Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
16/2 (2023), 103-113. DOI: http://dx.doi.org/10.21609/jiki.v16i2.1148

Knowledge Management for Electronic-Based Government System Using Semantic Thesaurus

Nuraisa Novia Hidayati[1], Agoeng Srimoeljanto[2]

[1] Research Center for Data and Information Science, National Research and Innovation Agency, Serpong, Indonesia
[2] Research Center for Artificial Intelligence and Cybersecurity, National Research and Innovation Agency, Serpong, Indonesia

nura017@brin.go.id [1], agoe003@brin.go.id [2]

Abstract

Sistem Pemerintahan Berbasis Elektronik (SPBE), the electronic-based government system, is Indonesia's e-government policy providing services to citizens through information and communication technology. Knowledge containing SPBE must be managed in various ways, one of which is the creation of an SPBE thesaurus to facilitate access and search for SPBE-related items using words or terms about it. In this study, we provide an overview of the thesaurus development process that has complied with the ISO 25964 standard and uses the Simple Knowledge Organization System (SKOS) as the application of the thesaurus in the web environment. Basic concepts or related terms and relationships between concepts have been linked with similar concepts in other thesauri that have existed before. This research also looks at the process of automating the recognition of related terms in internet articles using Word2Vec and Doc2Vec. In the process of adding terms, we discover challenges in filtering terms, determining relationships between terms, and determining reciprocal relationships between terms.

Keywords: *SPBE, thesaurus, ISO 25964, SKOS, semantic, e-government, word2vec, doc2vec*

1. Introduction

During the last four years, the Indonesian government has attempted to re-intensify efforts to promote the digital transformation to provide better services to Indonesian citizens through Presidential Regulation No. 95 of 2018 concerning Electronic-Based Government Systems (SPBE). SPBE strives for clean, effective, transparent, accountable governance, quality, and dependable public services. National SPBE governance and management are also required to improve the integration and efficiency of the electronic-based government system. One of the major initiatives in SPBE is implementing knowledge management to improve government services and decision-making.

The implementation of SPBE has closely related to business transformation, and the process is undoubtedly very complex. In the previous studies, implementing enterprise architecture based on several frameworks is necessary to achieve the transformation goal [1][2]. Besides enterprise architecture, a study to build digital enterprise architecture compared three models for managing digital repositories to make SPBE much more efficient and green-minded [3]. SPBE then became a fundamental issue regarding digitizing the government system. We discuss the SPBE architecture, especially the service and infrastructure domains.

Digitizing public services is currently a demand for many governments worldwide [4]. Managing organizational knowledge is one of the most critical aspects of digital transformation. Sharing and creating organizational knowledge is made more effective and efficient thanks to advancements in information and communication technologies. It is necessary to conduct knowledge management related to SPBE to develop and add value to the government's products, services, and administrative processes [5]. However, there seems to be a lack of knowledge management awareness in the public

sector [4].

One of the knowledge management processes can be through the creation of the SPBE semantic web. The Semantic Web is increasingly used in e-government to describe and specify e-government services to achieve semantic integration and interoperability. Within the layered structure of the Semantic Web, thesaurus finds a relevant role in supporting semantic searches and other value-added services [6]. Several applications are owned and operated by central and local government agencies to support the implementation of SPBE digital services. Semantic Web-based applications must be developed to support interoperability between applications that support SPBE services. The thesaurus can play a role in managing all that, like how the previous study assembled it to manage knowledge related to the smart city [7].

This paper introduces the semantic knowledge thesaurus model using the Simple Knowledge Organization System (SKOS) to organize knowledge in SPBE Knowledge Management System. A semantic thesaurus is a type of thesaurus that organizes and structures knowledge by utilizing semantic relationships between terms. Semantic relationships refer to the meaning and context of terms and how they relate to one another. Non-semantic thesauri, on the other hand, are used in specialized fields where precise terminology is required, and terms are organized alphabetically or numerically based on specific criteria.

We intend to incorporate this SPBE thesaurus into the knowledge management information system (SIMPAN) application search engine, developed and used at the local government level in version one last year. Implementing a standard vocabulary for metadata management is a prospective use of the SKOS thesaurus in Electronic-Based Government systems, especially in public service and IT infrastructure. Using the SKOS thesaurus, government agencies can establish a unified language for labelling and organizing digital assets such as documents, images, and videos. For instance, a government website that provides information on public services can use SKOS to establish a standardized vocabulary for labelling service-related content, such as forms, guides, and frequently asked questions. This thesaurus would enable citizens to quickly locate and access the information they need while facilitating the government's ability to manage and maintain its digital assets efficiently.

2. Related Work

Making a thesaurus is collecting information based on words or terms closely related to knowledge about the domain. Extraction of the right words or phrases and the relationship among them are the core of the thesaurus. Therefore, sources of information extraction with massive data, such as Wikipedia, are used [8][9]. Hyperlink relationships between articles referred to as structure mining is used to determine the relationship between words or terms identified as keywords [8].

The semantic pattern with various Natural Language Processing (NLP) methods can analyze words or term relations in a domain. A study applied sentence simplification, Part Of Speech tagging (POS), and Finite State Machine (FSM) to gain information about semantic relation patterns [9]. The relationships between words or terms first searched for are hypernyms and synonyms. With the speed of information related to a predetermined domain increasing, the thesaurus must continuously be updated. Searching for words or other terms in a particular domain in other sources of information using existing relationships can use the word embedding model. Previously, the exploration for synonyms and related words was expanded using word embedding, specifically doc2vec, with a determined seed word [10]. Research in Italy for cybersecurity thesaurus uses a term extraction process with Bidirectional Encoder Representations from Transformers (BERT) and a hypernym and synonym relation discovery process comparing two methods on the word2vec and fast text models [11]. The study of making a thesaurus in the cardiology domain identifies hypernym-hyponym relationships in words and terms and clusterized them to group concepts with the K-means method. Also, it involves an expert intervention to validate the results [12].

3. Methodology

This paper describes the proposed method for creating an SPBE thesaurus. The proposed method's stages are depicted in the Fig. 1.

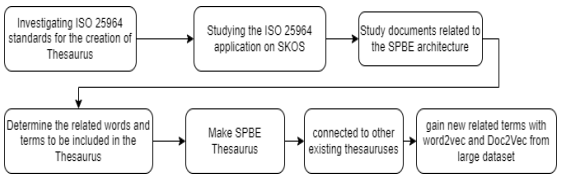

Fig. 1. Research Methodology

A. ISO 25964 vs SKOS

ISO 25964 is a standard for thesaurus development. Following ISO 25964 ensures that the thesaurus is designed and developed using industry-standard practices, making it easier to

share and integrate with other systems. It also contributes to the accuracy, comprehensiveness, and ease of use of the thesaurus, which can improve the effectiveness of information retrieval and knowledge management.

Meanwhile, the World Wide Web Consortium (W3C) recommends Simple Knowledge Organization System (SKOS) as a tool for creating and managing thesaurus in a web environment. This chapter explained creating a thesaurus with SKOS that adhere to the ISO 25964 standard. There are significant differences between ISO 25964 and SKOS, including the thesaurus construction and the integrity rules. The SKOS element category consists of several categories, each consisting of several subcategories, as shown in Table 1.

Table 1. SKOS Categories and Subcategories

Categories	Subcategories
Concepts	Concept, Concept Scheme, In Scheme, Has TopConcept,Top Concept Of
Labels & Notation	Pref Label, alt Label, hidden Label, notation
Documentation	Note, change Note, definition, editorial Note, example, history Note, scope Note
Semantic Relations	broader, narrower, related, broader Transitive, narrower Transitive, semantic relation
Mapping Properties	broad Match, narrow Match, related Match, close Match, exact Match, mapping relation
Collections	Collection, ordered Collection, member, member List

The equivalence between ISO 25964 and SKOS is discussed in [13]. Some of the mappings are direct, but some correspondences are more challenging to establish. There are differences in thesaurus construction in ISO 25964 and SKOS, where thesaurus construction based on ISO 25964 does not directly apply to SKOS construction [6][14]. Following are some of these differences:

Concept Groups: In the construction of SKOS, concept groups can be applied using a concept scheme or collection. However, they cannot represent the whole concept group as in ISO 25964.

Concept Group Nesting: As of ISO 25964, the hasSubGroup/hasSuperGroup relationship cannot be applied to SKOS. The solution is found in ISO 25964 SKOS extension (ISO-THES), where a new property is created to add a subgroup/supergroup.

Top Concept: Different representations of concept groups affect the top concept. Thesauri use concept schemes to represent micro thesauri. However, the relation cannot be used if collections are used as in ISO-THES to represent concepts. For this reason, information about top concepts often cannot be shared explicitly and depends on the interpretation of individual users and creators.

Effects of different types of NT/BT relationships: The representation of the diversity of Narrower Term/Broader Term (NT/BT) relationships does not have a definite solution on SKOS, except for introducing transitivity. Transitivity can be used if the existing relationship is generic (kind of), not mixed (part of).

According to ISO 25964, "When the scope of one concept completely overlaps the scope of the other, a hierarchical relationship is established. It must be based on degrees or levels of superordination and subordination, with the superordinate concept representing the class or whole and the subordinate concept referring to its members or parts. The following tags should be used in reverse order: The broader term, BT, is written as a prefix to the term superordinate; the narrower term, NT, is written as a prefix to the term subordinate". [15]. The reciprocal relationship required in ISO 25964 does not automatically happen.

B. Determine SPBE-Related Term

In this study, the information to be recorded is contained within the scope of the SPBE infrastructure. The related concepts are determined using the SPBE regulatory documents and the implementation of the SPBE evaluation on the Abacus application, which is a Ministry of State Apparatus Utilization, and Bureaucratic Reform used to carry out the SPBE Architecture preparation. Three experts who assisted in implementing the SPBE architecture in some organizations over the last three years carried out concept extraction and hierarchy as the basis for making this semantic thesaurus.

The terms that will constitute the top concept and its derived ideas and hierarchical relationships must be determined. The core concept grouping is accomplished by developing a concept Scheme called "Arsitektur SPBE" that contains all concepts related to SPBE architecture. The "Arsitektur SPBE" concept scheme includes six top concepts: "Proses Bisnis" (Business Processes), "Layanan" (Services), "Data dan Informasi" (Data and Information), "Keamanan" (Security), "Aplikasi" (Applications), and "Infrastruktur" (Infrastructure). The six top concepts are "themes", "domains", or "subject categories", which include the narrower concepts

of each of these six top concepts. Meanwhile, the implementation of the collection, which is a compilation representation in the ISO 25964 thesaurus terminology on "SPBE Thesaurus" by grouping several concepts under "Layanan Publik" (Public Service/ Public Administration)" as a top concept into Collection "G2C (Government to Citizen)" or "G2B (Government to Business) as shown in Fig. 2.

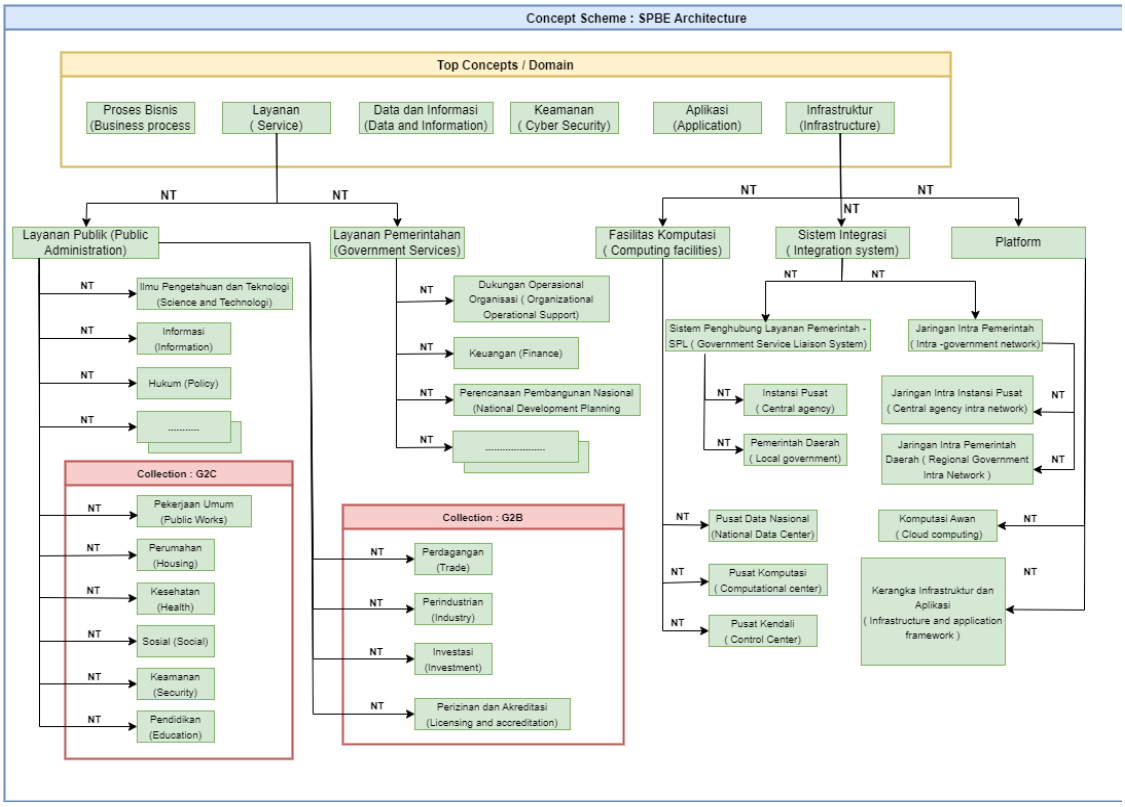

Fig. 2. Architecture SPBE Hierarchy Concept

C. Making Thesaurus

The two SPBE architecture domains of *layanan* (services) and *infrastruktur* (infrastructure) will receive the most attention in this study. We create thesaurus using VocBench, an open-source web-based collaborative platform developed and maintained by the Semantic Technology Laboratory of the National Research Council of Italy (CNR). It integrates easily with other semantic web tools and applications and supports several standard formats such as SKOS, RDF, and OWL. VocBench displays the domain's hierarchical concepts tree structure and the resource view of a certain concept. As its derived notion, each domain is provided with an initiating term. These ideas will serve as the foundation for the extensional search for similar terms. The term initiation as a concept is then applied to SKOS.

D. Connected to Existing Thesauri

There is the possibility of describing a process for analyzing the features of another current thesaurus, like previous studies that manage the knowledge organization systems used in cultural heritage collections published as Linked Data, focusing on the quality of SKOS-based models [16] . The existing thesauri for our research with a similar scope and related terms are listed below.

1) AGROVOC

FAO publishes AGROVOC, a thesaurus containing food, nutrition, and agriculture information. It is organized hierarchically into 25 concepts, such as "activity," "entity," and "location".

2) UNESCO

The thesaurus of the United Nations Educational, Scientific, and Cultural Organization (UNESCO) is available in SKOS Core and SKOS-XL versions, containing 4,417 concepts organized into seven domains and 88 microthesauri. It is downloaded as a SKOS dataset following Linked Open Data (LOD) principles and complies with ISO 25964.

3) ELSST

The thesaurus is owned and operated by the Consortium of European Social Science Data Archives (CESSDA) and is used for data discovery across Europe. It supports Danish,

Dutch, Czech, English, Finnish, French, German, Greek, Hungarian, Icelandic, Lithuanian, Norwegian, Romanian, Slovenian, Spanish, and Swedish. It covers over 3,300 concepts in politics, sociology, economics, education, law, crime, demography, health, employment, information, communication technology, and environmental science.

4) LCSH

The Library of Congress Subject Headings (LCSH) is a monolingual subject title thesaurus maintained by the US Library of Congress. It contains 30,000 concepts and is available as data linked to a database of data. Cross-references are used with headers for users of new terms, such as titles for available terms and related topics.

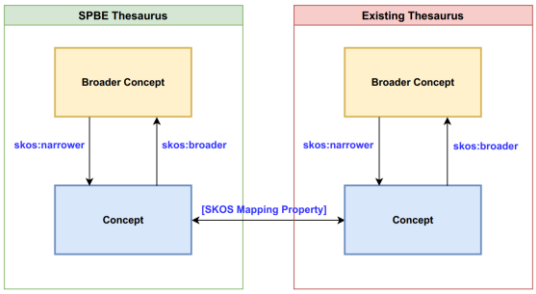

Fig. 3. Concepts Mapping Diagram between Different Concept Schemes

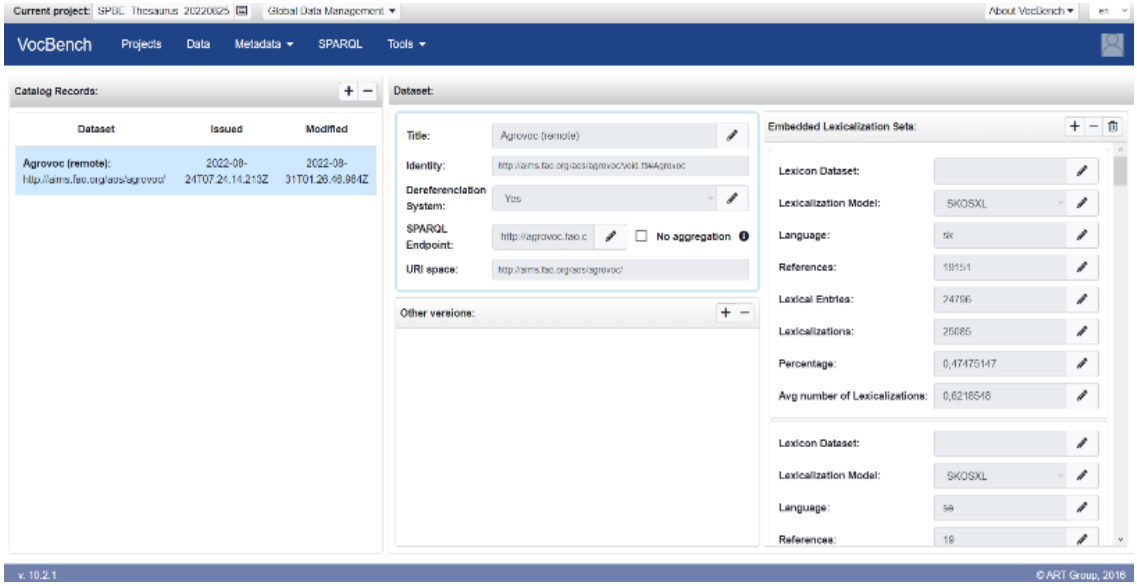

Fig. 4. Metadata Registry for SPBE-AGROVOC Concepts Mapping

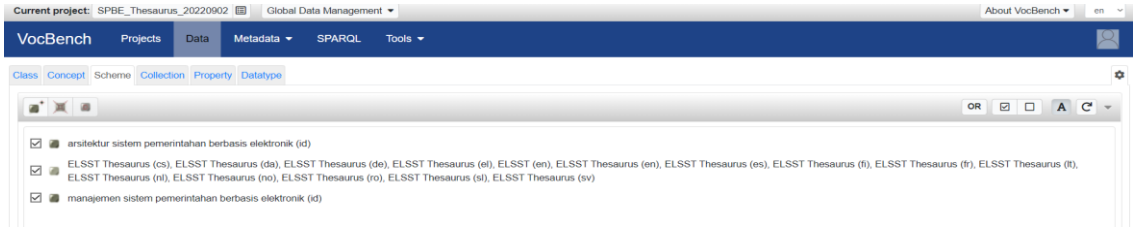

Fig. 5. Namespace and Import for SPBE-ELSST Concepts Mapping

The VocBench tool maps or aligns the SPBE thesaurus with existing thesauri. This mapping process is shown in Fig 3, which shows the mapping of two concepts from two different concept schemes. The Metadata Registry View can create a catalogue of known remote datasets using metadata such as DCAT, VoID/LIME, Friend of a Friend (FOAF), and Dublin Core Metadata Initiative (DCMI) Metadata Terms. AGROVOC is one of the cloud's remote datasets. Open datasets because AGROVOC is available in Dataset Catalogs of Linked Open Data (LOD)[1].

For example, we aimed to search the AGROVOC thesaurus by utilizing its Internationalized Resource Identifier (IRI)[2] in the VocBench tool. A successful search generated one AGROVOC Catalog Record, confirming the addition of AGROVOC metadata, as depicted in Fig. 4. The researchers aligned external resources

[1] https://lod-cloud.net

[2] http://aims.fao.org/aos/agrovoc

by mapping concepts from their SPBE thesaurus to those in AGROVOC. During the mapping process, they selected one of the five mapping properties: skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch, and skos:relatedMatch. In addition, they used the Namespaces and Imports feature of VocBench to import a concept scheme from the European Language Social Science Thesaurus (ELSST-CESSDA)[3] in the turtle format and URI. Adding the namespace as a new reference, a new concept scheme named "ELSST Thesaurus" appeared on the Scheme tab in the VocBench tool, as shown in Fig. 5, enabling the researchers to carry out the concept mapping process between SPBE Thesaurus and ELSST Thesaurus.

E. Obtain a new term by using Word2Vec and Doc2Vec.

Concepts were gathered from various SPBE infrastructure documents to identify potential new terms to add to the thesaurus and then compared to seed words that were defined as similar or related terms, as shown in Fig. 2. This research focuses on identifying potential new terms for the thesaurus by utilizing seed words from the service and infrastructure domain. The expert then manually compiles and verifies these seed words.

The SPBE thesaurus, which already contains seed words, will be used to generate keywords for articles or data obtained from Wikipedia scrapped by BeautifulSoup python library. Wikipedia is a collection of dispersed and constantly updated articles. We tried several keywords as filters when retrieving text data on Wikipedia for Indonesian language articles, such as *layanan public* (public services), *layanan pemerintahan* (government services), *layanan* (services), *ilmu pengetahuan dan teknologi* (science and technology), *infrastruktur* (infrastructure), and *pemerintahan* (government), due to the difficulty in filtering related terms. Meanwhile, articles in English contain public administration, infrastructure, government, services, science and technology, and infrastructure.

The total number of articles gathered is 100. We prioritize Indonesian articles because this thesaurus still uses Indonesian terms, with only a few terms translated into English for easy mapping into existing thesauruses. Seed words in the service and infrastructure domain will be checked for synonyms and related words. The search process for this SPBE thesaurus concept can be carried out using a two word-embedding method as AI helps to automate extraction terms, Word2vec and Doc2vec. The Word2Vec and Doc2Vec models will be trained with data from 100 Wikipedia articles and added with several SPBE documents on the service and infrastructure domain.

Word2Vec is an approach to learning high-quality word embeddings based on neural networks [17]. It is trained on a large corpus of text to learn word distributional patterns represented by the similarity of two vectors. The algorithm has two architectures: skip-gram and continuous bag-of-words (CBOW). The skip-gram architecture predicts context words when given a target word, whereas the CBOW architecture predicts target words when given a context [18]. To automatically expand thesauri, the algorithm uses the most similar() method to find similar words in the corpus and add them as synonyms to the thesaurus. The model was fine-tuned for a specific task or domain to improve performance.

The Doc2Vec algorithm is a Word2Vec extension that learns vector representations of documents, also known as document embeddings[19]. It represents each document in the corpus as a fixed-length vector that captures its contextual and semantic relationships with other documents. The distributed memory (DM) or distributed bag-of-words (DBOW) architectures can be used to train the algorithm [18]. The DM architecture predicts the next word in a document based on its context, whereas the DBOW architecture predicts a word's presence in a document based on its context. To automatically expand the thesauri using Doc2Vec, the algorithm searches the corpus for similar documents and adds them as related topics to the thesaurus. The algorithm accomplishes this by employing the most similar() method, which returns a list of similar documents based on the cosine similarity of their document vectors.

In this study, we train Word2vec with the following parameters were used:

- Train epochs: 6
- Vector size: 200
- Minimum count: 3
- Algorithm: Skip-gram
- Negative sampling: 5
- Window size: 5
- Downsampling: 1e-3

The train epochs parameter specifies how many iterations the model should perform on the training data. The vector size parameter determines the size of the vector representations learned by Word2vec. The minimum count parameter specifies how many times a word must

[3] https://elsst.cessda.eu/id/

appear in the corpus before it is included in the vocabulary. Infrequent terms are typically excluded because they do not contribute significantly to the model's performance. Negative sampling is a technique for increasing training efficiency while lowering computational costs. Instead of updating the weights of all words in the vocabulary for every training sample, negative sampling randomly selects a small number of words as negative examples and only updates their weights.

The number of negative samples is a parameter that determines how many negative examples are used for each training sample. The number of negative samples was set to 5 in this case. The downsampling parameter specifies how frequently words in the corpus are randomly downsampled during training. This parameter helps to prevent overfitting and improves model performance. The downsampling parameter was set to 1e-3 in this case. Doc2vec uses the same train epochs, vector size, minimum count, negative sampling, window size, and downsampling parameters as Word2vec. The only difference is that the distributed memory (DM) algorithm is used. The DM algorithm is a CBOW variant that can learn vector representations of documents and words. It predicts the next word based on the preceding terms and the document vector. The document vector is updated during training to capture the contextual and semantic relationships between documents.

4. Result and Discussion

A. Mapping of Terms Between Thesauri

We could not locate an Indonesian thesaurus with the same scope, but as stated previously, this SPBE thesaurus will be linked to several English thesauri with virtually the same scope. Because we are constructing our thesaurus exclusively in Bahasa Indonesia, we translate the terms into English; for instance, we translate *layanan publik* into public administration so that the terms public and administration can capture similar words or terms.

A team of experts was involved in this research, and we performed the translation procedure manually to ensure that the context was maintained. The term public administration, which resembles public service on ELSST and AGROVOC, is one example of a concept map, as shown in Fig. 6. As depicted in Fig. 7, the term ilmu *pengetahuan dan teknologi* on the ELSST is comparable to science and technology. The mapping process is performed on a term-by-term basis, with each term first being translated into English for recognition purposes, thereby rendering the SPBE thesaurus bilingual. This procedure necessitates vigilance and discernment of the same scope so that each term has the same meaning. The collected terms are then arranged according to the hierarchical structure created in Fig. 2, resulting in the initial thesaurus structure depicted in Fig. 8.

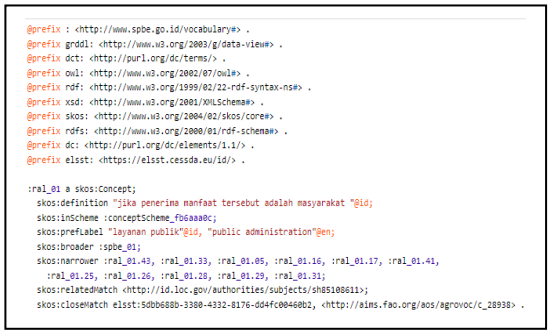

Fig. 6. Public Service Concept Relation

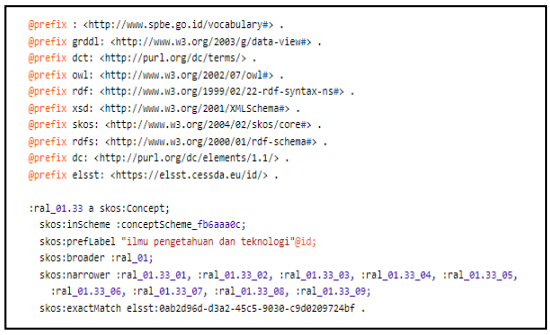

Fig. 7. Technology Mapping Term

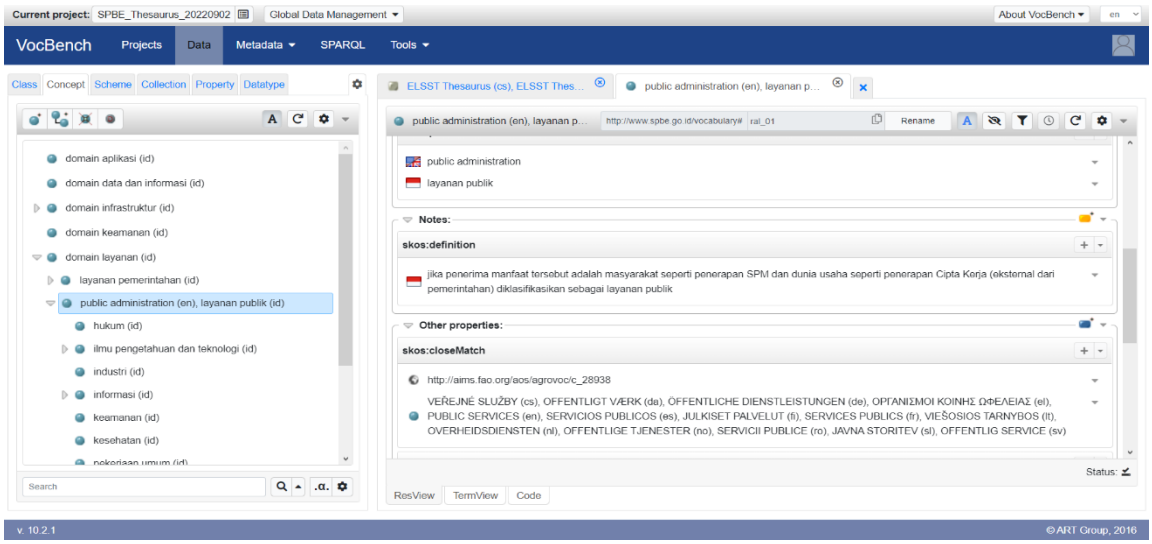

Fig. 8 Hierarchical Concepts Tree Structure of the Domains and Resource View of a Certain Concept

B. Word2Vec and Doc2Vec Related Term Generated

Some examples of words generated based on the Word2Vec and Doc2Vec models with several examples of keywords such as services, public, science and technology, infrastructure, and government can be seen in Table 2**.** We investigate the top 5 similarity scores of terms generated by Word2Vec and Doc2Vec algorithms, like the previous study that used Doc2Vec only [10]. Both algorithms are commonly used for natural language processing and are used to analyze text data. Word2Vec is a shallow neural network-based model that focuses on word embeddings. At the same time, Doc2Vec is an extension of Word2Vec that can handle variable-length texts such as sentences, paragraphs, or documents by creating document-level embeddings. Looking at the table, we can observe that both algorithms generate similar terms, but the similarity score varies.

For example, Word2Vec generates the term *pelayan* (waiter) with a similarity score of 0.80, while Doc2Vec generates *pelayanan* (service) with a similarity score of 0.77. Similarly, Word2Vec generates the term *komputer* (computer) with a similarity score of 0.80, while Doc2Vec generates *inovasi* with a similarity score of 0.81. One key difference between the two algorithms is the level of analysis. Word2Vec focuses on word-level embeddings, while Doc2Vec considers document-level embeddings. As a result, Doc2Vec can better capture the context of the entire document, while Word2Vec is better suited for analyzing individual words in isolation. Another difference is the nature of the similarity scores.

Word2Vec tends to generate higher similarity scores than Doc2Vec, as it focuses more on the relatedness of words, while Doc2Vec considers both the relatedness and similarity of the documents. In terms of the generated terms, both algorithms generate terms that are commonly used in academic and technical writing. For example, *pemerintah* (government), *teknologi informasi* (information technology), *riset* (research), and *kebijakan* (policy) are all commonly used in academic writing. However, there are some differences in the terms generated, with Word2Vec generating terms like *komputer* (computer) and data mining , which are more technical. In contrast, Doc2Vec generates terms like *publikasi* (publication) and *jasa* (services), which are more general and encompass a more comprehensive range of topics. In summary, the table shows that both Word2Vec and Doc2Vec algorithms generate similar terms, but the similarity scores and the nature of the generated terms differ between the two algorithms. Word2Vec is better suited for analyzing individual words in isolation, while Doc2Vec is better for analyzing documents.

The process of filtering words and terms that appear has its challenges. Because the scope set is SPBE, while the articles used as training data are articles with top concepts as keyword terms, these terms may have a general meaning and may not be related to SPBE. Words and terms found from Word2Vec and Doc2Vec modeling process also need to identify the relationships among them. The relationship between generated related terms and the top concept terms needs to be determined, whether it is NT, BT, or RT. In the end, the expert must manually determine the relationship.

It should be noted that ISO 25964 requires a reciprocal relationship between NT and BT, which SKOS does not automatically support. When the

top concept has the concept below it, there will only be a BT relationship between the concepts. Meanwhile, the NT process from the previous concept must be manually entered. We must keep this in mind when adding terms that every time terms are added, the reciprocity of each related term must be ensured. Despite their similarities, words not within the scope of SPBE appear during the training process for searching for synonyms and related terms in Indonesian. The term "layanan"(service) brings up words like "pelayan"(waiter) and "pelanggan"(customer), which are not in the SPBE scope. The expert has eliminated the words like that.

Table 2. Example of Words generated by Word2Vec and Doc2Vec

	Word2Vec		**Doc2Vec**	
	Word	**Similarity Score**	**Word**	**Similarity Score**
Layanan (service)	Pelayan (waiter)	0.80	Pelayanan (service)	0.77
	Pelanggan (customer)	0.78	Pelanggan (customer)	0.75
	Fasilitas (facility)	0.76	Konsumen (consumer)	0.75
	Inovasi (innovation)	0.75	Teknis (technical)	0.70
	Penyedia (provider)	0.74	Jasa (service)	0.70
Publik (public)	Masyarakat (people)	0.73	Publikasi (publication)	0.78
	Pemerintah (government)	0.64	Masyarakat (people)	0.77
	Layanan (service)	0.63	Umum (public)	0.74
	Administrasi (administration)	0.63	Pengunjung (visitor)	0.74
	Birokrasi (bureaucracy)	0.62	Pemerintah (government)	0.72
Ilmu Pengetahuan dan Teknologi (science and information technology)	Komputer (computer)	0.80	Inovasi (Inovation)	0.81
	teknologi informasi (information technology)	0.74	teknologi informasi (Information Technology)	0.72
	Sains (science)	0.74	Riset (research)	0.65
	Teknologi (technology)	0.74	Pengembangan (development)	0.65
	data mining	0.72	cybersecurity	0.64
Infrastruktur (Infrastructure)	Pembangunan (development)	0.84	Jaringan (network)	0.78
	Prasarana (infrastructure)	0.76	Transportasi (transportation)	0.67
	Proyek (project)	0.75	Komunikasi (communication)	0.65
	Modernisasi (modernization)	0.74	Teknologi (technology)	0.65
	Penyediaan (provision)	0.70	Fasilitas (facility)	0.63
Pemerintah (government)	Birokrasi (bureaucracy)	0.76	Lembaga (organization)	0.78
	Kebijakan (policy)	0.69	Negara (country)	0.72
	Daerah (regional)	0.69	Kebijakan (policy)	0.69
	Pusat (central)	0.68	Politik (political)	0.65
	Peraturan (regulation)	0.66	organisasi	0.64

C. New Terms Obtained as an Expansion of Seed Words

Further investigation revealed that the algorithms generated 210 unique words with multiple keyword associations in the services domain. However, only 75 of these words were approved after expert review, indicating that the remaining 135 words were either irrelevant to the services domain or needed to be clearer to be considered valid.

Similarly, the algorithms generated 150 unique words with several keyword associations in the infrastructure domain, but experts approved only 56. A precision score is a metric to assess a natural language processing system's ability to identify relevant information correctly [11]. In this case, the precision score is the ratio of approved words to total words generated by the algorithms. As a result, the precision score in the services domain would be 75/210 = 0.36, and in the infrastructure domain, it would be 56/150 = 0.37. The precision score provides valuable information about how well the Word2Vec and Doc2Vec algorithms generate relevant words for a given domain. A low precision score indicates that the algorithms generated many irrelevant or ambiguous words. In contrast, a high precision score suggests that the algorithms could capture the domain's relevant language. As a result, the precision score is an essential factor to consider when evaluating the performance of natural language processing systems in domain-specific applications.

This thesaurus will store all keywords related to the SPBE architecture to find related information quickly. However, expanding the SPBE thesaurus was difficult because the sentences caught were outside the scope of SPBE despite using keywords related to SPBE. The process of filtering and relations is still very manual and requires human power, in this case, experts and us as researchers, so creating a thesaurus takes a long time and is prone to errors. Another external issue is that the keywords in SPBE sometimes need to be more specific and are used for purposes other than explaining the SPBE architecture.

D. SPBE Thesaurus Maintenance Strategy

According to our findings, we propose a maintenance strategy for an SPBE SKOS thesaurus. The strategy entails establishing a system to automatically extract new terms from online articles using word2vec and doc2vec models, validating the context of the newly extracted terms, and integrating the thesaurus with the website's search engine. The system can be configured to run periodically to add new terms to the thesaurus.

Keeping track of changes in the digital government domain, maintaining consistency and organization in the thesaurus, regularly reviewing and updating scope notes and definitions, and providing clear documentation on how to use the thesaurus are additional considerations for ensuring the accuracy and utility of the thesaurus. This strategy ensures that the thesaurus remains useful and pertinent for users navigating the complex terminology of digital government.

5. Conclusion and Future Work

Thesaurus development in the web environment must adhere to the SKOS structure and the ISO 25964 standard. In this study, the SPBE thesaurus's SKOS construction applied concept schemes, top concepts, concepts, and the relationship between NT and BT. In this study, we look for related terms scattered throughout articles on the internet. The initial terms in the main hierarchies are in the SPBE scope and domain and have been identified as keywords in advance. The term search process employs Word2Vec and Doc2Vec, with data training articles containing Wikipedia keywords. Using this method to add terms to the thesaurus presents three challenges, one of which is that the related terms produced by the Word2vec and Doc2Vec models are outside the scope of SPBE.

The precision value for each generated word that can be entered into the thesaurus still needs to be higher in the focused domains, namely 0.36 for the service domain and 0.37 for the infrastructure domain. The relationship between terms, namely NT, BT, and RT, is then difficult to ascertain. Finally, due to SKOS limitations, the reciprocal relationship between NT and BT to follow the ISO 25964 standard must always be ensured.

In the future, we will concentrate on e-government, online services, and IT infrastructure-related content in Wikipedia and other online articles. Avoid extracting irrelevant keywords from irrelevant content to maintain precision. Identify frequently used phrases and terms in digital government, such as "open data," "digital citizenship," and "smart cities," and employ domain-specific knowledge to improve keyword extraction. Participation by experts may necessitate knowledge of digital government procedures, policies, and technologies, as well as assuring proper NT and BT relationship patterns among the terms as concepts.

References

[1] R. Rigin and I. Dua Reja, "Survey Paper tentang Enterprise Architecture di Sektor Publik," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 56–70, 2022, doi:

10.24002/konstelasi.v2i1.5365.

[2] S. Supriyanto, A. Ridwan, R. Tamam, M. I. Santoso, D. Satria, and A. I. S. Mutaqin, "Perancangan sistem pemerintahan berbasis elektronik (SPBE) yang berkelanjutan di Provinsi Banten," *J. Ind. Serv.*, vol. 7, no. 1, p. 171, 2021, doi: 10.36055/12952.

[3] I. S. Rozas, K. Khalid, N. Yalina, N. Wahyudi, and D. Rolliawati, "Digital Enterprise Architecture for Green SPBE in Indonesia," *CCIT J.*, vol. 15, no. 1, pp. 26–42, 2022, doi: 10.33050/ccit.v15i1.1366.

[4] A. Alvarenga, F. Matos, R. Godina, and J. C. O. Matias, "Digital transformation and knowledge management in the public sector," *Sustain.*, vol. 12, no. 14, 2020, doi: 10.3390/su12145824.

[5] H. Khanchel, "The Impact of Digital Transformation on Banking," *J. Bus. Adm. Res.*, vol. 8, no. 2, p. 20, 2019, doi: 10.5430/jbar.v8n2p20.

[6] M. M. Martínez-González and M. L. Alvite-Díez, "Thesauri and Semantic Web: Discussion of the Evolution of Thesauri Toward Their Integration with the Semantic Web," *IEEE Access*, vol. 7, no. October, pp. 153151–153170, 2019, doi: 10.1109/ACCESS.2019.2948028.

[7] K. Nataliia, "Application of the ' Smart City ' data domain thesaurus as the tool for representing knowledge while improving the problem-oriented Web search effectiveness," pp. 17–20, 2019.

[8] K. Nakayama, T. Hara, and S. Nishio, "A Thesaurus Construction Method from Large Scale Web Dictionaries," no. Aina, 2007.

[9] P. Arnold, P. Arnold, and E. Rahm, "Extracting Semantic Concept Relations from Wikipedia Extracting Semantic Concept Relations from Wikipedia," no. January 2016, 2014, doi: 10.1145/2611040.2611079.

[10] D. A. Koutsomitropoulos, "Subject Classification of Learning Resources Using Word Embeddings and Semantic Thesauri," pp. 3–8, 2019.

[11] C. Lanza and A. Hazem, "Towards Automatic Thesaurus Construction and Enrichment," no. May, pp. 62–71, 2020.

[12] N. Lagutina, I. Paramonov, I. Vorontsova, and N. Kasatkina, "An Approach to Automated Thesaurus Construction Using Clusterization-Based Dictionary Analysis."

[13] I. S. O. Tc, T. A. Isaac, and S. Recommendations, "Correspondence between ISO 25964 and SKOS / SKOS - XL Models Contributors," pp. 1–13, 2012.

[14] M. M. Martínez-gonzález, M. Alvite-díez, T. I. T. Edificio, and C. M. Delibes, "Computer Standards & Interfaces The support of constructs in thesaurus tools from a Semantic Web perspective : Framework to assess standard conformance," *Comput. Stand. Interfaces*, vol. 65, no. July 2018, pp. 79–91, 2019, doi: 10.1016/j.csi.2019.02.003.

[15] ISO 25964-1:2011(E), information and documentation — Thesauri and interoperability with other vocabularies — Part 1:Thesauri for information retrieval, First edition, 2011-08-15

[16] D. Díaz-corona, J. Lacasta, M. Á. Latre, F. J. Zarazaga-soria, and J. Nogueras-iso, "Profiling of knowledge organization systems for the annotation of Linked Data cultural resources," vol. 84, pp. 17–28, 2019, doi: 10.1016/j.is.2019.04.008.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.

[18] S. Martinčić-Ipšić, T. Miličić, and L. Todorovski, "The influence of feature representation of text on the performance of document classification," *Appl. Sci.*, vol. 9, no. 4, 2019, doi: 10.3390/app9040743.

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st Int. Conf. Mach. Learn. ICML 2014*, vol. 4, pp. 2931–2939, 2014.