DEVELOPING SEMANTIC ONTOLOGY FOR PRACTICAL DIGITAL BALINESE DICTIONARY

Cokorda Pramartha[1), IW Arka [2), Kevin Kuan[3), and IDMBA Darmawan [4)

[1,4) Net-Centric Computing Lab, Udayana University, Bali 80361, Indonesia
cokorda@unud.ac.id
[2) Australian National University, Canberra 2601, Australia
[3) The University of Sydney, Sydney 2006, Australia

ABSTRACT

There has been an increase in the number of young millennial Balinese who are bilinguals showing a shift with high competency in Indonesian at the concomitant expense of their Balinese language competency and related indigenous knowledge. The project takes advantage of the momentum with the surge of the enthusiasm and needs for online learning in the current context of COVID-19. The project focuses on digital dictionary research and development, addressing a clear gap in this space of Balinese digital humanity to have a learning resource in the broader literacy context of culture and language preservation and education. The digital dictionary online design is expected to benefit Balinese learners and other parties, including academic communities in Bali and across the globe. The language digital dictionary is a cross-lingual system to facilitate the translation of a source language to a target language based on keywords to query the translation. The use of keyword-based query for translation presents an ambiguity, which the use of ontology can help to minimise. The design, development, and evaluation of Balinese language dictionary ontology in this study was done in consultation with Balinese language experts.

Keywords: *Balinese Language, Digital Dictionary, Semantic Ontology, Balinese Dictionary, Digital Heritage*

1. INTRODUCTION

Bali is one of the thousands of islands in Indonesia, and is recognized internationally for its deep, rich, and diverse cultural heritage. According to the latest study by the Indonesian Education and Culture Ministry (Kemendikbud), around 250 million people live in Indonesia, with 718 indigenous languages spoken among 600 different ethnic groups (Hutapea, 2020). However, a study by the Summer Institute of Linguistics (SIL) indicated that 13 indigenous languages have been forgotten and lost due to the community no longer using them for daily communication (Widiyanto, 2018). Researchers believe that the extinction of an indigenous language means the loss of philosophical values, rules, beliefs, and historical and cultural knowledge related to that language (Dixon, 1997).

The Balinese language, spoken mainly on the islands of Bali and Lombok, is one of 718 living indigenous languages spoken in Indonesia. People in Bali practice different hierarchy levels of Balinese language, in which they use different word classifications to speak to different groups of the Balinese social system. The Balinese language hierarchy consists of two major classifications 1. *Basa Bali Singgih* (high tongue), this language usually utilise by Balinese when they communicate with someone with a different caste and 2. *Basa Bali Sor* (middle and low tongue), this language level is used when Balinese speak with someone that has the same or lower level of caste (Pramartha, Iswara, & Mogi, 2020; Suwija, 2019).

In the past, Balinese people utilized the Balinese language for daily communication within the family and community. Teaching of the Balinese language is also accommodated in formal education (elementary and high school). However, today it is understood that fewer and fewer Balinese practice their mother tongue, and the scope of usage has become narrow (Beratha, Sukarini, & Rajeg, 2017). With the spread of the COVID19 pandemic, all levels of school and social interactions and communication have become challenging, which has hindered the spread and growth use of Balinese language. The Balinese younger generation has been required to utilize information technologies for their learning, including of the Balinese language. Many schoolteachers provide students with Balinese language material with limited references, such as a dictionary. Many students have experienced difficulty understanding the Balinese language, and they try to source it from the traditional dictionary and the Internet with limited help. Using a traditional Balinese dictionary becomes challenging because many of the words are missing (*alus*, *madya*, and *sor*) and the order of words is quite complex when dealing with a non-basic word (*kata dasar*).

In Indonesia, the erosion of understanding of mother tongue languages, especially by the younger generation, may be due to the globalization of the use of English and use of the national language (Bahasa Indonesia). Also, the limited availability of Balinese content and digital dictionaries has decreased the enthusiasm of the Balinese younger generation to learn about their language. This has raised concerns in many groups, such as academics, local people and the international community. Losing the ability to use mother-tongue languages not only causes a loss of understanding of the structure of the language, but also a loss of understanding of cultural knowledge and history.

The aim and significance of the current initiative project is to increase the understanding of the Balinese language for Balinese people who live in Bali, and the Balinese community overseas. This knowledge is normally spread using daily communication within the Balinese social system and community. The development of a semantically practical digital Balinese dictionary will provide an opportunity for interested Balinese and others to learn and understand the Balinese cultural knowledge. This project also supports the Governor of Bali regulation number 80 of 2018 and Bali Province regulation number 4 of 2020 concerning the Protection and Use of Balinese Language, Literature and the Implementation of the Balinese Language Month. Currently, there is no reliable and comprehensive digital Balinese dictionary that can be used by people who want to learn, preserve and spread the use of Balinese language. Therefore, this initiative research project is expected to help the local government and Balinese people by utilizing information technology to protect, maintain and preserve the Balinese language, not only in the school domain but in all organizations and communities in Bali and in other countries.

2. RESEARCH DESIGN

Our study is based on a design science research methodology (DSRM) (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) that consist of five stages: 1. Problem identification; 2. Objective for solutions; 3. Design and development; 4. Demonstration and evaluation; and 5. Communication (see Figure 1). The DSRM focused on developing a proof-of-concept level prototype, and evaluating the resulting information technology artefact using a selected group of users to solve humanities and heritage problems (Pramartha & Davis, 2016). Due to the nature of our research problem, we start the sequence of DSRM from the problem-centered initiation phase, as previously mentioned that a limited digital resource is available for those who would like to learn about Balinese language and culture. Based on the problem identification we intend to develop a practical digital Balinese dictionary with a semantic web technology that can utilized manipulated by the computer-based program and human.

The design and development of our approach will be described in the next section, and will involve a Community-based Crowdsourcing (CBC), a specific type of user on the Internet with a personal engagement (Brambilla, Ceri, Mauri, & Volonterio, 2014) to populate data in our semantic ontology.

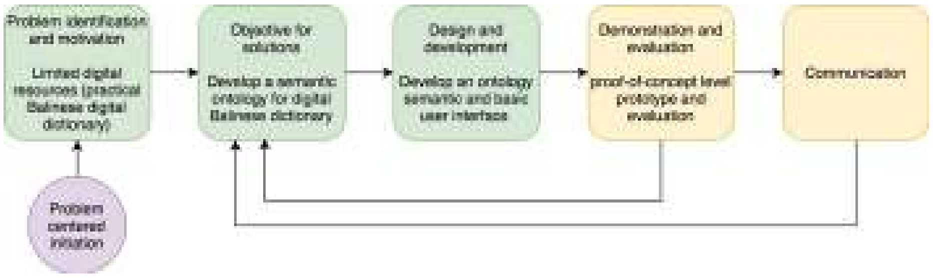
Problem identification and motivation
Limited digital resources (practical Balinese digital dictionary)
Objective for solutions
Develop a semantic ontology for digital Balinese dictionary
Design and development
Develop an ontology semantic and basic user interface
Demonstration and evaluation
proof-of-concept level prototype and evaluation
Communication
Problem centered initiation

Figure 1. Research phases

3. RELATED WORK

The use of dictionaries in the school domain aims to improve student vocabulary and develop their language skills. The use of proper vocabulary by a student is dependent on the way they acquire and transform the information in the dictionary into their daily habits. Dictionaries can be classified as either printed or digital. A study by Thenmozhi and Aravindan (2018) suggests that the users (middle school student) tend demonstrate a better attitude toward a digital dictionary compared to a printed version.

A language digital dictionary is a cross-lingual system that allows users to translate a source language to a target language based on keywords to query the translation. The use of keyword-based query for translation presents an ambiguity, which the use of ontology can help to minimize. A study by Thenmozhi and Aravindan (2018) shows the development of an ontology – specifically, for a bilingual Tamil-English dictionary – provides a positive result in terms of recall precision.

The semantic web makes use of an ontology to represent the knowledge base and web resources. Ontology is the backbone of the semantic web: it connects symbols that humans understand with forms that can be processed by machines; thus, ontology becomes a bridge between humans and machines. An ontology is a mechanism to represent a set of knowledge based on the relationship between concepts contained in a particular domain. Thus, an ontology can be used to present information semantically and to organize and map a collection of information resources in a systematic and structured manner. This is very useful in terms of data interoperability because it can be done in a more effective and efficient manner (Lombardo & Pizzo, 2016). Moreover, ontology also has the benefit of increasing accuracy in the process of searching for information on the web.

4. RESULTS AND DISCUSSION

The ontology design and development method utilized a Methontology (Fernández-López, Gómez-Pérez, & Juristo, 1997; Pramartha, 2020), which comprises of three events: 1. Knowledge acquisition; 2. Evaluation; and 3. Documentation. Moreover, this approach involves of six stages (specification, conceptualization, formalization, integration, implementation, and maintenance). We did not strictly follow the project's determined process and stages and have instead used a adjusted iterative and incremental method due to the nature of our project.

In this stage, we design and develop a basic ontology to model the domain of Balinese language. The well-known Simple Knowledge Organization System (SKOS) that support for basic ontologies for language resources (Grévisse & Rothkugel, 2020; Miles & Bechhofer, 2009) and the Lexicon Model for Ontologies (lemon) as an outcome research of the Ontology Lexicon community group (OntoLex) (Bosque-Gil, Gracia, Montiel-Ponsoda, & Aguado-de-Cea, 2016) studied and integrated to the basic ontology model developed. The large semantic electronic lexical database for English (WordNet) that minimize the ambiguity between word (Vial, Lecouteux, & Schwab, 2019) also being studied that will later be used to merge the populated data in Balinese or Indonesia language to English.

In the first stage, we model the domain in the Balinese language or in the Indonesian language. Also, we focus to include the basic vocabulary items within the set called the Swadesh list, to begin with. Selected Balinese community members are involved to give an initial understanding of the language being studied. For each resource we utilized the Resource Description Framework (RDF) localization using (`rdfs:label`) to represent the resource in different languages, such as English. We modelled the Balinese language ontology (see Figure 2) as follows:

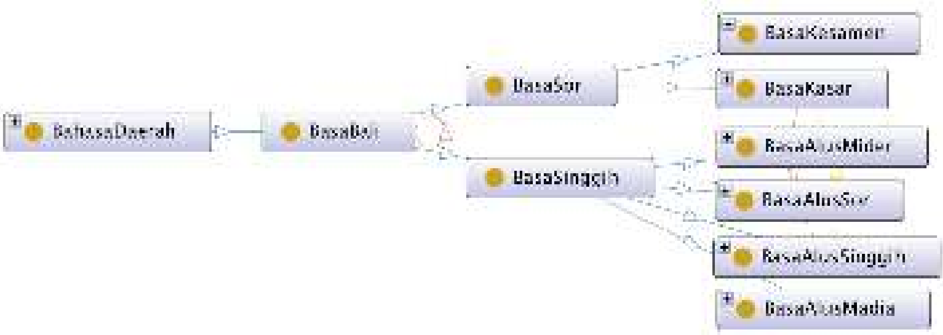

Figure 2. Classes of levels of Balinese language

- To be modelled as concepts or classes:
    - All entities that are families of Balinese language levels or classes; for example: BasaSor, BasaSinggih
- To be modelled as individuals:
    - All types of words
    - All entities that cannot have subclasses or instances

This in-progress study adopts the ontology development workflows used by Pramartha, Davis, and Kuan (2017) that comprise five stages (specification, conceptualization, formalization, implementation, and maintenance) and while similar to them, due to the nature of our project, we did not exactly follows each of the steps in sequence.

We utilize OWL2 language representation and the protégé ontology editor (Musen & Protégé, 2015) to create the classes, relationship, and example instances as shown in Figure 3 and the representation in the computer code shown in Example 1.

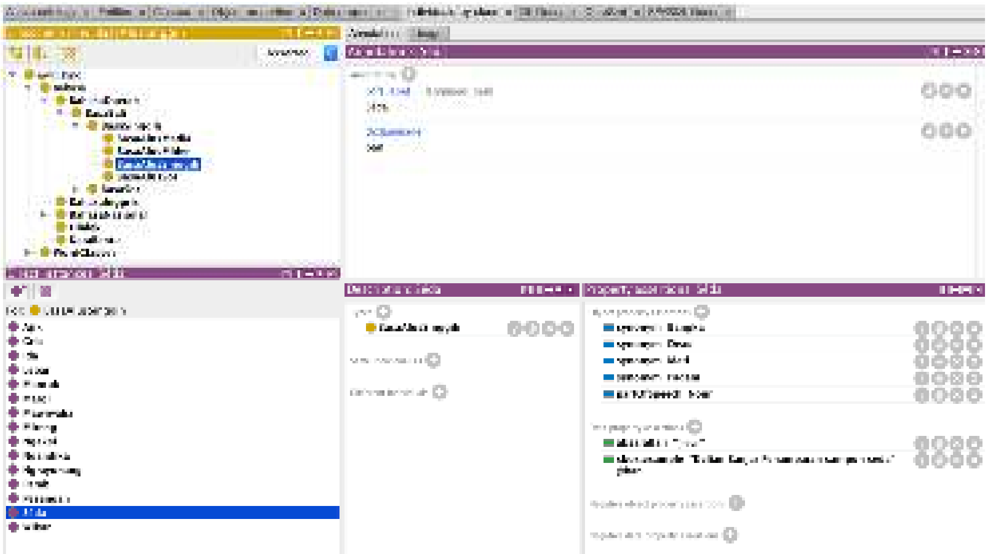

Figure 3. Entities developed on the protégé ontology editor

The draft of basic ontology Balinese dictionary is available at http://dpch.oss.web.id/Bali/BalineseDictionary.owl#. Moreover, here is a sample (see Example 1) of the ontology representation in Turtle triple syntax serialization (ttl) to represent the semantic relationship of the word *séda* (dead). The syntax shows the synonyms of *séda* (Balinese high tongue) (`:Séda`) to *bangka* (Balinese low tongue) (`:BasaKasar :Bangka`), antonym *maurip* `(lexinfo:antonym :Maurip`), and synonym *mati* (Indonesian language) (`:synonym :Mati`) and dead (English) `(lexinfo:synonym :Dead`). Also, it shows how to write the word in Unicode Balinese script (:aksaraBali "□□□"). Moreover, the syntax also uses the SKOS vocabulary to write an example of usage in Balinese language (`skos:example "Kelian Banjar Penamparan sampun séda"@ban`)

Example 1: Turtle triple syntax *séda* (dead)

```
   # Core model
   @prefix : <http://dpch.oss.web.id/Bali/BalineseDictionary.owl#> .
   # Other model
   PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
   PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
   PREFIX owl: <http://www.w3.org/2002/07/owl#>
   PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
   PREFIX thk: <http://dpch.oss.web.id/Bali/TriHitaKarana.owl#>
   PREFIX kamus:
<http://dpch.oss.web.id/Bali/BalineseDictionary.owl#>
   PREFIX vcard: <http://www.w3.org/2006/vcard/ns#>
   PREFIX dc: <http://purl.org/dc/elements/1.1/>
   PREFIX lexicog: <http://www.w3.org/ns/lemon/lexicog#>
   PREFIX lexinfo: <http://www.lexinfo.net/ontology/3.0/lexinfo#>
   PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
   PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
   PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

```
http://dpch.oss.web.id/Bali/BalineseDictionary.owl#Séda
:Séda a owl:NamedIndividual , :BasaAlusSinggih ;
    lexinfo:synonym :Bangka , :Dead , :Mati , :Padem ;
    :aksaraBali "□□□" ;
    lexinfo:antonym :Maurip ;
skos:example  "Kelian Banjar Penamparan sampun séda"@ban ;
lexinfo:partOfSpeech lexinfo:noun .
```

Ontograph plugins in the protégé ontology editor can be used to visualize the relationship of the sample ontology above (see Figure 4).

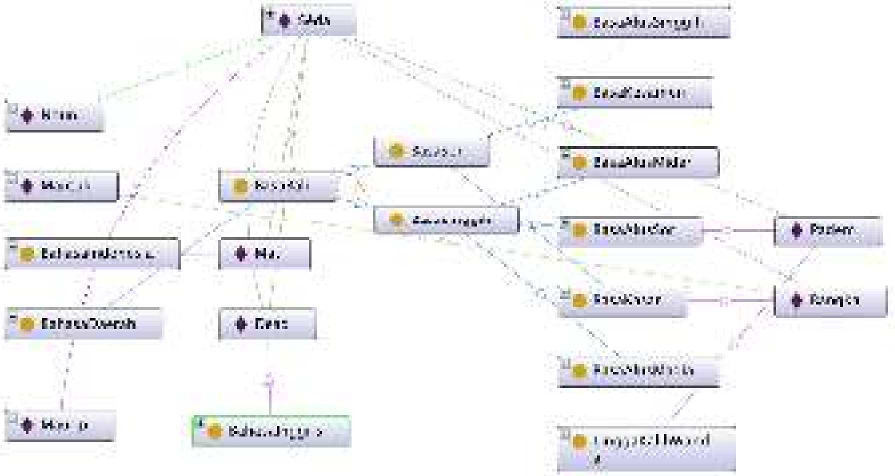

Figure 4. Example of the semantic of word dead (séda)

5. CONCLUSION

We present the initial phase of our study dealing with the humanity and heritage domains of the Balinese language. Our contributions include the development of a basic ontology, and the relationships of words in Balinese language to represent this knowledge, that can be utilized by computer-based systems. Currently, we are working to develop a web-based user interface to allow the Community-based Crowdsourcing populating data in the practical digital Balinese Dictionary.

.

ACKNOWLEDGEMENT

This project received funding from the 2022 Udayana University International Research Collaboration Grant [grant no. B/78.191/UN14.4.A/PT.01.03/2022]

REFERENCE

Beratha, N. L. S., Sukarini, N. W., & Rajeg, I. M. (2017). Balinese Language Ecology: Study about language diversity in tourism area at Ubud village. *Jurnal Kajian Bali, 7*(2), 121-134.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., & Aguado-de-Cea, G. (2016). *Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case.* Paper presented at the GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme.

Brambilla, M., Ceri, S., Mauri, A., & Volonterio, R. (2014). *Community-based crowdsourcing*. Paper presented at the Proceedings of the companion publication of the 23rd international conference on World wide web companion, Seoul, Korea.

Dixon, R. M. (1997). *The rise and fall of languages*. Cambridge: Cambridge University Press.

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). *Methontology: from ontological art towards ontological engineering*. Paper presented at the AAAI-97 Spring Symposium Series, Stanford University, EEUU. http://oa.upm.es/5484/

Grévisse, C., & Rothkugel, S. (2020). *An SKOS-Based Vocabulary on the Swift Programming Language*, Cham.

Hutapea, E. (2020, 22 February). Indonesia Punya 718 Bahasa Ibu, Jangan Sampai Punah! *Kompas*. Retrieved from https://edukasi.kompas.com/read/2020/02/22/21315601/indonesia-punya-718-bahasa-ibu-jangan-sampai-punah?page=all#page2

Lombardo, V., & Pizzo, A. (2016). Multimedia tool suite for the visualization of drama heritage metadata. *Multimedia Tools and Applications, 75*(7), 3901-3932. doi:10.1007/s11042-014-2066-3

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation*.

Musen, M. A., & Protégé, T. (2015). The Protégé Project: A Look Back and a Look Forward. *AI matters, 1*(4), 4-12. doi:10.1145/2757001.2757003

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems, 24*(3), 45-77.

Pramartha, C. (2020). Pengembangan Ontologi Tujuan Wisata Bali Dengan Pendekatan Kulkul Knowledge Framework. *SINTECH (Science and Information Technology) Journal, 3*(2), 77-89. doi:10.31598/sintechjournal.v3i2.592

Pramartha, C., & Davis, J. G. (2016). Digital Preservation of Cultural Heritage: Balinese Kulkul Artefact and Practices. In M. Ioannides, E. Fink, A. Moropoulou, M. Hagedorn-Saupe, A. Fresa, G. Liestøl, V. Rajcic, & P. Grussenmeyer (Eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection: 6th International Conference, EuroMed 2016, Nicosia, Cyprus, October 31 – November 5, 2016, Proceedings, Part I* (pp. 491-500): Springer International Publishing.

Pramartha, C., Davis, J. G., & Kuan, K. K. Y. (2017, 4-6 December). *Digital Preservation of Cultural Heritage: An Ontology-Based Approach.* Paper presented at the The 28th Australasian Conference on Information Systems, Hobart, Australia.

Pramartha, C., Iswara, I. B. A. I., & Mogi, I. K. A. (2020). Digital Humanities: Community Participation in the Balinese Language Digital Dictionary. *Jurnal Sistem Informasi (Journal of Information System), 16*(2), 18-30. doi:10.21609/jsi (jis).v16i2.956

Suwija, I. (2019). Tingkat-Tingkatan Bicara Bahasa Bali (Dampak Anggah-Ungguh Kruna). *Sosiohumaniora, 21*(1), 90-97.

Thenmozhi, D., & Aravindan, C. (2018). Ontology-based Tamil–English cross-lingual information retrieval system. *Sādhanā, 43*(10), 1-14.

Vial, L., Lecouteux, B., & Schwab, D. (2019). *Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation.* Paper presented at the the 10th Global WordNet Conference - GWC 2019.

Widiyanto, N. (2018, 24 July). Badan Bahasa Petakan 652 Bahasa Daerah di Indonesia. Retrieved from https://www.kemdikbud.go.id/main/blog/2018/07/badan-bahasa-petakan-652-bahasa-daerah-di-indonesia