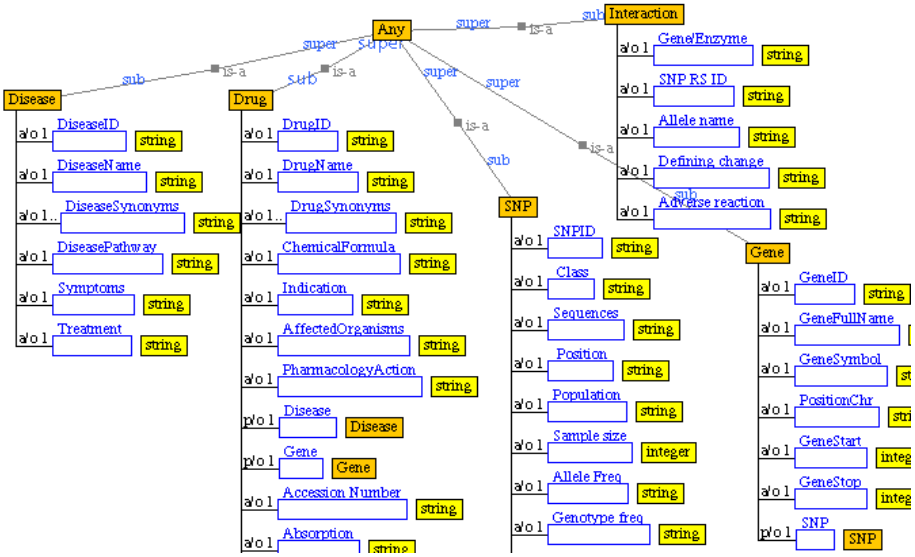

Fig. 2. Some parts of the designed ontology using Hozo Editor

3.3 Schema mapping

With ontology as a schema to gather and relate concepts from the datasets, the ontology can be mapped to all data in the datasets. An ontology mapping from Ontology Application Management (OAM) [15] tool is chosen to help us in mapping. The mapping of ontology class and dataset field was exemplified in Table 2.

Table 2. Examples of an ontology-data mapping table

Ontology property	**Data Field**			
	DB1	**DB2**	**DB3**	**DB4**
Drug name	Drug name	-	-	name
Drug Chemical Structure	Structure	-	-	chemical structure
Drug Indication	Indication	-	-	indication
Genotype	Gene	gene	genotype	-
Disease Description	-	indication	description	-

3.4 Data Linking

To combine these data, we have to focus on commonness of data and concepts. In this work, we found three types of integration of different dataset as 1. Same head concept, different data, 2. Same head concept, different properties and 3. One of the concept is a property of another.

For the first integration, this results in gaining more instances. This helps in expanding a variety of data. It should benefit users in giving more data to look through. The second integration is about linking more data attributes for the same data. This widens more aspects of data and includes more relevant information to the concepts. This type gives users for more broad view of data properties. Last, the integration of different tables. This is to give in-depth details of the information since the information will be added with information of whole table. This is a change of single value or text of one attribute to more details. The three types of integration are summarized in Table 3.

Table 3. A summary of data linking from different datasets

Data Integration Type		Linking	Ontology Representation
1	Same head concept, different data	Adding of instance roll	none
2	Same head concept, different properties	Adding of attributes to data table	More properties of a concept
3	one of the concept is a property of another	Changing from value required field to foreign key for linking to another table	Changing of attribute-of property to part-of property

4 Usage Scenario

In this section, we demonstrate a case scenario of the data linking types to exemplify actual cases with the dataset chosen for integration in this work.

The scenario is an integration of Drugbank and CTD. In these datasets, they have common data about a drug. Both of these datasets are about drug. For Drugbank, given data are relevant to drug chemical compound and affected organ. Pharmaceutical action of drug, however, is mentioned in CTD. These attributes do not exist in another set, and it is recognized as the second type mentioned in Table 3. With ontology mapped to these data, we realize that the table is the same concept since they are both mapped to the same ontology class although table labels are different. After integration, we obtained a combination of more properties to widen more aspects of the same data.

5 Conclusion and Future Work

This paper presents a data integration of pharmacology data from several sources using ontology as a central schema. Five datasets are gathered in which provides data about drug, disease, gene and interaction. To integrate data, cleaning process and uniform of data format are initiated. Ontology is created with concepts given in datasets and is

mapped to the data via OAM framework. With ontology mapped to data, data from those five sources are linked with semantic. In this work, the linked data contain three types of links that are addition of instances, addition of attributes and changing of variable data field to a link to another table. In the future, we plan to include more relevant dataset to link more pharmacology data. An automatic method to map data to ontology class will be researched to reduce human burden and time consuming in mapping process. Lastly, we will apply semantic search with the obtained linked data.

References

[1] Bruijn, J., Welty, C., RIF RDF and OWL Compatibility: available online at https://www.w3.org/TR/2013/REC-rif-rdf-owl-20130205/

[2] Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, Hastings J, Chen B and DJ Wild DJ.: The ChEMBL database as linked open data. Journal of cheminformatics. 5(1), 1–12 (2013).

[3] Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey J.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668-D672 (2006).

[4] Wysocki K, Ritter L.: Diseasome: an approach to understanding gene-disease interactions. Annu Rev Nurs Res. 29, 55–72 (2011).

[5] Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI.: DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. The journal of biological Database and Curation. 2015, 1-17 (2015).

[6] Samwald M., Jentzsch A., Bouton C., Kallesøe CS., Willighagen E., Hajagos J., Marshall MS., Prud'hommeaux E., Hassenzadeh O., Pichler E., Stephens S.: Linked open drug data for pharmaceutical research and development. J Cheminform. 3, (2011).

[7] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355-D360 (2010).

[8] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457-D462 (2016).

[9] Kanehisa, M. and Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[10] Hattirat, S., Ngamphiw, C., Assawamakin, A., Chan, J., and Tongsima, S.: Catalog of Genetic Variations (SNPs and CNVs) and Analysis Tools for Thai Genetic Studies. Computational Systems-Biology and Bioinformatics. 115, 130-140 (2010).

[11] Davis, A., Grondin, C., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B., Wiegers, T, Mattingly, C.: The Comparative Toxicogenomics Database's 10th year anniversary: Update 2015. Nucleic Acids Res. 43, D914-D920 (2015).

[12] Rogers, Frank B.: "Communications to the Editor." Bulletin of the Medical Library Association. 70, 5131–5136 (1963).

[13] Hozo Ontology Editor: available online at http://www.hozo.jp

[14] Mizuguchi, R.: Tutorial on ontological engineering - part 1: Introduction to ontological engineering. In: New Generation Computing. 21, 365–384 (2003).

[15] Buranarach, M., Thein, Y., Supnithi, T.: A Community-Driven Approach to Development of an Ontology-Based Application Management Framework. Semantic Technology. 7774, 306–312 (2013).