Towards an ontology model for Malay Manuscripts

M.N. Zahila[1], A. Noorhidawati[2], M.K. Yanti Idaya Aspura[2]
[1]IIUM Library
International Islamic University Malaysia
Jalan Gombak, 53100 Kuala Lumpur, MALAYSIA
[2]Department of Library and Information Science,
Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, MALAYSIA
e-mail: zahila@iium.edu.my; noorhidawati@um.edu.my; yanti@um.edu.my

ABSTRACT

This paper proposes an ontology-based conceptual framework for descriptive and content knowledge of Malay manuscripts in the University of Malaya Library. The purpose of this ontology framework is to provide a knowledge base that represent invaluable information contained in Malay manuscripts and enhance semantic search and retrieval. Currently manuscripts are kept in digital format. The digitized manuscripts in the image format can only be downloaded and read through the digital library without any mechanism to access knowledge contained in it. The current approach to make them accessible is using metadata. It is insufficient to describe the semantics of documents and to search knowledge by using metadata alone. The library also provides very little information regarding the semantics and subjective aspect of the manuscript. It is infeasible to describe content simply using words. Manuscripts contains more important information that are invaluable for retrieval purposes such as the transliteration work and annotation added by experts or users and how each manuscript link with others across subjects such as history, folklore, and legends. Additionally, the data of the historical documents are often heterogeneous, semantically rich and highly interlinked. Users may desire knowledge from a combination of facts found in multiple sources. Therefore, there is a demand for a powerful and efficient system which allow users to access and explore the content of the manuscript in depth and make all the information data searchable by direct queries. The main reference of the model is, Simple Event Model (SEM) and BIO ontologies. The scope of this study is concerned with a Malay manuscript information retrieval system which exploits a domain-specific ontology and open knowledge-based sources. The domain is restricted to the Malay manuscripts domain, and the ontology was built by referring to existing manuscripts ontologies. This study contributes to the development of semantic technology model for Malay manuscripts. The model supports organizing and integrating diverse knowledge contained in the manuscripts and could enable better understanding of the domain and provide more effective ways of discovering useful and unique knowledge contained in the manuscripts.

Keywords: ontology; Malay manuscript; digital library; cultural heritage

INTRODUCTION

Malay manuscripts are invaluable documentary records of the past which must be preserved because they are irreplaceable. These handwritten documents reflect the rich cultural heritage and documentation of high intellectual accomplishments of the Malays. The manuscripts cover a wide range of subjects, such as history, religion, law, culture, folklore, and legends. Some of these works were later republished in printed format when printing was introduced first in Java then Penang, Malacca, and Singapore in the 19th century.

Today, manuscripts are presented in digital format using metadata. Although the use of extended Dublin Core (DC) metadata scheme for manuscripts collection is useful to represent the semantic resources of the manuscript such as title, category, dimension, language material, and condition (Zainab, Abrizah, & Hilmi, 2009), manuscripts however contains many more important information that are invaluable for retrieval purposes such as the transliteration work, annotation added by experts or users and how each manuscript link with the others across subjects such as history, folklore and legends. Therefore, knowledge representation of the manuscript content could be enhanced using ontology-based schema taking advantage of semantic technology to improve content descriptions, knowledge organization and later retrieval of the digital collection.

There is a need to discover the knowledge contain in the manuscript. Ontology is said to be the best way in representing, organizing and sharing the knowledge. According to Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999), ontologies are "… content theories about the sorts of objects, properties of objects, and relations between objects that are possible in a specified domain of knowledge. They provide potential terms for describing our knowledge about the domain."(Chandrasekaran, Josephson, & Benjamins, 1999). According to Sowa (2000), ontology is a discipline that is part of the knowledge representation field. It is an important tool for the organization and contextualization of knowledge, particularly in well-bounded contexts, such as scientific research, or within individual organizations (Brewster & O'Hara, 2007). It also can provide an interface to the content, and the combination of a concept ontology and associated content can be used to generate a separate content representation (Boyce & Pahl, 2007). This ontology is needed to improve access to a mass of information and for development of better search, retrieval, and organization (Lin & Liang, 2005).

This paper presents the conceptual framework on event ontology-based approach to Malay manuscript content annotation and retrieval. An event is used to represent the content of the manuscript because the event contains information about people, place and time.

RELATED RESEARCH

There are a lot of studies that have been done in developing ontology models for various fields. The ontology-based retrieval can be classified into 4 which are: vector space model, probabilistic model, context-aware model, and semantic-based approach. The

semantic model can be further divided into semantic similarity, semantic association, and semantic annotation. Among these three models, semantic association and Semantic similarity method retrieve more relevant documents (Sakthi Murugan, Bala, & Aghila, 2013). The main objective of any semantic annotation activity should be to produce an annotation of the resources in the underlying digital collection that satisfies all the requirements of accuracy, completeness, and adequacy posed by the intended uses of the collection (Juan Cigarrán-Recuero, 2014). Semantic annotation has been used in retrieving Digitized Museum Artifacts by combining ontological concepts, visual and textual features automatically extracted from images and their textual descriptions. (Sharma & Siddiqui, 2016).

There are two types of knowledge representation which are descriptive knowledge and content knowledge. Descriptive knowledge is information about the characteristics of the document which do not involve its content (or its aboutness) (Diakite & Markhoff, 2015). It is also known as formal metadata or bibliographic metadata (Weller, 2010) or standard ontology (Noah et al., 2010).

Content knowledge is knowledge about the content of the document or what the document is talking about, also known as domain ontology (Noah et al., 2010). There are several major domain ontologies which are general concept ontologies, actor ontologies, place ontologies, time and period ontologies, event ontologies and domain nomenclatures or terminologies (Hyvönen, 2012).

According to Hyvonen (2012), event ontology is useful for indexing historical cultural heritage content. Events are the semantic glue that associates actors, objects, places, and time together. He found that the event-based approach can be used in annotations for interoperability, as a search object of their own and provide the end-user with insightful semantic recommendations with explanations. (Hyvönen, Alm, & Kuittinen, 2007).

Ramli, Azman, and Noah (2016) have developed an event ontology for the historical domain. They constructed the competency questions to verify whether sufficient information is available in order to achieve the goals and scope of ontology and reused the existing Simple News and Press ontologies (SNaP). They followed the 101 Method as their guide in developing the ontology and used METHONLOGY to perform the analysis in the conceptualization process.

In 2016, Zhitomirsky_Geffet and Prebor proposed an event-based ontological model for Hebrew Manuscripts. This model includes 4 main classes, which are: i) manuscript biographic events (e.g. creation, printing, acquisition, copying, storing, censoring, dedication etc.), ii) Historical manuscript, Manuscript agent, and Historical figure. The model was built based on the existing ontologies in the field of cultural heritage and facilitate the classes and properties. The model was analyzed by constructing several queries in SPAQL (Zhitomirsky-Geffet & Prebor, 2016).

OMOS, Ontology for Western Saharan Manuscripts was proposed by Diakite and Markoff (2015). This ontology has two different levels of knowledge which are descriptive knowledge and content knowledge. The descriptive knowledge is derived

from existing metadata in the library, whereas the content knowledge is about what the manuscript is talking about (Diakite & Markhoff, 2015).

Event information is important for cultural heritage because it is central to understanding heritage information (Doerr, 2009). Even entities can also represent the relationship between object and person (Kakali et al., 2007) and be regarded as a type of concept or type of relationship, for example event "teach" is always regarded as a relation between "teacher" and "student" (Liu, Liu, Fu, Hu, & Zhong, 2010).

PROPOSED FRAMEWORK

The proposed framework uses the event ontology domain to create a knowledge base for manuscript content. Figure 1 shows the proposed framework for Malay Manuscripts ontology. The framework consists of four modules; knowledge development, knowledge enhancement, knowledge evaluation, and knowledge enrichment. Below are details about each module.

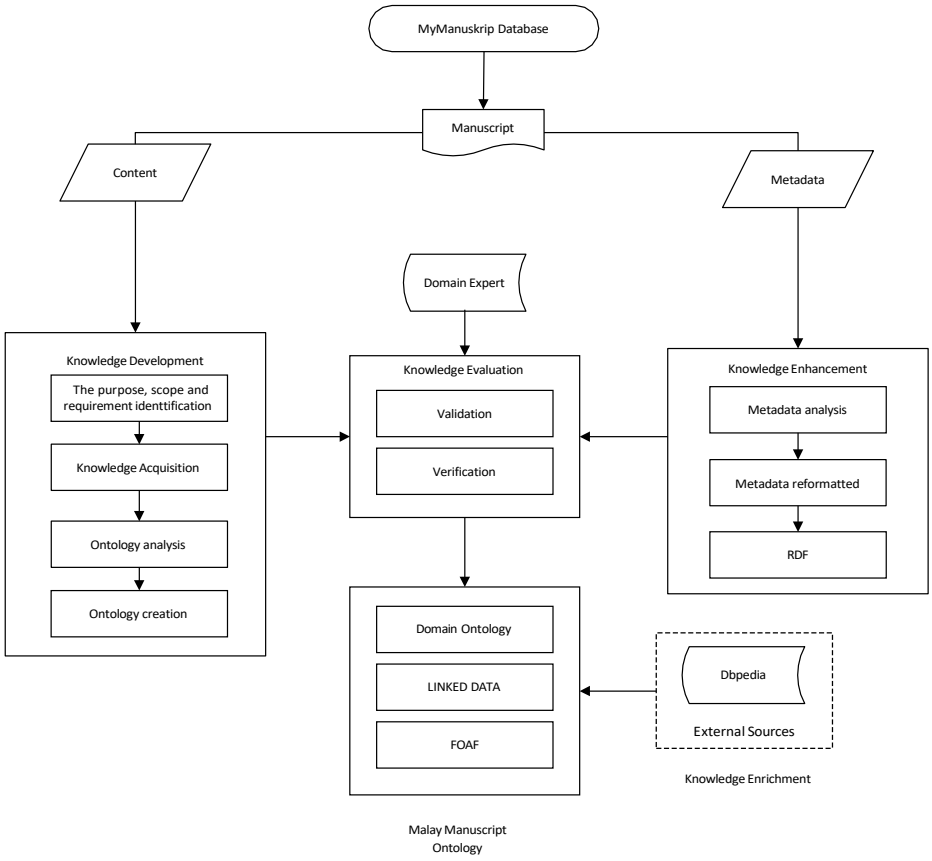

Figure 1: Proposed Knowledge Base Framework for Malay Manuscripts Ontology

Knowledge development. The knowledge development refers to building the domain knowledge for Malay manuscript content. Manuscript content covers a wide range of

subjects, such as history, religion, law, culture, folklore, and legends. There are various methods proposed by scholars in developing the domain ontology.

Knowledge enhancement. The knowledge enhancement is to develop descriptive knowledge for Malay manuscript. This is done by reformatting the existing Malay manuscript metadata into Resource Description Framework (RDF). This metadata comes from the MyManuscript Database records. The metadata will be reformatted into subject-predicate-object statements.

Knowledge Evaluation. The knowledge evaluation is to prove the correctness, consistency, completeness, and conciseness of the ontology. Knowledge evaluation is done by domain experts.

Knowledge Enrichment. The knowledge enrichment refers to integrating domain ontology with an external source which is Dbpedia.

ONTOLOGY CONSTRUCTION

The Malay manuscripts ontology will be built based on descriptive knowledge and content knowledge as shown in Figure 2. The descriptive knowledge is metadata of the manuscript and content knowledge is what the manuscript is talking about. The development of the Malay Manuscripts ontology is to design and implement an ontology representing knowledge embedded in the manuscript and to improve information discovery. This process requires a formal representation of concepts and their relationship.

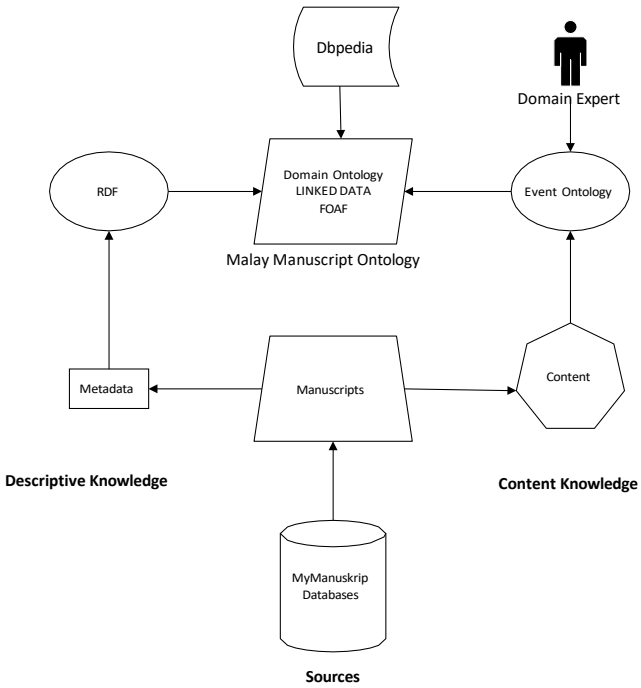

Figure 2: Conceptual architecture of Malay Manuscript Ontology Model

Content Knowledge

Content knowledge is knowledge coming directly from the manuscript's content. Content knowledge is considered an unstructured document. In this research, content knowledge will be represented using an event ontology. As mentioned above, an event ontology is important to understand the information of historical documents, because it can represent the relationship between object, person, time and place. According to Doerr (2009), information contained in cultural heritage is event centric, things, people and ideas connect and relate via events.

There are several models that have been proposed for representing event such as Event Ontology (Raimond, Yves, Abdallah, 2007), LODE (Shaw & Hardman, 2009), Simple Event Model (Van Hage, Malaisé, Segers, Hollink, & Schreiber, 2011), CIDOC CRM, BIO, and Event Model-F (Scherp, Franz, & Staab, 2010). In this study, we reused the existing SEM and BIO ontologies as our main reference.

During the experiments, we used a Malay manuscript which was already published in romance script Sulalatus Salatin. In this study, researcher manually wrote down all events mentioned in the text of the manuscript. For example, “mangkat”, “kelahiran”, “berperang”, “membuka negeri baru” etc. We also listed all persons’ name, time, objects and places which were related to the event as well as the causes of the event.

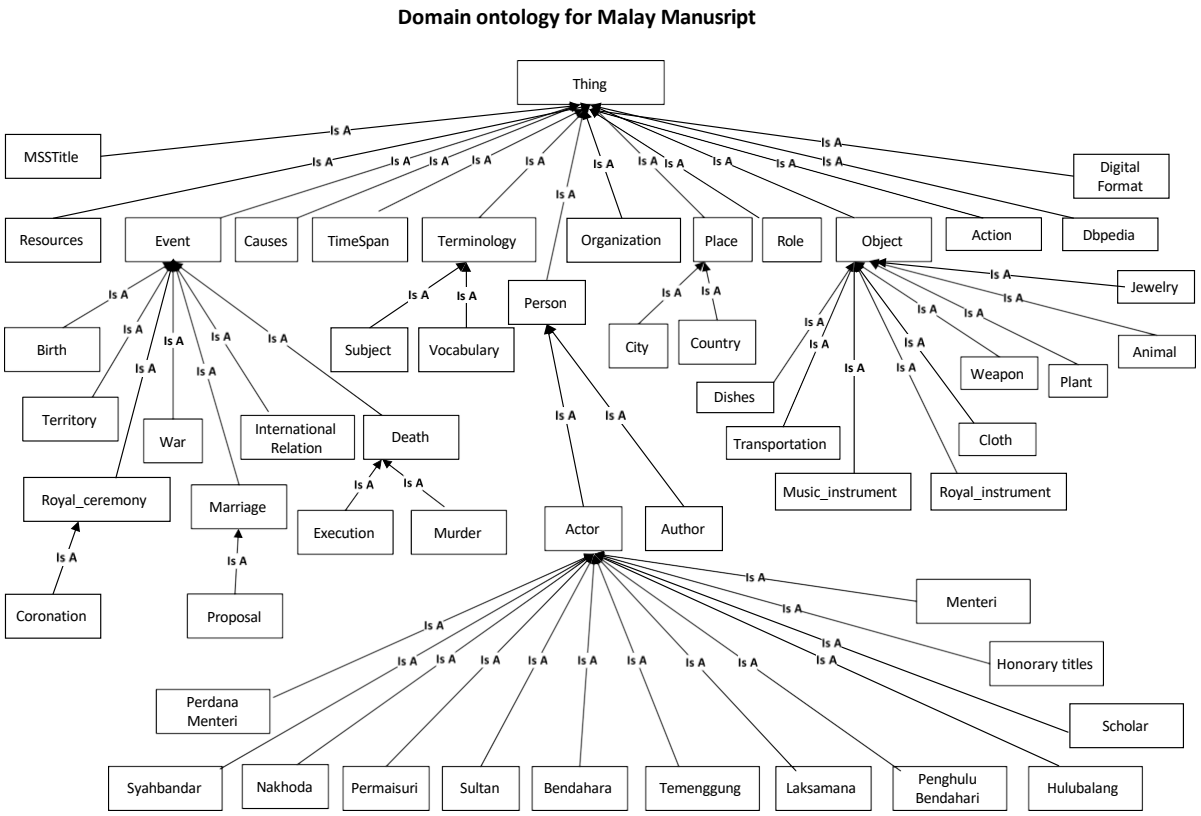

Figure 3: Event ontology for manuscript contents

After all events and related terms have been listed, we clustered and identified all these events into concepts. These concepts are identified based on basic classes in Simple Event Model ontology. Among the basic classes that are matched and appropriate for Malay manuscript ontology are event, actor, object, place, role, and time. Then we expanded the model by adding some classes which are causes and actions. The class causes are added to define factors or causes that make the event happen and action class is to define actions taken by a person in the event. In this research, an "is-a" relation is used to represent the class hierarchy. This means that class A is a subclass of B, and every instance of A is also an instance of B. This relationship indicates a generalization/specialization relationship between two concepts. Below are details for each class and subclasses.

1. Events are something that happens and to describe "what" is happening. This class has seven subclasses:
    i. Birth: The event of a person entering into life.
    ii. Territory: The event of opening new country by war or agreement
    iii. War: The event of an armed conflict between different countries or different groups within a country.
    iv. Royal ceremony. It has one subclass:

v. Coronation: The event of crowning
vi. International relations: The event of making international relations with other states.
vii. Marriage. It has one subclass:
viii. Proposal: The event where one person in a relationship asks for the other's hand in marriage
ix. Engagement: The event of engagement
x. Death. Death divided into two:
xi. Execution: The event of a person's life ending.
xii. Murder : The event of murder

2. Causes. Causes or factors that make events happen
3. Object. It has nine subclasses:
    1. Animals
    2. Plants
    3. Cloth
    4. Weapon
    5. Dishes
    6. Jewelry
    7. Music instrument
    8. Royal instrument
    9. Transportation
4. Person: It has two subclasses:
    1. Author: Author name for the manuscript
    2. Actor. Refers to actor involved or participate in the events. Actors are divided into thirteen subclasses based on their positions and honorary titles.
        i. Perdana Menteri
        ii. Syahbandar
        iii. Nakhoda
        iv. Permaisuri
        v. Sultan
        vi. Bendahara
        vii. Temenggung
        viii. Laksamana
        ix. Penghulu Bendahari
        x. Hulubalang
        xi. Scholars
        xii. Honorary titles
        xiii. Menteri
5. Terminology. Divided into two:
        i. Vocabularies: Vocabularies for old Malay language
        ii. Subject: Subject Heading for manuscript

6. Place. Place of events are divided into two subclasses:
    1. City
    2. Country
7. Time Span
8. Organization
9. Role:
10. Action
11. Dbpedia: A link of ontology to Dbpedia
12. Digital Format: Page of the source in digital format
13. Resources: Name of resources
14. MSS Title: Title of the manuscript

Each class will have its own instances. The relationship between instances is called properties. These properties represent the relationship and define the characteristics of the class. There are two types of properties; object properties and datatype properties. Object properties link between two individuals, whereas datatype properties link an individual to an XML Schema Datatype value or an RDF literal. They describe relationships between individual and data values. There is also annotation properties, which is used to add information to classes, individuals and object/datatype properties. For object properties, it may have a corresponding inverse property.

In the context of Malay manuscript ontology, properties are developed in order to provide detailed information for each class. We analyzed several relevant event ontology model object properties to learn whether and how they can be reused and matched with Malay manuscript ontology. These ontology model are BIO, CIDOC-CRM, SEM, LODE, and Event Ontology.

Descriptive knowledge

Descriptive knowledge is developed from the exploitation of the existing metadata in MyManuskrip database. The manuscript description is intended to facilitate the recording of the important physical features of a manuscript. The existing metadata is Dublin Core. In order to make the existing metadata to be presented as a linked data, a description of resources must be presented as a set of statements, with each statement giving a value which describes a specific aspect of the resources (Dunsire, 2012). The value is as objects, the aspects as its predicates or property and the resource itself as a subject, also known as triple form: subject-predicate-object. This set of statement expresses the semantic relationship (predicate) between two concepts (subject and object).

The description of a manuscript might include information about title, writer, subject, and information about the physical characteristics of the manuscript. The description of a digital manuscript usually stored as fields as shown in Table 1.

Table 1: Fields and values for a simple description of a manuscript

Field	Value
Record ID	MSS1
Title	Sulalatus Salatin
Subject	Melayu--Sejarah
Shelf Mark	MS93

Each record usually required a unique value to act as an identifier for the set of statements. The identifier for this record is MSS1. This information was then reformatted into a subject-predicate-object statement as shown in Table 2.

Table 2: Description formatted as a set of statement

Record ID	Attribute	Value
MSS1	hasTitle	Sulalatus Salatin
MSS1	hasSubject	Melayu--Sejarah
MSS1	Shelf Mark	MS93

In this research, DM2E model is analyzed to learn whether and how their properties can be reused and matched with Malay manuscript ontology. The DM2E model is a specialization of the EDM for the domain manuscript.

Table 3 below shows a list of properties for descriptive knowledge which is used to convert and extend the data catalog records representing the manuscripts into linked data and integrate them with domain ontology.

Table 3: Properties used for descriptive knowledge

Properties	Definition	Property type
hasTitle	Title of manuscript	Object property
alternativeTitle	Alternative Title of manuscript	Object property
Identifier	Unique number for the manuscript	Data property
Creator	Creator of the manuscript i.e. Author	Object property
Copyist	Name of the copyist	Object property
Date	Date of the manuscript	Data property
Descriptions	A summary of the content and topics of the collection	Data property
hasPart	Any other document contained within the current manuscript i.e. pages of manuscript in digital format	Object property
Language	Language of the manuscript, Malay, Java	Data property
hasSubject	Subject of the manuscript based on LCSH	Object property
shelfMark	Shelf mark number	Data property
callNumber	Call number of manuscript LCC	Data property
Incipit	First 3 lines text of manuscript	Data property
explicit	Last 3 lines text of manuscript	Data property
Script type	Type of script if manuscript written in Jawi, e.g Naskh,	Data property
inkColor	Color of ink written	Data property
noOfLine	Average line of text per page	Data property
pageNumber	Number of pages	Data property
Watermark	Description of watermark if any	Data property
manuscriptDimension	Dimension of manuscript	Data property
writtenAreaSize	Dimension of written area	Data property
pageDimension	Dimension of page	Data property
writtenAt	Place where the manuscript was written	Object property
publishedAt	Place where the manuscript was published	Object property
currentLocation	Holding institution	Object property

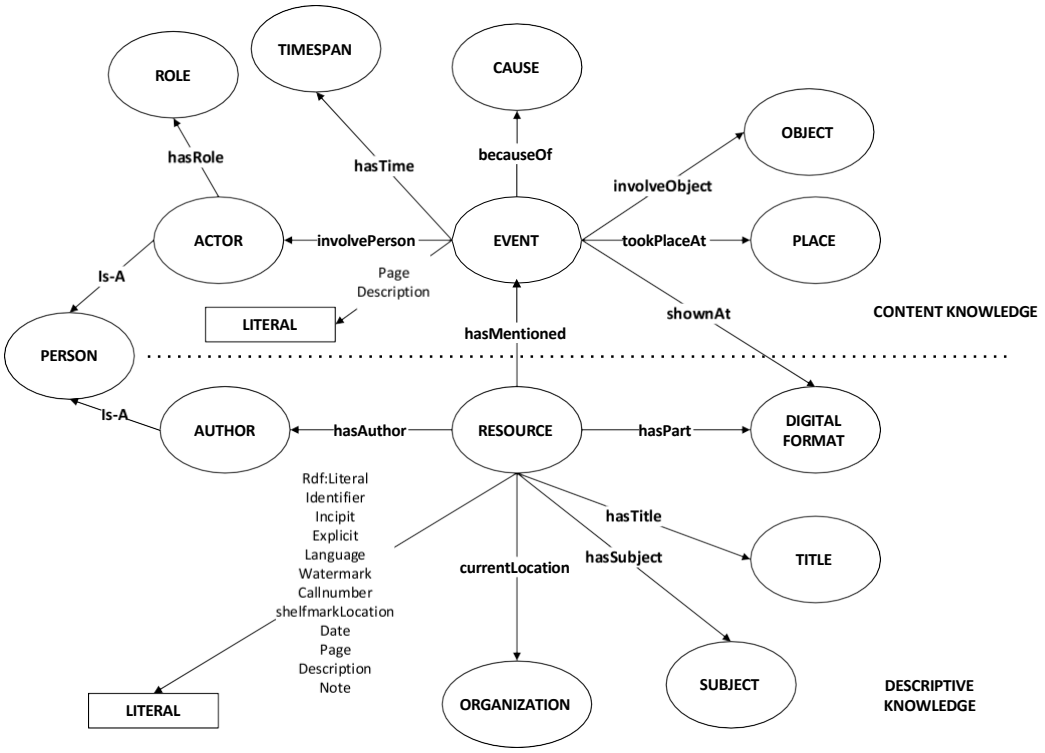

Figure 4: Linkage between Content and Descriptive Knowledge

Figure 4 shows the linkage between content and descriptive knowledge. Descriptive and content knowledge are integrated into building the Malay manuscript ontology.

Input from domain expert is essential to ensure the correctness of the ontology, consistency, completeness, and conciseness. Finally, the Malay manuscript ontology will be integrated with DBpedia to enhance and enrich the domain ontology. There are two activities in this process: first, we will collect and extract automatically external knowledge from Dbpedia such as person, places, event, and object. Second, the external knowledge will be integrated with Malay manuscript ontology by using owl: sameAs. For example, the person with the name Tun Perak is an instance of an actor, and he participated in one event in the resources. There is limited information about Tun Perak in this resource, so in order to give more details about him, we enrich by adding owl: sameAs: http://dbpedia.org/page/Tun_Perak. As a result, the information about Tun Perak will be enhanced and enriched.

CONCLUSION

In this paper, the conceptual framework for Malay manuscripts ontology was proposed. The proposed framework uses an Event ontology to represent the content of Malay manuscripts and uses Dublin Core metadata to represent the descriptive knowledge. This framework will contribute to the creation of knowledge-based information that contains data on Malays history, culture, and civilization and interlinking different individual content in Malay manuscripts using RDF. Development of a prototype model

of an ontology-based system with a high level of semantic granularity which is an ontology that reflects the various cultural riches and intellectual aspect stored in Malay manuscripts. This will enable systematic research of the knowledge embedded in the manuscripts and make it widely and easily accessible by everyone.

Acknowledgement

This research has been funded by Fundamental Research Grant Scheme (FRGS) FRGS/1/2018/ICT04/UM/02/8.

REFERENCES

Boyce, S., & Pahl, C. 2007. Developing domain ontologies for course content. Educational Technology and Society, 10(3), 275–288.

Brewster, C., & O'Hara, K. 2007. Knowledge representation with ontologies: Present challenges-Future possibilities. International Journal of Human Computer Studies, 65(7), 563–568.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. 1999. What are ontologies, and why do we need them? IEEE Intelligent Systems and Their Applications, 14(1), 20–26.

Diakite, M. L., & Markhoff, B. B. 2015. OMOS: Ontology for Western Saharan Manuscripts, 7606.

Doerr, M. 2009. Handbook on Ontologies, 463–486. http://doi.org/10.1007/978-3-540-92673-3

Doerr, M. 2009. Ontologies for cultural heritage. Handbook on Ontologies.

Gruninger, M., & Fox, M. S. 1994. The Role of Competency Questions in Enterprise Engineering. IFIP WG5 - 7 Workshop on Benchmarking - Theory, and Practice, 1–17.

Hyvönen, E. 2012. Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology, 2(1), 1–159.

Hyvönen, E., Alm, O., & Kuittinen, H. 2007. Using an ontology of historical events in semantic portals for cultural heritage. Proceedings of the Cultural Heritage on.

Kakali, C., Lourdi, I., Stasinopoulou, T., Bountouri, L., Papatheodorou, C., Doerr, M., & Gergatsoulis, M. 2007. Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies. International Conference on Dublin Core and Metadata Applications, (Epeaek Ii), 128–139.

Leung, N. K., Lau, S. K., Fan, J., & Tsang, N. 2011. An integration-oriented ontology development methodology to reuse existing ontologies in an ontology development process. Proceedings of the 13th International Conference on Information Integration and Web-Based Applications and Services, 174–181.

Lin, H.-F., & Liang, J.-M. 2005. Event-based Ontology design for retrieving digital archives on human religious self-help consulting. 2005 IEEE International Conference on E-Technology, E-Commerce, and E-Service, EEE-05, (1), 522–527.

Liu, W., Liu, Z., Fu, J., Hu, R., & Zhong, Z. 2010. Extending OWL for modeling event-oriented ontology. CISIS 2010 - The 4th International Conference on Complex, Intelligent, and Software Intensive Systems, 581–586.

López, M. F., Gómez-Pérez, A., Sierra, J. P., & Sierra, A. P. 1999. Building a chemical

ontology using methontology and the ontology design environment. IEEE Intelligent Systems and Their Applications, 14(1), 37–46.

Noah, S. A., Alias, N. A. R., Osman, N. A., Abdullah, Z., Omar, N., Yahya, Y., & Yusof, M. M. 2010. Ontology-driven semantic digital library. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6458 LNCS, 141–150.

Noy, N. F., & McGuinness, D. L. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory, 25.

Ramli, F., Azman, S., & Noah, M. 2016. Building an Event Ontology for Historical Domain to Support Semantic Document Retrieval, 6(6), 1154–1160.

Sakthi Murugan, R., Bala, P. S., & Aghila, G. 2013. Ontology-Based Information Retrieval-An Analysis. International Journal of Advanced Research in Computer Science and Software Engineering, 3(10).

Sowa, J. F. 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations, (February), 594.

Weller, K. 2010. Knowledge Representation in the Social Semantic Web.

Zainab, A. N., Abrizah, A., & Hilmi, M. R. 2009. What a digital library of Malay manuscripts should support: An exploratory needs analysis. Libri, 59(4), 275–289.

Zhitomirsky-Geffet, M., & Prebor, G. 2016. Toward an Ontopedia for Historical Hebrew Manuscripts. Frontiers in Digital Humanities, 3, 3.