University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal)

Libraries at University of Nebraska-Lincoln

9-10-2019

Evaluating the Quality of the Indonesian Scientific Journal References using ParsCit, CERMINE and GROBID

Ariani Indrawati
Center for Scientific Data and Documentation, Indonesian Institute of Sciences, Jakarta, Indonesia, indrawati.ariani@gmail.com

Ambar Yoganingrum
Research Center for Informatics, Indonesian Institute of Sciences, Cibinong, Indonesia, ambaryoganingrum@gmail.com

Pradipta Yuwono
Center for Scientific Data and Documentation, Indonesian Institute of Sciences, Jakarta, Indonesia, pradipta.yuwono.temp@gmail.com

Indrawati, Ariani; Yoganingrum, Ambar; and Yuwono, Pradipta, "Evaluating the Quality of the Indonesian Scientific Journal References using ParsCit, CERMINE and GROBID" (2019). *Library Philosophy and Practice (e-journal)*. 2852.
https://digitalcommons.unl.edu/libphilprac/2852

Evaluating the Quality of the Indonesian Scientific Journal References using ParsCit, CERMINE and GROBID

Ariani Indrawati, Pradipta Yuwono

Center for Scientific Data and Documentation, Indonesian Institute of Sciences, Jakarta, Indonesia

E-mail: indrawatiariani@gmail.com, pradipta.yuwono.temp@gmail.com

Ambar Yoganingrum

Research Center for Informatics, Indonesian Institute of Sciences, Cibinong, Indonesia

E-mail: amba002@lipi.go.id, ambaryoganingrum@gmail.com

Abstract

There are several open-source tools available to extract the bibliographic references of the Pdf. Those tools based on the various approaches including rule-based approach, knowledge-based approach, machine learning-based approach, and the combination. To improve the services of the Indonesian Scientific Journal Database (ISJD), Center for Scientific Data and Documentation – Indonesian Institute of Sciences (PDDI-LIPI) intends to have an automatic bibliographic references extraction tool. The paper aims to analyze the quality of the reference metadata of the local journals with the three open-source tools, namely ParsCit, CERMINE and GROBID. The accuracy test of the three tools are poor. Those are 0.555, 0.633, and 0.605 for ParsCit, CERMINE, and GROBID respectively. It caused by many authors do not use a reference manager when they write the bibliography section. On the such condition this paper proposed to build an application to identify and correct errors in the bibliographic references of paper in ISJD. This application become a liaison between ISJD and open source tool for the bibliographic reference extraction. This paper proposed the combination of building software and using an open source.

Keywords: *Automatic extraction bibliography, Indonesian scientific journals, open-source tools, application architecture*

1 Introduction

Currently, scientific document networking is an important issue for the discipline of the library and information science. The networking is employed for special services of a digital library such as providing related works to the customer along with the main papers in a database. It is also as the main source of information for conducting analysis to assess the journal impact and author/institution performance as well as to depict the knowledge trend, authors/institutions collaboration, and knowledge history.

The network of scientific document is naturally occurred because a paper cited by others. Accordingly, the automatic extraction of the bibliographic references for the scientific articles becomes a challenging task. It is caused by the various referencing styles such as APA, Harvard, Chicago, IEEE etc. In addition, each style has variations in using separators, between fields can be separated by spaces, periods, commas, or parentheses. Moreover, each referencing style has a variety in writing the name of the author. Furthermore, occasionally the references are written not in accordance with the rules, such as missing or errors in using punctuation.

There are various approaches have been suggested to solve those problems, namely rule-based approach [1], knowledge-based approach [2], machine learning-based approach [3][4][5][6][7][8][9][10][11], and the combination [12]. Several open-source tools have been built among others ParsCit [3], CERMINE [4], GROBID [5], Anystyle-Parser [13], and Biblio [14] developed by using machine learning approach; INFOMAP [2] created by using knowledge-based approach; BibPro [15], Citation [16] and Citation-Parser [17] developed by a rule-based approach. Meanwhile, some databases built their own application such as Scopus and Web of Science.

Employing an open-source or developing institution’s own application for automatic bibliographic references extraction is a decision that must be taken. Center for Scientific Data and Documentation – Indonesian Institute of Sciences (PDDI LIPI) intends to employ the application with the intention to improve the services of Indonesian Scientific Journal Database (ISJD). ISJD is a database of Indonesian journals consist of metadata and full text paper on a PDF format. All journals in ISJD are graded by the Indonesian accreditation standards called Science and Technology Index (SINTA).

This paper aims to propose which should be done by PDDI LIPI. Since the writing errors of the bibliographic references in a paper become a problematic, at the initiate stage the quality of the bibliographic references writing of the paper in ISJD must be identified. To reach the goals we create following research questions (1) How good is the quality of the bibliographic references writing in the Indonesian journals? (2) What type of application should be developed by the institution?

This paper provides a selection technique for references quality determination using an open source or developing an own application. The decision is taken based on the quality of the writing reference metadata from the papers in ISJD. Thus, three open-source applications namely CERMINE, ParsCit, and GROBID are applied to test the metadata quality of the Indonesian journal references. The selection of the applications is based on Tkaczyk et al. [4] opinions that the three tools are able to extract more fields compared to others. In addition, the tools allow for analyzing and extracting the input scientific publication in PDF format considering all articles in ISJD are available in a PDF format.

2 Related Works

2.1 Evaluating Tools

This section explains the three applications i.e. ParsCit, CERMINE, and GROBID.

A. ParsCit

Councill, Giles, and Kan [3] introduced ParsCit, an open-source tool for locating reference string, then parsing them and retrieving their citation context. The core of ParsCit is a trained Conditional Random Field (CRF) model used to label the token sequences in the reference string. A heuristic model wraps this core with added functionality to identify reference string from a plain text, and to retrieve the citation context. In ParsCit, 23 features are extracted, including capitalization, punctuation, numeric type, the length of the word, the location of the world within the reference string, the presence of substring in the word, and a pair of n-grams per word. ParsCit also uses external knowledge bases from six dictionary features, publisher names, place names, months, surnames, male and female names. ParsCit achieved 95.66% accuracy on CORA datasets. Figure 1 shows the extraction reference workflow architecture of ParsCit. First, ParsCit converting an input file (e.g. PDF) to a plain UTF-8 text file. Then, ParsCit finds the reference string using a set of heuristic by searching for a labeled reference section in given a plain UTF-8 text. The next phase is segmenting complete reference section to individual reference line. The list of individual reference line that applied to CRF++ model. Final step is normalizing each tagged field output from CRF++.

B. CERMINE

CERMINE (Content ExtRactor and MINEr), introduced by Tkaczyk et al [4], is a tool for extract metadata and structure, including parsed bibliography of scientific articles in PDF format. CERMINE workflow architecture is provided in Figure 2. This application applies Support Vector Machine (SVM) to classify an article to 4 zones (metadata, body, references, and other). Next dividing the

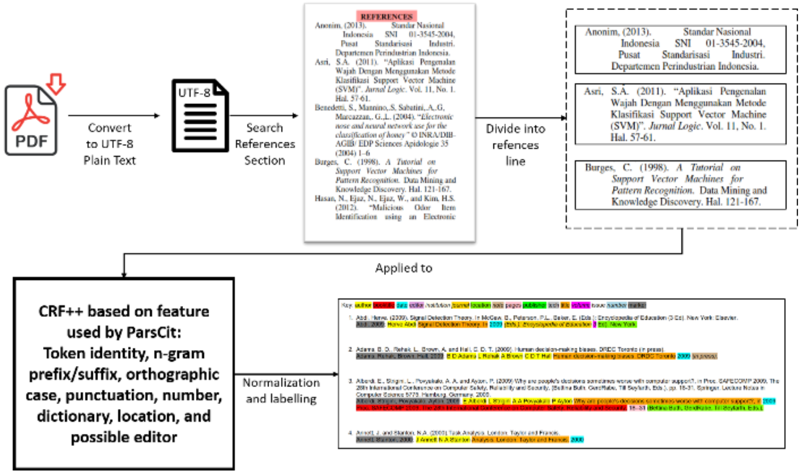

PDF
Convert
to UTF-8
Plain Text
UTF-8
Search
References
Section
REFERENCES
Divide into
refences
line
Anonim, (2013). Standar Nasional Indonesia SNI 01-3545-2004, Pusat Standarisasi Industri. Departemen Perindustrian Indonesia.
Asri, S.A. (2011). "Aplikasi Pengenalan Wajah Dengan Menggunakan Metode Klasifikasi Support Vector Machine (SVM)". Jurnal Logic. Vol. 11, No. 1. Hal. 57-61.
Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Hal. 121-167.
Applied to
CRF++ based on feature used by ParsCit: Token identity, n-gram prefix/suffix, orthographic case, punctuation, number, dictionary, location, and possible editor
Normalization
and labelling

Figure 1: The ParsCit extraction references workflow architecture (Modified from [3])

content of references zone into individual reference string using K-Means algorithm with Euclidean distance metric. Finally, extracting the metadata from references string using CRF built on top of GRMM and MALLET packages. For citation test they use 4000 parsed citation, 2000 from CiteSeer and CORA, 2000 from PubMed Central documents. The parser achieved precision 92.9%, recall 93,8%, and F-Score 93.3%.

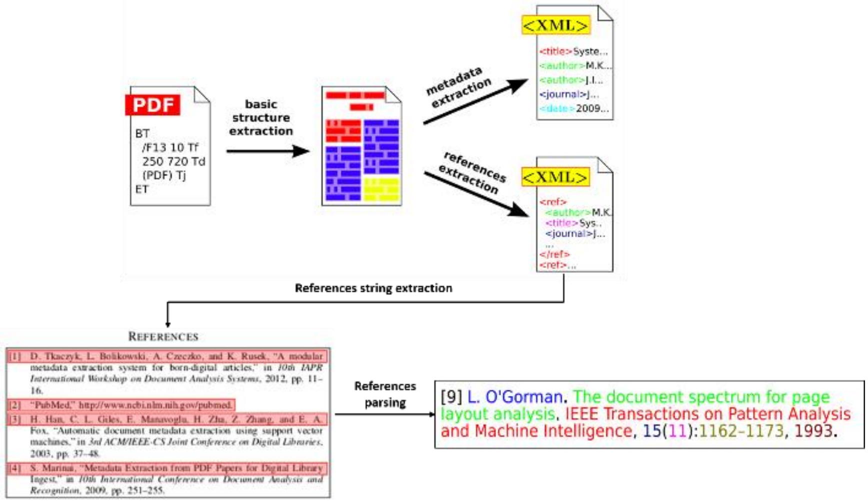

PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET
basic
structure
extraction
metadata
extraction
references
extraction
<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
<XML>
<ref>
<author>M.K.
<title>Sys..
<journal>J...
...
</ref>
<ref>...
References string extraction
REFERENCES
[1] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek, "A modular metadata extraction system for born-digital articles," in 10th IAPR International Workshop on Document Analysis Systems, 2012, pp. 11–16.
[2] "PubMed," http://www.ncbi.nlm.nih.gov/pubmed.
[3] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," in 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, 2003, pp. 37–48.
[4] S. Marinai, "Metadata Extraction from PDF Papers for Digital Library Ingest," in 10th International Conference on Document Analysis and Recognition, 2009, pp. 251–255.
References
parsing
[9] L. O'Gorman. The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11):1162-1173, 1993.

Figure 2: The CERMINE extraction references workflow architecture (Modified from [4])

C. GROBID

GROBID (GeneRation Of BIbliographic Data), proposed by Lopez [5], is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications. They have applied CRF implemented with MALLET toolkit for extracting various common fields from the headers and citation of research papers. An evaluation with the reference CORA dataset, GROBID shows a reliable level of accuracy 95.7% per citation field and 78.9% per citation instance. GROBID extraction references workflow architecture is illustrated on Figure 3. First convert PDF format to XML use pdf2xml/Xpdf. Then divide them into cover, header, body, footnotes, headnotes, biblio, and annexes. The next phase is bibliography segmentation model using CRF. The final step is parsing the reference.

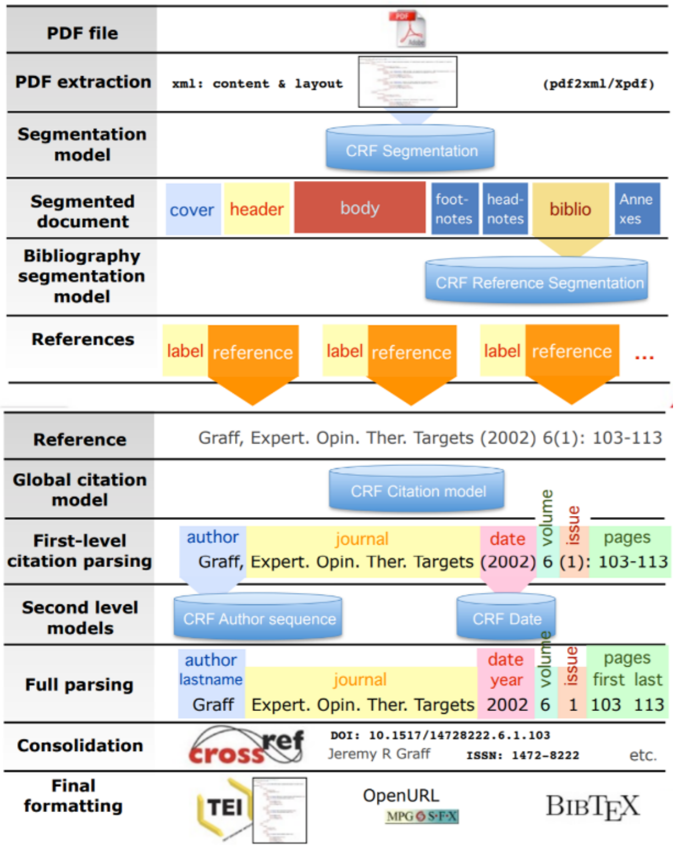


Figure 3: The GROBID extraction references workflow architecture (Source: [18])

2.2 ISJD and SINTA

ISJD integrates the management of scientific journals in Indonesia. There have been uploaded more than 283,000 articles to ISJD between 2009 to 2018 [19]. In the meantime, SINTA (Science and Technology Index), initiated by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia (*Kemenristek dikti*) in 2016, is a portal for measuring the performance of science and technology, which includes the performance of authors, journals, and institutions. Journals in SINTA are divided to 6 clusters, those are S1 to S6. The clustering is based on the score in national journal accreditation system called *Arjuna*. Table 1 shows the SINTA score.

Table 1: The SINTA Score

Cluster	Criteria
Sinta-1 (S1)	Journals with a value between 85 – 100 and/or indexed by Scopus
Sinta-2 (S2)	Journals with a value between 70 – 85

Sinta-3 (S3)	Journals with a value between 60 – 70
Sinta-4 (S4)	Journals with a value between 50 – 60
Sinta-5 (S5)	Journals with a value between 40 – 50
Sinta-6 (S6)	Journals with a value between 30 – 40

2.3 Software Development in PDDI LIPI

There are three choices to improve or develop a software. Those are using an open-source software, buying a product, or building a software from scratch. Before deciding an option, the organization should evaluate the advantages and disadvantages each option also know the organization capabilities.

- Using an open-source software

The open-source option is great for the organization that have less money but strong in technical staff support. An open-source software allows the organization to freely use and modify it based on the needs. To modify and integrate it into the current infrastructure, the organization needs the technical skills and spends more time to test the software for functionality and security [20].

- Buying a product

Buying a software is an option if your organization has money. The great advantage with purchasing a commercial product is technical support, updates, maintenance, and reliable backup from your vendors. But sometimes that software does not integrate into the organization's current infrastructure [20].

- Building a software from scratch

Building a software from scratch will allow the organization to have full control and design it really fits the needs and infrastructure of organization. However, this option requires more cost and good technical skills of the staff [20].

PDDI LIPI has a task to manage scientific data and information in Indonesia. Some computer programs that has been successfully developed are ISJD and LARAS, stands for library and archive analysis system. However, PDDI LIPI has several problems in developing computer programs for their own needs, among others the limited budget for program development as well as the limited number of human resources. Moreover, the available software engineers do not focus on developing the computer program but also have other assignments. Thus, it often has poor planning and management in the software development projects.

3 Methodology

Figure 4 shows the steps in assessing the references metadata quality in Indonesian journal paper.

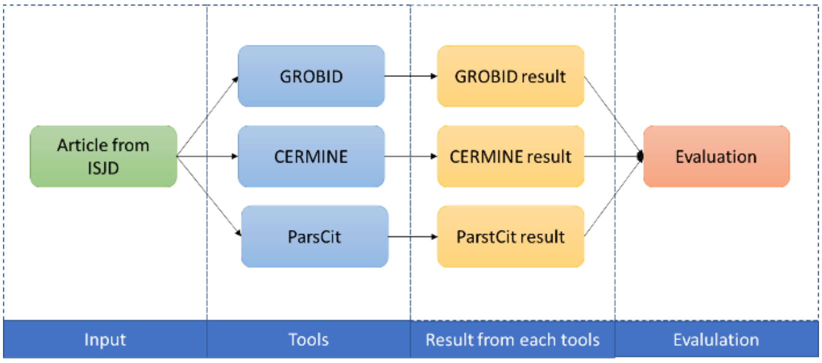
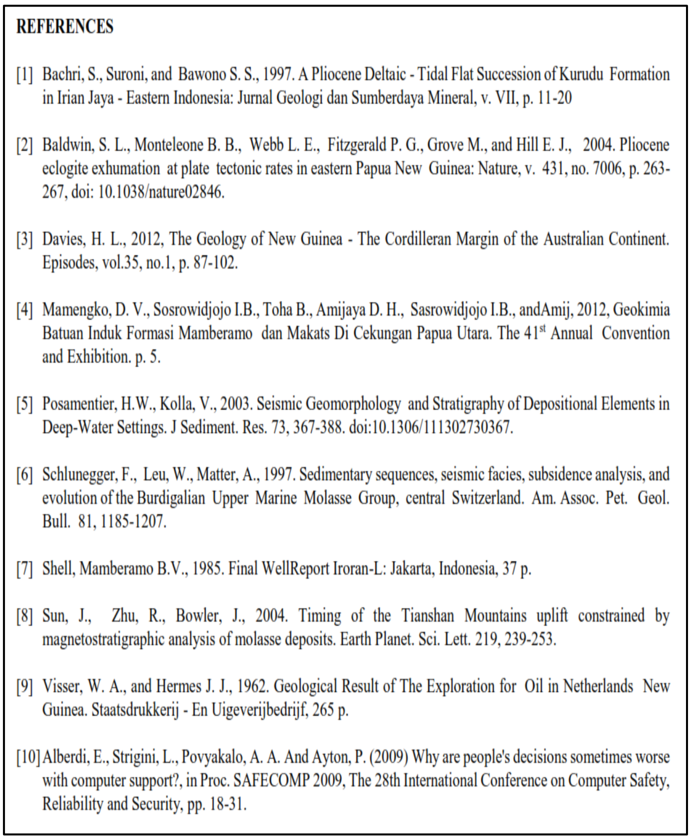


Figure 4: The steps of the bibliographic reference quality assessment of Indonesian journal paper

We employed articles from ISJD for training and testing data. To allow a fair comparison, all tools were retrained using 100 references taken randomly and corrected from the writing errors. For testing data, we selected one article per-cluster (S1-S6) randomly, which is consisted of 88 references. The testing data was divided to two groups namely group of S1 and group of S2-6. We assumed that the S1 group has good bibliographic references, therefore we test the group separately from the other groups. Figure 5 is an example of the reference from an Indonesian scientific article.



REFERENCES

[1] Bachri, S., Suroni, and Bawono S. S., 1997. A Pliocene Deltaic - Tidal Flat Succession of Kurudu Formation in Irian Jaya - Eastern Indonesia: Jurnal Geologi dan Sumberdaya Mineral, v. VII, p. 11-20

[2] Baldwin, S. L., Monteleone B. B., Webb L. E., Fitzgerald P. G., Grove M., and Hill E. J., 2004. Pliocene eclogite exhumation at plate tectonic rates in eastern Papua New Guinea: Nature, v. 431, no. 7006, p. 263-267, doi: 10.1038/nature02846.

[3] Davies, H. L., 2012, The Geology of New Guinea - The Cordilleran Margin of the Australian Continent. Episodes, vol.35, no.1, p. 87-102.

[4] Mamengko, D. V., Sosrowidjojo I.B., Toha B., Amijaya D. H., Sasrowidjojo I.B., andAmij, 2012, Geokimia Batuan Induk Formasi Mamberamo dan Makats Di Cekungan Papua Utara. The 41st Annual Convention and Exhibition. p. 5.

[5] Posamentier, H.W., Kolla, V., 2003. Seismic Geomorphology and Stratigraphy of Depositional Elements in Deep-Water Settings. J Sediment. Res. 73, 367-388. doi:10.1306/111302730367.

[6] Schlunegger, F., Leu, W., Matter, A., 1997. Sedimentary sequences, seismic facies, subsidence analysis, and evolution of the Burdigalian Upper Marine Molasse Group, central Switzerland. Am. Assoc. Pet. Geol. Bull. 81, 1185-1207.

[7] Shell, Mamberamo B.V., 1985. Final WellReport Iroran-L: Jakarta, Indonesia, 37 p.

[8] Sun, J., Zhu, R., Bowler, J., 2004. Timing of the Tianshan Mountains uplift constrained by magnetostratigraphic analysis of molasse deposits. Earth Planet. Sci. Lett. 219, 239-253.

[9] Visser, W. A., and Hermes J. J., 1962. Geological Result of The Exploration for Oil in Netherlands New Guinea. Staatsdrukkerij - En Uigeverijbedrijf, 265 p.

[10] Alberdi, E., Strigini, L., Povyakalo, A. A. And Ayton, P. (2009) Why are people's decisions sometimes worse with computer support?, in Proc. SAFECOMP 2009, The 28th International Conference on Computer Safety, Reliability and Security, pp. 18-31.

Figure 5: An Example of the Bibliography in an Indonesian scientific article

The labelling result of each tool is compared with the ground truth of the metadata fields. Figure 6 shows the example of the labelling conducted by the tools. The evaluation was done also by measuring the precision, recall and F1-measure for each tool. Precision is the ratio of the number of correctly extracted fields to the number of all extracted fields. Recall is the fraction of correctly extracted fields to the number of expected fields. F1 measure is a measure of a test's accuracy of precision and recall.

4 Result and Discussion

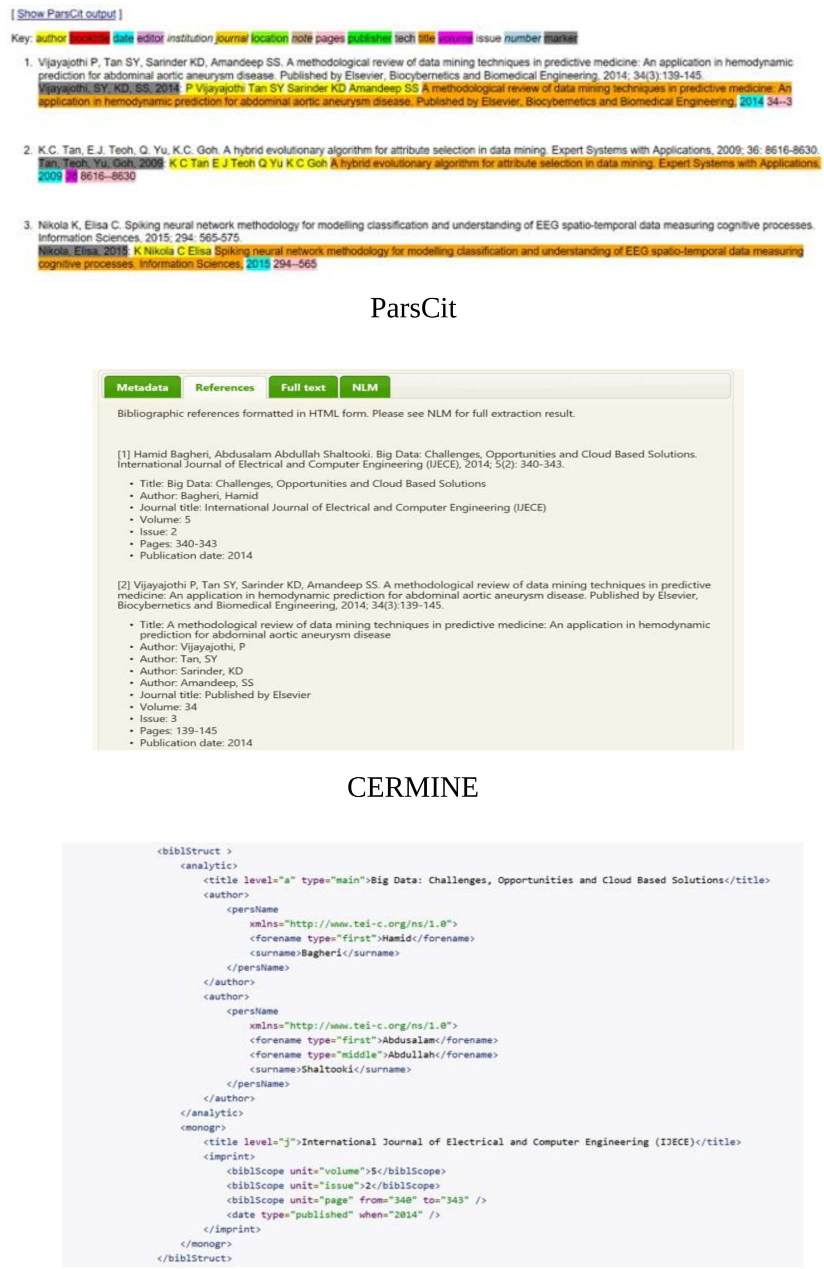
Table 2 shows an example of the S2-6 group that was error in the labeling. There are a lot of mislabeled bibliography metadata. It is indicated also by the low value of recall and F1-measure in the Table 3, 4 and 5. Those table presents the evaluation results for ParsCit, CERMINE, and GROBID respectively. It looks that ParsCit has the best result in the field of author. CERMINE provides the best results in the fields of year, title, volume, and pages. GROBID shows the best for source, volume, and issue.

Figure 7 shows the overall results of the group of S1. ParsCit achieved 0.942, 0.817, and 0.870 respectively. Meanwhile, GROBID achieved 0.976, 0.821, and 0.890 respectively. Next, CERMINE achieved 0.990, 0.905, and 0.944 respectively. Meanwhile Figure 8 presents the overall results of the group S2-6. ParsCit achieved the following results, precision: 0.739, recall: 0.468, and F1 measure: 0.555. Next, GROBID achieved the precision: 0.831, recall: 0.522, and F1 measure: 0.633. In the meantime, CERMINE achieved the precision: 0.811, recall: 0.525, and F1 measure: 0.605.

Table 2: An example of the mislabeled of the group S2-6

Original Source	Actual	ParsCit	CERMINE	GROBID
Vijayajothi P, Tan SY, Sarinder KD, Amandeep SS	Author	Author	Author	Author
A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease	Title	Title	Title	Title
Biocybernetics and Biomedical Engineering	Source	Title	Other	Source
2014	Year	Year	Year	Year
34	Vol	Other	Vol	Vol
3	Issue	Other	Issue	Issue
139-145	Pages	Pages	Pages	Pages
Pribadi, Agus	Author	Author	Author	Title

Perencanaan Sistem Informasi Spasial Program pembangunan Kabupaten Lombok Barat	Title	Detected as title but the value is not correct	Detected as title but the value is not correct	Detected as title but the value is not correct
Proceeding Seminar Nasional, SeNTIA, Politeknik Negeri Malang,	Source	Title	Title	Title
2012	Year	Not detected	Title	Title



[Show ParsCit output]

Key: author booktitle date editor *institution* *journal* location *note* pages publisher tech title volume issue *number* marker

1. Vijayajothi P, Tan SY, Sarinder KD, Amandeep SS. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. Published by Elsevier, Biocybernetics and Biomedical Engineering, 2014; 34(3):139-145.
   Vijayajothi, SY, KD, SS, 2014: P Vijayajothi Tan SY Sarinder KD Amandeep SS A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. Published by Elsevier, Biocybernetics and Biomedical Engineering, 2014 34--3

2. K.C. Tan, E.J. Teoh, Q. Yu, K.C. Goh. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 2009; 36: 8616-8630.
   Tan, Teoh, Yu, Goh, 2009: K C Tan E J Teoh Q Yu K C Goh A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 2009 36 8616--8630

3. Nikola K, Elisa C. Spiking neural network methodology for modelling classification and understanding of EEG spatio-temporal data measuring cognitive processes. Information Sciences, 2015; 294: 565-575.
   Nikola, Elisa, 2015: K Nikola C Elisa Spiking neural network methodology for modelling classification and understanding of EEG spatio-temporal data measuring cognitive processes. Information Sciences, 2015 294--565

ParsCit

Bibliographic references formatted in HTML form. Please see NLM for full extraction result.

[1] Hamid Bagheri, Abdusalam Abdullah Shaltooki. Big Data: Challenges, Opportunities and Cloud Based Solutions. International Journal of Electrical and Computer Engineering (IJECE), 2014; 5(2): 340-343.

- Title: Big Data: Challenges, Opportunities and Cloud Based Solutions
- Author: Bagheri, Hamid
- Journal title: International Journal of Electrical and Computer Engineering (IJECE)
- Volume: 5
- Issue: 2
- Pages: 340-343
- Publication date: 2014

[2] Vijayajothi P, Tan SY, Sarinder KD, Amandeep SS. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. Published by Elsevier, Biocybernetics and Biomedical Engineering, 2014; 34(3):139-145.

- Title: A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease
- Author: Vijayajothi, P
- Author: Tan, SY
- Author: Sarinder, KD
- Author: Amandeep, SS
- Journal title: Published by Elsevier
- Volume: 34
- Issue: 3
- Pages: 139-145
- Publication date: 2014

CERMINE

```
<biblStruct >
    <analytic>
        <title level="a" type="main">Big Data: Challenges, Opportunities and Cloud Based Solutions</title>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0">
                <forename type="first">Hamid</forename>
                <surname>Bagheri</surname>
            </persName>
        </author>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0">
                <forename type="first">Abdusalam</forename>
                <forename type="middle">Abdullah</forename>
                <surname>Shaltooki</surname>
            </persName>
        </author>
    </analytic>
    <monogr>
        <title level="j">International Journal of Electrical and Computer Engineering (IJECE)</title>
        <imprint>
            <biblScope unit="volume">5</biblScope>
            <biblScope unit="issue">2</biblScope>
            <biblScope unit="page" from="340" to="343" />
            <date type="published" when="2014" />
        </imprint>
    </monogr>
</biblStruct>
```

GROBID

Figure 6: The example of the labelling by the tools

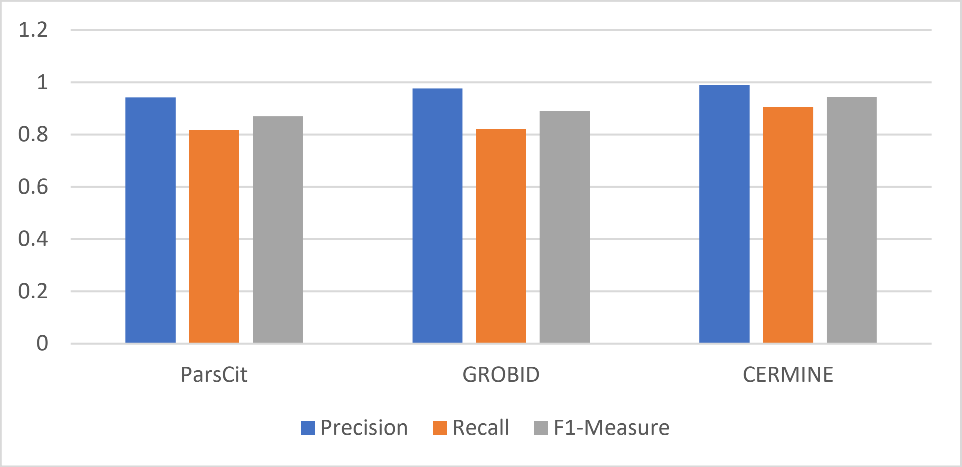


Figure 7: The overall results of the group of S1

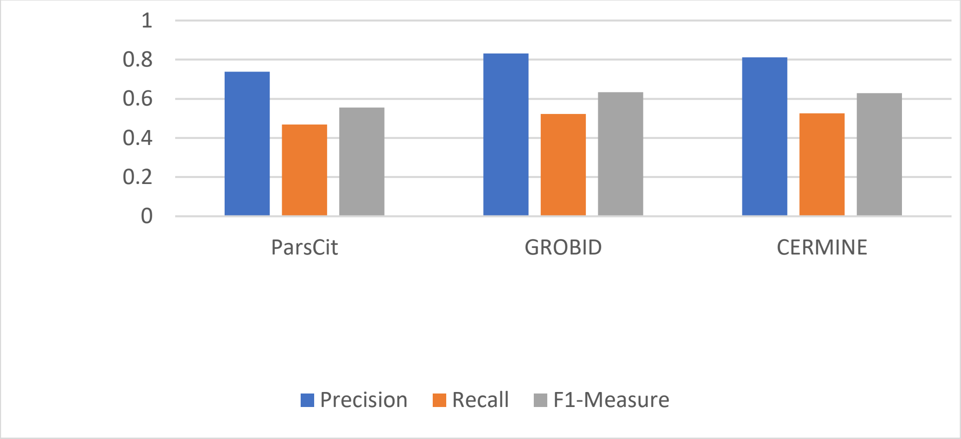


Figure 8: The overall results of the group of S2-6

Table 3: Evaluation results of ParsCit with the Indonesian scientific article SINTA score S2-S6

Field	Precision	Recall	F1-Measure
Author	0.921	0.761	0.833
Year	0.941	0.696	0.800
Title	0.520	0.283	0.366
Source	0.692	0.281	0.400
Volume	0.800	0.400	0.533
Issue	0.500	0.111	0.182
Page	0.800	0.750	0.774

Table 4: Evaluation results CERMINE with the Indonesian scientific article SINTA score S2-S6

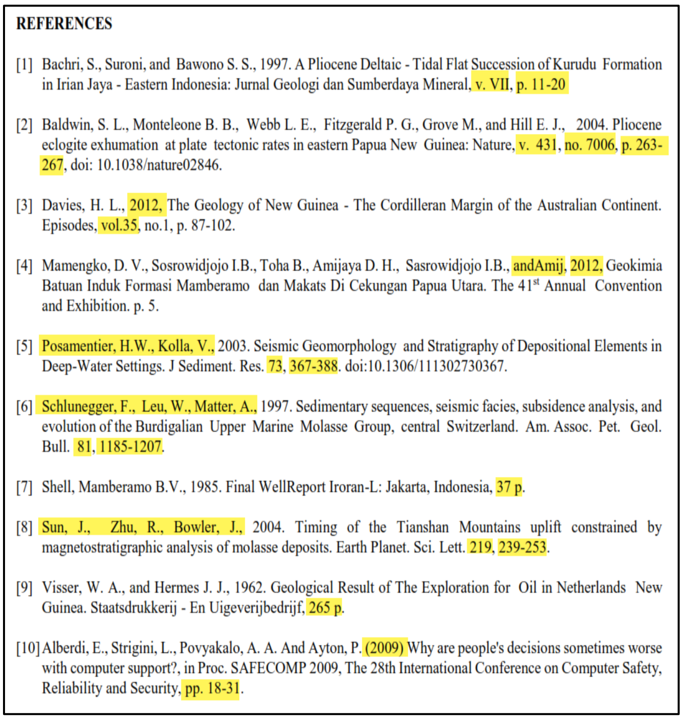
Field	Precision	Recall	F1-Measure
Author	0.935	0.630	0.753
Year	0.971	0.739	0.840
Title	0.828	0.522	0.640
Source	0.588	0.313	0.408
Volume	0.833	0.500	0.625
Issue	0.667	0.222	0.333
Page	0.857	0.750	0.800

Table 5: Evaluation results of GROBID with the Indonesian scientific article SINTA score S2-S6

Field	Precision	Recall	F1-Measure
Author	0.879	0.630	0.734
Year	0.944	0.739	0.829
Title	0.714	0.435	0.541
Source	0.818	0.281	0.419
Volume	0.833	0.500	0.625
Issue	0.800	0.444	0.571
Page	0.833	0.625	0.714

The poor results are caused by the errors of the bibliographic references writing. Figure 9 shows the example of the errors in the Indonesian journals.

- References number 3 and 4 use commas (,), but reference number 10 uses parentheses and the other references use period (.) as a separator between year and title.
- Reference number 4 consists a writing error between ‘and’ and the last author name without space character.
- References number 5, 6, and 7 do not use ‘and’ before the last author like others.
- Volume, issue and page also are written in the different style.



REFERENCES

[1] Bachri, S., Suroni, and Bawono S. S., 1997. A Pliocene Deltaic - Tidal Flat Succession of Kurudu Formation in Irian Jaya - Eastern Indonesia: Jurnal Geologi dan Sumberdaya Mineral, v. VII, p. 11-20

[2] Baldwin, S. L., Monteleone B. B., Webb L. E., Fitzgerald P. G., Grove M., and Hill E. J., 2004. Pliocene eclogite exhumation at plate tectonic rates in eastern Papua New Guinea: Nature, v. 431, no. 7006, p. 263-267, doi: 10.1038/nature02846.

[3] Davies, H. L., 2012, The Geology of New Guinea - The Cordilleran Margin of the Australian Continent. Episodes, vol.35, no.1, p. 87-102.

[4] Mamengko, D. V., Sosrowidjojo I.B., Toha B., Amijaya D. H., Sasrowidjojo I.B., andAmij, 2012, Geokimia Batuan Induk Formasi Mamberamo dan Makats Di Cekungan Papua Utara. The 41st Annual Convention and Exhibition. p. 5.

[5] Posamentier, H.W., Kolla, V., 2003. Seismic Geomorphology and Stratigraphy of Depositional Elements in Deep-Water Settings. J Sediment. Res. 73, 367-388. doi:10.1306/111302730367.

[6] Schlunegger, F., Leu, W., Matter, A., 1997. Sedimentary sequences, seismic facies, subsidence analysis, and evolution of the Burdigalian Upper Marine Molasse Group, central Switzerland. Am. Assoc. Pet. Geol. Bull. 81, 1185-1207.

[7] Shell, Mamberamo B.V., 1985. Final WellReport Iroran-L: Jakarta, Indonesia, 37 p.

[8] Sun, J., Zhu, R., Bowler, J., 2004. Timing of the Tianshan Mountains uplift constrained by magnetostratigraphic analysis of molasse deposits. Earth Planet. Sci. Lett. 219, 239-253.

[9] Visser, W. A., and Hermes J. J., 1962. Geological Result of The Exploration for Oil in Netherlands New Guinea. Staatsdrukkerij - En Uigeverijbedrijf, 265 p.

[10] Alberdi, E., Strigini, L., Povyakalo, A. A. And Ayton, P. (2009) Why are people's decisions sometimes worse with computer support?, in Proc. SAFECOMP 2009, The 28th International Conference on Computer Safety, Reliability and Security, pp. 18-31.

Figure 9: The errors in the bibliographic references writing of a paper in an Indonesian journal

We assumed that authors, who publishing paper in the journal on the group of S2-6 do not employ application for managing references. One of examples, BACA, a journal categorized of S2 published by PDII LIPI, although the managing editor requires the usage of the application, but most authors choose to write references manually. It is because they are not familiar in using the application[1].

Based on the such circumstances in PDDI LIPI described on the related works aforesaid such as a limited budget as well as the number of human resources, we propose to create an application as a liaison between ISJD and references metadata extraction open-source tools. The application serves to identify and correct the mistake in the bibliographic references of paper in ISJD. Output of this application will be input to the open-source applications of the bibliographic reference extraction. This proposal will minimize the need of cost and time. The workflow architecture of the proposed application is provided in Figure 10.

[1] Personal communication to the managing editor of the journal of BACA

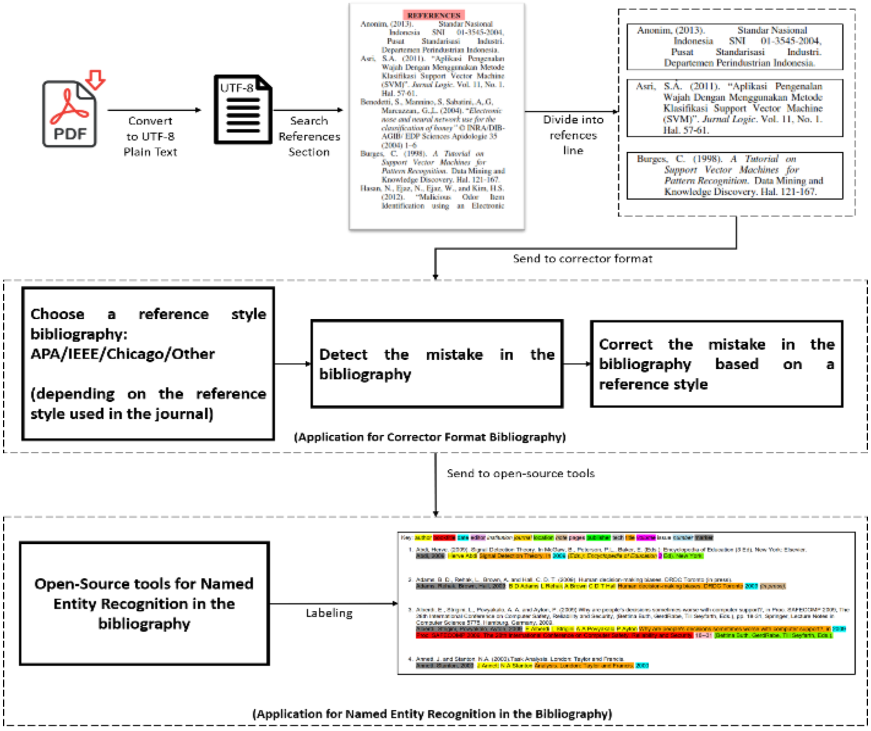

PDF
Convert to UTF-8 Plain Text
UTF-8
Search References Section
REFERENCES
Divide into refences line
Anonim, (2013). Standar Nasional Indonesia SNI 01-3545-2004, Pusat Standarisasi Industri. Departemen Perindustrian Indonesia.
Asri, S.A. (2011). "Aplikasi Pengenalan Wajah Dengan Menggunakan Metode Klasifikasi Support Vector Machine (SVM)". Jurnal Logic. Vol. 11, No. 1. Hal. 57-61.
Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Hal. 121-167.
Send to corrector format
Choose a reference style bibliography: APA/IEEE/Chicago/Other
(depending on the reference style used in the journal)
Detect the mistake in the bibliography
Correct the mistake in the bibliography based on a reference style
(Application for Corrector Format Bibliography)
Send to open-source tools
Open-Source tools for Named Entity Recognition in the bibliography
Labeling
(Application for Named Entity Recognition in the Bibliography)

Figure 10: The Proposed Workflow Architecture

5 Conclusion

PDDI LIPI wants to complete the ISJD with the bibliographic references extraction application to improve the services of the database. Based on the quality of the bibliographic metadata of the paper in ISJD, we propose to build liaison application that can serve the identifying and correcting the bibliography metadata. Thus, the database can employ the open source of bibliographic references extraction. The combination between building and using an open source software would minimalize the need of the cost and time. Next, we will build the application.

Acknowledgment

Thank you for the head of Center for Scientific Data and Documentation – Indonesian Institute of Sciences (PDDI LIPI), Hendro Subagyo, M.Eng, who supports this paper by providing data about ISJD.

References

[1] Z. Guo and H. Jin, "Reference Metadata Extraction from Scientific Papers," in *12th International Conference on Parallel and Distributed Computing, Applications and Technologies*, 2011, pp. 45–49.

[2] M.-Y. Day, R. Tzong-HanTsai, and Cheng-Lung Sung, "Reference metadata extraction using a hierarchical knowledge representation framework," *Decis Support Syst*, pp. 152–167, 2007.

[3] I. Councill, C. Giles, and M.-Y. Kan, "ParsCit: an open-source CRF reference string parsing package," in *International Conference on Language Resources and Evaluation*, 2008.

[4] D. Tkaczyk, S. Paweł, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: automatic

extraction of structured metadata from scientific literature," *Int J Doc Anal Recognit*, vol. 18, no. 4, pp. 317–335, 2015.

[5] P. Lopez, "GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications," *Res Adv Technol Digit Libr*, pp. 473–474, 2009.

[6] D. Matsuoka, M. Ohta, A. Takasu, and J. Adachi, "Examination of Effective Features for CRF-Based Bibliography Extraction from Reference Strings," in *The Eleventh International Conference on Digital Information Management (ICDIM 2016)*, 2016, pp. 243–248.

[7] D. Namikoshi, M. Ohta, A. Takasu, and J. Adachi, "CRF-Based Bibliography Extraction from Reference Strings Using a Small Amount of Training Data," in *The Twelfth International Conference on Digital Information Management (ICDIM 2017)*, 2017, pp. 59–64.

[8] M. Ohta, D. Arauchi, A. Takasu, and J. Adachi, "CRF-based Bibliography Extraction from Reference Strings Focusing on Various," in *10th IAPR International Workshop on Document Analysis Systems*, 2012, pp. 276–281.

[9] M. Ohta, D. Arauchi, A. Takasu, and J. Adachi, "Empirical Evaluation of CRF-Based Bibliography Extraction from Reference Strings," in *11th IAPR International Workshop on Document Analysis Systems*, 2014, pp. 287–292.

[10] B. Ojokoh, M. Zhang, and J. Tang, "A trigram hidden Markov model for metadata extraction from heterogeneous references," *Inf Sci (Ny)*, pp. 1538–1551, 2011.

[11] X. Zhang, J. Zou, D. Le, and G. Thoma, "A Structural SVM Approach for Reference Parsing," in *Ninth International Conference on Machine Learning and Applications*, 2010, pp. 479–484.

[12] E. Suryawati and D. Widyantoro, "Combination of Heuristic, Rule-Based and Machine Learning for Bibliography Extraction," in *5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, 2017, pp. 276–281.

[13] "Anystyle-Parser." [Online]. Available: https://www.rubydoc.info/gems/anystyle-parser. [Accessed: 26-May-2019].

[14] M. Carl Staelin, "Biblio: automatic meta-data extraction," *Int J Doc Anal Recognit*, vol. 10, no. 2, pp. 113–126, 2007.

[15] C.-C. Chen, K.-H. Yang, C.-L. Chen, and J.-M. Ho, "BibPro: A Citation Parser Based on Sequence Alignment," *IEEE Trans Knowl Data Eng*, vol. 24, no. 2, pp. 236–250, 2012.

[16] T. Nishimura, "Parse Citation List in Paper," 2016. [Online]. Available: https://github.com/nishimuuu/citation.

[17] M. Romanello, "citation-parser 0.4.1," 2017. .

[18] P. Lopez, "GROBID from PDF to structured documents," 2015. [Online]. Available: https://grobid.readthedocs.io/en/latest/grobid-04-2015.pdf. [Accessed: 20-May-2019].

[19] PDDI, "Statistik Jumlah Artikel," 2018. [Online]. Available: http://isjd.pdii.lipi.go.id/.

[20] J. Fagan and J. Keach, "Build, buy, open source, or web 2.0? making an informed decision for your library," *Comput Libr*, pp. 8–11, 2010.