

VISUALIZATION OF ONTOLOGY-BASED DATA WAREHOUSE FOR MALARIA SPREAD INCIDENCES USING PROTEGE

Aan Kardiana dan Nova Eka Diana

YARSI E-Health Research Center, Faculty of Information Technology
YARSI University, Jakarta, Indonesia

Email: aan.kardiana@yarsi.ac.id

Abstract

Malaria is a communicable disease caused by a plasmodium parasite and transmitted among human by Anopheles mosquitoes. Late medication of this disease can cause a death of patients. Indonesia has many endemic areas with a high volume of patients diagnosed by Malaria. Currently, this incidences data is stored in Microsoft Excel files. We need to build a data warehouse to easily manage these data. Here, we create ontology of Malaria's incidence data to figure out the important information in Malaria data warehouse that we want to build. We identify entities, classes, subclasses, and relationships between these entities. We employed Protégé to build and visualize the ontology of Malaria incidence data.

Keywords: ontology, data warehouse, malaria, visualization, Protégé

Abstrak

Malaria adalah penyakit menular yang disebabkan oleh parasite plasmodium dan dipindahkan ke tubuh manusia oleh Nyamuk Anopheles. Penanganan yang lambat terhadap penyakit malaria dapat menyebabkan kematian pasien. Indonesia adalah salah satu negara yang memiliki banyak wilayah endemik dengan volume yang cukup tinggi terhadap pasien didiagnosis penyakit malaria. Saat ini, data dari kasus-kasus malaria di berbagai wilayah endemik di Indonesia masih disimpan dalam banyak file excel. Akibatnya, terdapat kesulitan untuk memperoleh informasi yang cepat dalam pengambilan keputusan untuk penanganan kasus malaria. Oleh karena itu, perlu dibangun sebuah data warehouse untuk mengatur data tersebut secara terpusat. Pada artikel ini, dibuat ontology untuk mengidentifikasi informasi dan parameter penting dari elemen-elemen yang harus ada dalam data warehouse, seperti: entitas, kelas, sub-kelas, dan hubungan antar entitas. Protégé digunakan untuk membangun visualisasi dan memudahkan pemahaman terhadap hubungan antar entitas dalam data kasus malaria.

Kata Kunci: ontologi, data warehouse, malaria, visualisasi, Protégé

1. Introduction

Malaria is an infectious disease spread by a mosquito of the genus Anopheles. This animal carries out a plasmodium parasite and spread it into human blood circulation through a bite. According to WHO data, Malaria causes a death in children for every 30 seconds. In every year, there are about 300-500 million people infected and 1 million people died caused by this disease. Most of 90% Malaria incidence happened in Africa with children as the most victims. To overcome this incidence, some medicines have been proved able to alleviate Malaria disease. But, it still cannot totally cure this disease because of plasmodium parasite that hidden in human liver. Hence it will be very difficult for those medicines to attack it.

In Indonesia, Malaria endemic areas spread from west to east part of the country. However, most of these endemic data are not organized well.

Therefore, it was very difficult for the ministry of health to adapt the effective way for reducing and preventing Malaria spread. A very niche representation and visualization of Malaria incidences data over the years will facilitate the ministry to quickly take action based on the data.

Data warehouse gives a multidimensional view about a big amount of historical data from operational data sources, in order to provide useful information for decision maker to improve their organization business process [1]. The representation of this multidimensional view of data, the relationship and data flow among entities in the organization can be easily visualized using a specified tool for data warehousing purpose. One research department in Stanford University specialized on Ontology, has developed software called Protégé. Protégé is an application used to easily process data with a specified format so it can be easily understand and transformed into a more useful form of

data. Then, this data can be utilized in supporting decision making process.

In this paper, we create a visualization of Malaria's cases data warehouse in Indonesia using various variables, such as class, entity, object, data property, and individuals. Our data contain of many variables and data types from Malaria spread to the detail of the patients diagnosed with Malaria.

In the past years, a big amount of data can be classified into many categories based on the specification criteria. Data warehouse has become the most voted option for storing data in a very big capacity, such as academic, business, or health data. Data warehouse is usually developed to put together data from various sources and then process it for decision making purpose. A big amount of data and complex relationship among entities in data warehouse make it difficult to understand and generate meaningful information which represents those data. Therefore, we need a clear and detail representation that visualize all of these entity relationship. One way to easily represent the connectivity among entities is by creating the ontology visualization of these entities.

Usually, data warehouse is mainly used for effectively managing and analyzing a huge amount of data. Therefore, the structure of data warehouse should consist of multidimensional models, Online Analytical Processing (OLAP) model for analysis task [2]. Inputs for data warehouse are data that coming from operational database or even data from various web pages. The bigger amount of data processed in data warehouse, then the multidimensional representational would not be a good option because of the complex visualization.

Web semantic technology, such as ontology, is very useful in representing data on the webs into coherent information. Ontology can also be utilized in design process to formally represent business requirements. In research field, ontology usually employed as a tool for visualizing a logical thinking and designing the research output. Ontology also can clearly explain the multidimensional model of data warehouse and break it down into a more detail information. In building the ontology of data warehouse, algorithm scenario is introduced. Automatic scenario engagement is very important to eliminate the dependency toward an expert when designing or analyzing data sources. Many approaches have been proposed to automate the designing task for finding multidimensional elements that facilitate data warehouse development.

Many researches had utilized ontology for visualizing, digitizing process works, and describing detail of variables and data type in data warehouse. Khouri et al [3] explained that ontology can be categorized as Conceptual Ontology (CO) and Linguistic Ontology (LO). Conceptual Ontology is a category of objects and properties in domain, whereas Linguistic Ontology is terms used in the given domain. In their research, Talebzadeh et al [4]

was focusing on creating data warehouse for unstructured data type, e.g. text file, web data, etc. It is called unstructured because these data have a very high volume of capacity and resource level which are really complicated. Therefore, it takes a very long time to process those data. In this context, ontology should be utilized to better model the relationship among entities in the data resource. Thus, more detail information can be extracted and easily understood by everyone.

Romero and Abello in their paper stated that most of data warehouse has been traditionally developed using reengineering process which start from end user requirements and move to what data sources can provide. Here, they introduced a user-centered approach to support design tasks of multidimensional data warehouse that comply with the incoming requirements. They created ontology, a conceptual formalization of the domain, by fully analyzing the data sources to capture multidimensional knowledge. Later, this knowledge can be exploited to assist user requirement elicitation tasks [5]. Thenmozhi and Vivekanandan proposed a framework for designing a multidimensional schema data warehouse using ontologies. Their approach employed a hybrid method which conciliated user requirements and data sources at the early stage of design. They adopted ontology reasoning to automatically generate multidimensional components, e.g. facts and dimensions [6].

Insight of our research is quite similar with the notion of the all mentioned research works above, which is to present entities relationship in the data warehouse with a better visual that can be easily understood and used by users. Here, we visualize the relationship of entities about Malaria Incidence Spread in Indonesia.

2. Research Method

Ontology

Neches et al defined ontology as "Ontology defines the basic terms and relations comprising the vocabulary of an area as well as the rules for combining terms and relations to define extensions of the vocabulary" [7]. It gives a meaning and definition of an object and relation among objects in knowledge domain. Basically, ontology is a concept that

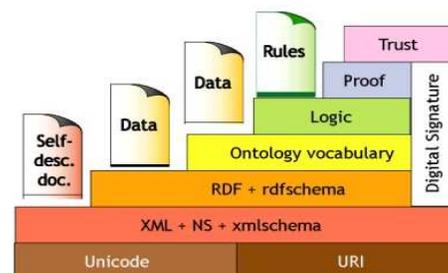


Fig. 1. Ontology Layers [9]

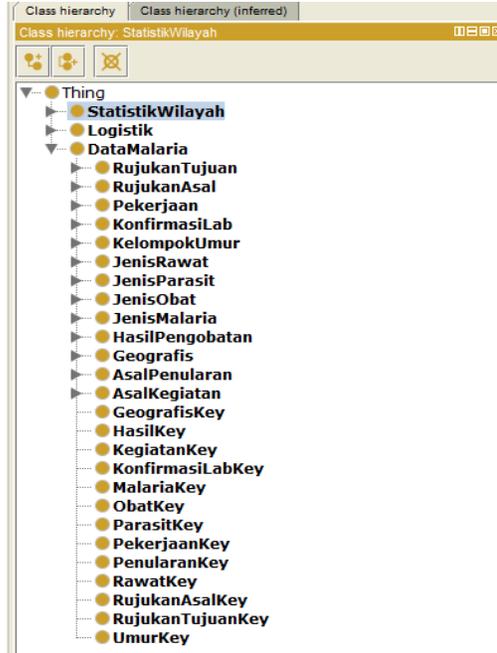


Fig. 3. Classes of Malaria's Incidences Data

Below are some components that usually employed to form the structure of ontology:

- XML (*Extensible Markup Language*) provides output syntax for a structured document, but has not been enforced for XML documents using semantic constrains.
- XML *Schema* is used to confine the structure of XML document.
- RDF (*Resource Description Framework*) gives a simple semantic of data model of objects or resources and their relationship. It can be expressed using XML syntax.
- RDF *Schema* is a vocabulary to describe the properties and classes of sources by using a semantic to spread the hierarchies.
- OWL (*Ontology Web Language*) adds some vocabularies to explain about properties and classes.
- Fig. 1 shows the structure of ontology layer. Each layer has its own function: XML layer is developed to store web page contents; RDF layer represents the semantic of web page contents, Ontology layer describes vocabularies of the domain, and Logic layer is used to access the specified data.

Protégé

Protégé is a free, open-source tool editor that has been widely used to build an ontology of domain. Its plugin architecture can be adapted to develop ontology-based application. Moreover, we can integrate the output of Protégé with rule system to build an intelligent system. Protégé support va-

rious format to save the data such as OWL, RDF, XML, and HTML. Protégé gives a conceptual basic of integrated knowledge and provides a visualization features to easily model the knowledge bases.

Many researchers have been actively utilizing Protégé to represent the ontology of data warehouse. Awad et al used Protégé to implement ontology of genetic neurological disease [12]. Here, they used OWL and SWRL language to implement ontologies. Prat et al also utilized Protégé to transform multidimensional models data warehouse into OWL-DL ontologies [13].

3. Result and Discussion

In this research, we are using Malaria data incidences in Tanah Bumbu regency, South Kalimantan, one of the most endemic areas in Indonesia. Currently, Public Health office in this region is still using Microsoft Excel files to record all the incidences data of Malaria within a year. At the end of each year, they must report all of the data to the Health Department Headquarters, in which different file for each month. By saving these data into different files, it is difficult for the staff to manage and understand the relationship among entities in the data. Hence, it is also hard to generate information and make conclusion about the data.

Fig. 2 shows the structure of the current Microsoft Excel file used to store Malaria's incidence data. This document records historical information about patient who diagnosed with Malaria. Furthermore, it also saves the details of data including the starting point of Malaria spread to the information about patient data recovery. In conclusion, these data consist of many variables that need a clearer representation so it will be easier to be understood.

Here, we are building a visualization of entities relationship in Malaria's incidences data. We are using Protégé, a popular and open-source platform to model, visualize, and build knowledge-based application with ontologies. Before working with Protégé, we need to identify entities and their relationship that involved in current Malaria's incidence data. At first, we identify three main classes of Malaria's incidence data. Fig. 3 shows the list of these classes: StatistikWilayah (information about the spread of Malaria's incidence), Logistik (information about Malaria's medicine logistic in each health office), and DataMalaria (records about Malaria's incidence). Next, we define subclasses for each class, such as DataMalaria's subclasses as shown in Fig. 3. We define the relationship between two classes, class and its subclasses, by using primary key of each subclass. These primary keys identify unique property for each subclass which later will be converted into dimension table in data warehouse. All of these classes and subclasses have disjoint property, in which they are independent

- [3] S. Khouri, I. Boukhari, L. Bellatreche, and E. Sardet, "Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool," no. 14.
- [4] S. Talebzadeh, M. A. Seyyedi, and A. Salajegheh, "Automated Creating a Data Warehouse from Unstructured Semantic Data," vol. 88, no. 10, pp. 19–25, 2014.
- [5] O. Romero and A. Abello, "A framework for multidimensional design of data warehouses from ontologies," *Data & Knowledge Engineering* Vol. 69, pp. 1138-1157, 2010
- [6] M. Thenmozhi and K. Vivekanandan, "An ontology based hybrid approach to derive multidimensional schema for data warehouse," *International Journal of Computer Applications* Vol. 54, No. 8, pp. 0975-8887, 2012
- [7] R. Neches, R. E. Fikes, T. Finin, T. R. Gruber, T. Senator, and W. R. Swartout, "Enabling technology for knowledge sharing," *AI Magazine*, 12(3): 36-56, 1991
- [8] B. Chandrasekaran and J.R. Josephson, "The Ontology of Tasks and Methods," *Symposium on Ontological Engineering, AAAI Spring Symposium Series*, Stanford, CA. 1997
- [9] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "Ontologies: What are ontologies, and why do we need them?," *IEEE Intelligent Systems and Their Applications*, Special Issue on Ontologies, 14(1): 20-26, 1999
- [10] A. G. Perez and V. R. Benjamins, "Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods," *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*, Stockholm, Sweden, 1999
- [11] York Sure and Rudi Studer. *Towards the Semantic Web: Ontology driven Knowledge Management*, 2003.
- [12] D. Awad, H. Tout, V. Courboulay, and A. Revel, "Ontology-based solution for data warehousing in genetic neurological disease," *Proceedings of the World Congress on Engineering*, Vol. 1, 2012
- [13] N. Prat, J. Akoka, and I. C. Wattiau, "Transforming multidimensional models into OWL-DL ontologies," *Proceedings of Research Challenges in Information Science (RCIS)*, pp. 1-12, 2012

Appendix 1

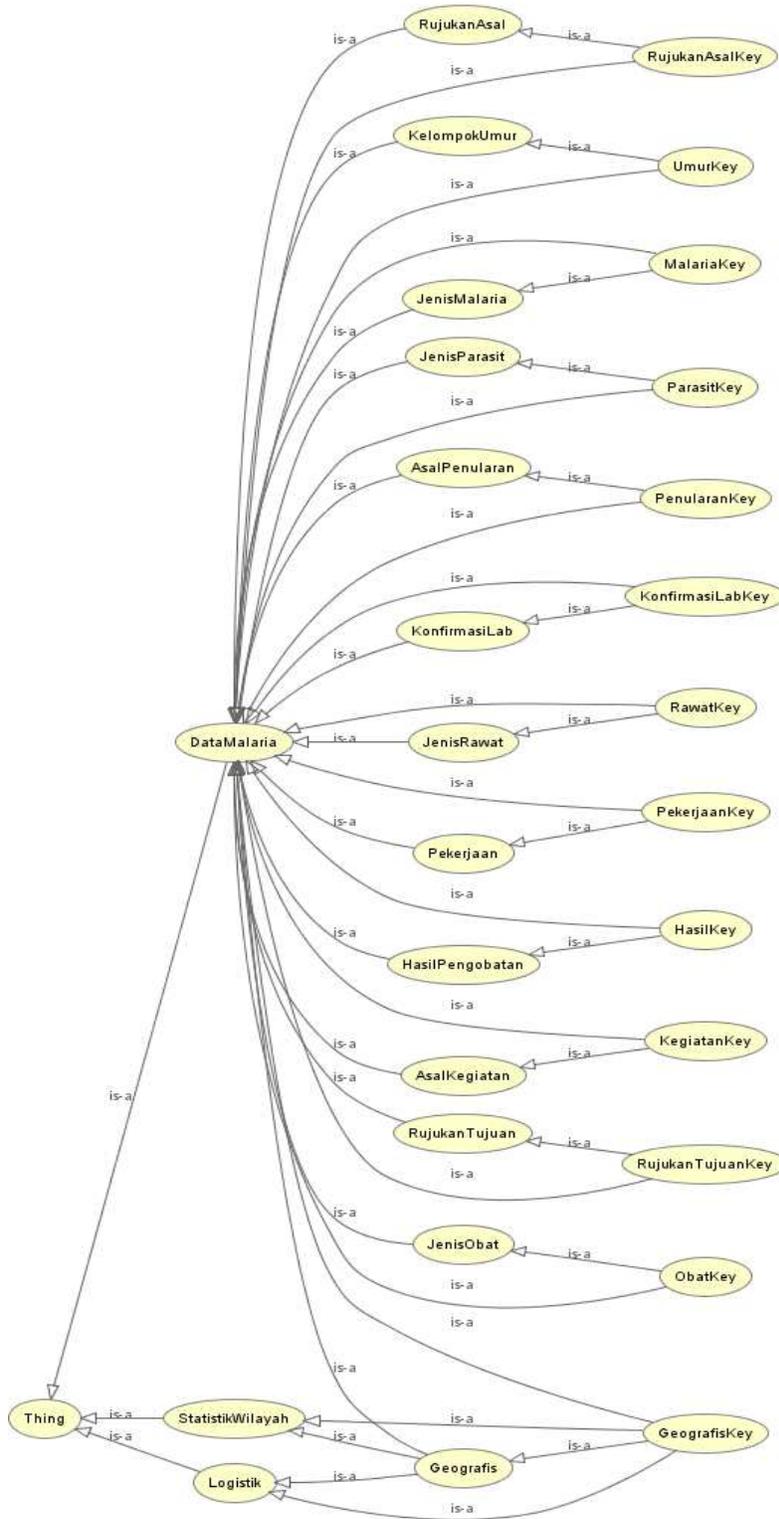


Fig. 1. OWL Visualization of Malaria Data