*Legal Knowledge and Information Systems*
*E. Schweighofer (Ed.)*


*doi:10.3233/FAIA210324*

# Analyze the Usage of Legal Definitions in Indonesian Regulation Using Text Mining Case Study: Treasury and Budget Law

Bakhtiar AMALUDIN [a] Fitria Ratna WARDIKA [a]
Putu Jasprayana MUDANA PUTRA [a] and I Gede Yudi PARAMARTHA [b]
[a] *Legal Bureau, Ministry of Finance of Indonesia*
[b] *Inspectorate General, Ministry of Finance of Indonesia*

**Abstract.** Legal definitions are an integral part of legal drafting practice to understand legal documents easily and prevent ambiguity. This research aims to describe how legal definitions are used among regulations in the domain of Indonesian Treasury and Budget. Simple text mining techniques are used to perform and deliver the process. We extracted definitions from more than 1.362 related regulations enacted through the period 2003-2020. We found that legal definitions were used in many variations which may lead to inconsistencies.

**Keywords.** legal definition, legal term, consistency, harmonization, text mining.

## 1. Introduction

Do the definitions in regulations need to be consistent? Gauci points out that there are situations where regulation has defined a legal term, but in another, the legal term is given a different definition [4]. This situation could trigger different interpretations and thus, it is not a mistake when Lucius Priscus said that every definition in law is dangerous [4]. Hence, as an essential part of legal drafting practice, having harmonized legal definitions is not merely for precise and effective communication [2]. In this situation, the challenge is how legal drafters formulate harmonized definitions and, more importantly, do not potentially contradict each other.

In Indonesia, there are several rules in drafting legal definitions, including consistency in defining terms, particularly in similar fields; and definitions in lower regulations must be in line with higher regulations [7]. However, learning from Gauci's findings [4], definitional inconsistencies seem unavoidable in Indonesia. For example, since State Finance Law[1] and State Treasury Law enacted [2], there have been numerous implementing regulations comprises of Government Regulations, Presidential Regulations, and Minister of Finance Regulations. Thus, the legal definitions also increase as the number of regulations grows. We often found that the definition of some legal terms within two or more regulations are varied and lead to confusion.

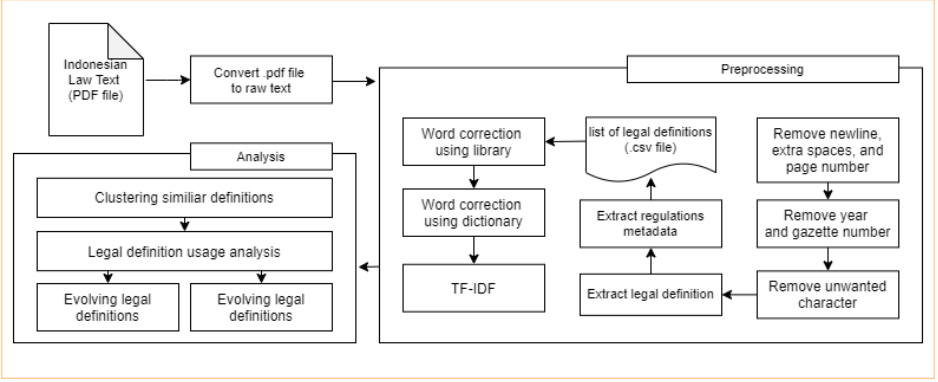
[1] Indonesian Law No. 17 of 2003 on State Finance
[2] Indonesian Law No. 1 of 2004 on State Treasury

To perform an in-depth analysis of this issue, we propose text mining to explore the use of definitions across regulations in the domain of Indonesia's treasury and budget. Finally, the results of this study are expected to be an input for legal drafters to make a consistent definition to prevent legal problems due to the existence of a term that is defined differently.

## 2. Methodology

Figure 1 presents an overview of the proposed framework for analyzing legal definition usage in Indonesian regulations. In Step 1 (Data Collection), regulations in the domain of treasury and budget law are collected. Step 2 (Preprocessing) emphasizes noise removal and data transformation. Step 3 (Analysis) aims to analyze legal definition usage in the regulations and find potential mismatch.

**Figure 1.** Legal Definition Usage Analysis Framework




### 2.1. Data Collection

We gathered 1362 regulations in the treasury and budget domain from Indonesian state gazette stored as PDF files. It comprises 4 Laws (UU), 15 Government Regulation (PP), 2 Presidential Regulation (PERPRES), and 1,313 Regulation of Minister of Finance (PMK). We transformed all the PDFs documents into machine-readable form (i.e. raw text) using the pdfminer library in Python.

### 2.2. Pre-processing

The purpose of this step is to extract information we need for data analysis (i.e. legal terms contained in regulation and its metadata). Initial text cleaning performed on the raw text to remove noises such as new lines, extra spaces, page number, year and gazette number, and unwanted character in the raw text.

### 2.3. Extract Legal Definitions

To extract the legal definitions from regulations, our legal drafting experts analyzed examples of legal definitions appearing in regulations. These patterns then transformed into regular expressions that form specific kinds of text patterns for a faster searching [5].

Indonesian regulations have a standardized structure that legal terms are always defined in the general provision of the regulations. Thus, we first identified the general provision part of each regulation (i.e. in the first article of the regulation). After that, we split general provision text into the segmentation of sentences using a sentence tokenizer. We analyzed each sentence to see whether it meets the legal definition pattern (Table 1) and extracted three parts from each of these (terms, alias, and definitions).

**Table 1.** Legal definition pattern in general provision.

| Type | Sentence Pattern | Regular Expression |
|---|---|---|
| Direct definition | *term* is [...] | ^(.+)(adalah)(.+) |
| Acronym | *term* hereinafter referred to as *alias* is [...] | ^(.+)(yang selanjutnya disebut)(.+)(adalah)(.+) |
| Abbreviation | *term* hereinafter abbreviated as *alias* is [...] | ^(.+)(yang selanjutnya disingkat)(.+)(adalah)(.+) |

After automatically extracting legal keywords from regulation documents, we were able to extract 8,202 legal terms and the number of unique legal terms is 2,546 which means some legal terms appear in more than one regulation. However, not all legal definitions are extracted correctly by these regex patterns. From manual inspection we found the 117 incorrect legal definitions were captured and 11 records must be discarded because they are not considered as legal definitions. For the rest, although the regex were able to identify legal definitions component (terms, alias and definition) correctly, many of them have misspelled words, missing letters, incorrect word order, and mixed words. Thus, further data was cleaning performed to handle this problem. We used symspell library complemented with Bahasa Indonesia Frequent Words Dictionary[3] and dictionary-based spelling correction to fix several misspell words.

### 2.4. Extract Regulation Metadata

We also extracted some relevant information in the header part of each regulation. It was related to the source of legal definitions such as the regulation number, the type of regulation and year of enactment. The regular expression pattern for this was "REPUBLIK\s+INDONESIA \s+NOMOR\s+(\d+)\s+TAHUN\s+(\d{4})". We then captured the first group as the regulation number and the second group as the year of enactment.

#### 2.4.1. Cluster Legal Definition

In this step, each legal term from the previous process clustered according to its similarity in definition. However, to work with clustering algorithms, we need to transform text into numerical representation. In this case, we implemented TF-IDF to transform legal definition text into number of matrix [5]. Nothing excessive in this transforming

[3]https://github.com/hermitdave/FrequencyWords

process, the only intervention is regarding tokenization. Indonesian cases are different where special cases occur such as not treating hyphen (-) as signs of word segmentation.

We used the most popular density based clustering [9] in particular DBSCAN clustering algorithm with euclidean distance to group similar words in separate clusters [9]. We set the best parameter that is given by silhouette score 0.56418 (i.e. epsilon=0.01 and min_samples=1). It produced 4,691 clusters which were then used for labeling each legal definition. The final result of these processes described in Table 2.

**Table 2.** Final dataset.

| Terms | Alias | Description | Source File | Year | Reg.Type | Label |
|---|---|---|---|---|---|---|
| General Allocation Fund | DAU | General Allocation Fund, [...] | 68/PMK.02/2016.pdf | 2016 | PMK | 809 |

### *2.5. Legal Definition Variation Analysis*

Based on the dataset illustrated in Table 2 above, we identified some potential conditions that may cause variation on legal definitions as follows.

- **Evolving Definitions**: a condition when the same legal terms appear as different cluster labels but used sequentially in time order according to the number and year of enactment.
- **Potential Mismatch** : a condition when same legal terms appears as different cluster in same or different type of regulations

These condition will be used as baseline to subset legal definitions for further analysis. However variation here can not be judged as unlawful practice since our approach in detecting variation is limited only on syntactical differences in definition text.

## 3. Analysis

### *3.1. Legal Definition Usage and Variety*

Initial analysis goes into an insight how legal definitions are used repeatedly in several regulations. As depicted in Figure 1, the more frequent the legal terms used, the more varied they are defined across regulations. For example, terms ”DIPA” has more than 40 different definition that spread within in two different type of regulations (i.e. PP and PMK) totaling more than 100 documents.

### *3.2. Evolving Definitions*

We found that there are 403 definitions that can be considered as evolving definitions. Evolving definition is a common aspect that causes the diversity of definitions which legal drafter make enhancement or improvement carried out in accordance with the current situation and conditions faced by policymakers. Laws are often revised several times and it is a necessary part of the legal process that may be modified or extended [1]. Legal rules are more general in the present and for the future scenarios such rules must be applied [11].

**Figure 2.** Legal Definition Usage and Variety

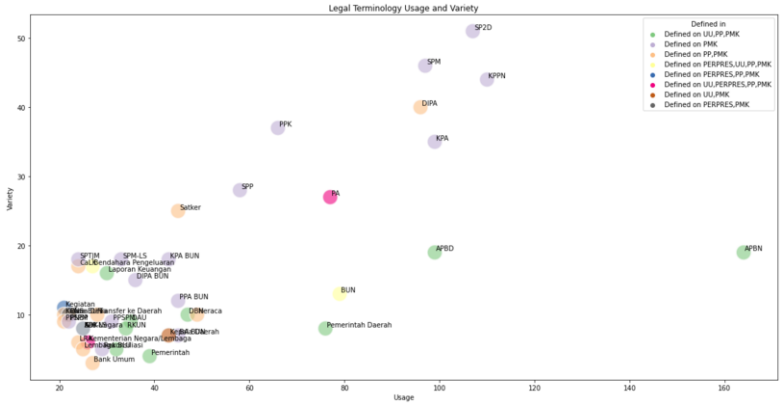



**Table 3.** Example of Term ”SIKD” as evolving definitions

| Definition | Year | Appear In |
|---|---|---|
| Regional Financial Information System, hereinafter abbreviated as SIKD, is a system that documents, administers, ***and processes other related data*** into information that is presented to the public and as [...] | 2016-2017 | 93/PMK.07/2016; 18/PMK.07/2017; |
| Regional Financial Information System, hereinafter abbreviated as SIKD, is a system that documents, administers, and ***processes regional financial management data and other related data*** into information that is presented to the public and as [...] | 2018-now | 189/PMK.05/2018; 24/PMK.07/2020; 231/PMK.07/2020; 233/PMK.07/2020; |

### *3.3. Potential Mismatch*

Potential mismatch legal definition occurs when legal terms are defined differently across regulations on different level of regulations. We found 135 legal definitions that meet this condition. For instance the term ”DIPA” has different definitions between those listed in the PP and the PMK as depicted in Table 4. This difference should not have occurred as it is clearly stated in Legal Drafting Guidance [7] that the general definitions in lower regulations cannot contradict with the definition of the higher regulation particularly in the same domain.

**Table 4.** Example of Potentialy Mismatch Definitions of Term ”DIPA”

| Definition | Appear In |
|---|---|
| DIPA is a ***budget execution (allotment) document*** which is used as ***reference*** for Budget Users in carrying out government activities as the implementation of the state revenue and expenditure budget | PP 45 of 2013; |
| DIPA is a ***budget execution (allotment) document*** prepared by the Budget User or Budget User Proxy | 36/PMK.02/2015; 94/PMK.05/2016; etc. |

Another example regarding potential conflict is when legal terms are defined differently across regulation on the same level. We found 658 legal definitions that meet this condition. For instance, inconsistency occurred in the definition of the term ”SPTJM” (Table 5). This is not considered as evolving regulation since initially they used definition

A, then used definition B and using definition A again. Based on Legal Drafting Law [7] if a definition is reformulated in a new regulation, the formulation must be the same with the definition of the previous enforced regulation.

**Table 5.** Example of potential mismatch in definition of term "SPTJM"

| **Definition** | **Appear In** |
|---|---|
| **A.** SPTJM is a statement letter which among other things contains a statement that all consequences of an official/person's actions that may result in state losses are the full responsibility of the official/person who took the action. | 212/PMK.05/2020<br>156/PMK.05/2019 |
| **B.** SPTJM is a statement of responsibility from the official for all expenses for payment of meal allowances and to return it to the state when overpayment and state loss. | 110/PMK.05/2020 |

Therefore, to make regulations harmonized with each other as well as reduce risk of misinterpretation, a legal term must have consistent definition in every regulation, regardless of its type/level.

## 4. Conclusions

We presented a framework for exploring the consistency and to find potential mismatch of the use of legal definitions in regulations. The use of text mining for this purpose can be extended not only to other legal domains in Indonesia, but also by other jurisdictions or non-governmental organizations. Furthermore, the result also can be used as a baseline for building a legal terms dictionary that can be used by legal analyst/drafter.

## References

[1] Ajani G, Boella G, Di Caro L, Robaldo L, Humphreys L, Praduroux S, Rossi P, Violato A. The European legal taxonomy syllabus: a multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. Applied Ontology. 2016 Jan 1;11(4):325-75.

[2] Chiocchetti E. Harmonising legal terminology. EURAC research; 2008.

[3] Culy C, Chiocchetti E, Ralli N. Visualizing conceptual relations in legal terminology. In 2013 17th International Conference on Information Visualisation 2013 Jul 16 (pp. 333-338). IEEE.

[4] Gauci, Gotthard MarkİIs It a Vessel, a Ship or a Boat, Is It Just a Craft, Or Is It Merely a Contrivance? Journal of Maritime Law and CommerceOctober 2016, 47(4): 479

[5] Goyvaerts J. & Levithan S. Regular Expression Cookbook. BIM Publishing Servies. 2012.

[6] Hwang R.H., Hsueh YL, Chang YT. Building a Taiwan law ontology based on automatic legal definition extraction. Applied System Innovation. 2018 Sep;1(3):22.

[7] *Law No. 12 of 2011 on the Legislation Making as amended by Law No. 15 of 2019* (Indonesia).

[8] M. F. L. Schmitt and E. J. Spinosa, "Outlier detection on semantic space for sentiment analysis with convolutional neural networks," in 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Jul. 2018, ppİ-8, doi:10.1109/IJCNN.2018.8489200

[9] Mohammed S. M., Zeebaree S. R. M., & Jacksi K. A state of the art survey on semantic similarity for document clustering using GloVE and density based algorithms. Indonesian Journal of Electrical Engineering and Computer Science. April 2021.

[10] Mommers L, Voermans W. Using Legal Definitions to Increase the Accessibility of Legal Documents. In JURIX 2005 May 15 (pp. 147-156).

[11] Rawls J. Two concepts of rules. The philosophical review. 1955 Feb 1;64(1):3-2.

[12] Šavelka J, Ashley KD. Legal information retrieval for understanding statutory terms. Artificial Intelligence and Law. 2021 Jul 8:1-45.

[13] Winkelsm R, Hoekstra R. Automatic Extraction of Legal Concepts and Definitions. In Legal Knowledge and Information Systems: JURIX 2012: the 25th Annual Conference 2012 (Vol. 250, p. 157). IOS Press.