Comparative Studies of Ontologies on Sarawak Gazette

Fatihah Ramli, Bali Ranaivo-Malançon, Stephanie Chua and Mira Shumiza Mohammad
Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia.
rfatihah@unimas.my

***Abstract*—This paper presents a discussion on experience and process during initial stage of ontology building in history. The objective of this paper is to create a manual semantic annotation process to determine the concepts that will be used in the historical news ontology. It will describe the tasks of facilitating the analysis of missing concepts existing in Sarawak Gazette (SAGA) documents. Semantically annotating SAGA documents enable to enrich the element of concepts and relations taken from existing ontologies. Furthermore, an initial result is provided to observe the performance gain due to domain-specific annotations. Finally, we conclude on the importance of semantic annotations process in the construction of an ontology.**

***Index Terms*—Semantic Annotation; Ontology; SAGA Document.**

I. Introduction

Ontology is a formal and explicit specification of a shared conceptualization. An ontology consists of a set of concepts, set of relations, set of rules, and instances of concepts. In the past few years, there has been increasing concern on ontology for its ability to explain data semantics in the usual manner independent of the data source characteristics, providing a schema that allows interchanging data between heterogeneous information systems and users. The ontology development in some areas is not expected due to a large amount of information, particularly in history, making it semantically impossible. One example of historical documents is the Sarawak Gazette (SAGA) historical newspaper. SAGA documents are considered as one of the important repositories of Sarawak history, containing government and politics news, people and their way of life, landscape, flora and fauna [1]. It consists of ten volumes of documents. In this initial stage, we considered only noun entities in the annotation process. Due to a large amount of information in SAGA documents, we have to embed the semantic process for enabling each text to be linked to a specific meaning. Semantic annotation is an approach to link ontologies to the original information sources [2]. Annotation is the extra information associated with a particular point in a document or other piece of information. For semantic annotation, the extra information is the meaning definitions of the concepts used in a document.

In the initial stage, to build the ontology, some assumptions were raised. One of the assumptions was how do we determine the concepts that will be used in the ontology or why did we want to have this particular concept to be included in the ontology. In our case, ontologies in General Architecture for Text Engineering (GATE) support semantic annotations. Therefore, the semantic annotation approach is seen to potentially solve this assumption. This paper will discuss this assumption in detail in the next section.

In this work, an existing ontology is used to do the semantic annotation. The materials and methods are described in Section II on existing historical and news ontologies as well as the semantic annotation process. In Section III, the result on semantic annotation will be discussed and finally, the paper is concluded in Section IV.

II. Materials And Methods

The objective of this section is to discuss the process of creating a semantic annotation for building historical news ontology. This section will detail two steps in creating semantic annotations. They are: 1) select ontology and 2) design and implement semantic annotation process [3].

A. Step 1: Select Ontology

In the first step, we considered reusing existing ontology developed by others for semantic annotation. Available resources had to be checked whether they could improve and expand our particular domain and task. For our work, ontology reuse was very helpful as there was a time constraint in developing a new ontology from scratch especially in adapting and updating the necessary concept in a new ontology. There were several existing historical and news ontologies that were reviewed as follow:

1) STOLE Ontology

STOLE is a reference ontology which provides a vocabulary of terms and relations to clearly model the domain specific. STOLE ontology used the history of Italian Public Administration as domain specific. The main aim of the STOLE Ontology is to have a clear design model on historical concepts and seek views on specific areas. STOLE aims to gather information about the most relevant journals on history of public administration legislation in Italy that published between 1848 and 1946. The STOLE ontology's construction consists of three main phases: 1) Identification of key concepts, 2) Identification of the proper language and Tbox implementation, 3) Ontology population [4]. In the first phase, the key concepts involved in specific domain must be defined by the domain expert. The domain experts provide manual semantic annotations that would be added to the ontology by means of JAVA program. Next, they classified all the data that are related to historical documents and the results of all the concepts would be viewed in the form of a taxonomy that consist of three elements as shown in Table 1. Table 2 shows the size of the STOLE ontology that was computed by PROTEGE. Finally, ontology populations are

carried out to automatically fill in missing entities in Abox with semantic annotations. STOLE ontology is accessible to public and can be considered as an expandable ontology.

Table 1
Taxonomy of STOLE ontology

Elements	Examples
Data on the author of the article	Name, surname, biography
Data on the journal and the article	Article title, journal name, date and topic raised in the article.
Data on the relevant facts and persons cited in the article.	Persons, historical events, institutions

Table 2
Tbox statistics about STOLE ontology

Classes	14
Axioms	440
Object Properties	30
Data properties	29

2) Event Ontology

Hyvonen et al. [5] stated that a semantic portal for cultural heritage required event ontology because of three reasons: 1) events need ontological identifiers (URIs) to build a metadata collection, 2) events are important in creating a semantic relationship between cultural content and 3) Historical events are important to shape the backbone of chronological history. Hyvonen et al. [5] developed event ontology using Finnish history as a domain specific. The historical event ontology was based on the timeline that was created by Agricola network and being utilized as part of the semantic portal "CultureSampo—Finnish Culture on the Semantic Web", a cross-domain follow-up system of Museum Finland. The classifications of events were based on temporal timeline and other dimension such as event types i.e. war, coronation or branch history i.e. political history, history of science. They annotated manually 220 events between the years 1850–1920 utilizing the SAHA annotation tool combined with ONKI Ontology library servers for utilizing shared domain ontologies. As a result, history ontology defines URIs for events can be utilized for annotating other cultural objects and relating them with each other. However, the event ontology is not accessible to public and cannot be considered as an expandable ontology.

3) The FDR Historical Ontology

The main goal of FDR/Pearl Harbor project was developing applications that could help to improve searching and retrieving information from a set of documents taken from the Franklin D. Roosevelt Presidential Library (FDRL). This project used a set of documents that referred to situations and events over the ten-year period which was before the bombing of Pearl Harbor. The FDR/Pearl Harbor Project built the historical ontology based on the model presented using the entities and events in its document collection [6]. The FDR temporal ontology included only clearly defined endowment entities in the collection of documents, which comprised the following general categories: geopolitical entities, geopolitical organizations, military organizations, military vehicles, geographical objects, geographical artifacts, documents, agreements, persons and political organizations. Event and entity annotation of these documents used General Architecture for Text Engineering (GATE) to complete the manual semantic annotation. Next, automatic annotations are carried out using machine learning based on hand validated annotation. However, the FDR temporal ontology is not accessible to public and cannot be considered as an expandable ontology.

4) RDF/OWL Ontology on Henry III Fine Rolls

The Henry III is a collaborative project between King's College London and the National Archives (UK). The main aim of this project was to represent the complexity of historical documents known as the Fine Rolls [7]. FRH3 ontology consists of several classes such as authority (Person, Place, and Subject) and Factoid (Role, Relationship and Role_Relationship). The RDF/OWL had been chosen to do authority list based on several reasons: 1) It is a W3C standard for the Semantic Web; 2) The number of existing tools is greater for the RDF/OWL; 3) It can be expressed as XML, simplifying the process of data delivery and this makes it easy to index people, places and subjects using XSLT; 4) It can create the expression of relationship among the instances explained in the fine rolls source materials [7]. However, this ontology is not accessible to public and cannot be considered as an expandable ontology.

5) Ontology Driven Access to Museum Information

Ontology driven access to museum information can be represented as "core ontology" that combines basic entities and relationship across the various metadata vocabularies [8]. The core ontology is useful in helping to integrate information from multiple vocabularies and uniform processes across multiple sources of information. Core ontology is the basic core formal model for tools that integrate source data and perform a variety of functions [8]. There are several classes in this ontology such as E2 Temporal Entity, E52 Time-span, E3 Condition State, E4 Period and E5 Event. The ontology process was also helping in enriching knowledge. Hence, higher levels of complexity are acceptable and the design should be more motivated by logical correctness and completeness than human understanding. However, this core ontology is not accessible to public and cannot be considered as an expandable ontology.

Table 3
Summary of the features of existing historical and news ontologies

Ontology	STOLE	Event Ontology	FDR Historical Ontology	RDF / OWL Ontology	Ontology Driven Access to Museum Information	SNAP
Number of concept	14	None	10	8	5	22
Tool for annotation	None	SAHA	GATE	None	None	None
Availability	Yes	No	No	No	No	Yes

6) Simple News and Press Ontologies (SNAP)

SNaP ontology is a news ontology that consists of multiple ontologies, which describe assets (text, images, video) and the events as well as entities (people, places, organisations, abstract concepts, etc.). There are two categories of entities in SNaP ontology: simple entities i.e. stuff and complex entities i.e. event. The term stuff can be represented as abstract and intangible concepts as well as tangible things. The total

numbers of concepts that are involved in event and stuff ontologies are 22 concepts. While it is intended for news documents, it is found to be appropriate in our case as it contains detailed representation of events, people, organizations, locations, tangible and intangible things as well as documents [9]. SNaP ontology is accessible to public and can be considered as an expandable ontology.

7) Table of Comparison

In conclusion, based on the above studies we have identified several important features for selecting an appropriate ontology to be expanded. The most important feature is availability whereby existing ontology must be accessible for reuse and subsequently developed based on domain specific. For instance, only STOLE and SNaP ontologies are accessible to public. In addition, we also need to know the size and content of an ontology to facilitate the development of ontology. For example, SNAP and STOLE ontologies have the most number of concepts compared to other ontologies. With this, both ontologies have the potential to be reused for this study. Finally, we also study if there are appropriate tools to use in implementing the annotation process. For example, most studies do not clarify appropriate annotation tool except FDR historical ontology and event ontology. Therefore, based on the study of all these features, STOLE and SNaP ontologies have great potential to be reused in step 2. A summary of the features of existing historical and news ontologies is shown in Table 3.

B. Step 2: Design and Implement Semantic Annotation Process

In this step, we started downloading STOLE and SNaP ontologies from available resources. Then, we imported STOLE and SNaP ontologies into GATE and started running it together with A Nearly-New Information Extraction system (ANNIE) using SAGA documents. General Architecture for Text Engineering (GATE) is use for language processing task including semantic annotation. Semantic annotation is an important process to represent semantic relations in the SAGA documents. In this work, we used an ontology editor tool in GATE to provide basic viewing of ontologies which allows the linking to texts via semantic annotation as well as some editing functionalities of new concepts, instances and properties. All the basic concepts for both ontologies in GATE can then be viewed in Figure 1 and Figure 2. We have listed only the important concepts for both ontologies in Table 4 because most of the concepts are repetitive. Semantic annotation can be created manually or automatically. In our case, we created semantic annotation manually using Ontology Annotation Tool (OAT). Figure 1 and Figure 2 show the example of semantic annotation process for SNaP and STOLE ontologies.

Logically, the semantic annotation process in Figure 1 and Figure 2 shows the annotation of texts or entities related to ontology concept is a process that is carried out directly where if the certain entity is found to match the concept of ontology, a new triplet will be added to the database during the ontology development stage.

Table 4
Concept in STOLE and SNaP ontologies

Ontology	STOLE	SNaP
Concepts	Subject Place Event BeginPublication Death Birth EndPublication	Agent Organization Instant Stuff Event Intangile Person Location Tag Image Identifiable Asset

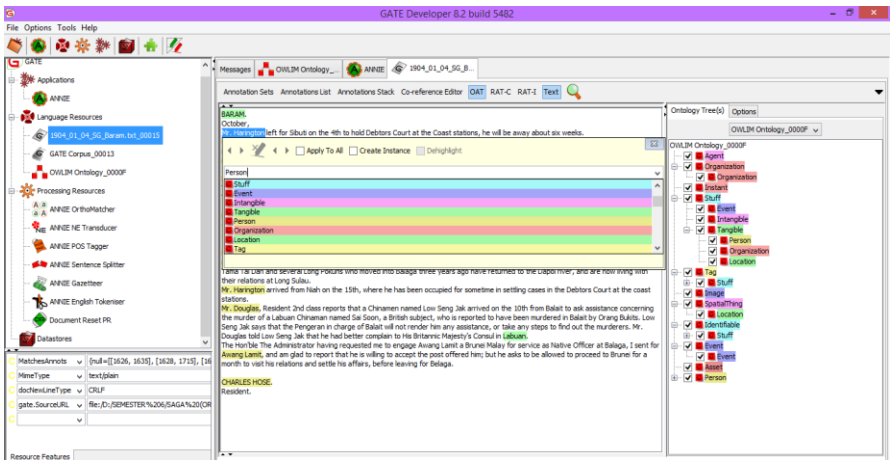

Figure 1: Example of Semantic Annotation Process for SNaP ontologies

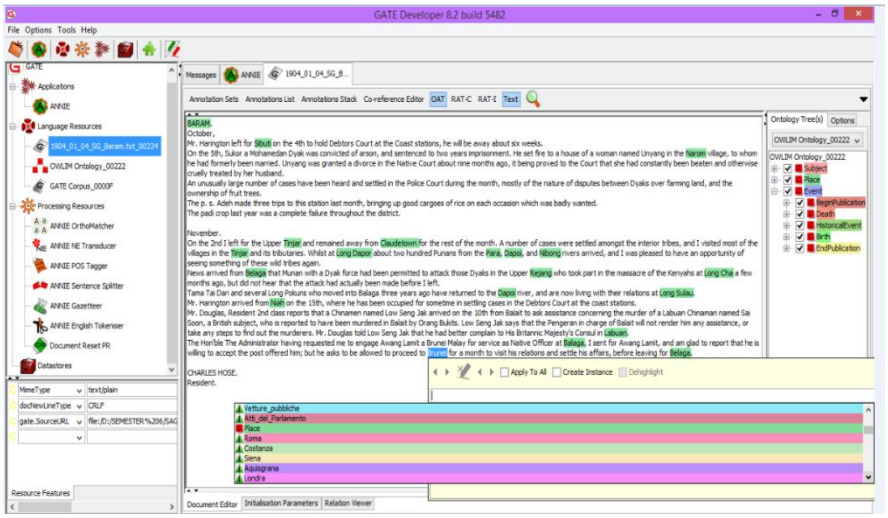

Figure 2: Example of Semantic Annotation Process for STOLE ontologies

III. Results And Discussion

Based on Table 4, we found that the STOLE ontology has eight main concepts and only two (i.e. Place and Event) of the concepts are noun concepts, while SNaP ontology has twelve main concepts and four (i.e. Organization, Event, Person and Location) of them are noun entities. Therefore, SNaP ontology contributes more existing noun concept to be matched with noun entities compared to STOLE ontology. Next, we checked the semantic annotation result manually for both ontologies and listed down the entire missing entities as well as proposed concepts in Fig 3. The total number of noun entities that matched the noun concept in SNaP ontology is 1014, while 266 of the person concepts were mismatched and we could not match with the existing concepts in the ontology i.e. missing concept. For example, based on Figure 3, we have listed case 1 – case 4 as below to discuss on missing concept. Meanwhile, for STOLE ontology, 321 noun concepts could match to the noun concept in ontology directly. For example, Singapore is a place. Therefore, we matched Singapore concept with the existing place concept in the STOLE ontology. As a conclusion, there were no missing concepts in STOLE ontology, and no concept will be added into it.

There are some cases that we have identified, which was mismatched with the existing concepts in SNaP ontology as follow:

CASE 1: For example, P. C. Ram Singh can be categorised as person name, but we do not know the meaning of words P and C. Therefore, we could not put this entity as person concept. Figure 3 shows the proposed concept that we suggested for this case.

	SNaP ontology		STOLE ontoloy	
No.	Manually annotate	Proposed concept	Manually annotate	Proposed concept
1	P. C. Ram Singh	Unknown, Person	Singapore	Place
2	Mrs. Deshon	Salutation, Person	Bau	Place
3	Major Selwyn-Payne	Ranks, Person	Kuching	Place
4	Dr. Woolrabe	Salutation, Person	Trusan	Place
5	Pangeran Miah	Title, Person	Blacksmith's and Datu's Road	Place
6	Mr. Gillan	Salutation, Person	Matang Road	Place

Figure 3: Example of some missing concepts with the proposed concepts

CASE 2: We had identified another entity such as Mrs Deshon is supposed to be in the person concept but the word "Mrs" could not categorise as a person. Thus, we have to create a new concept for the "Mrs" word, which is a salutation concept (refer Figure 3).

CASE 3: Pangeran Miah can be categorised as a person. But only "Miah" can put as a person concept, while "Pangeran" is only a title for the prince. Therefore, we have to create a new concept for "Pangeran", which is a title concept (refer Figure 3).

CASE 4: Major Selwyn-Payne is a person. But only "Selwyn-Payne" can put as a person concept, while "Major" is referring to the military rank of commissioned officer. Thus, we have to create a new concept for "Major", which is a ranks concept (refer Figure 3).

Due to the missing concepts issues, we propose to reuse the SNaP Ontology for building an historical news ontology for SAGA documents. All the missing concepts in SNaP ontology will go through the semantic annotation process automatically to create new concepts in the new ontology. The implementation of automatic semantic annotation process will use Ontogazetteer and rules. The list in Figure 3 will be used as a guideline for creating new concepts in the new ontology. We will expand this semantic annotation process to determine the date, verbs and terms in the next stage to support the conceptualization process of the new ontology.

In conclusion, this paper demonstrated how we achieved semantic annotation process manually on nouns. We claimed that the use of semantic annotation can connect the text mention to knowledge about the concept that was mentioned. Therefore we recommend semantic annotation as our solution to access additional data about the ontology.

IV. CONCLUSION

In this paper, the main contribution that we showed was how to use semantics to link annotation to the concept in ontology. This approach was used to determine the missing concepts in the ontology before we build the historical news ontology for SAGA documents. The contribution of this paper is the manual semantic annotation process of historical news ontology which was improved and expanded from SNAP ontology by using GATE tool.

ACKNOWLEDGEMENTS

This research was supported by the Malaysia Ministry of Education Grant F08/SpSTG/1363/16/5 awarded to the Faculty of Computer Science and Information System at the Universiti Malaysia Sarawak.

REFERENCES

[1] M. O. Rosita, R. Fatihah, K. M. Nazri, A. W. Yeo, and D. Y. Tan, "Cultural Heritage Knowledge Discovery: An Exploratory Study of the Sarawak Gazette," *In and Knowledge Engineering Conference (STAKE 2010)*, p. 20, Jul. 2010.

[2] Y. Lin, *Semantic annotation for process models: Facilitating process knowledge management via semantic interoperability*. Fakultet for informasjonsteknologi, matematikk og elektroteknikk, 2008.

[3] Y. Liao, M. Lezoche, H. Panetto, and N. Boudjlida, "Semantic annotation model definition for system interoperability," *On the Move to Meaningful Internet Systems: OTM 2011 Workshops*, pp. 61-70, 2011. Springer Berlin/Heidelberg.

[4] G. Adorni, M. Maratea, L. Pandolfo, and L. Pulina, "An Ontology for Historical Research Documents," *In International Conference on Web Reasoning and Rule Systems*, pp. 11-18, Aug. 2015.

[5] E. Hyvonen, O. Alm, and H. Kuittinen, "Using an Ontology of Historical events in semantic portals for cultural heritage," *In Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, vol. 12, Nov. 2007.

[6] N. Ide, and D. Woolner, "Historical Ontollogies," In K. Ahmad, C. Brewster, & M. Stevenson, *In Words and Intelligence II*, pp. 137-152, 2007. Springer Netherlands.

[7] J. M. Vieira and A. Ciula, " Implementing an RDF/OWL Ontology on Henry III Fine Rolls," *In OWLED*, 2007.

[8] D. O. Signore, "Ontology Driven Access to Museum Information," *In Annual Conference of CIDOC Documentation and Users CIDOC*, May 2005.

[9] F. Ramli and S. M. Noah, "Building an event ontology for historical domain to support semantic docment retrieval," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1154-1160, 2016.