

## Evaluation of the Implementation of Indonesian Electronic Journals Citation System Using Regex Technique and PDF Extraction Tool

Riri Fitri Sari and Agung Kurniawan  
Department of Electrical Engineering, Faculty of Engineering, University of Indonesia,  
Kampus Baru UI, 16424 Depok, Indonesia

**Abstract:** All research papers produced by researchers worldwide now are based on previous academic publication written by other researchers. Many research papers published in electronic media and new media also referred to previous publications. The advancement of technology makes internet become the most widely used media. Research papers are published in many formats such as in the journal. Relation among journals can be traced through their citations. The number of citation to a journal study can also be calculated to show the contribution of that particular study. In order to know the relation among journal articles published on the internet, a system was designed which can automatically produce a relationship information between articles from different journals which are located in different websites. Therefore, in this research we created a mashup in order to extract the web pages and then pick required files automatically. This system produced a database to save the extracted files and then find the relations. The results of the process are shown in a web portal. The interface has functionalities for searching by using the key words inputted by users. As a result, the whole system forms a Mashup. We created an automatic extraction for Indonesian electronic journals system using data from fourteen university e-journal sites. We built the system using PHP language and MySQL database, after carefully studied the algorithm in Openkapow Robomaker. The system can successfully extract information from journal provider's web pages which include special type of PDF pages then save them in database. The system generated finally shows the connection and the relation among the journals. The test result shows the processing time and memory usage evaluation for a random number of files. The evaluation results show that the execution time is dependent on the number of journal series, volumes and number of articles on related e-journal sites. The system has been complemented with some functionalities for the user interface to report the number of the total journal articles extracted automatically from different sites. Some approach such as the use of DOM tree, Regular Expression techniques and PDF extraction tools have been used to improve the system in extracting web pages and getting full journal articles to be processed.

**Key words:** Web extraction, citation index, journal, regular expression technique, PDF extraction tools

---

### INTRODUCTION

Development of science and technology now can not be separated from many studies previously conducted by individuals or institutions i.e., education institution, private and government. Many publication of research studies, articles, news and journals are published in new media or electronic format. Many publications of research results in the form of technical report documents, study, journals, articles are widely available on the internet. Those documents are easy to access, many of them are used as reference material or reference for further research of other researchers. Therefore, reference relationship between documents on the can be evaluated automatically. A new document will refer to an older

documents published previously. In order to know the relationship between documents, we need to evaluate the reference part of the full text. Subsequently, we can count how many times a document have been cited. Based on the above mentioned research problem in order to enable the extraction and to implement easier document processing from the web, it is necessary to build a system that automatically extract and process data obtained from the document to search for a relationship between documents. Due to the variation of document on the Internet, the document processed in this research are journal articles in PDF format. The electronic journals used in this study are free access journals which are issued by several educational institutions in Indonesia. The goal of this research is to create a system that can

help to search relationship between documents obtained from the internet. We implemented the system using PHP language for application program and MySQL database. Some improvement of methodology such as the use of Document Object Model (DOM) tree, Regular Expression extraction and PDF Extraction tools have been conducted. The system can be used to extract information from different sites with different designs automatically. The final version of the system has been put alive in the web.

## LITERATURE REVIEW

Web scrapping, web harvesting or web data extraction is a technique to extract the data or information from a website using a particular software application program. Usually application program simulates human exploration of the web using a low-level HTTP or uses full-fledged web such as internet Explorer and Mozilla (Palmer, 2001).

Web Scraping is also associated with a web automation which simulate web browsing activities of people using computer software. Obtaining relevant data from a web page can be performed with scanning effective part of a document such as reference, author, book title, date, etc. Variation of document is possible. For documents with different document format, an automatic identification and process will be conducted by the web extraction system.

One of the methods used in this web extraction technique is the Document Object Model (DOM) which is a breakdown of how an object (text, image, headers, links, etc.) of a web page displayed. DOM defines what attributes are associated with each each object is and how an object can be used. Dynamic HTML (DHTML) relied on DOM to display dynamic web page in the browser after being downloaded.

DOM will convert (parse) an XML document into a hierarchical format (Chen, 2010). In addition, there is another technique to extract the web i.e., Regular Expression or often referred to as Regex. Regex is a formula for finding the pattern of a sentence/string. At low levels regex is usefull to search for a word fragment. At higher level, regex is able to control the data to find, delete and modify it (Li *et al.*, 2008).

**Mashup:** Mashup is a web application that combines data from one or multiple sources into one integrated system or device. Mashup consists of two main parts, namely web applications that provides new services using various sources of available data or data from other

sources. An example of Mashup is a cartographic use of data from Google maps to add location information to real-estate data to create a new web service which differs from the existing one (Palmer, 2001).

**Citation and citation index:** Citation is a reference to a book, article, journal, web page forms or other publications with sufficient details to identify the source of the reference. Citation usually consists of the researchers name, book title or article, publisher, publication year and the URL of the study. There are some standard in writing citation, created and published by various associations or individuals (author) i.e., the Chicago style and Turabian style that is used for all fields. Modern Language Association (MLA) which is used for arts, literary and humanism and American Psychological Association (APA) which is used for psychology, education and other social sciences (Pressmann, 2005). Citation on a scientific study or document have been performed if a particular journal article uses particular findings in that document and this is used as the base for future contribution. Therefore, we will be able to know how often a document is cited, the contribution of the information and its influence or impact to other research. The size of the influence or impact of a document to provide information depends on a number of the citation to the document. Some paid and free citation index services currently exist i.e., Web of Science, Scopus, Citeseer, RepeC, Google Scholar and Publish or Perish.

**Web data extraction tools:** There exists some open source and proprietary tools that can be used for web data extraction process. Currently there are some tools available to create a web data extraction program with a visual programming environment with features such as reducing script writing for web data extraction i.e., Kapow Mashup Server 6.3 Robomaker, Lixto Visual Developer.

**Portable document format:** Portable Document Format (PDF) is a file format used for exchanging digital documents, created in 1993 by Adobe systems. PDF is used for representing two dimensional documents in the application software, hardware or operating system independent which is a PDF file consists of a set of descriptions for the two-dimensional layout such as text, type, letters, images and two-dimensional vector graphics. PDF file consists of several objects, like Boolean which represent the value of true or false, numbers, strings, names, array consisting of several objects, dictionaries which are the collection of objects indexed by the name, streams that contain large amounts of data, the null and the object.

Text in PDF is represented in a stream of text element. In the study by Kaplan and Tokunaga (2008), Nanba *et al.* (2000) and Kaplan *et al.* (2009) introduced their research in automatic extraction of citation contexts for research study summarization using Cite-sum. The systems accept a study title as a query and find the citing studies and also classifies extracted information. In this research, we also also processed Microsoft Word documents, a word processing software (word processor) from Microsoft. Some free text publications are provided in this format. Word document format (.Doc) is divided into several parts: the header, the storage property and the file information.

**DESIGN**

The design is important when creating a system (Pressmann, 2005). Good design will produce a system which perform the function and the purpose of the system. Design generally consist of several stages such as clarifying specifications and functions of the system, determining how the system researchs, identifying the requirement the system and determining the tool used to build the system.

The system is a web-based system that functions to search and display citation index of the electronic journals of institutions in Indonesia. The system works by collecting data in terms of articles journals acquired from the higher education institutions portal in Indonesia. For the users, the system will be similar to search engines. User will enter a keyword based on the title or author in the field.

Then, the system displays the search results accordingly. The general frameresearch of the system is shown in Fig. 1. Stages of system design is a development from the previous system which is a web-based system that functions to locate and display the citation index of journals from the institutions in Indonesia. The system researchs by collecting journal articles data derived from 14 Indonesian institutions. From the user perspective, the system will look like search engines where users will enter a keyword based on the title or author on the field. Then the system displays the search results accordingly. In addition from the user perspective the system will display some of the resume data obtained from extraction result.

**Extraction system:** The new extraction system generally works the same way with the previous system as shown in Fig. 2. The system has been added with some methods for extracting HTML pages and also the extraction of the PDF or the full article page. Some web extraction system using the selection of string and cutting existing data in an HTML page using PHP operation string have been implemented. The system is more efficient in processing and the result is more precise. DOM parsing and

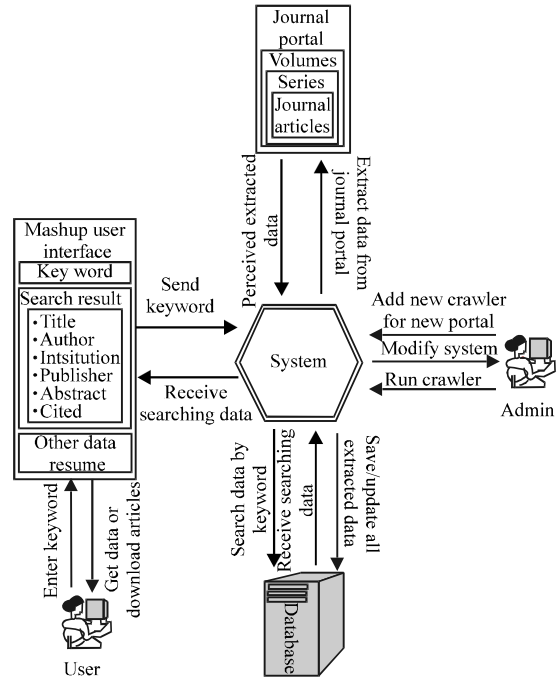


Fig. 1: Frame research of the system

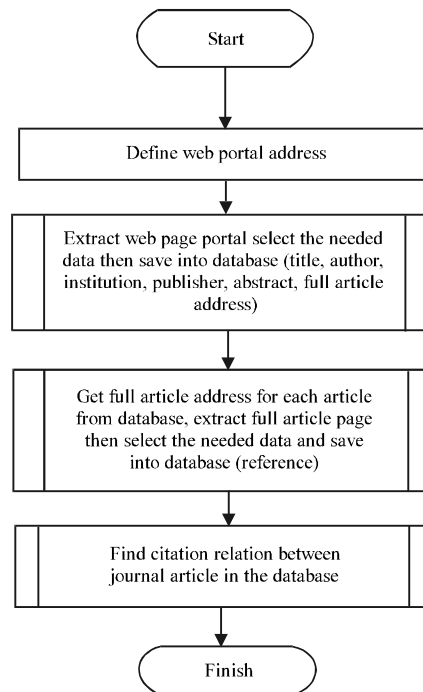


Fig. 2: Activity diagram of the extraction system

regular expression implementation to extract an HTML page have been generating a more precise results using the tagging information. Previously we used PHP script to convert PDF files into a string/text tan can be read and identified. Subsequently we use a tool called pdftotext

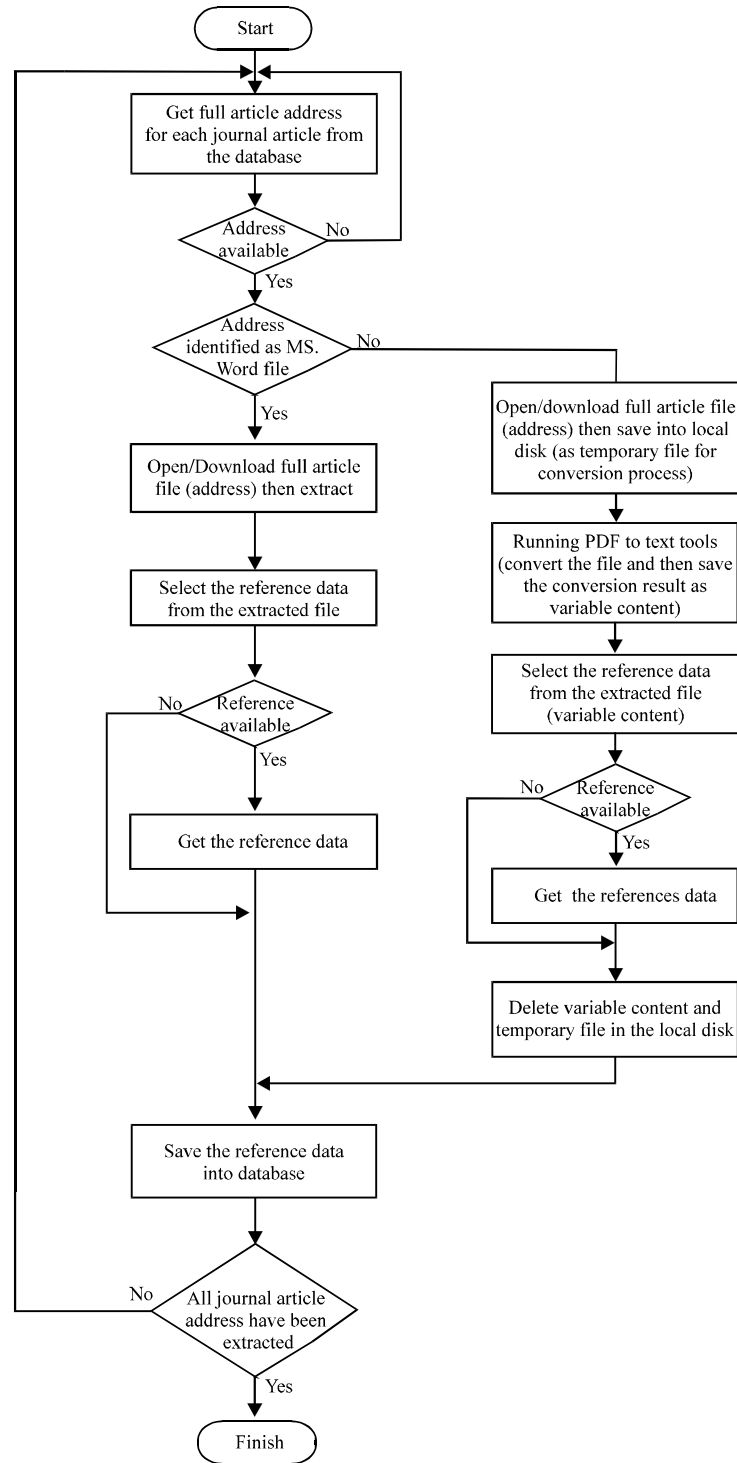


Fig. 3: Activity diagram of the full article extraction

to convert the PDF file better. The file should be saved before the conversion into a text format. Considering that some of the electronic journal full articles are provided in

Microsoft Word format, we make an additional script to allow the extraction of the Microsoft word file. The procedure can be found in Fig. 3.

**User interface system:** For the user interface system, we designed a space in which additional link to journal providers. Some functions to view the summary of the extracted information has been provided. The data can be viewed based on the list of researchers, their researches and the total number of researches cited, list the amount of research of each institution per year and the number of articles cited, the graphical representation of the the number of articles from each institution and graph the number of articles cited for each institution.

## EVALUATION

In the implementation to create a Mashup, the previous system researchs to extract information from the e-journals sites in Indonesian higher education institutions. Testing were conducted by calculating crawler application execution time. The research flow of the crawler can be shown in the flowchart on Fig. 4. Figure 5 and 6 shows some differences in execution time for each site of the journals. To execute the entire script program, the time difference due to the difference in the number of series, number of volume and journal articles for each volume. The difference between the combination of number of series, volume and the journal article may lead to differences in execution time because there will be a difference in the number of web pages opened. In terms of the execution time, the volume and abstract address, there is the time difference caused by the number of web pages that must be opened to get the relevant data that is required which is different for each site. The execution time for the separation of the title, author, institution, publisher, full address of the article and abstract are caused by differences in the format of the display of abstract web pages processed. Other functions in the application system is to obtain the reference from a journal article in which the address has been stored in the database resulting from the application process of the web page extraction functions. Experiment result shows that there are differences in the execution time which is for the time in reading the full page article dependent on the size of the file processed.

Ideally, the larger the file size, the more time required for execution with a note that the data transfer speeds used are the same. The situation in which an execution time is longer for a smaller file size can be caused by the decreased in the speed of data transfer when opening a file. The time for data conversion of the PDF file to get data referenced will depend on the number of streams in the PDF file that is processed or converted until reference data can be obtained. The more file stream, the longer the

execution time. The use of memory and file size will affect the amount of memory used. This is related to the use of the variables used as buffers to process the data. The larger the size of the files processed will also increase the memory needed by the application.

Concurrent execution of other application will require memory allocation which will cause the addition of the memory usage. This is done by using the command `memory_get_usage()` in PHP.

**User interface performance testing:** Testing the performance of the system is conducted by running an application program on localhost. It is the same with the extraction system on the test. The resulted Mashup user interface is shown in Fig. 7. We have performed the extraction of the portal provider of full text online journals, including:

- University of Indonesia (<http://journal.ui.ac.id>)
- Bandung Institute of Technology (<http://proceedings.itb.ac.id>)
- Udayana University (<http://ejournal.unud.ac.id>)
- Petra Christian University (<http://puslit2.petra.ac.id/ejournal>)
- Gadjah Mada University (<http://i-lib.ugm.ac.id/jurnal>)
- Diponegoro University (<http://eprints.undip.ac.id>)
- Islam University of Indonesia (<http://journal.uin.ac.id>)
- Sriwijaya University (<http://digilib.unsri.ac.id/jurnal>)
- Muhammadiyah University of Purwokerto (<http://jurnal.ump.ac.id>)
- Bogor Agricultural Institute (<http://journal.ipb.ac.id/index.php>)
- Yogyakarta State University (<http://journal.uny.ac.id/index.php>)
- Andalas University (<http://ffarmasi.unand.ac.id>)
- University of North Sumatra (<http://ejournal.usu.ac.id>)
- University Mercubuana (<http://research.mercubuana.ac.id>)

**New extraction system result testing:** Testing extraction system is done by counting the number of journal articles that successfully extracted from each portal providers journal. From extraction of the journal provider portals extraction is obtained following data. From the extracted HTML page, it can be found that not all of the journal

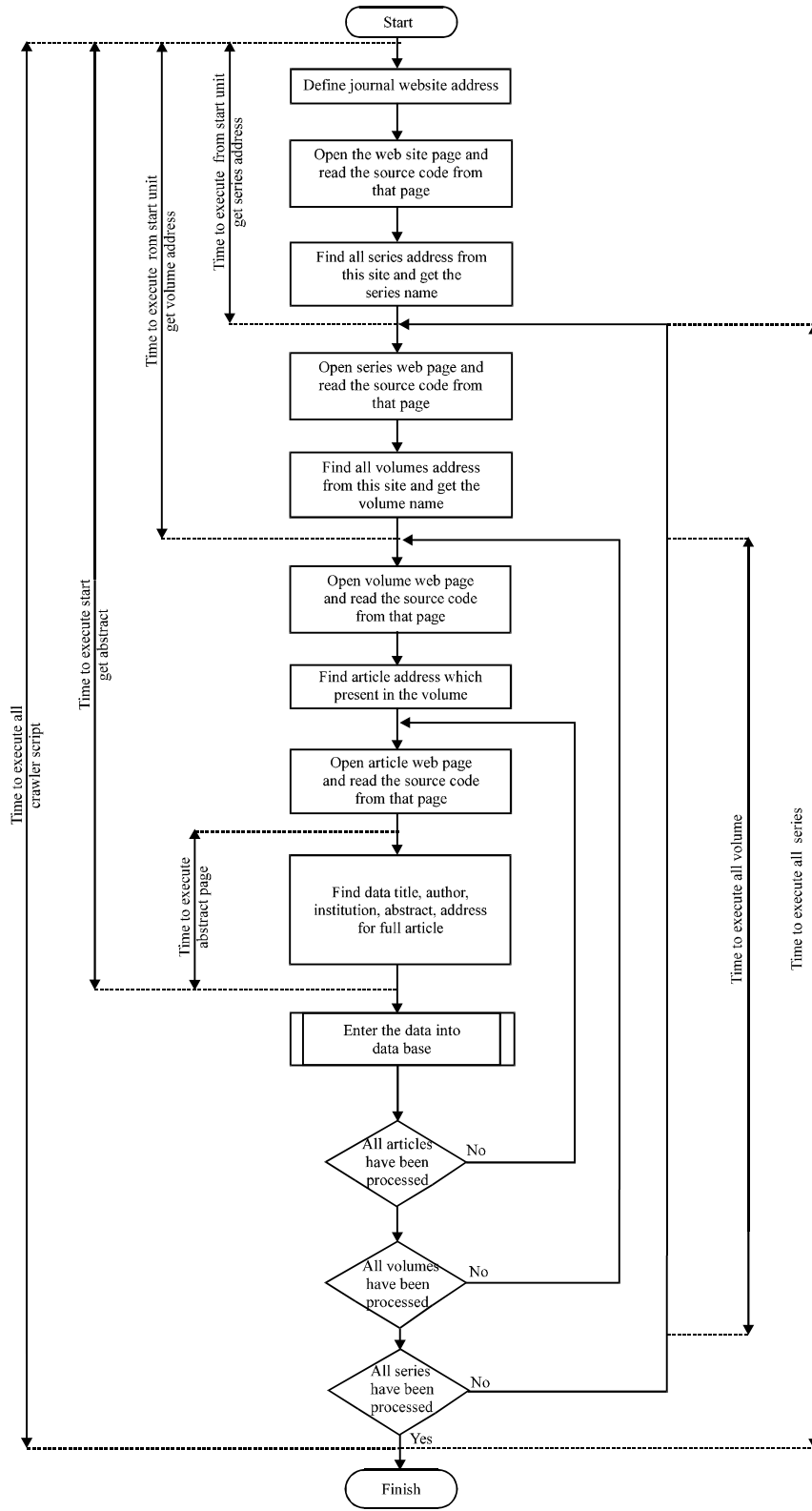


Fig. 4: Web page extraction function flow chart

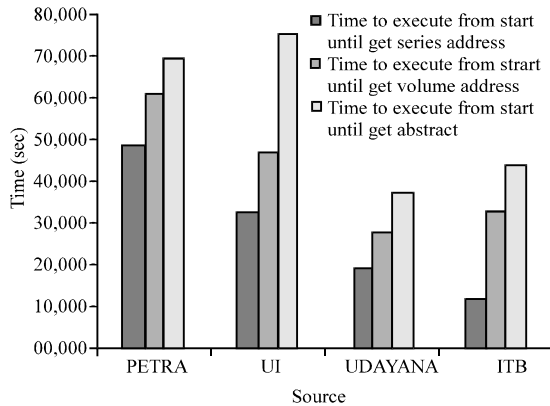


Fig. 5: Execution time of the extraction from web pages home page to get first abstract of the article journal

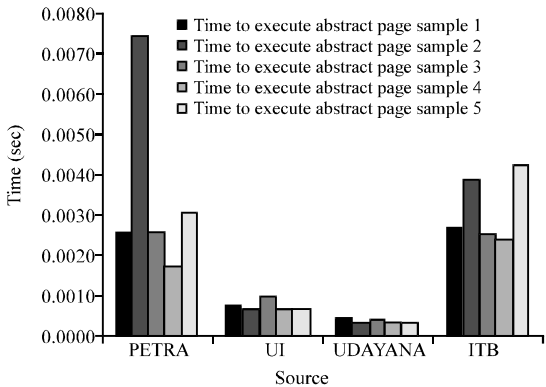


Fig. 6: Execution time of extraction scripts of web page information to obtain title, author, institution and abstract from article journal abstract page

articles can be extracted and inserted into the database. This is because there are some web pages that have different tags (Table 1). Some existing data on the page can not be extracted if obtained data do not match the specification. The extraction process of full page articles in PDF and Microsoft Word is conducted to get the reference from the articles. Table 2 shows the results. It can be noted that some universities provide abstracts of journal articles only and therefore no referencing can be found.

Some of the PDF pages result that is not all full page article (reference) can not be extracted. This is because not all PDF files can be converted by the existing tools (PDFtotext). Some files are encrypted. Some are scan results. Others required payment for full text access. Some files can not be opened even though the address to get a full article has been obtained from the extraction of HTML pages. In addition, not all the pieces of reference data can be extracted well, this is because in one file a full journal article contained more than one reference word. A word searching process starts from the beginning of the file until the end of the file. The system will crop the data from the first word discovered until the last word of the file. This does not happen in the previous system because it starts from the end of the file to the beginning of the file.

The extraction process or reading the contents from Microsoft Word files (under 2007 version) is easier than accessing the PDF file format. This is because there is no data compression. Thus the content of the file can be directly readable with a simple PHP command with the addition of several processes to remove unnecessary characters. In the process to find relationship between journal articles with one another based on data obtained from the extraction HTML pages and full page article that



Fig. 7: User interface of system

has been stored in a database, obtained the number of citations for each institution as shown in Table 3. The relationship of searching process between one journal articles (citations) to another, indicates that a journal article can refer to the other journal articles which previously published, beside that it is possible a journal article cited by more than one other article. From the data

Table 1: Number of articles extracted HTML

Institution name	Total
University of Indonesia	524
Bandung Institute of Technology	258
Udayana University	1315
Petra Christian University	2176
Gadjah Mada University	10287
Diponegoro University	43
Islam University of Indonesia	271
Sriwijaya University	186
Muhammadiyah University of Purwokerto	45
Bogor Agricultural Institute	176
Yogyakarta State University	105
Andalas University	35
University of North Sumatra	701
University Mercubuana	96

Table 2: Number of PDF full text articles extracted

Institution name	Total
University of Indonesia	490
Bandung Institute of Technology	212
Udayana University	944
Petra Christian University	2054
Gadjah Mada University	0
Diponegoro University	45
Islam University of Indonesia	233
Sriwijaya University	126
Muhammadiyah University of Purwokerto	44
Bogor Agricultural Institute	165
Yogyakarta State University	85
Andalas University	0
University of North Sumatra	681
University Mercubuana	84

found know that most of the citations are made to journal articles published by their institution with the same series by series of journal articles that refer.

**New user interface system testing:** Testing on the user interface system is done by looking at the results/appearance of additional functions that run and compare it with the data in the database. Testing results show that the additional functions which are added in the user interface system can research well to display the specified data. The result of the user interface system can be shown in Fig. 8.

**New extraction process testing:** Testing the extraction process is done by comparing the time of extraction operation process between the new system that have been developed and the previous system. The calculation of operating times for HTML extraction system can be shown in Fig. 9-13. After the testing process is conducted, we obtained the comparison of the graph of the operating time as shown in Fig. 14 and 15. From these

Table 3: Total citation of each institution

Institution name	Total
University of Indonesia	0
Bandung Institute of Technology	3
Udayana University	34
Petra Christian University	100
Gadjah Mada University	68
Diponegoro University	1
Islam University of Indonesia	1
Sriwijaya University	11
Muhammadiyah University of Purwokerto	0
Bogor Agricultural Institute	2
Yogyakarta State University	1
Andalas University	0
University of North Sumatra	21
University Mercubuana	6



Fig. 8: Display of the main page



**Sitasi Jurnal Elektronik Indonesia**

**TABEL DAFTAR PENULIS**

PENULIS	JUMLAH KARYA	DISITASI
nalik wenediyani	4	0
teguh nurata sidiqga	1	0
yoel sulaheidi	1	0
devi fitriani	1	0
budi salsono	1	0
Isdar Idris sudirman	4	0
julis dvi magrobo	1	0
irdika mannan	1	0
agus purwito	1	0
endang subandiang	1	0
joko prayitno	2	0
harry roffe	1	0
agus kurnia	1	0
shachi satoh	1	0
himmawati puji	1	0

[Lihat Tabel Daftar Penulis](#) [Lihat Tabel Jumlah Artikel Tiap Institusi](#) [Lihat Grafik Institusi VS Jumlah Artikel](#) [Lihat Grafik Institusi VS Sitasi](#) [Home](#)

Sumber : UI ITR UNUD PETRA USM UNDP UII UNSRI UMP IPB UNY MERCUBUANA UNAND USU

Fig. 9: Display page list of the author and the amount of research

**Sitasi Jurnal Elektronik Indonesia**

**TABEL JUMLAH ARTIKEL JURNAL TIAP INSTITUSI**

INSTITUSI	TAHUN	JUMLAH KARYA	DISITASI
Institut Pertanian Bogor	2010	20	0
Institut Pertanian Bogor	2009	30	0
Institut Pertanian Bogor	2008	30	0
Institut Pertanian Bogor	2007	32	0
Institut Pertanian Bogor	2006	32	0
Institut Pertanian Bogor	2005	32	0
Universitas Andalas	2008	9	0
Universitas Andalas	2007	18	0
Universitas Andalas	2006	8	0
Universitas Negeri Yogyakarta	2010	25	0
Universitas Negeri Yogyakarta	2009	68	0
Universitas Negeri Yogyakarta	2008	9	0
Universitas Negeri Yogyakarta	2007	1	0
Universitas Negeri Yogyakarta	2004	1	0
Universitas Negeri Yogyakarta	2003	1	0
Universitas Sumatera utara	2003	8	0

[Lihat Tabel Daftar Penulis](#) [Lihat Tabel Jumlah Artikel Tiap Institusi](#) [Lihat Grafik Institusi VS Jumlah Artikel](#) [Lihat Grafik Institusi VS Sitasi](#) [Home](#)

Sumber : UI ITR UNUD PETRA USM UNDP UII UNSRI UMP IPB UNY MERCUBUANA UNAND USU

Fig. 10: Display page list of the institutions and the number of articles



Fig. 11: Graph of the number of page views of each institution's article

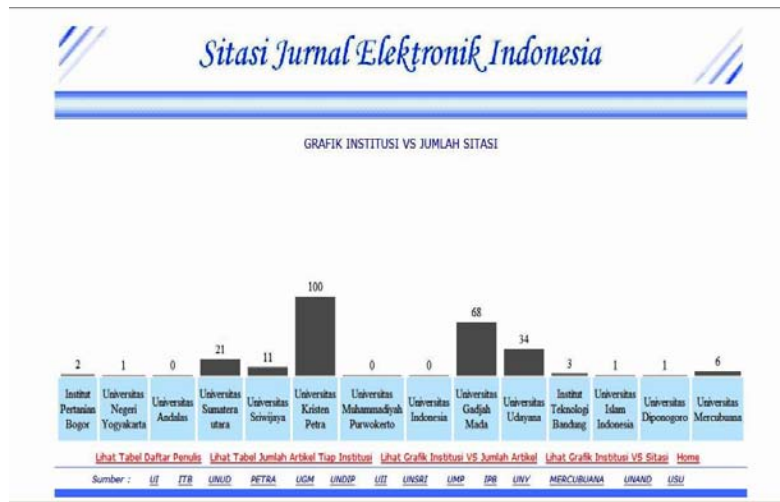


Fig. 12: The number of citations of each institution

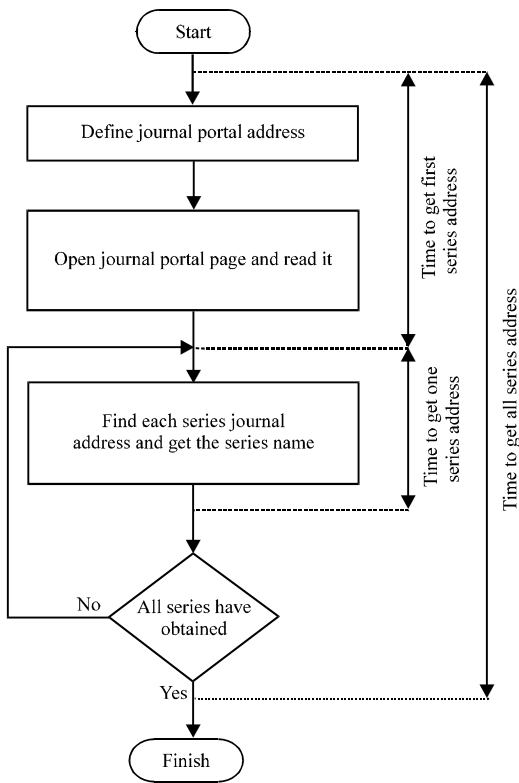


Fig. 13: Flow diagram of the extraction system testing of HTML

figures it can be found that the extraction using DOM methods have a longer processing time than the manual method (string operation). This occurs because by using DOM method there are more calls functions and also there are searching processes and distribution processes of

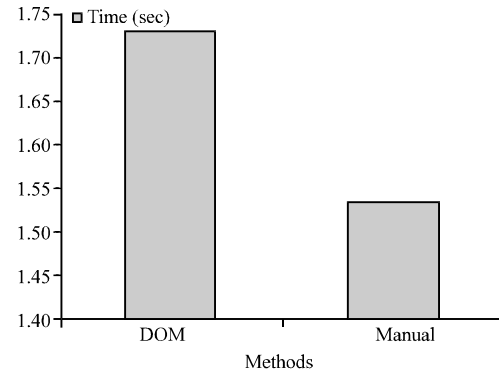


Fig. 14: The average time to obtain address of the first series

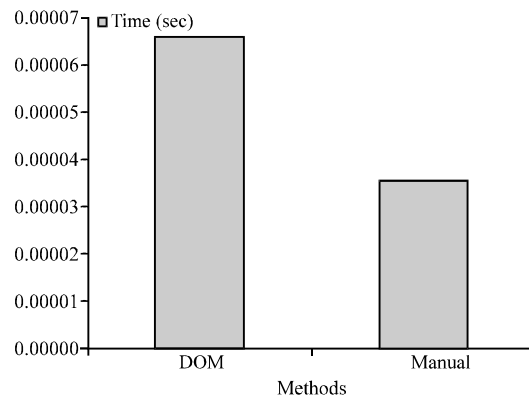


Fig. 15: The average time to obtain a series of addresses

objects displayed by a web page. When using the manual method (string operation) there is no call function required because data process cutting is done directly Fig. 16. Extraction system of fullarticle (PDF) testing it

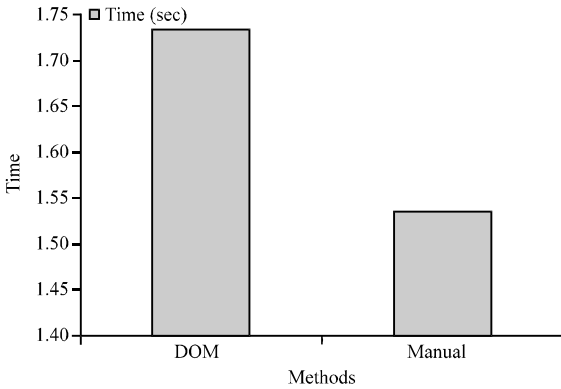


Fig. 16: The average time to obtain entire series addresses

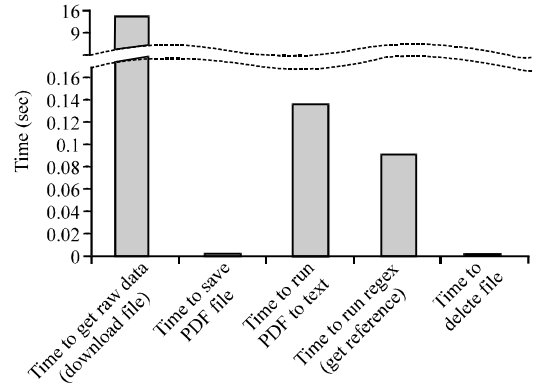


Fig. 19: The extraction process time used for PDF to text conversion

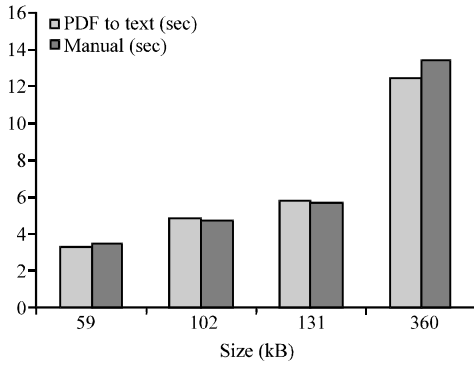


Fig. 17: Comparison of the average time for PDF extraction

done by comparing the time of the extraction process between pdftotext and manually using PHP script by executing four files full article with a different size. Comparison of the time of the operating process can be shown in the the Fig. 17. From the Fig. 18 it is shown that there is no significant operating time difference between using pdftotext and manually (script). Where conversion in the manual system (script) done by uncompress on each stream which is on file reading while for systems using pdftotext process sequence operation can be viewed at the following flow chart. After testing the PDF extraction system using pdftotext, known that execution time of each process as shown in the Fig. 19. From Fig. 19 it is shown that the longest operation time of extraction process using PDF to text (except time to downloading files) is the process of converting files PDF to text (running PDF to text) then the searching process reference data using a regular expression.

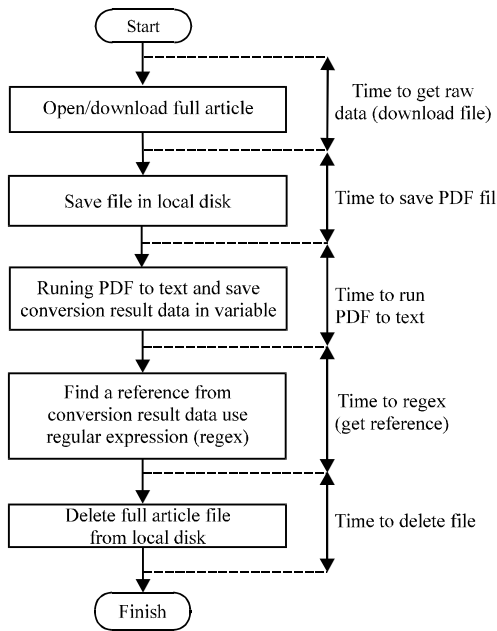


Fig. 18: Flow chart PDF extraction system testing using PDFtotext

**CONCLUSION**

After the design, implementation, testing, analysis the applications which has been conducted with the information extraction from electronic journal portal and built a mashup some conclusions can be drawn. Each portal electronic journals universities have different characteristic, so it requires PHP based script for each portal. Data from each web page conducted with identify location of data and used tag. The results test and measurement system show that the more number of articles journal and number of combinations of series and volumes on the site of the journal will require more time to extract all needed data from that site. The bigger file size extracted in PDF extraction the application will need more memory. Larger data size will entered into database will require more time and memory. In general, the system can

research well and can show the relation between journal articles. Future development of the system can be conducted to add the ability to extract PDF file below version 1.4 and adding the number of electronic journals site to be extracted.

The new develop system show that extraction system which is made by using Regular expression syntax and DOM parsing produce better required piece of data than using string operations although, with a longer operating time. Use of Regular Expressions and DOM makes it easier and less time in the development extraction scripts than using string operations. Use PDFtotext tool for extracting full article files needed a temporary storage place to save the downloaded file then converted it into text format by the tool. The longest time in PDFtotext extraction process is the conversion of PDF to text (running PDF to text) and search for reference data using regular expression. And type PDFtotext tool will be different for different OS. The process of reading Microsoft Word files (under 2007) is more easily compared with the PDF format file because no data compressed. Not all journal articles which have been extracted have all the required datas, like as title, authors, institutions, publishers, abstracts and full address of the article. Not all reference from full article journal can be extracted because the files is encryption and scan results. Most of the citations are made to journal articles that published by author institution.

#### **ACKNOWLEDGEMENT**

Researchers thank the Directorate General of Higher Education, Ministry of National Education of Republic of Indonesia for funding this Project through Hibah Kompetensi 2010 Research Grant.

#### **REFERENCES**

- Chen, S.C., 2010. PHP Simple HTML DOM Parser. <http://simplehtmldom.sourceforge.net/index.htm>.
- Kaplan, D. and T. Tokunaga, 2008. Sighting Citation Sites: A Collective-Intelligence Approach for Automatic Summarization of Research Papers using C-Sites. Proceedings of ASWC 2008 Workshop. <http://tanaka-www.cs.titech.ac.jp/publication/archive/633.pdf>.
- Kaplan, D., R. Iida and T. Tokunaga, 2009. Automatic extraction of citation context for research paper summarization: A coreference-chain based approach. Proceedings of the Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP, Aug. 7, Suntec, Singapore, pp: 88-95.
- Nanba, H., N. Kando and M. Okumura, 2000. Classification of research papers using citation links and citation types: Toward automatic review article generation. Proceedings of 11th SIG/CR Workshop, pp: 117-134.
- Palmer, S.B., 2001. The semantic web: An introduction. <http://infomesh.net/2001/swintro/>.
- Pressmann, R.S., 2005. Software Engineering: A Practitioners Approach. 6th Edn., McGraw-Hill, New York.
- Li, Y.R. Khrisnamurty, S. Raghavan and S. Vaithyanathan, 2008. Regular expression learning for information extraction. Proceedings of the Empirical Methods in Natural Language Processing Conference, Oct. 08, Honolulu, HI USA., pages 21-30.