The Observation of Bahasa Indonesia Official Computer Terms Implementation in Scientific Publication

D Gunawan¹, A Amalia², M S Lydia², M I Muthaqin¹

¹Department of Information Technology, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155
²Department of Computer Science, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

Email: danigunawan@usu.ac.id

Abstract. The government of the Republic of Indonesia had issued a regulation to substitute computer terms in foreign language that have been used earlier into official computer terms in Bahasa Indonesia. This regulation was stipulated in Presidential Decree No. 2 of 2001 concerning the introduction of official computer terms in Bahasa Indonesia (known as Senarai Padanan Istilah/SPI). After sixteen years, people of Indonesia, particularly for academics, should have implemented the official computer terms in their official publications. This observation is conducted to discover the implementation of official computer terms usage in scientific publications which are written in Bahasa Indonesia. The data source used in this observation are the publications by the academics, particularly in computer science field. The method used in the observation is divided into four stages. The first stage is metadata harvesting by using Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH). Second, converting the harvested document (in pdf format) to plain text. The third stage is text-preprocessing as the preparation of string matching. Then the final stage is searching the official computer terms based on 629 SPI terms by using Boyer-Moore algorithm. We observe that there are 240,781 foreign computer terms in 1,156 scientific publications from six universities. This result shows that the foreign computer terms are still widely used by the academics.

1. Introduction

The era of the early Internet in Indonesia brought new computer terms in foreign language that had no suitable translation in Bahasa Indonesia. In 2001, the government of Republic of Indonesia responded to these computer terms by issued the Presidential Decree No. 2 of 2001 concerning the introduction of official computer terms in Bahasa Indonesia. The official computer terms are also known as Senarai Padanan Istilah (SPI). There are 629 terms to substitute the computer terms in foreign language. These terms were generated from the Computer Generation Specific Formatting Guidelines.

Sixteen years after being officially published by the government of Republic of Indonesia, the official computer terms in Bahasa Indonesia should have been implemented widely in scientific publications which are written in Bahasa Indonesia. A research that conducted in 2015 had analyzed the implementation of official computer terms in Bahasa Indonesia. The implementation of official computer terms in Bahasa Indonesia was analyzed by conducting questionnaire to the students. The research shows that the students only use as much as 33% of the official computer terms in Bahasa Indonesia [1].
Recent research in 2016 conducted the observation of the implementation of official computer terms in Bahasa Indonesia for the government websites [2]. The research observed 1,084 Indonesian government websites. It focused on the government websites’ template. It means the research only observed the repeated part in the websites, not the content. This research found that there are 47% of Indonesian government websites which implement 88.4% official computer terms in Bahasa Indonesia compared to the foreign language. This result does not depict the condition of the whole websites because the observation was done for the template only, not included the content of the website.

This research will analyze the implementation of official computer terms in Bahasa Indonesia in scientific publications which are written by the academics, especially by the students who majoring computer science (or related to computer). We use the publications written by the higher education students instead of questionnaire to analyze the implementation of official computer terms in Bahasa Indonesia. The research stages include data harvesting by using Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH), converting harvested documents (in pdf format) to plaintext, conducting text pre-processing and the last stage is comparing the computer terms by using a string matching algorithm. The OAI-PMH was used in previous research to harvest scholarly research journal’s content [3]. The result of this research is proposed to be the starting point for the government to evaluate the implementation of official computer terms in Bahasa Indonesia.

The challenges of this research are the terms that have more than one meaning and the term with suffix. The terms that have more than one meaning, for example “address”, might lead to the miscalculation of the number of computer terms. Because the term “address” can be used as personal address or computer address. The term that should be calculated is the one which has computer related meaning. The next challenge is the term with suffix. However, this can be covered by implementing word stemming or a dictionary. As an alternative, the dictionary can be replaced by lexical database. The recent research to build a lexical database for Bahasa Indonesia is still in design phase [4]. Another challenge in natural language processing is multiword expressions (MWE). Multiword expressions are the combination of two or more words that form a meaning. Current research is still determining the candidates of the multiword expressions for Bahasa Indonesia [5]. However, because of this research used the official computer terms which listed in the Presidential Decree No. 2 of 2001 (the list includes the terms with two or more words), then the multiword expressions handling is sufficiently done by implementing string matching. This research uses Boyer-Moore algorithm to implement string matching.

The organization of this paper is as follow: the first section is the introduction about the conducted research. Then, section two discusses the observation approach used in this research. Section three discusses the result of this research. The last section is conclusion.

2. Research Methodology
The research was done by conducting several steps. The first step is collecting the documents from several major universities in Indonesia. As we research the usage of official computer terms, we limit the documents to the fields which are related to computer, such as computer science, information technology and informatics. Second, we extracted the documents into tokens to prepare for the next process. According to the ordinance of writing in Bahasa Indonesia, the foreign language must be written in italics format. As the format has no effect on the calculation of the official computer terms, this research does not consider the format of the writing. Therefore, we implemented string matching algorithm to compare the extracted tokens with the official computer terms for Bahasa Indonesia in the third step. The last but not least, we observed the result and draw conclusion. The details of the research are as follow:

2.1 Harvesting the Documents
The aim of this research is to demonstrate the usage of official computer terms for Bahasa Indonesia in academics, especially in computer field. We harvested the thesis documents which are published by the universities. We utilize Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH) to
harvest all the documents in each university. OAI-PMH provides services to publish metadata based on HyperText Transfer Protocol (HTTP) and eXtensible Markup Language (XML). To enable data interoperability, the harvested metadata should be available in agreed format. Usually, metadata is available in Dublin Core standard. The OAI-PMH provides data interoperability between two providers namely service providers and data providers. The service providers harvest the metadata published by the repositories. These repositories are called the data providers. Both service providers and data providers communicate using the same protocol called OAI-PMH.

There are six different types of protocol request and response in OAI-PMH, such as GetRecord, Identify, ListRecords, ListIdentifiers, and ListMetadataFormats. GetRecord is a verb to retrieve single metadata record from a data provider (repository). Identify is a verb to retrieve information about a data provider (repository). Usually, part of the retrieved information is required for other request. ListRecords is used to obtain or harvest records from a repository. ListIdentifiers is similar to ListRecords. However, ListRecords only harvests headers instead of records. The last, ListMetadataFormats will response metadata formats available in a data provider (repository).

As mentioned previously, data interoperability in OAI-PMH will work if both data providers and service providers communicate with the same format. In this research, we use simplified Dublin Core standard because this format is common standard to create digital web catalog. It consists of 15 metadata elements, such as title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. These metadata elements describe the published document. Figure 1 shows the example of Dublin Core implementation in XML. All the Dublin Core elements are part of the dc scheme. This scheme is pointed to http://purl.org/dc/elements/1.1/ (by using xmlns attribute). Therefore, as shown in figure 1, each tag of Dublin Core elements should be followed by dc. For example, <dc:title>Revitalisasi Perpustakaan</dc:title> means the title elements belongs to the dc (Dublin Core) scheme and has the value: “Revitalisasi Perpustakaan”. All the metadata do not have to be published or available. As shown in figure 1, the data provider only publishes 9 metadata elements.

![Dublin Core implementation in XML](image)

Initially, we observed that there are 63 universities in Indonesia which have computer related study program. However, not all of the universities support OAI-PMH for their repositories. In addition, not all universities open the access to the documents. Some of them do not allowed guest access. This will prevent the service provider to harvest all the available documents. Therefore, this research only used 10 universities in Indonesia. The list of universities and their OAI-PMH addresses are shown in table 1.

The universities listed in table 1 provides open access to their documents. Those documents can be downloaded by using GetRecord protocol. In addition, this research limits the documents to PDF format and written in Bahasa Indonesia only.
Table 1. The list of universities

| University                        | OAI-PMH Address                           |
|-----------------------------------|--------------------------------------------|
| Universitas Airlangga             | http://repository.unair.ac.id/cgi/oai2    |
| Universitas Andalas               | http://teknosi.fti.unand.ac.id/index.php/teknosi/oai |
| Universitas Diponegoro            | http://jtsiskom.undip.ac.id/index.php/jtsiskom/oai |
| Universitas Sumatera Utara       | http://repository.undip.ac.id/oai/request  |
| Universitas Sultan Syarif Kasim  | http://repository.uin-suska.ac.id/cgi/oai2 |
| Universitas Sriwijaya             | http://eprints.unsri.ac.id/cgi/oai2       |
| Institut Teknologi Sepuluh November | http://iptek.its.ac.id/index.php/index/oai |
| Universitas Gadjah Mada          | http://i-lib.ugm.ac.id/oai-ilib/oai21.php |
| Universitas Bengkulu             | http://jamal.ub.ac.id/index.php/index/oai  |
| Universitas Malikussaleh         | http://ejurnal.tif.unimal.ac.id/index.php/index/oai |

2.2 Text Pre-processing
This research requires single word (token) to be compared with official computer terms in Bahasa Indonesia. Therefore, after the metadata and document harvesting process, we extracted all the text in the documents and applied text pre-processing such as tokenizing, stop-word removal, case folding and stemming. Tokenizing is used to extract all the text into single word. Stop-word removal is intended to remove all the words that cannot be used as a single word. This research uses stop-word list provided by Tala [6]. Case folding will convert all the text to lowercase. This step will avoid case sensitive issue when comparing each token. The last text pre-processing is stemming. As the official computer terms are available in root word format, hence the token should be available in root word format also. Stemming process will extract the token to its root word. We used confix-stripping approach to provide words stemming [7]. In addition, this research does not consider the abbreviation of English terms. Therefore, the abbreviation of English terms is considered as the other terms and will not be calculated.

2.3 Comparing Token by Using String Matching Algorithm
The string matching algorithm has two types, namely: 1) exact string matching and 2) approximate string matching. This research uses extract string matching type. We utilize Boyer-Moore algorithm to match the token with the official computer terms for Bahasa Indonesia. This algorithm is known as the most famous and efficient algorithm in model matching [8].

2.4 The Result Observation
After comparing tokens with the official computer terms for Bahasa Indonesia, we observed the result for each university. Then we observed the result from the all available universities.

3. Results and Discussion
3.1 Harvesting Document Result
As mentioned earlier, the data are obtained by harvesting documents via OAI-PMH. This research limits to the documents that related to computer field only. To obtain only the specific documents, we should know the category structure in each university repository. The category structure can be observed via ListSets protocol. For example, we can observe the categories in Universitas Sumatera Utara repository by accessing http://repository.usu.ac.id/oai/request?verb=ListSets. This requires manual approach to observe each repository as there is no convention by name or code to classify specific document type. After determining the category which will be harvested for each repository, the metadata is ready to be
harvested by specifying SetSpec attribute. All the documents’ location can be extracted by using ListRecords protocol.

There are several issues in harvesting process. There are three universities that are registered as data provider but they do not provide required metadata. Therefore, we exclude those three universities from this research. They are Universitas Gadjah Mada, Universitas Bengkulu and Universitas Malikussaleh. In addition, we exclude Institut Teknologi Sepuluh Nopember from this research because it uses English instead of Bahasa Indonesia.

3.2 Data Extraction

Figure 2 shows the example of harvested XML from Universitas Sumatera Utara. We extract several fields such as published date (represented by <atom:published>), file name (represented by <atom:title>), and link to file download (represented by <atom:link>). For example, the extracted data are shown in figure 2 (a) (b) (c) respectively are published date, title and link. We notice that there is difference between Universitas Sumatera Utara and the other universities in providing link to file download. Universitas Sumatera Utara separates a document to several files according to its chapter. However, the other universities provide only one file for each document. These XMLs are extracted by using a DOM Parser.

![Figure 2. XML from Universitas Sumatera Utara.](image)

| University                    | Number of documents from the XML | Number of downloaded files |
|-------------------------------|---------------------------------|---------------------------|
| Universitas Airlangga         | 130                             | 123                       |
| Universitas Andalas           | 43                              | 43                        |
| Universitas Diponegoro        | 186                             | 185                       |
| Universitas Sumatera Utara    | 325                             | 325                       |
| Universitas Sultan Syarif Kasim | 654                          | 395                       |
| Universitas Sriwijaya         | 99                              | 85                        |

3.3 Document Download

After extracting required field from the XML, the document will be downloaded to be processed for the next step. We found that some repositories do not provide file download links for the certain documents. Therefore, the number of downloaded files are fewer than the number of documents available in the XML data.
3.4 Results
Before implementing the algorithm, the downloaded files are required to be converted to plain text. Next, we implement text pre-processing such as tokenizing, stop-word removal, and case folding. To search the official computer terms in Bahasa Indonesia we utilize a string matching algorithm called Boyer-Moore. By utilizing the algorithm, we found ten most used computer terms in English such as user, file, input, database, server, output, text, image, error, and password.

The result of the searching official term for Bahasa Indonesia is shown in table 3. The column Year describes the year when the documents are published. This column is followed by UNAND, UA, UNDIP, UNSRI, UIN-SUSKA, USU which represent Universitas Andalas, Universitas Airlangga, Universitas Diponegori, Universitas Sriwijaya, Universitas Islam Negeri – Sultan Syarif Kasim and Universitas Sumatera Utara respectively.

| Year | UNAND | UA | UNDIP | UNSRI | UIN-SUSKA | USU | Sub Total |
|------|-------|----|-------|-------|-----------|-----|-----------|
| 2001 | -     | 9  | -     | -     | -         | -   | 9         |
| 2002 | -     | 0  | -     | -     | -         | -   | 0         |
| 2003 | -     | -  | -     | -     | -         | -   | 0         |
| 2004 | -     | 21 | -     | -     | -         | -   | 21        |
| 2005 | -     | 0  | -     | 243   | -         | 243 | 243       |
| 2006 | -     | 0  | -     | 291   | -         | -   | 291       |
| 2007 | -     | 4  | -     | -     | -         | -   | 4         |
| 2008 | -     | 0  | -     | -     | -         | -   | 0         |
| 2009 | -     | 39 | -     | 83    | 1,974     | -   | 2,096     |
| 2010 | -     | 85 | -     | 334   | 3,925     | -   | 4,344     |
| 2011 | -     | 85 | -     | 547   | 25,619    | -   | 26,251    |
| 2012 | -     | 28 | -     | 1,218 | 7,751     | -   | 8,997     |
| 2013 | -     | 38 | 1,945 | 1,900 | 28,797    | 5,331 | 38,011   |
| 2014 | -     | 45 | 2,802 | 5,893 | 3,437     | 32,634 | 44,811   |
| 2015 | 564   | 510| 5,181 | 1,406 | -         | 63,581 | 71,242   |
| 2016 | 3,309 | 1,477| 4,458| -     | 1,094     | 25,044 | 35,382   |
| Total| 3,873 | 2,341| 14,386| 11,915| 72,597    | 126,590|

| Number of documents | 43 | 123 | 185 | 85 | 325 | 395 |

Most of the repositories do not have any documents from 2001 except Universitas Airlangga. However, we cannot convert the documents from UA which are published in 2005, 2006 and 2008 because the PDF files are scanned using image format. According to the table 3, the usage of the computer terms in foreign language is still very high. This result is proposed be the starting point for the government to do profound research to evaluate the implementation of official computer terms in Bahasa Indonesia.

4. Conclusion
Sixteen years after being published by the government of Republic of Indonesia, the official computer terms in Bahasa Indonesia should have been implemented widely, especially by the academics. This
research observed the usage of official computer terms in scientific publications by the academics in computer related field. We observed 1,084 scientific publications which are published by six universities in Indonesia. We found that the usage of computer terms in foreign language is still widely used by the academics in their scientific publications.

References
[1] Sari C A 2015 Tanggapan Mahasiswa di Kota Surakarta Terhadap Pengindonesiaan Istilah Asing Bidang Pengkomputeran (Kajian Sosiolinguistik) (Universitas Sebelas Maret)
[2] Amalia A, Gunawan D, Lydia M S, and Charlie C 2017 IOP Conf. Ser. Mater. Sci. Eng. vol 180 p 12060
[3] Simek P, Jarolimek J, Vanek J, and Stoces M 2011 Proc. of the 5th International Conference on Information and Communication Technologies for Sustainable Agri-production and Environment (HAICTA 2011) p 827–834
[4] Gunawan D and Amalia A 2017 IOP Conf. Ser. Mater. Sci. Eng. vol 180 p 12052
[5] Gunawan D, Amalia A, and Charisma I 2016 6th IEEE Intl. Conf. on Control System, Computing and Engineering (ICCSCE) p 304–309
[6] Tala F Z 2003 A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia (Universiteit van Amsterdam)
[7] Adriani M, Asian J, Nazief B, and Tahaghoghi S M M 2007 ACM Trans. Asian Lang. Inf. Process. vol 6 p 1–33
[8] Yuan L 2011 Intl. Conf. on Computer Science and Service System (CSSS) pp 182–184