Resource Description Framework Generation for Tropical Disease Using Web Scraping

Amalia Amalia
Departement of Computer Science
Universitas Sumatera Utara
amalia@usu.ac.id

Rizky Maulidya Afifa
Departement of Computer Science
Universitas Sumatera Utara
rizky.maulidya.afifa@students.usu.ac.id

Herriyance Herriyance
Departement of Computer Science
Universitas Sumatera Utara
herriyance@usu.ac.id

***Abstract* — Tropical diseases are diseases that commonly occur in tropical areas like Indonesia. The most frequent diseases plaguing Indonesia community consists of common diseases that are categorized as tropical diseases, e.g., malaria, leprosy, and lymphatic filariasis. As first aid, Indonesian people usually rely on a search engine to search for information about diseases and medication especially of general illnesses like coughs, colds, and fever. However, some result of conventional search engine still generates un-related articles. Another drawback of a conventional search engine, sometimes it cannot handle synonym meaning of disease in searching especially for the non-medical term in Bahasa Indonesia. For example, for lymphatic filariasis is better known as kaki gajah (In English: Elephant foot). One solution is to build a search engine based on semantic web technology. The semantic web can show the relationships between the data, including synonym of disease name in Bahasa Indonesia in Resource Description Framework (RDF) serialization form. Unfortunately, based on our initial survey, the websites that provide health information in Bahasa Indonesia do not provide RDF files. This research aims to generate an RDF serialization that describes the relationship ontology in tropical diseases and treatments terms. We implemented a web scraping method to harvesting and extracting vocabularies from some popular health websites, creating the relationship between terms with ontology and convert this information into an RDF serialization.**

Keywords: web scraping, semantic web, Resource Description Framework (RDF), medication, ontology, RDF Schema

I. INTRODUCTION

Many people in Indonesia rely on the internet search engine to find out about a disease and its medicine as first aid. However, some result of this searching information still contains un-related articles, and the search engine sometimes cannot handle synonym meaning of disease in searching especially for the non-medical term in Bahasa Indonesia. For example, for lymphatic filariasis is better known as kaki gajah (In English: Elephant foot), Rubeola is better known as campak or kerumut. The drawbacks of this conventional search engine because it uses keywords to generate the result. This search engine generates information that depends on the keywords without knowing the meaning of the terms. Most of the information circulating on the internet are only provided for human consumable and not yet available for machine-readable. One solution is to build search engine based on semantic web technology, where semantic web can identify information or meaning of the data, such as disease, drug classes, the synonym name of the disease both in a medical term or in the term that better known in Bahasa Indonesia, e.g., “nyctalopia” or “rabun ayam” (in English: myopic chicken).

Semantic Web is the thought of Sir Tim Berness-Lee, as well as the inventor of the WWW, URLs, HTTP, and HTML. Sir Tim Berness-Lee defines the semantic web as a web of development with information that has a well-defined meaning so that humans and machines can work together so that information processing can be optimized [1]. In addition, the five-star open data requires semantic web to provide the flexibility in processing information. However, Indonesia has not implemented semantic web yet to provide the open data [2]. Previous research also suggests semantic web as the protocol for e-commerce data harvesting [3].

The advantages of semantic web technology, when compared to the relational database, is relational databases can store data for information, but it not able to show the relationships between the data, so the constraints of the relational database is not capable of displaying the results with the desires of the user if the query is not matched with the field on the database. On semantic web, relationships between terms are described in an RDF serialization. RDF is a standard model [4] for data exchange on the internet. RDF consists of a subject, predicate, and object triples, in which subject and object are used to represent the two things in the world and a predicate to determine the relationships between the data [5]. For example, to state that “type of skin disease is panu (Tinea Versicolor)”, can use the triple with the subject shows “skin disease”, the predicate that indicates the “type of”, and an object that shows “panu”. The subject and the object in one triple can have many relationships and created other triples. For example, "panu" as a subject, "same as" as a predicate, and "Tinea Versicolor" is an object. So if the user searches "panu" the system will generate both of term "panu" and "Tinea Versicolor”.

This data model can make data interchange or linked data between websites and machine automatically. Some websites already provide their RDF serialization to describe their metadata. Unfortunately, based on our initial survey, the sites

978-1-5386-6717-0/18/$31.00 ©2018 IEEE

that offer health information in Bahasa Indonesia do not provide RDF files.

According to the background, this study aims to create an RDF serialization for tropical disease domain. The challenge of this study is, we need a set of vocabularies as infrastructure formation to generate an RDF. One technique to collect vocabulary that is by extracting from web pages known as web scraping method. Web scraping is a technique for automatically extracting data from HTML in a specific website using a special program to manipulate, such as changing web pages into other formats like CSV and Extensible Markup Language (XML) [6]. This study is an essential preparation as a first step to establish a method to make data interchange automatically. Based on this RDF serialization, we can create a further broader relationship like ontology.

The paper is divided into five sections. Section 2 describes several of the related studies to this research. Section 3 describes the methodology, Section 4 presents result and discussion and Section 5 draws conclusions of this study.

A. Related Work

Research conducted by [7] provides information about tropical diseases that can be used as a reference for the data. Previous studies to crawling and scraping data from websites are already done by [8] and [9]. The study of [8] is shown the method to collect health information from many websites with a focused crawler that implements multithreaded programming, Larger-Sites-First algorithm and also Naïve Bayes classifier. Research by [9] that Breadth First Search (BFS) can perform crawling in accordance with domains that want to extract, as well as on research [10] mention that data extraction with the semantic web can provide adequate information because information is displayed only information required users without a lot of spam. Then in the research [11], ontology is used for structuring and filtering the knowledge repository approach: Ontology-based web crawlers use ontological concepts for improving their performance. So, it will be effortless to get relevant data as per the user requirements. The differences in this research are system uses data extraction techniques for drugs and tropical diseases and use an RDF so that data has meaning and relationships between data.

II. METHODOLOGY

In this study, implementation of web scraping on the ontology-based semantic web for tropical diseases drug data has several steps. First, the user collects a dataset in a variety of tropical diseases. Then the dataset is extracted from the valid websites. Step by step will be described using the architecture diagram in Figure 1.

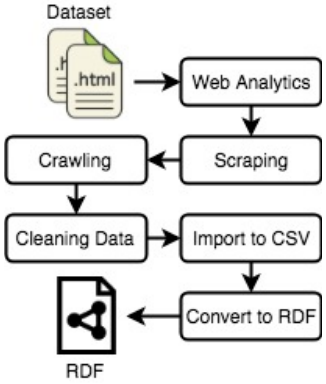


Figure 1. Architecture diagram

A. Dataset

This research utilizes a dataset retrieved from several trusted and valid information about drugs, such as drugs.com (www.drugs.com), alodokter (www.alodokter.com), and Badan Pengawas Obat dan Makanan-BPOM (in English: National agency of drug and food control of Republic Indonesia) in URL address www.pom.go.id). For tropical diseases, we retrieve information from an interview with a student of the Faculty of Medicine of the Universitas Sumatera Utara, whose data is extracted from the lecturer learning material on tropical diseases.

B. Web analytics

Web analytics is a process to extract the critical feature from the dataset. We retrieve drugs information, i.e. registration number, product name, brand, composition, packaging, date, registrars and other preparations. We extract this information from BPOM. The information about a disease and its medication is retrieved from drugs.com. For additional the data complementing like disease synonym in Bahasa Indonesia, we extracted from alodokter.com.

C. Crawling

Crawling is a process to collect relevant pages from drugs.com, alodokter and BPOM websites. The method of the crawling using Breadth First Search algorithm. Breadth First Search (BFS) is a search algorithm from left to right, visit each link on a page before processing to the next page. The algorithm is to visit the links on the first page, then visit each link on the first page on the first link until the last visited links [12]

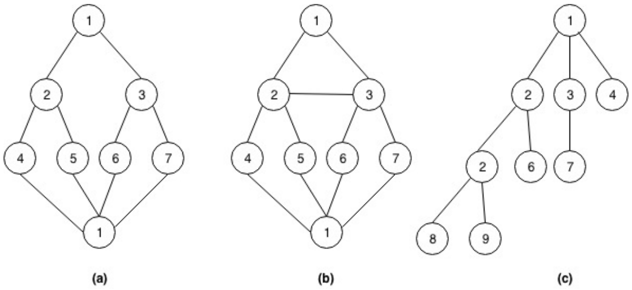


Figure 2. Breadth-first search algorithm graph [11]

Description:
Figure (a), solution : 1, 2, 3, 4, 5, 6, 7, 8
Figure (b), solution :1, 2, 3, 4, 5, 6, 7, 8
Figure (c), solution : 1, 2, 3, 4, 5, 6, 7, 8, 9

Breadth-first search algorithm used to check the data that will be taken on a particular link if retrieved data then this algorithm will continue the search process until the last link. After the information is retrieved, the data is stored and then will be taken using the technique of scraping.

D. Scraping

Web scraping is the process of retrieving a document from the internet, generally like a web page in a markup language such as HTML or XHTML document and analyze the specific data to be taken from that page to be used for other interests. Web scraping only focuses on obtaining data through the retrieval and extraction of data with different data sizes.

E. Cleaning data

Data extraction has the HTML tags so that the data should be in the cleaning to remove HTML tags and generates essential data as needed

F. Import to CSV

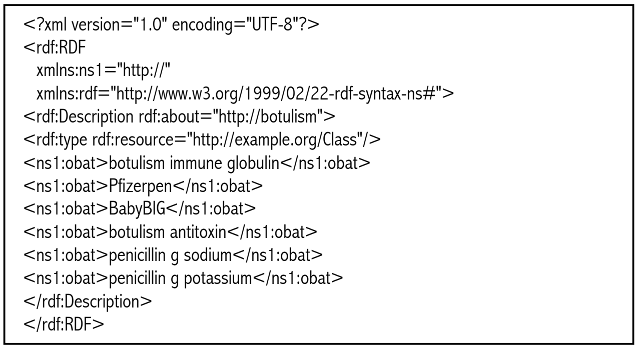
After cleaning data from HTML tag, then it will import the data into a different format, i.e., CSV, the format used to ease the process of conversion to RDF so that it can be made a relation between data with other data.

G. Convert to RDF

After getting a CSV format, then the data is converted into data structures to describe an RDF metadata from any data. Converting CSV to RDF using open source tools that work in python platforms i.e. CSV2RDF. In addition to generating RDF documents, this tool also creates an RDF Schema.

H. RDF

Resource Description Framework (RDF) is a W3C standard is used to describe a metadata. Metadata is information about information. RDF is designed to be understood by the computer and is not intended to be shown to the user. RDF uses a Uniform Resource Identifier (URI) to identify the resource in which the resource is created with the properties and property values. RDF Schema is a representation that can be used to develop a vocabulary for describing the class, subclass, and property of an RDF [13]. This RDF can be seen in Figure 3. RDF Schema to define the relationship between class and property, as well as the relationship between subclass and class properties to property between one another, have specific relationships. In figure 4 it is shown RDF graph. The figure shows the relationship of each object, and triples of data. The subject is botulism; the object is the name of the drug and the predicate that is the drug to determine the relationship between the disease and the name of a drug.



```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:ns1="http://"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:about="http://botulism">
<rdf:type rdf:resource="http://example.org/Class"/>
<ns1:obat>botulism immune globulin</ns1:obat>
<ns1:obat>Pfizerpen</ns1:obat>
<ns1:obat>BabyBIG</ns1:obat>
<ns1:obat>botulism antitoxin</ns1:obat>
<ns1:obat>penicillin g sodium</ns1:obat>
<ns1:obat>penicillin g potassium</ns1:obat>
</rdf:Description>
</rdf:RDF>
```

Figure 3. Resource Description Framework

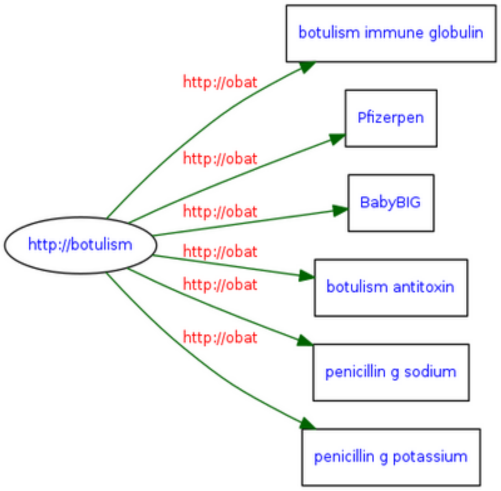


Figure 4. RDF graph

III. RESULT

Table 1 and Table 2 showed the result of tropical disease and general disease. Table 3 showed the relationship between diseases and drugs. Experiments result will generate the structure of RDF triples (subject, predicate, object) so that it can be seen a correlation between data with each other. Figure 5 will explain the disease and treatment as well as the sense of allergies and the meaning of disease it. RDF data structure, then do a query

Table 1. Tropical disease

Tropical Disease		
Anthrax	Leptospirosis	Tetanus
Botulism	Meningitis	Tuberkulosis (TBC)
Difteri	Kusta	Cacar Air
Ebola	Campak	Virus Zika
Malaria	Batuk Rejan	Cytomegalovirus
Brucellosis	Rakitis	Penyakit Lyme
Filariasis	Rubella	Tifus
Herpes genital	Sepsis	Demam Kuning
Influenza	Toksoplasmosis	

Table 2. General Disease

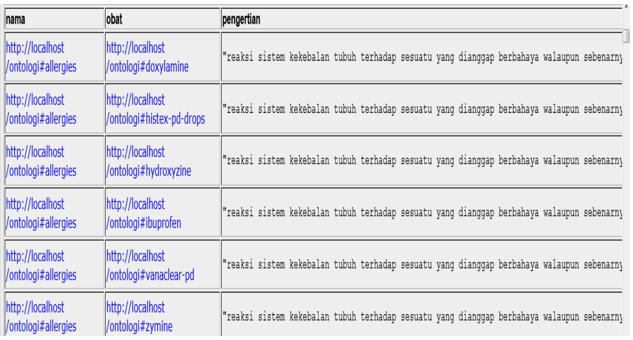
General Disease		
Demam	Alergi	Eksim Atopik
Batuk-batuk	Mata Kering (Mata Merah)	Mata Merah

Table 3. Disease and Drugs Relationship

Disease	Drugs
MALARIA	Artcleaemether/Lumefantrine
	Coartem
	Malarone
	Plaquenil
	Doxycycline
	Mefloquine
	Chloroquine
BOTULISME	Botulism immune globulin
	BabyBig
	Penicillin g potassium
	Pfizerpen
	Botulism antitoxin
	Penicillin g sodium

```
<rdfs:Description rdf:about="http://localhost/ontologi#allergies">
 <rdf:type rdf:resource="http://localhost/ontologi#penyakitUmum"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#doxylamine"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#histex-pd-drops"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#hydroxyzine"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#ibuprofen"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#vanaclear-pd"/>
      <ontologi:pengobatan
rdf:resource="http://localhost/ontologi#zymine"/>
      <ontologi:pengertian>reaksi sistem kekebalan tubuh terhadap
sesuatu yang dianggap berbahaya walaupun sebenarnya tidak berbahaya.
Ini bisa berupa substansi yang masuk atau bersentuhan dengan
tubuh</ontologi:pengertian>
      <rdfs:label>allergies</rdfs:label>
   </rdfs:Description>
```

Figure 5. Structure of RDF



nama	obat	pengertian
http://localhost /ontologi#allergies	http://localhost /ontologi#doxylamine	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny
http://localhost /ontologi#allergies	http://localhost /ontologi#histex-pd-drops	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny
http://localhost /ontologi#allergies	http://localhost /ontologi#hydroxyzine	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny
http://localhost /ontologi#allergies	http://localhost /ontologi#ibuprofen	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny
http://localhost /ontologi#allergies	http://localhost /ontologi#vanaclear-pd	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny
http://localhost /ontologi#allergies	http://localhost /ontologi#zymine	"reaksi sistem kekebalan tubuh terhadap sesuatu yang dianggap berbahaya walaupun sebenarny

Figure 6. The result of triples data

IV. CONCLUSION

This research aims to generate an RDF serialization that describes the relationship ontology in tropical diseases and treatments terms especially in relevancy to terms in Bahasa Indonesia. We implemented a web scraping method to collected relevant vocabularies from websites that provide health information in Bahasa Indonesia. The results on the application of semantic web for drug and disease data generate RDF structure as a triple (subject, predicate, object) that has the relationships between the data, so the data has meaning. This RDF result can be implemented to build a searching system infrastructure based an ontology.

V. REFERENCES

[1] P. Hitzler, M. Krotzsch, and S. Rudolph, *Foundations of semantic web technologies*. 2009.

[2] D. Gunawan and A. Amalia, “The Implementation of open data in Indonesia,” in *2016 International Conference on Data and Software Engineering (ICoDSE)*, 2016, pp. 1–6.

[3] D. Gunawan, “Protocol for E-Commerce data harvesting,” in *Proceedings of the 2015 International*

Conference on Technology, Informatics, Management, Engineering and Environment, TIME-E 2015, 2016, pp. 1–5.

[4] S. K. R. Madula, “Archiving Relational Databases using a Semantic Web Representation,” Uppsala University, 2007.

[5] H. Asghar, Z. Anwar, and K. Latif, “A deliberately insecure RDF-based Semantic Web application framework for teaching SPARQL/SPARUL injection attacks and defense mechanisms,” *Comput. Secur.*, vol. 58, pp. 63–82, May 2016.

[6] S. Munzert, C. Ruoba, P. Meiboner, and D. Nyhuis, *Automated data collection with R : a practical guide to Web scraping and text mining*. .

[7] A. V. Vitianingsih, D. Cahyono, and A. Choiron, “Analysis and design of web-geographic information system for tropical diseases-prone areas: A case study of East Java Province, Indonesia,” in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2017, pp. 255–260.

[8] A. Amalia, D. Gunawan, A. Najwan, and F. Meirina, “Focused crawler for the acquisition of health articles,” in *Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016*, 2017.

[9] N. Pawar, K. Rajeswari, and A. Joshi, “Implementation of an efficient web crawler to search medicinal plants and relevant diseases,” in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, 2016, pp. 1–4.

[10] K. Sundaramoorthy, R. Durga, and S. Nagadarshini, “NewsOne — An Aggregation System for News Using Web Scraping Method,” in *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, 2017, pp. 136–140.

[11] C. Saini and V. Arora, “Information retrieval in web crawling: A survey,” in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2635–2643.

[12] M. Turland, *php|architect’s Guide to Web Scraping*. Marco Tabini & Associates, , 2010.

[13] L. Yu, *A Developer’s Guide to the Semantic Web*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.