# Identifying and Correcting the Indonesian Bibliography Metadata using Regular Expression

Ariani Indrawati[1], Ambar Yoganingrum[2], and Hendro Subagyo[3]

[1,3]*Center for Scientific Data and Documentation, Indonesian Institute of Sciences, Jakarta, Indonesia*
[2]*Research Center for Informatics, Indonesian Institute of Sciences, Cibinong, Indonesia*

**ABSTRACT**
*Bibliographical reference extraction is essential to make networking of the scientific document. However, most of the bibliographic references of Indonesian journals are written not according to the rules. Therefore, the usage of the open-source automated bibliographic reference extraction tools gives the unwell results. This paper proposes an instrument to improve the quality of the bibliographical reference metadata of the articles in Indonesian journals. We apply regular expressions (commonly known as "RegEx") to find writing errors in the bibliographic references then correct them according to the rule. The experimental results show that the tool performs well, with the correct percentage of 85%.*

**Keywords :** *Automatic extraction bibliography, Indonesian scientific journals, Regular Expression.*

## I. INTRODUCTION

Currently, the development of a tool to extract scientific references metadata automatically becomes an essential issue. The device can develop a bibliographical references networking, which is useful to analyze citation and provide the related recommendations as a service of a scientific database. Some researchers have created several applications for references metadata extraction, such as CERMINE, ParsCit, and GROBID, the three open-source tools discussed the most. CERMINE had capacity to extract a large dataset of the most metadata types with the average F score of 77.5% (Tkaczyk et al., 2015). Meanwhile, ParsCit demonstrated a significant advantage for the aspects of language and the multilingual data (Prasad, Kaur and Kan, 2018). In the meantime, GROBID showed a decent level of accuracy of 95.7% per citation field and 78.9% per citation instance based on the CORA dataset (Lopez, 2009).

Using open-source tools is a suitable choice for libraries that has less money but dependable in technical staff support (Fagan and Keach, 2010). However, there is also a barrier in implementing the open-source tool, especially the usage of the automated bibliographical references extraction tools. The application of the tools to extract the reference metadata of Indonesian journals showed low performances. It is caused by inconsistency in writing the references. Most authors, who published a paper in the Indonesian journal, seem writing the references manually. A reference manager has been socialized since 2012 by the Indonesian Ministry of Research, Technology, and Higher Education (Kemenristekdikti). However, the quality of the most references metadata is still low.

This paper proposes a tool built by Regular Expression or often referred to as RegEx, to improve the references metadata that not follow the rules. By improving the quality of the metadata, then the institution can apply the open-source reference metadata extraction tool. Even though the libraries want to build the appliance by themselves, they still have to repair the quality of the reference metadata, especially the old collections. This paper offers an integrating approach between an open-source and a self-built tool for software development in the organizations. This approach has a benefit, namely fitting the need of the organizations with less money and time.

The contributions of this paper are showing that RegEx can build an intermediary tool for increasing the quality of the references metadata. Others are inspiring to construct an intermediate device that needs less money and time in building and offering an integrating approach for the component sourcing options in a library.

A reference list is the sources referred to in an article. Each reference list has a specific format, for example, American Psychological Association (APA) style that has the following pattern: Author, A. (Publication Year). Article title. Periodical Title, Volume(Issue), pp-pp. We choose RegEx since it is the most extensive tool used in pattern analysis

## II. RELATED WORKS

Stephen Cole Kleene, a mathematician, created the concept of Regular Expressions to follow up on the idea of McCulloch and Pitts in developing a model to studying the behavior of the nervous systems (Leung, 2010). The definition of Regular Expressions explains about the regular events, which is called regular languages in modern textbooks. Regular Expressions describe natural languages using mathematical notation.

RegEx is a formula for searching patterns of a sentence or string. Many instruments, such as word processors, text editors, and other tools, used RegEx to find and manipulate sentences based on a specific pattern. RegEx is very powerful; at a low level, RegEx can search for a fragment of words. At high levels, RegEx can control over data management, like editing, input, or delete. Many programming languages support regular expressions such as PHP, Perl, VB, Java, Python, and many more.

Currently, Some researchers are continuing developing RegEx. Chang, Li, and Chen (2015) proposed the techniques of compression and pattern segmentation for memory usage efficiency when processing the multiple regular expressions jointly. Meanwhile, Wang et al. (2014) proposed the method of rooted that can transform the given large-scale set of complex RegEx into a compact and fast matching engine. In the meantime, Medeiros, Mascarenhas, and Ierusalimschy (2014) optimized the RegEx to parsing expression grammars

## III. METHODOLOGY

The tool developed only detects the reference lines using APA style. The style is the reference and citation format most widely used by local journals. Figure 1 shows the illustration of identifying and correcting the Indonesian bibliography metadata process. It starts from taking an incorrect form of a reference line as input to the corrector tool.

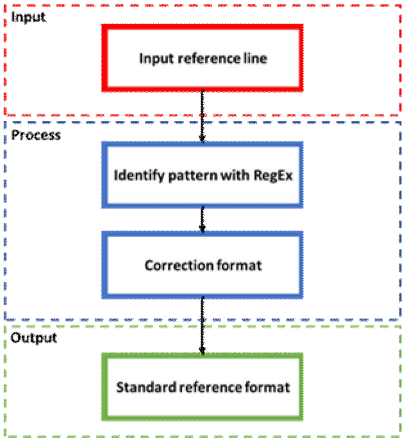
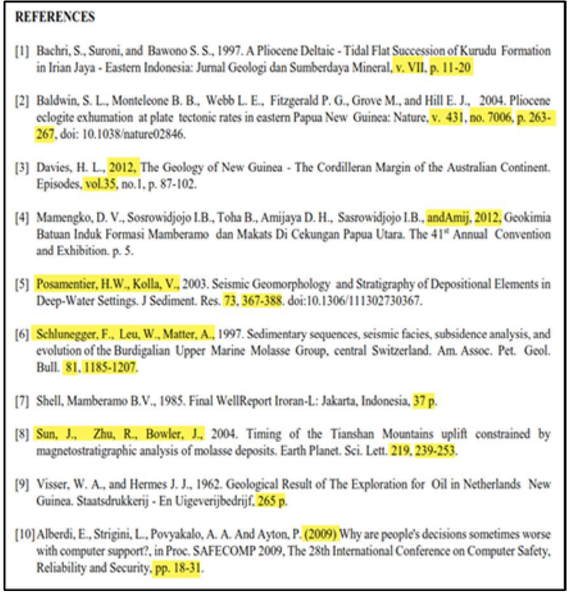


**Fig. 1**. The steps in identifying and correcting the Indonesian bibliography metadata

Figure 2 shows examples of input reference lines, which consist of following various kinds of mistakes.

- 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 of a writing error between 'and' and the last author's name without space character.
- References numbers 5, 6, and 7, do not use 'and' before the last author, like others.
- Volume, issue, and page also are written in a 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.

**Fig. 2.** The example of various kinds of mistakes of the input

Those inputs will proceed to identify patterns with the RegEx pattern. The tools will identify and extract author, year, article title, journal title, volume number, issue number, and page number from each reference line. Then, the tool will change into the correct APA format. The illustration of this process shown in Figure 3 using a without separator reference line example.

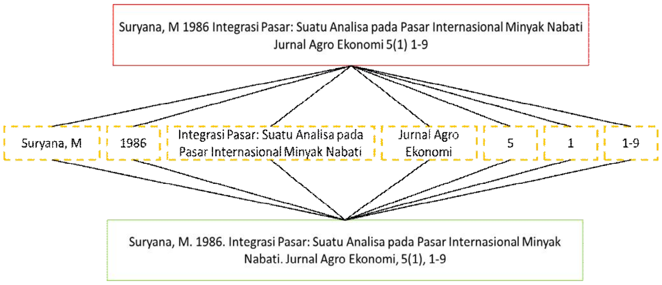

Suryana, M 1986 Integrasi Pasar: Suatu Analisa pada Pasar Internasional Minyak Nabati Jurnal Agro Ekonomi 5(1) 1-9
Suryana, M
1986
Integrasi Pasar: Suatu Analisa pada Pasar Internasional Minyak Nabati
Jurnal Agro Ekonomi
5
1
1-9
Suryana, M. 1986. Integrasi Pasar: Suatu Analisa pada Pasar Internasional Minyak Nabati. Jurnal Agro Ekonomi, 5(1), 1-9

**Fig. 3**. The illustration process of identification and correction

This paper built a tool on Python and Regular Expression (RegEx) to identify the pattern from the input. If there are error formats in reference lines, then the device will automatically correct those reference lines to a standard reference format or style.
As a starting test, we tried 100 reference lines taken randomly from several Indonesian local journals that contain various kinds of mistakes. A formula (1) was employed to evaluate reference lines.

$$Percentage\ Correct = \frac{correct\ output}{total\ reference\ lines}\ x\ 100\% \qquad (1)$$

## IV. RESULTS AND DISCUSSION

Figure 4 shows that the tool managed to fix 85 of 100 reference metadata lines:

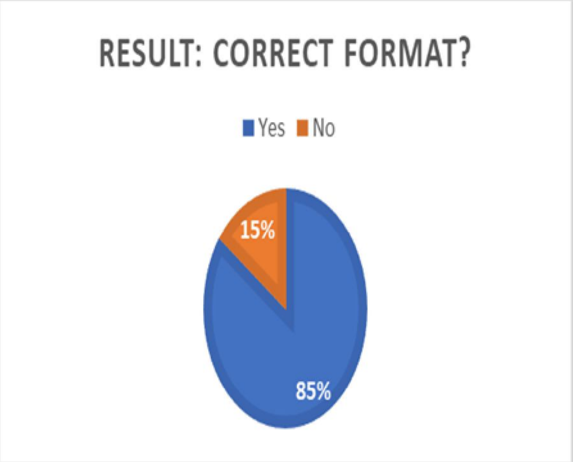
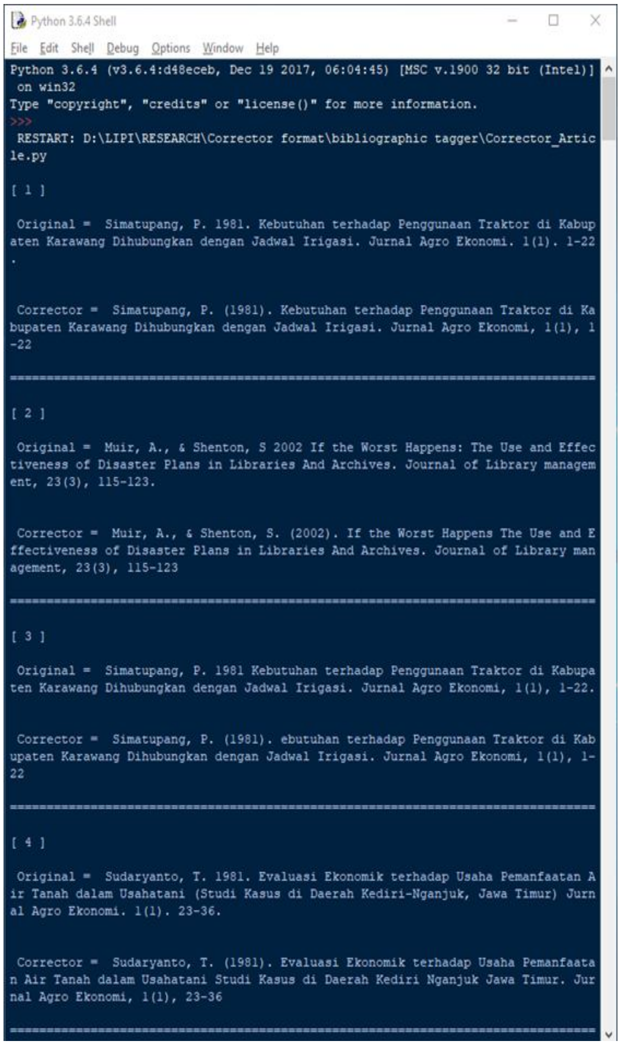


**Fig.4**. The tool performance

Meanwhile, figure 5 shows the process of the correction reference format system in Python.



```
Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)]
 on win32
Type "copyright", "credits" or "license()" for more information.
>>> 
 RESTART: D:\LIPI\RESEARCH\Corrector format\bibliographic tagger\Corrector_Artic
le.py

[ 1 ]

 Original =  Simatupang, P. 1981. Kebutuhan terhadap Penggunaan Traktor di Kabup
aten Karawang Dihubungkan dengan Jadwal Irigasi. Jurnal Agro Ekonomi. 1(1). 1-22
.


 Corrector =  Simatupang, P. (1981). Kebutuhan terhadap Penggunaan Traktor di Ka
bupaten Karawang Dihubungkan dengan Jadwal Irigasi. Jurnal Agro Ekonomi, 1(1), 1
-22

================================================================================

[ 2 ]

 Original =  Muir, A., & Shenton, S 2002 If the Worst Happens: The Use and Effec
tiveness of Disaster Plans in Libraries And Archives. Journal of Library managem
ent, 23(3), 115-123.


 Corrector =  Muir, A., & Shenton, S. (2002). If the Worst Happens The Use and E
ffectiveness of Disaster Plans in Libraries And Archives. Journal of Library man
agement, 23(3), 115-123

================================================================================

[ 3 ]

 Original =  Simatupang, P. 1981 Kebutuhan terhadap Penggunaan Traktor di Kabupa
ten Karawang Dihubungkan dengan Jadwal Irigasi. Jurnal Agro Ekonomi, 1(1), 1-22.


 Corrector =  Simatupang, P. (1981). ebutuhan terhadap Penggunaan Traktor di Kab
upaten Karawang Dihubungkan dengan Jadwal Irigasi. Jurnal Agro Ekonomi, 1(1), 1-
22

================================================================================

[ 4 ]

 Original =  Sudaryanto, T. 1981. Evaluasi Ekonomik terhadap Usaha Pemanfaatan A
ir Tanah dalam Usahatani (Studi Kasus di Daerah Kediri-Nganjuk, Jawa Timur) Jurn
al Agro Ekonomi. 1(1). 23-36.


 Corrector =  Sudaryanto, T. (1981). Evaluasi Ekonomik terhadap Usaha Pemanfaata
n Air Tanah dalam Usahatani Studi Kasus di Daerah Kediri Nganjuk Jawa Timur. Jur
nal Agro Ekonomi, 1(1), 23-36

================================================================================
```

**Fig. 5**. The output correction tool

The tool fixes the various kinds of mistakes, such as remove or add period, add parentheses, and correct the words. Table 1 shows the types of corrections of the reference metadata lines.

**Table I**: The Kinds Of Correction By The Tool

| Types of Corrections | Original | Correction Output |
|---|---|---|
| Add parentheses in year | Sutardji. 2003. Pola Sitiran dan Pola Kepengarangan pada Jurnal Penelitian Pertanian Tanaman Pangan. Jurnal Perpustakaan Pertanian, 12(1),1–9. | Sutardji. (2003). Pola Sitiran dan Pola Kepengarangan pada Jurnal Penelitian Pertanian Tanaman Pangan. Jurnal Perpustakaan Pertanian, 12(1),1–9. |

| | | |
|---|---|---|
| Change colons to commas after issue number | Soehardjan, M. 2000. Pengertian tentang mutu karya tulis ilmiah. Jurnal Perpustakaan Pertanian, 9(1): 18-21. | Soehardjan, M. 2000. Pengertian tentang mutu karya tulis ilmiah. Jurnal Perpustakaan Pertanian, 9(1), 18-21. |
| Add period after year | Rusydi, I. 2014 Pemanfaatan E-Journal Sebagai Media Informasi Digital. Jurnal Iqra', 8(2), 200–210. | Rusydi, I. 2014. Pemanfaatan E-Journal Sebagai Media Informasi Digital. Jurnal Iqra', 8(2), 200–210. |
| Change period to comma after issue number | Sudaryanto, T. 1981. Evaluasi Ekonomik terhadap Usaha Pemanfaatan Air Tanah dalam Usahatani (Studi Kasus di Daerah Kediri-Nganjuk, Jawa Timur) Jurnal Agro Ekonomi. 1(1). 23-36. | Sudaryanto, T. (1981). Evaluasi Ekonomik terhadap Usaha Pemanfaatan Air Tanah dalam Usahatani Studi Kasus di Daerah Kediri Nganjuk Jawa Timur. Jurnal Agro Ekonomi, 1(1), 23-36. |
| Add comma after journal title | Drestya, & Dyane, A. 2013. Motif Menggunakan Sosial Media Path pada Mahasiswa di Surabaya. Jurnal Commmonline Departemen Komunikasi 3(3), 530–536. | Drestya, & Dyane, A. 2013. Motif Menggunakan Sosial Media Path pada Mahasiswa di Surabaya. Jurnal Commmonline Departemen Komunikasi, 3(3), 530–536. |
| Add a comma after issue number | Briggs, J., &Ferrucci, M.T. 2009. The development, cost, and impact of a current awareness service in an industrial organization. Journal of Chemical Documentation, 11(2) 72-75. | Briggs, J., &Ferrucci, M.T. 2009. The development, cost, and impact of a current awareness service in an industrial organization. Journal of Chemical Documentation, 11(2), 72-75. |
| Add a period after author-name | Wilson, T.D 1981. On User Studies and Information Needs. Journal of Documentation, 37 (1), 3 – 15. | Wilson, T.D, 1981. On User Studies and Information Needs. Journal of Documentation, 37 (1), 3 – 15. |
| Change the comma to period after a year | Georgas, H. 2015, Google Vs the Library (Part III): Assessing the Quality of Sources Found by Undergraduates. Portal: Libraries and the Academy, 15 (1), 133–161. | Georgas, H. 2015. Google Vs the Library (Part III): Assessing the Quality of Sources Found by Undergraduates. Portal: Libraries and the Academy, 15 (1), 133–161. |
| Remove the word 'Vol' | Bryan, J. E. 2016. The Preparation of Academic Librarians Who Provide Instruction: A Comparison of First and Second Career Librarians. Journal of Academic Librarianship, Vol. 42(4), pp. 340–354. | Bryan, J. E. 2016. The Preparation of Academic Librarians Who Provide Instruction: A Comparison of First and Second Career Librarians. Journal of Academic Librarianship, 42(4), 340–354. |
| Change the comma to period after the title | Bodic, V.B. 2015 A computerized current awareness service using Chemical-Biological Activities (Cbac), Journal of Chemical Documentation 9(3):158-161. | Bodic, V.B. 2015. A computerized current awareness service using Chemical-Biological Activities (Cbac). Journal of Chemical Documentation, 9(3),158-161. |
| Change period to a comma after journal title | Bustamam, M., Reflinur, R., Agisimanto, D., & Suyono, S. 2004. Variasi genetic padi tahan blas berdasarkan sidik jari DNA dengan markah gen analog resisten. Jurnal Bioteknologi Pertanian. 9(2): 56-61. | Bustamam, M., Reflinur, R., Agisimanto, D., & Suyono, S. 2004. Variasi genetic padi tahan blas berdasarkan sidik jari DNA dengan markah gen analog resisten. Jurnal Bioteknologi Pertanian, 9(2), 56-61. |
| Remove comma after author-name | Suhairi, K., & Gaol, F. L., 2013. The Measurement of Optimization Performance of Managed Service Division with ITIL Framework using Statistical Process Control. Journal of Networks, 8(3), 518-529. | Suhairi, K., & Gaol, F. L. 2013. The Measurement of Optimization Performance of Managed Service Division with ITIL Framework using Statistical Process Control. Journal of Networks, 8(3), 518-529. |

| | | |
|---|---|---|
| Remove word 'pp' | Bryan, J. E. 2016. The Preparation of Academic Librarians Who Provide Instruction: A Comparison of First and Second Career Librarians. Journal of Academic Librarianship, Vol. 42(4), pp. 340–354. | Bryan, J. E. 2016. The Preparation of Academic Librarians Who Provide Instruction: A Comparison of First and Second Career Librarians. Journal of Academic Librarianship, 42(4), 340–354. |

The tool cannot fix 15 of 100 reference lines. Table 2 shows samples from each similar original format.

**Table II**: The List Of Unfixed Reference Lines

| Original | Output |
|---|---|
| Ukachi, N. B. 2010. Library and information science professionals and skills for the electronic information environment. Journal of Library and Information Science, 7 (1 & 2), 160-168 | Error (cannot read the pattern) Containing two issues |
| Aliyu, Murtala. 2011. Author Productivity and Colaboration Among Academic Scientists in Modibbo Adama University of Technology, Yola. The Information Manager, 11(1&2): 32-35. | Error (cannot read the pattern) Containing two issues |
| Sooryamoorthy, Radhamany. 2013. Scientific Research in The Natural Sciences in South Africa: A Scientometric Study. Scientific research in natural sciences, 109 (7/8), 1-11. | Error (cannot read the pattern) Containing two issues |
| Simatupang, P., & Ariani, M. 1987. Analisa Permintaan Waktu Luang Keluarga Petani PIR-Karet NES I Talang Jaya Sumatera Selatan. Jurnal Agro Ekonomi, 6(1-2), 83-93. | Error (cannot read the pattern) Containing two issues |
| Chen, Shih-chuan. 2014. Information Needs and Information Sources of Family Caregivers of Cancerpatients. Journal of Information Management, 66:6, 623-639. | Error (cannot read the pattern) There is colon before issue number |
| McKenzie, P.J. 2003. A Model of Information Practices in Accounts of Everyday-Life Information Seeking. Journal of Documentation, Vol.59, No.1, 19 – 40 | Error (cannot read the pattern) 'no' before issue number |
| Arianto, A., Budiman, N., & Nurhaedah, N. 2014. Analysis of Acid Content of Cyanide (HCN) at Koro Sword Beans (Canavalia ensiformis) Using Different Old Immersion NaCL. J. Galung Tropika. | Error (cannot read the pattern) There is no information about volume, issue number, and pages |
| Soep. 2011. Penerapan Edinburgh Post-Partumdepression Scale sebagai Alat Deteksi Risiko Depresi Nifas pada Primipara dan Multipara. Jurnal Keperawatan Indonesia, Vol.14, pp. 95 – 100. | Error (cannot read the pattern) There is no information about issue number |
| Todorinova, L. 2015. Wikipedia and Undergraduate Research Trajectories. New Library World, 201–212. | Error (cannot read the pattern) There is no information about the volume and issue number |

We studied that the following reasons cause errors in the tool:

- The information in the references line is not completed, for example, not availability the info about volume number, issue number, or pages.

- The tool does not work, especially if there is an error in writing the volume number and issue number. The device works well in correcting information about the author, year of publication, publication title, and journal name.

## V. CONCLUSION

This paper shows that the tool powered by RegEx is a potential tool to improve the quality of reference lines. This research has a weak, only use a few data to test the device. Based on the Indonesian Scientific Journal Database (ISJD) there were 283,000 articles published between 2009 to 2018 in local journals (PDDI, 2018). In the future, we will use many more data in testing and increase the performance of the tool. We also will develop the capability of the instrument. Therefore it can analyze other references format.

## REFERENCES

[1] Chang, Y.-K., Li, Y.-S. and Chen, Y.-T. (2015) "*A Memory Efficient DFA Using Compression and Pattern Segmentation*", Procedia Computer Science, 56, pp. 292–299. doi: 10.1016/j.procs.2015.07.211.

[2] Fagan, J. and Keach, J. (2010) "*Build, buy, open source, or web 2.0? making an informed decision for your library*", Computer in Libraries, pp. 8–11.

[3] Leung, H. (2010) Regular Languages and Finite Automata. Available at: https://web.archive.org/web/20131205193130/https://www.cs.nmsu.edu/historical-projects/Projects/kleene.9.16.10.pdf (Accessed: 26 August 2019).

[4] Lopez, P. (2009) "*GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications*", Research and Advanced Technology for Digital Libraries, pp. 473–474.

[5] Medeiros, S., Mascarenhas, F. and Ierusalimschy, R. (2014) "*From RegExes to parsing expression grammars', Science of Computer Programming*." Elsevier B.V., 93, pp. 3–18. doi: 10.1016/j.scico.2012.11.006.

[6] PDDI (2018) Statistik Jumlah Artikel. Available at: http://isjd.pdii.lipi.go.id/.

[7] Prasad, A., Kaur, M. and Kan, M.-Y. (2018) "*Neural ParsCit: a deep learning-based reference string parser***"**, International Journal on Digital Libraries, 19(4), pp. 323–337. doi: 10.1007/s00799-018-0242-1.

[8] Tkaczyk, D., Paweł, S., Fedoryszak, M., Dendek, P. J. and Bolikowski, Ł. (2015) "*CERMINE: automatic extraction of structured metadata from scientific literature*", International Journal on Document Analysis and Recognition, 18(4), pp. 317–335.

[9] Wang, K., Fu, Z., Hu, X. and Li, J. (2014) "*Practical regular expression matching free of scalability and performance barriers*", Computer Communications. Elsevier B.V., 54, pp. 97–119. doi: 10.1016/j.comcom.2014.08.005