Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/ 10.33395/sinkron.v7i4.11576
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Web Scraping for Journal Data Collection on the SINTA Website

Nelawati Adila[1)*], Falentino Sembiring[2)], Wisuda Jatmiko[3)]
[1)2)3)]Nusaputra University, Indonesia
[1)] nelawati.adila_si18@nusaputra.ac.id, [2]falentino.sembiring@nusaputra.ac.id,
[3)] wisuda.jatmiko@nusaputra.ac.id

Submitted : July 23, 2022 | **Accepted** : | **Published** :

Abstrak: SINTA is a website portal pioneered by the Director General of R&D Improvement, Ministry of Research Technology, and Higher Education, Republic of Indonesia to make it easier for researchers to search for journals for publication. However, in its implementation, many researchers have encountered problems, one of which is in the form of searching for publication duration and ranking, as well as searches that are carried out manually, making it difficult for researchers to find a place to publish. In this study, the author took journal data based on the SINTA website with Web scraping techniques using the Python programming language and then stored the data in the SINTA database, then scheduling a Cron job so that the data in the database was always updated. It is hoped that the results of this study can help researchers in searching for journals for publication. In this study, the results obtained as many as 7412 data, then after filtering using a MySQL query the data became 977 data by displaying information on the month of publication of the journal.

Keywords: Publications, Journals, SINTA, Web scraping, Python

INTRODUCTION

The development of the times has made the pace of information increasingly rapid, information that is now sought after is information about publications for scientific journals. A scientific journal is a place or forum for the publication of scientific papers, where the quality and substance of the content lies entirely in written words, sentences, and paragraphs. (Yanti et al., 2020)

In academia, scientific publications in accredited national journals and reputable international journals have become an inevitable necessity. The goal is to increase the competitiveness of Indonesian universities to publish scientific papers that are currently still low. At the beginning of 2012, the Director of Higher Education No. 152 / ET / 2012 concerning the obligation of undergraduate, postgraduate, and doctoral students to publish scientific works / articles in accredited national and international scientific journals as one of the graduation requirements.(Saputra, 2020).

This aims to improve the quality of student research so that it can improve the quality of research at Indonesian research institutions by providing research resources from an early age. To facilitate the data collection and mapping of scientific publications by Indonesian scholars and researchers, the Director General of Research and Development Strengthening of the Ministry of Research, Technology and Higher Education initiated the establishment of the Indonesian Indexing and Citation System (SINTA) (Lukman et al., 2019).

SINTA is a forum for exchanging scientific and technological works with mankind in Indonesia in the form of a web-based research information system perfected by the Director General of Research and Development of the Ministry of Research Technology and Higher Education of the Republic of Indonesia 2016. This portal measures the performance of Indonesian institutions, researchers, and journals. SINTA indexes into six ranking categories of all accredited national journals published by ARJUNA, the body appointed to assess the quality assurance of scientific journals by screening manuscripts for fairness, administrative feasibility and timeliness of publication of scientific journals, which includes SINTA 1 to SINTA 6(Saputra, 2020).

SINTA is very helpful in making it easier for researchers to find journals for publication. In addition, not all accredited national journals in the field of library science present information on the website properly, such as inactive websites, focus and scopes, templates, and manager contacts that are not available, making this an obstacle for researchers when publishing scientific journals. From the background above, a program is needed that presents journal data, be it information containing journal names, journal websites, to information

*nelawati adila

published in journals every year. The program developed applies web scraping techniques as data retrieval based on the SINTA website to obtain journal information.

Web scraping is the process of extracting content in the form of data or information from a website. Web scraping is used because the required data is not available in RSS or API. In addition to extracting content, data and information, this technique is also used to automate the data retrieval process or called a robot (Sembiring & Sari, 2019). In addition, how to develop web scraping techniques, namely the maker of the scraping script must first learn the HTML document from a website that will be enclosed in information to use HTML tags .(Sahria, 2020)

The results of this web scraping will be stored in the form of a database that will be scheduled for cron jobs to update data so that the data generated is always updated and up to date. It is hoped that it can be developed into a program and can be useful for researchers in seeking scientific publications.

LITERATUR REVIEW

Research conducted by Lucky Metha Purnomo and Mewati Ayub with the title Analisis Data Hasil Web scraping untuk Menentukan Kualitas Jurnal Ilmiah, This research discusses programs that can assist researchers in determining the quality of scientific publications in the portal, or choosing the results of scientific publications based on the field of science. Data collection in this study was carried out by applying web scraping techniques from the SINTA and SCImagoJR websites. The results of this study are in the form of an application that can display scientific journal information for the field of computer science studies with journal quality from the SINTA indexing portal and SCImagoJR.(Purnomo & Ayub, 2021)

Furthermore, the previous research conducted by Endah Ratna Arum and Pristi Sukmasetya with the title *Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm.* This study discusses taking online news data based on 3 different sites, namely Detik.com, Liputan6, and CNN Indonesia using web scraping techniques with the DFS algorithm to assist in the search because this method can check the date data sent and then track the destination URL. This research was conducted within 1 year with the results of comparing news data that is relevant or not with the keywords used. The results of such data will be stored in a NoSQL database.(Arumi & Sukmasetya, 2020)

Research with the title Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar written by A Rahmatulloh dan R Gunawan, The purpose of this study is to create a data recapitulation application in scientific publications based on Google Scholar that displays a list of researcher profiles, a list of affiliates, a list of citations, and a list of article titles in the form of data output in the form of *.pdf or *xlsx which have been completed with sort features. Data collection using web scraping method on google scholar.(Rahmatulloh & Gunawan, 2020)

Further research conducted by Ufi Mufidah, Manasse Siahaan with the title PERANCANGAN APLIKASI PERBANDINGAN HARGA PRODUK (HISTORICAL DATA) MENGGUNAKAN TEKNIK *SCRAPING WEB*, This research was conducted to find out the daily price changes of the products of one of the e-commerce. The information obtained through the URL address of the Zalora product, which then the data will be extracted to get the necessary data. The data that has been stored is visualized into a graph to display the price changes that occur every day. The output produced in this study is a website that can display information about changes in the price of a product every day. It is hoped that this information can help buyers in monitoring changes in the price of Zalora products without the need to check the main Zalora Indonesia website.(Mufidah, 2021)

Research conducted by Yoga Sahria about Implementasi Teknik *Web scraping* pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia. The discussion in this journal on Indonesian health research requires an analysis of information used to map indonesian research that is faster and more efficient to obtain health research topics. Implementing scraping techniques in the SINTA journal, by taking the title of the health journal, research title, author, affiliation and then analyzed the results of the data collection. This research was made using Python language with modules that support to be applied The data obtained is then analyzed so that trends in health research topics in Indonesia can be known in the SINTA Journal. Extraction of information from the results of collecting such data by structuring previously unstructured data. further analyzing trends in health research topics in Indonesia.(Sahria, 2020)

The research previously conducted by A Priyanto dan M R Ma'arif yang berjudul Implementasi *web scraping* dan *text mining* untuk Akuisisi dan Kategorisasi Informasi Laman *web* Tentang *Hidroponik.* The problem discussed in this study is about the number of web pages that present information about hydroponics, so that people must provide more time to choose and access as many web pages as possible to get complete and accurate information. This research was conducted in order to automatically acquire information from web pages that contain information about hydroponics and categorize it according to a more specific topic than the hydroponics article contained in the web page. From the experiments that have been carried out, web scraping

*nelawati adila

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022 e-ISSN : 2541-2019
DOI : https://doi.org/ 10.33395/sinkron.v7i4.11576 p-ISSN : 2541-044X

and text mining have been successfully implemented to acquire hydroponic-related articles from the internet and automatically group them into several categories based on the topic of the article.(Priyanto & Ma'arif, 2018)

METHODS

Data and information collection is carried out by collecting several literature studies related to the object under study both in books and journals related to research. The authors collected several literature studies regarding journal publications, scientific journals, SINTA website, web scraping and cron job scheduling. The following are the stages used in this study, explained in figure 1 below

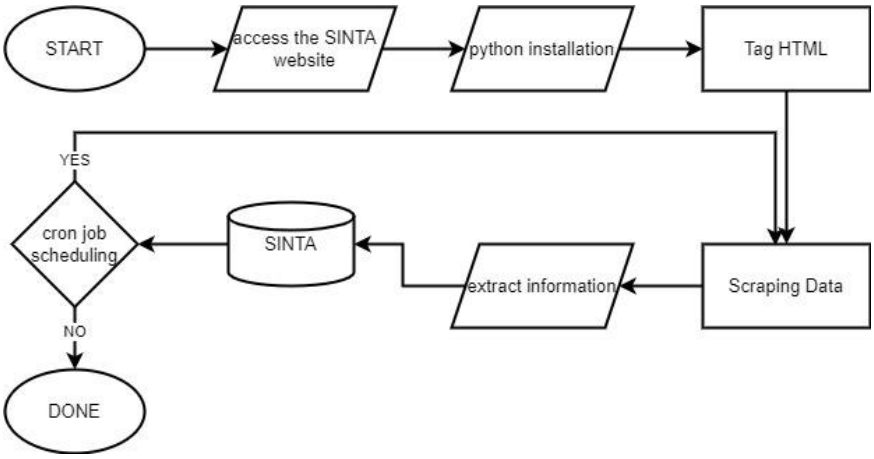

Figure 1. Research stage

Initial Stage of research

The initial stage of this study begins with accessing the SINTA website, analyzing what data will be taken information, to facilitate the data collection process.

Installing a programming language using the Python programming language, Python is a high-level programming language, to describe the operation of the system and can be used for various types of purposes (general-purpose) (Sembiring & Erfina, 2020)

This research uses the Python programming language with the beautifulsoup library. BeautifulSoup is a Python data extraction library developed by Leonard Richardson and other open source developers. It is licensed under a simplified BSD license for use with Python 2.7+ and Python 3. Can parse HTML and XML documents and provides an easy way to interact with DOM models (Uzun et al., 2018)

Data Collection Stage

After accessing the SINTA website and determining the programming language that will be used, at this data collection stage it will use data from the journal page with the amount of data that will be taken as much as 7412 data following the amount of journal data on the SINTA website. The data collection method uses the application of web scraping techniques based on a website. In theory, web scraping is the practice of collecting data through any other means than programs that interact with APIs (or, of course, through humans using web browsers)(Mitchell, 2018)

Web scraping has several steps, including: 1) Create Scraping Template: The creator of the scraping script program learns the HTML document information of the website to be retrieved, the HTML tags used to surround the information to be retrieved, 2) Explore Site Navigation, which is The creator of the scraping script program learns navigation techniques on the website where information is taken to be sorted in the scraper web application to be created, 3) Automatic Navigation and Extraction, Based on the information obtained in stages1 and 2, a scraper web application is created to automate the retrieval of information from the specified website, and, 4) Extracted Data and Package History: Information obtained from stage 3 is stored in a specific format, for example database and CSV (Comma Separated Values). (Josi & Andretti Abdillah, n.d.) How it works is depicted in figure 2

*nelawati adila

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/ 10.33395/sinkron.v7i4.11576
e-ISSN : 2541-2019
p-ISSN : 2541-044X

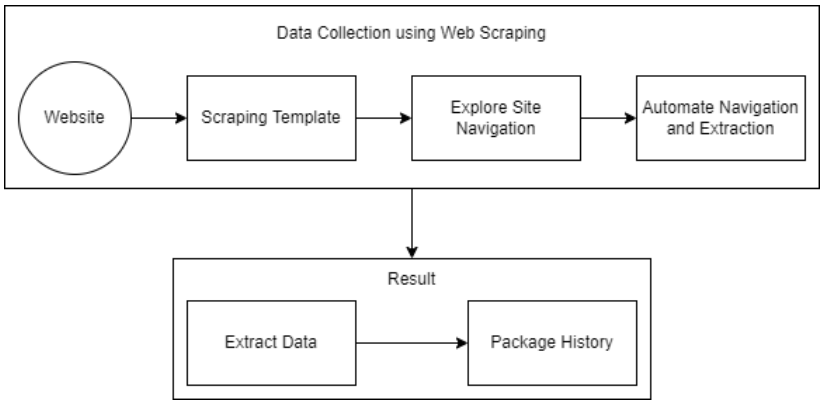

Figure 2. Web scraping procedure

Data Processing Phase

Datasets that have been taken from the website using web scraping techniques will be stored in a database. A database is a standalone software that stores a collection of data. Each database has one or more different APIs for creating, accessing, managing, searching, and replicating data that has been stored.(Sembiring et al., 2020)

The next step is to create an update function script on the database. Then schedule the cron job automatically using the task scheduler tool. A cron job is a rule on scheduling for commands that are executed periodically (Wirawan, 2020). Cron job can be used as a function to periodically retrieve data between integrated systems as a web server (Christanto & Rudiyanto, 2020)

Task scheduler is a component that provides the ability to schedule scripts or programs after a certain interval or at a time that has been scheduled. The task scheduler will run the program when windows is running, and will run any tasks that have been scheduled at the specified time at the time of creating the task .(Al hadi et al., 2019)

RESULT

The implementation using web scraping on the SINTA website produced a dataset of 7412 data. In this study, accessing the SINTA version 3 website

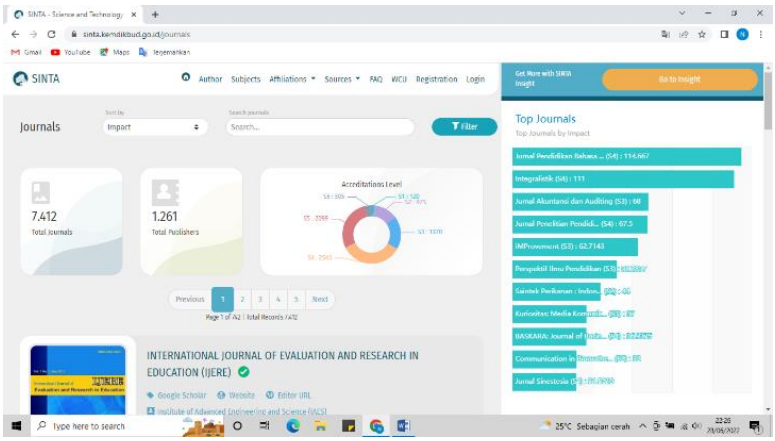

Figure 3. Website SINTA

After accessing the SINTA website, analyze the WEBSITE HTML tag to flank the information to be retrieved. At this stage, the author is looking for tags that display the entire journal data, by analyzing the HTML tags and then taking the HTML data from the domain name after that parsing the data to get the target information as in figure 4 which contains the HTML tag on the SINTA website.

*nelawati adila

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/ 10.33395/sinkron.v7i4.11576
e-ISSN : 2541-2019
p-ISSN : 2541-044X

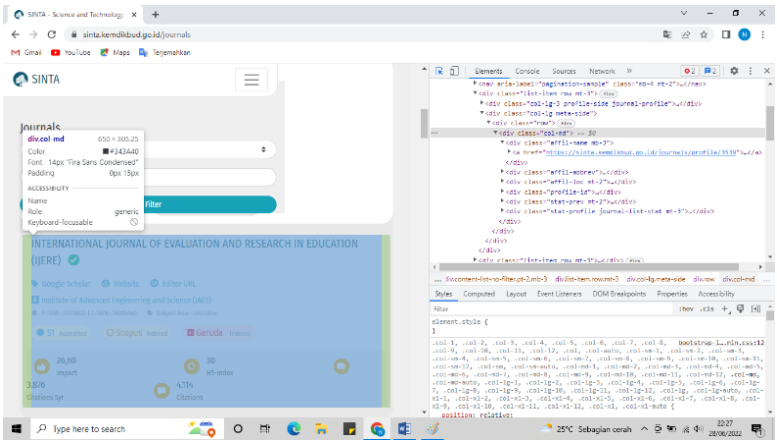

Figure 4. HTML Tag Inspection

After getting the HTML tag, the next process is to write a scraping script based on the HTML tag above using the Python programming language that has been installed first. Figure 5 illustrates the script library that the author uses to create a web scraping script.

```
import requests
from bs4 import BeautifulSoup
import mysql.connector
from mysql.connector import errorcode
```

Figure 5. Library script scraping

Table 1 is the result of data collection carried out using the web scraping stage, it produces 7412 data, and in table 2 it is the result of scraping which contains journal information after filtering become 977 data.

Table 1.SINTA Scraping Results

id	journal name	location_jurnal	profile_jurnal	accredited	link_jurnal
1	International Journal of Evaluation and Research in Education (IJERE)	Institute of Advanced Engineering and Science (IAES)	P-ISSN : 22528822 | E-ISSN :26205440 Subject Area : Education	S1 Accredited	http://ijere.iaescore.com/index.php/IJERE
125	Jurnal Teknologi dan Sistem Komputer	Department of Computer Engineering, Engineering Faculty, Universitas Diponegoro	P-ISSN : 23380403 | E-ISSN :23380403	S2 Accredited	https://jtsiskom.undip.ac.id/index.php/jtsiskom/index
1130	Jurnal Riset Biologi dan Aplikasinya	Universitas Negeri Surabaya	P-ISSN : 26559927| E-ISSN :26559927	S3 Accredited	https://journal.unesa.ac.id/index.php/risetbiologi
3028	Diakom : Jurnal Media dan Komunikasi	Kementerian Komunikasi dan Informatika	P-ISSN : 26231212 | E-ISSN :26231212	S4 Accredited	https://jurnaldiakom.kominfo.go.id/index.php/mediakom
4876	JITMI (Jurnal Ilmiah Teknik dan Manajemen Industri)	Universitas Pamulang	P-ISSN : 26205793 | E-ISSN :26856123	S5 Accredited	http://openjournal.unpam.ac.id/index.php/JITM
6801	Jurnal Ilmu	Universitas	P-ISSN :|	S6 Accredited	http://ejournal.unp.a

*nelawati adila

	Informasi Perpustakaan dan Kearsipan	Negeri Padang	E-ISSN :23023511		c.id/index.php/iipk

Table 2. Moonrise Scraping Results

id	journal name	Publication month
1	International Journal of Evaluation and Research in Education (IJERE)	September , June , March , December
9	TELKOMNIKA (Telecommunication Computing Electronics and Control)	June , April , February , December , August
18	IJAL (Indonesian Journal of Applied Linguistics)	May , January , September , July
331	DUNAMIS: Jurnal Teologi dan Pendidikan Kristiani	April , Oktober
2430	Journal of Architectural Design and Urbanism	May , March , September

The results obtained from web scraping will be accommodated in the database system. This container makes it easier for authors to create multiple report functions in the system layout so that the mysql.connector library can connect between the database and Python scraping. After that, the data is classified using a database query by generating the graph in Figure 7 to find out the number of journal publications each month.

nama_jurnal	location_jurnal	profile_jurnal	acredited	link_jurnal	bulan_terbit
International Journal of Evaluation and Research in Education (IJERE)	Institute of Advanced Engineering and Science (IAES)	P-ISSN : 22528822 | E-ISSN :26205440 Subject Area : Education	S1 Accredited	http://ijere.iaescore.com/index.php/IJERE	December , September , June , March
MAKARA of Science Series	Universitas Indonesia	P-ISSN : 23391995 | E-ISSN :23560851 Subject Area : Science	S1 Accredited	http://journal.ui.ac.id/index.php/science	June , March , December , September , August , April
Gadjah Mada International Journal of Business (GamaIJB)	Universitas Gadjah Mada	P-ISSN : 14111128 | E-ISSN :14111128 Subject Area : Economy	S1 Accredited	https://jurnal.ugm.ac.id/gamaijb	May , August , January , April , September , December
Jurnal Pendidikan IPA Indonesia (Indonesian Journal of Science Education)	Science Education Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang	P-ISSN : 23391286 | E-ISSN :20894392 Subject Area : Education	S1 Accredited	http://journal.unnes.ac.id/nju/index.php/jpii	June , March , December , September , April
TELKOMNIKA (Telecommunication Computing Electronics and Control)	Universitas Ahmad Dahlan (UAD) in collaboration with Institute of Advanced Engineering and Science (IAES)	P-ISSN : 23029293 | E-ISSN :16936930 Subject Area : Science, Engineering	S1 Accredited	http://journal.uad.ac.id/index.php/telkomnika	June , April , February , December , August
Jurnal Cakrawala Pendidikan	Lembaga Pengembangan dan Penjaminan Mutu Pendidikan UNY	P-ISSN :| E-ISSN :24428620 Subject Area : Education	S1 Accredited	http://journal.uny.ac.id/index.php/cp	June , February , November

Figure 6. SINTA Database View

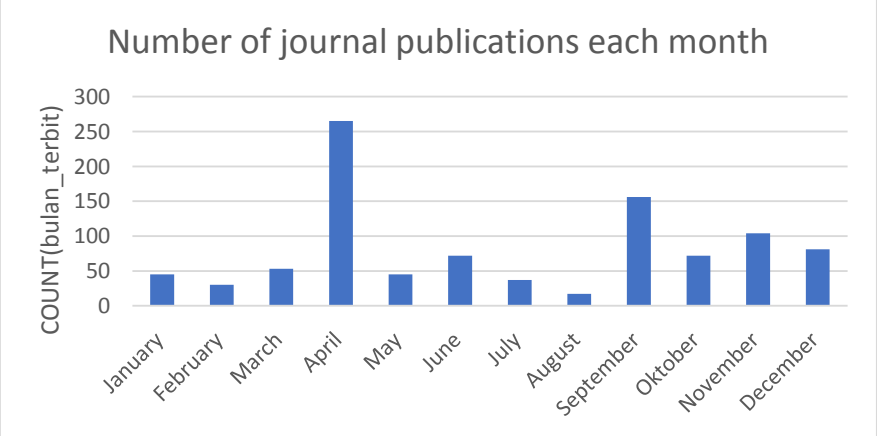

Figure 7. Graph of publication classification results every month

At the final stage after the data is stored in the database is to schedule a cron job using the task scheduler tool. By setting the time on the trigger menu, the author schedules the script to be run every Monday at 9 pm. Then

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/ 10.33395/sinkron.v7i4.11576
e-ISSN : 2541-2019
p-ISSN : 2541-044X

write down the action that will be done by the task scheduler, namely by running the web scraping script program and adding a database update function so that when the script is run, it will automatically make changes to the database if there are changes to the SINTA website. In figure 7, the process of creating scheduling on the task scheduler is depicted.

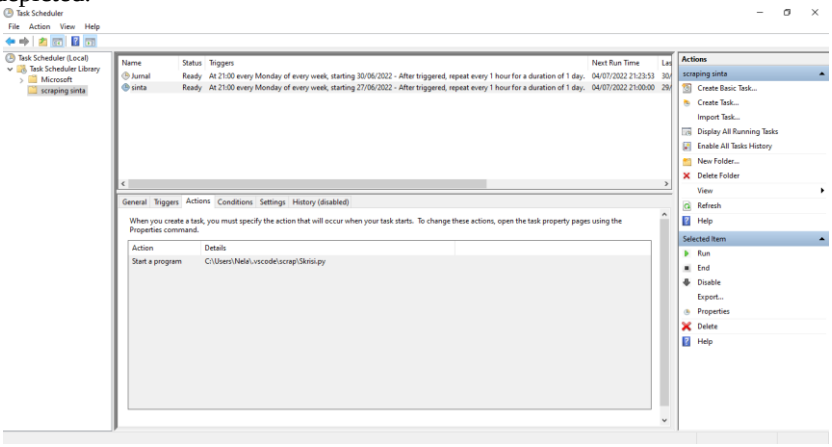

Figure 8. Task scheduler process

CONCLUSION

Research using web scraping techniques as a method of collecting journal data, the resulting data is 7412 data from the SINTA website, then after filtering using a Mysql query the data becomes 977 data by displaying information on the month of publication of the journal. Not all journals include information on the month of publication on the website and some journals have inactive websites, it is hoped that publishers can update the latest version of OJS so that it can facilitate this research.

REFERENCES

Al hadi, I. F., Chusna, C., Ilham, S., & Fauzan, A. C. (2019). Implementasi Penjadwalan Round Robin pada Task Scheduler untuk Pembaruan Aplikasi Otomatis. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(1), 11–14. https://doi.org/10.28926/ilkomnika.v1i1.7

Arumi, E. R., & Sukmasetya, P. (2020). Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm. *JOIN (Jurnal Online Informatika)*, *5*(1), 19–26. https://doi.org/10.15575/join.v5i1.548

Christanto, F. W., & Rudiyanto. (2020). Cron Job Technique pada Integrasi WLAN Controller Device dan Google Maps API Berbasis Website dalam Jaringan Indonesia Wifi. *Matrix : Jurnal Manajemen Teknologi Dan Informatika*, *10*(2), 50–57. https://doi.org/10.31940/matrix.v10i2.1477

Josi, A., & Andretti Abdillah, L. (n.d.). *PENERAPAN TEKNIK WEB SCRAPING PADA MESIN PENCARI ARTIKEL ILMIAH*.

Lukman, Ahmadi, S. S., Manalu, W., & Hidayat, D. S. (2019). Publikasi. *Kementerian Riset, Teknologi, Dan Pendidikan Tinggi*, 1–214. http://risbang.ristekdikti.go.id

Mitchell, R. (2018). *Web Scraping with Python and BeautifulSoup*. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 9.

Mufidah, U. (2021). Perancangan Aplikasi Perbandingan Harga Produk (Historical Data) Menggunakan Teknik Web Scraping. *Skripsi*, *1*(1), 1–14.

Priyanto, A., & Ma'arif, M. R. (2018). Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik). *Indonesian Journal of Information Systems*, *1*(1), 25–33. https://doi.org/10.24002/ijis.v1i1.1664

Purnomo, L. M., & Ayub, M. (2021). Analisis data hasil web scraping untuk menentukan kualitas jurnal ilmiah. *Jurnal STRATEGI-Jurnal Maranatha*, *3*(1), 122–132. http://strategi.it.maranatha.edu/index.php/strategi/article/view/237

Rahmatulloh, A., & Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar. *Indonesian Journal of Information Systems*, *2*(2), 95–104. https://doi.org/10.24002/ijis.v2i2.3029

Sahria, Y. (2020). Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia. In *URECOL (Unversity Research Colloqium)*. http://repository.urecol.org/index.php/proceeding/article/view/1079

Saputra, A. (2020). *Pemanfaatan Science and Technology Index (SINTA) untuk Publikasi Karya Ilmiah dan*

Pencarian Jurnal Nasional Terakreditasi.

Sembiring, F., & Erfina, A. (2020). *Bahasa Ular untuk Pemrograman Python* (R. Aminah (ed.)). Insan Cendekia Mandiri.

Sembiring, F., Fergina, A., Saepudin, S., Erfina, A., & Gustian, D. (2020). *Fundamental_Basis_Data*. Media Sains Indonesia.

Sembiring, F., & Sari, D. P. (2019). *INTEGRATED (Information Tecknology and Vocational Education) Volume Design Process Data Storage and Organize Dat... Design Process Data Storage and Organize Data Scraping*. *1*(1), 22–26.

Uzun, E., Yerlikaya, T., & Kırat, O. (2018). Comparison of Python Libraries used for Web Data Extraction. *Journal of the Technical University - Sofia Plovdiv Branch, Bulgaria*, *24*(May), 87–92. https://erdincuzun.com/wp-content/uploads/download/plovdiv_journal_2018_01.pdf

Wirawan, A. (2020). Sistem Scheduling Pelaporan Data Akademik di UIN Sunan Kalijaga ke Pangkalan Data Pendidikan Tinggi (PDDikti) dengan Menggunakan Fitur Cron Job di Linux. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *5*(3), 177–184. https://doi.org/10.14421/jiska.2020.53-05

Yanti, G., Z, Z., & Megasari, S. W. (2020). Pelatihan Penulisan Artikel untuk Publikasi E-Jurnal bagi Researcher Club. *Dinamisia : Jurnal Pengabdian Kepada Masyarakat*, *4*(3), 461–469. https://doi.org/10.31849/dinamisia.v4i3.4107

*nelawati adila