The Development of Meta data Extractor Plugin for Open Journal System

[1]Muhammad Isyak Rizqi , [2]Moch. Zawaruddin Abdullah , [3]Muhammad Afif Hendrawan

Department of Information Technology
Politeknik Negeri Malang
Malang, Indonesia
[1]1841720054@student.polinema.ac.id, [2]zawaruddin@polinema.ac.id,
[3]afif.hendrawan@polinema.ac.id

Abstract Scientific articles are scholarly publications that present objective and verifiable information pertaining to study findings and literature reviews. These articles are typically disseminated through reputable scientific journals. The Open Journal System (OJS) is an open-source platform that facilitates the online publication of scholarly journals. The Open Journal Systems (OJS) platform offers a systematic process for submitting articles, which includes a specific stage dedicated to inputting metadata. The author manually conducts the process of completing the various stages involved in filling in the meta data. This results in a decrease in the overall effectiveness of this stage and introduces a potential danger of inaccuracies in the appropriateness of the entered information. In this scenario, there is a requirement for the implementation of novel functionalities on the OJS platform to facilitate the automated extraction and population of meta data pertaining to scientific papers that have been submitted. The rule-based approach is a viable method for extracting information from scientific journals. The extracted metadata comprises several elements such as the title, author's name, author's affiliation, author's email, abstract, and keywords. The literary type and structure employed may include the utilization of typographical enhancements such as bold text, varying text sizes, text prefixes, and similar elements. The approach yields a precision rate of 95% in the extraction of articles. The developed plugin has demonstrated a significant improvement in the efficiency of populating meta data for scientific papers, achieving a time reduction of 3.72 times faster compared to the manual approach.

Keywords: meta data, ojs, rule-based, development

1. INTRODUCTION

Scientific articles are scholarly publications that present information derived from research reports and literature studies, typically disseminated through scientific journals. The dissemination of scientific articles can occur through either print or internet

M. U. H. Al Rasyid and M. R. Mufid (eds.), *Proceedings of the International Conference on Applied Science and Technology on Engineering Science 2023 (iCAST-ES 2023)*, Advances in Engineering Research 230,
https://doi.org/10.2991/978-94-6463-364-1_90

mediums. Based on the statistics obtained from Scopus, it is evident that the publication of scientific publications, particularly in Indonesian journals, has been consistently increasing on an annual basis [1]. The Open Journal System (OJS) is an open-source platform that facilitates the online publication of scientific journals [2]. The OJS article delivery channel facilitates the process for journal authors or administrators to electronically submit scientific publications, while also providing a platform to input the necessary details for subsequent online publication in the journal. The essential data comprises the meta data of the article encompassing the article's title, the author's identity, abstracts, and keywords [1].The scientific article completion phase and the submission of the article to OJS are still handled manually [2]. This occurrence reduces the effectiveness of the phase and makes it susceptible to errors in the completion of metadata by authors, journal management, or journal editors. Meta data extraction and filling can be automated for the purpose of optimizing the information sifting process [1]–[4]. The process of automatic meta data extraction can be categorized into two categories based on their methodologies. The first one is machine learning method [1], [5], and the second one is other method which mostly use rule-based approach [2], [3], [6]–[8]. The machine learning approach may have great potential in term of accuracy. However, it requires a lot of computation power to achieve desired result. Therefore, the rule-based approach mostly chosen to achieve low computational power that suitable for real time meta data extraction. A rule-based approach is one that can be used to extract meta data information from a document. One example of a rule-based approach is the rule - based text parsing method. The rule – based text Parsing method is used to convert articles in Portable Document Format (PDF) format to plain text format. The properties present on the scientific article text pattern will then be mapped to create a set of meta data extraction rules [2]. A similar study conducted by [7], carried out a meta extraction of a PDF-format book by converting a PDF document into a Raw text form using the PDFBox. After that, a rule was compiled based on lines, pages and prefixes of the information contained in the PDF Book. In the study by [3], scientific articles were converted into a Rich-text document to preserve the attributes present in the text. Rules were created in the form of association rules based on the writing format as well as the attribute of the text found in the scientific article. Further research by [9], proposed feature centric techniques for extraction of the logical layout structure of an article based on various types of composition styles from the publisher. Then to extract meta data from each article placed on the logical layout structure (LSS), developed the four-step approach “FLAG-PDFe”. This approach takes advantage of textual and geometric information contained in scientific articles. These rule-based approaches have relatively high accuracy and performance values, but the accurate result of the extraction. process depends on the rule set as well as the one used in the meta data extraction [2], [3]. The process of extraction of meta data on scientific articles on the OJS platform is then carried out automatically by the system. It does the process of extraction of meta data on scientific articles because of simplifying and accelerating the reading process and filling in the details of articles uploaded to OJS. So the burden of journal editors in determining the reviewer can be easier and faster. Methods that can be used to extract information using heuristic-based methods where rules are structured in the form of association rule-base. A rule-based approach is an approach by conducting critical analysis of data sets to find patterns or relationships that exist in such data sets [10]. The pattern used based on the writing format as well as the writing sequence in a scientific article to compile a set of rules. This makes it easier for writers, journal managers or journal editors to submit articles and reduces the possibility of errors or mismatches. Development of a new feature for extracting meta data on a

plugin-based scientific article that can be applied to the Open Journal System (OJS). This feature automatical ly extracts the meta data or information contained in the up

loaded scientific article. The information collected included the title, abstract, author's name, email, affiliate, and keyword of the uploaded scientific article.

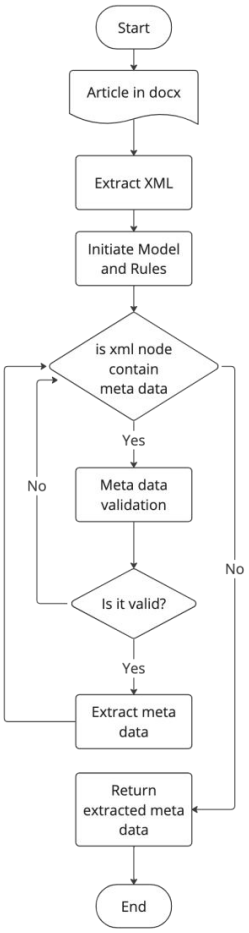

Fig. 1. Proposed meta data extraction process

II . **Methodology**

A. Data Collection

The data is obtained by downloading a collection of scientific articles contained in the Jurnal Informatika Polinema (JIP). The data collection process here is aimed at obtaining examples of articles to be submitted by the journal administrator or author. The data used in this study is a scientific paper taken from the period from 2016 to 2021. All of the documents written in Bahasa Indonesia.

B. Data Processing

The data processing process begins with extracting the XML document contained in the DOCX document. In this case, the extraction process can be done using the ZipArchive extension available in the PHP programming language. XML document extraction aims to maintain the writing format of the scientific article so that the writing style as well as the writing sequence of a scientific c article can be known based on the use of attributes or tags on each XML document node.

The XML document that has been extracted is subsequently subjected to metadata extraction. The process of metadata extraction involves employing a rule-based method to search for certain textual content inside a scientific publication, with the aim of ascertaining whether the identified text corresponds to the desired information. During the metadata extraction stage, the process verifies each line of node that is loaded in the XML document based on the defined rule or rules. The recognized text or information is subsequently processed to extract text based on the discovered metadata. Fig. 1 depicts the sequential flow of the process involved in extracting meta data.

Rules are formed based on writing characteristics that include the writing sequence as well as the format used in the text. The rules are structured in the form of association rules and then written in programming languages. The meta data to be extracted includes title, author name, author affiliate, author email, abstract, and keyword. Table 1 shows the proposed rules to extract meta data.

TABLE 1 . PROPOSED RULES

Meta data	Characteristics
Title	Written in the largest size and printed thick
Author's Name	The author's name is written flat in the middle of paper layout and printed thick.
Author's Affiliation	It's written medium and there's the word "institut", "universitas" or "politeknik".
Author's email	There is an @ character followed by a domain name and written in the middle of paper layout
Abstract	begins with the words "abstrak", "abstracts" or "initisari" and ends with the word "keyword", or "kata kunci" with thick prints.
Keywords	Begins with the word "keyword" or "kata kunci"; each keyword is separated by a comma sign (,); ends with a word "Pendahuluan" with a thick print.

C. Plugin Implementation

The implementation of the scientific article metadata extraction system on the Open Journal System (OJS) platform takes the form of a plugin. The plugin that has been built will subsequently be implemented within the scientific paper submission process. The plugin functions by introducing an additional step in the workflow, specifically the extraction of meta data following the sender's completion of the article upload procedure. Afterwards, the outcomes of the meta data extraction will be transmitted back to

the OJS platform for presentation during the metadata completion procedure. Fig. 2 illustrates the workflow of the process plugin that has been developed.

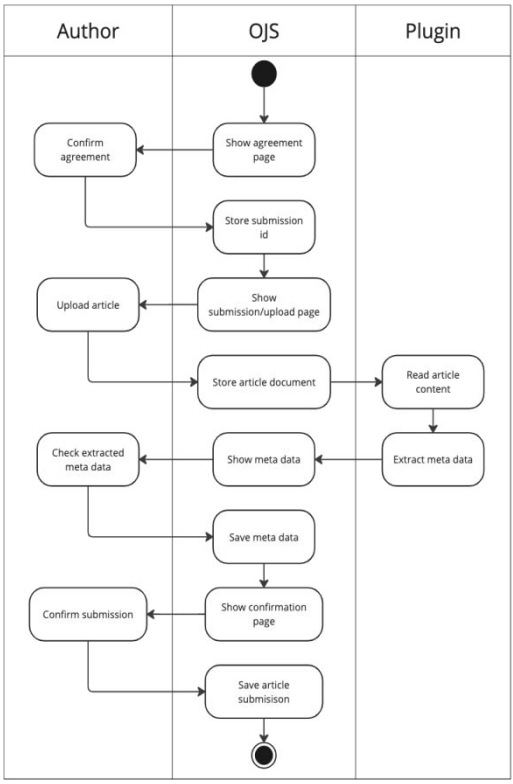

Fig. 2. Meta data extractor plugin workflow

TABLE II TIME EFFICIENCY TEST RESULT

No	Time	
	Automatic	**Manual**
1	00.11.19	00.47.25
2	00.07.18	00.29.54
3	00.14.35	00.53.47
4	00.10.48	00.46.28
5	00.20.08	01.00.58
Mean	00.12.50	00.47.42

III Result

A. System Accuracy

The test data used was a scientific article taken from the JIP. The data used amounted to 50 articles with a total of 300 meta data to be extracted. Then based on the results obtained, the accuracy value of the meta data extraction can be calculated Equation (1).

$$Accuracy = \frac{Correctly\ Extracted\ Meta\ data}{Total\ Extracted\ Metda\ data} \#(1)$$

It found that 285 of meta data is correctly extracted out of 300 meta data in total. Therefore, the test resulted in an accuracy of 95.0%. The failure of the meta data extraction process was due to the non-compatibility of the writing of the scientific article with the template set on the JIP polynema. Some examples of affiliate writing that uses acronyms and writing email addresses that do not meet the general standards of email writing.

B. Time Efficiency

The test was conducted on five people as OJS users. In this test each person submitted 10 articles without and using the created meta data extraction system. The articles used are taken randomly with the criteria of having a difference in the number of authors, keywords, and paragraphs on the abstract. It aims to know the time it takes the user to submit articles in different forms. The timing starts when the user presses the send button on the first page of article submission and ends with the user clicking the send key on the third page or metadata filling. Table 2 shows the result of the time efficiency test.

From the test results can be known the average time it takes users to submit articles with and without the developed plugin. Then a comparison is done to find out the efficiency of time. Based on time comparison testing, this plugin can speed up the process of filling meta data on article submission process by 3.72 times faster than without a meta data extraction system

IV. Conclusion

Based on the test results obtained, it can be concluded that the development of the meta data extraction system of scientific articles can be done using rule-based using the format and writing structure of JIP as a rule. This approach can result in a 95% accuracy value in extracting JIP article meta data. The resulting system is capable of speeding up the process of filling meta data on article submission process by 3.72 times faster than manually. A built-in meta data extraction system in the form of a plugin can automatically extract meta data from the

uploaded article and is able to reduce errors in the filling of meta data performed by the author or article sender.

Further research for the development of this system is the extraction of meta data using several sets of rules or using a more intelligent system to extract article meta data with different writing formats and structures.

References

[1] A. Kovačević, D. Ivanović, B. Milosavljević, Z. Konjović, and D. Surla, : "Automatic extraction of metadata from scientific publications for CRIS systems," *Program*, vol. 45, no. 4, 2011, doi: 10.1108/00330331111182094.

[2] F. Rahutomo, D. A. Irawati, and M. A. E. Pramudita, : "Pengembangan Sistem Ekstraksi Metadata Artikel ilmiah secara Otomatis," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 2, 2019, doi: 10.25126/jtiik.2019621227.

[3] Y. S. Soekamto, : "Ekstraksi Judul dan Abstrak Artikel Ilmiah Berbasis Rule," *Journal of Information System,Graphics, Hospitality and Technology*, vol. 2, no. 01, 2020, doi: 10.37823/insight.v2i01.69.

[4] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski,: "CERMINE: Automatic extraction of structured metadata from scientific literature," *International Journal on Document Analysis and Recognition*, vol. 18, no. 4, 2015, doi: 10.1007/s10032-015-0249-8.

[5] H. Yang and W. Hsu,: "Automatic metadata information extraction from scientific literature using deep neural networks," 2022. doi: 10.1117/12.2623554.

[6] T. Huynh and K. Hoang, : "GATE framework based metadata extraction from scientific papers," in *ICEMT 2010 - 2010 International Conference on Education and Management Technology, Proceedings*, 2010. doi: 10.1109/ICEMT.2010.5657675.

[7] A. Alamoudi, A. Alomari, S. Alwarthan, and Atta-Ur-rahman, : "A rule-based information extraction approach for extracting metadata from pdf books," *ICIC Express Letters, Part B: Applications*, vol. 12, no. 2, 2021, doi: 10.24507/icicelb.12.02.121.

[8] M. Zhu and J. M. Cole, : "PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format," *J Chem Inf Model*, vol. 62, no. 7, 2022, doi: 10.1021/acs.jcim.1c01198.

[9] M. W. Ahmed and M. T. Afzal, : "FLAG-PDFe: Features Oriented Metadata Extraction Framework for Scientific Publications, :" *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2997907.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Pearson, 2009.