The Proposed Algorithm for Semi-Structured Data Integration: Case Study of Setiu Wetland Data Set

Mustafa Man and Ily Amalina Ahmad Sabri
School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
ilylina@yahoo.com

***Abstract*—Recent evolutions in web technology and computer science provide environmental community in expanding resources for data collection and analysis. Today, people are facing challenges to the design of analysis methods, workflows, and interaction with data sets. Data integration is one of older research fields in database area. It is consists of three types of data; structured data, semi-structured data and unstructured data. Web pages is a part of semi-structured data. In this paper, we briefly introduce the problem of data extraction from web pages focus on images. We also discuss the evolution of extraction images from semi-structured to structured format using WEIDJ (Wrapper for extraction Images using Document Object Model (DOM) and JavaScript Object Notation Data (JSON) approach). An experiment was conducted on same website using different approach JSON and DOM to show the comparison of time performance.**

***Index Terms*—Document Object Model, JSON, Semi-Structured Data, WEIDJ.**

I. Introduction

The Big data represents a new era in data exploration. It is a large of data set that need tools or applications for data processing purpose. Data is tamed and understood using computer and mathematical models. These models are useful for understanding, but they have their limitations. Today, the world is experiencing a dynamic change in technologies, economies and societies. Empowering rural women in Setiu Wetlands by promoting women's entrepreneurship using e-business has been keyed out as an important approach to improve living standards and further sustainable development. There is a large volume of information available to be mined from the empowerment of women's activities via their web pages. The information on the web is contained in the form of structured, semi-structured and unstructured data. Those data can be transformed into meaningful information records. Such information records like demographic profile, economic activities in each village are important to be managed properly and store in a central database. It is necessary to extract such information records to provide relevant information needed by the decision makers for developing new policy about Women's activities. Certain information needs to be combined or integrated to gain a more comprehensive and meaningful. The integrated information also can be used for querying and reporting a comparative business activity.

The goal for data integration is aiming combination of multiple information systems in similar domain to provide users the illusions of interacting in single information systems. Generally, information systems are not designed to be integrated and combined. Data from different sources may not fit together. Certain techniques and approaches need to be consider while the goal is always to provide a homogenous, unified view on data from different sources.

Data extraction is an emerging area of data integration that offers various theories, techniques and tools for processing large volumes of data. The growing volume of semi-structured data that is available in world wide web has enhanced this curiosity. Every webpage has different images in different domain. There are several issues that need to be focused in web data extraction; i. how to extract images automatically from web pages without failure? ii. the difficulty to extract images from different unified resource locator (url) locations in one extraction process, iii. What approach is suitable to extract images in large volumes data and speedy?

There have been lots of techniques have been proposed in web data extraction [1-3]. The relevant information can be extracted by processor engine for web pages. Manjaramkar and Lokhande [4] agree that extraction information must be presented in structured format for record management. The performance of search engine can be increased due of the time consuming and memory allocation in storing irrelevant content. There exist other techniques that make use of the DOM structure [5, 6]. Many researches use Document Object Model (DOM) to clean, improve or transform the original web page but our research using DOM to identify the page level. This approach is different from others. Our approach will be discussed further in WEIDJ.

This research aims to establish a data integration middleware which can extract multimedia data types from web and export to multimedia database. This research embarks on the following objectives: i.to study the concept of multiple types semi-structured data for data integration process, ii.to develop a significant multiple types semi-structured data integration framework, iii.to develop specific algorithm for high availability semi-structured data integration processing and iv.to test and evaluate the accuracy and performance of the newly proposed framework. This paper propose algorithm for semi-structured data integration to extract images from single web page.

In this paper, Section II will discuss the related work. Then, we will show the perspective of JSON in Section III and the design of our model WEIDJ in detail in section IV. The experimental results will be discussed in Section V. Finally, Section VI will conclude the paper.

II. Related Works

Commercial websites such as portal sites, news and e-commerce contain much irrelevant and redundant information. It has been approved that around 50% of the content web page is generally irrelevant [7]. Irrelevant information mainly logo, table of contents, footers and headers are known as noisy of web page. There is a need to find a potential method for detecting main content of a web page without focusing on irrelevant information. Content extraction is a process of identifying main content blocks depends on user request. There have been so many studies about extracting data from web documents and numerous methods that have been developed [2, 4, 8, 9]. Many applications use the JSON approach to exchange data [10, 11]. People nowadays more comfortable return valuable information in simple way. Previous work shows people is convenient to extract data using DOM technique [5, 6], algorithm [12, 13], and wrapper [14-17]. JSON is knowns as schema less data format that is easy for computer to parse and use data [18].

Vagač, et al. [19] present automatic processing of traceology objects (APTO) to extract images from Internet. The main purpose of this work is to identify shoe manufacturer and brand according to foot print obtained at crime scene.

Kanaoka, et al. [20] propose a tool called Ducky, a semi-automatic web wrapper, which acts as a mediator tool that can extract data from web sources and translates them into structured data. The extraction process based on simple data extraction rule and consists of several parameters: unified resource locator (*url*) of the webpage, CSS selectors and so on. This study focus on text extraction not only single page but also hierarchical structured web sites.

Abidin, et al. [21] state that the unstructured data such as multimedia files, documents, spreadsheets, news, emails, memorandums, reports and web pages are difficult to capture and store in the common of this work is to database storage. Even there are many tools and techniques that proved to be successful in transforming unstructured data to valuable information but it simply do not work when it comes to unstructured or semi-structured data.

Derouiche, et al. [22] discuss that techniques or approaches for automatic extraction and integration of complex structured data from web pages is very important so that the process can be done fast and effectively. Ferrara, et al. [23] summarize many extractor and wrapper for web data extraction but there are lack tools to extract semi-structured data focusing on image from web pages then store it into database using JSON approach. Performance of extracting web page content can be increased by only processing relevant information [1]. These applications make users feel easier to access the relevant information in timely manner.

III. JavaScript Object Notation (JSON)

The development of web applications has become attractive disciplined in the web environment. The use and composition of different of API technology is very important and influent applications. This issue need to be dealt to discover JSON approach based on the web. In recent years, a new technology, JSON based on web applications has been spreading the web environment. JSON stands for JavaScript Object Notation, is a lightweight data-interchange format. It is self-describing and easy to understand. It is easy for humans to read and write. It is also easy for machines to parse and generate data. Attractive interface of website is very important to be competitive among web applications but their contents also attract user attention for further beneficial purpose.

Today, JSON is applied in web applications for various purpose include data extraction. Webpage consists of large volume of multimedia data [24, 25]. That is the reason why web data extraction becomes challenging task to mine the data from web page. The performance of data extraction should be speedy and efficient in handling multimedia data. A suitable approach is required to deal with this issue. JSON, DOM and XML are different technologies that have been developed to solve different problems. They are designed for different purpose. Table 1 discussed different technologies of JSON, DOM and XML.

Table 1
Units for Magnetic Properties

JavaScript Object Notation (JSON)	Document Object Model (DOM)	eXtensible Markup Language (XML)
• JSON is a lightweight, text based format can be used for data interchange format. • it is human readable	• DOM is used for manipulating and representing html and xml documents.	• XML is designed to store and transport data. • It is readable for human and machine. • XML is more complex compare to JSON.

This research proposes the information integration algorithm for data extraction focus on semi-structured data such as image, video, audio and text by using Document Object Model (DOM) and JavaScript Object Notation (JSON). Based on the proposed integration models, a mediator tool called as a wrapper will be developed as experimental to extract semi-structured data from heterogeneous source like web pages. Experiments will be conducted on Setiu Wetlands web site and biodiversity web pages dataset for testbed [26].

We apply DOM as a part of our method. Our objective is to detect images from web page, extract images in structured format such as in tabular form and store them into single multimedia database. The implementation is also different. The pre-processor will filter the similarity of images either there is a similar image or not. So, we construct repetitive algorithm to filter the similarity file of images. We focus on extracting all images and mine image details into tabular format using JSON environment. Then user can select the relevant images to be stored in database. All stored images will be indexed before stored in database to avoid similarity file name and redundancy records. Based on this process, we already extract semi-structured data (images) to structured format (tabular data). After that, we can manipulate data using Database Management System (DBMS). Details of our model will be discussed in next section, WEIDJ Model.

IV. WEIDJ Model

In this paper, we propose a mediator tool call WEIDJ (Wrapper for Extraction Images using Document Object Model (DOM) and JavaScript Object Notation Data (JSON) approach). This tool aims to extract images according to unified resource locator (*url*) and mine image details then presents images in structured format before storing them into multimedia database. Figure1 shows the flowchart of the WEIDJ algorithm [27].

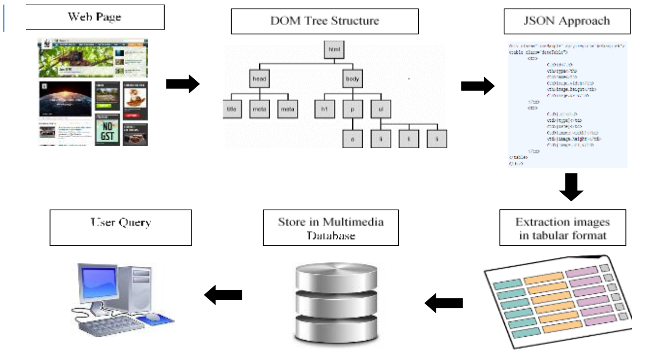


Figure 1: The WEIDJ flowchart

Figure 2 shows WEIDJ algorithm that has been proposed in our research. This algorithm applies DOM to structure the html documents in hierarchical structure. Then visual segmentation blocks are developed for html page to check available element of images in each block. This approach is important to make sure all required images can be extracted without failure. JSON environment approach is applied in extracting images. Other rules like filtering similar filename and removing noisy images such as logo and button are to be considered to make sure images that has been extracted are valuable information. At last, final images and their details will be display in tabular format before users can store them into multimedia database.

```
Algorithm 1: Extraction Images
INPUT     Web pages
STEP 1    Create DOM tree structure for each web page
STEP 2    Create visual segmentation block for each web page
          based on pattern rules.
STEP 3    Apply JSON approach in extracting multimedia element.
STEP 4    Avoid extracting similar image.
STEP 5    Remove noisy images
STEP 6    List all extracted images with details in tabular format
OUTPUT    Store images in multimedia database.
```

Figure 2: The WEIDJ Algorithm

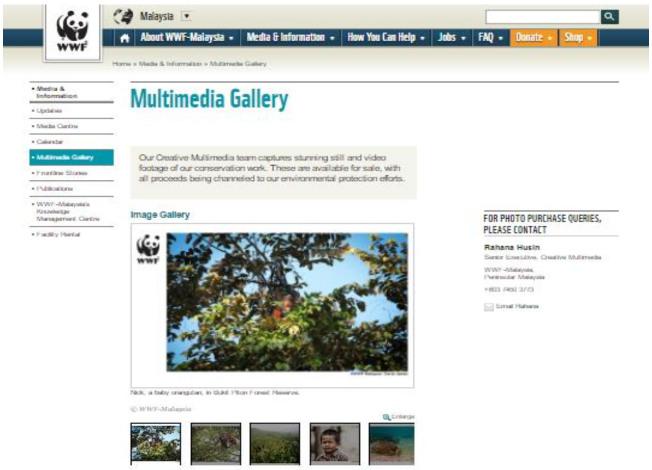
Some JSON formats use nested structures to simply group data together. Figure 3 shows an example of web API images extraction and their properties using JSON environment.

```
 $json_url = $json_url_path.$_REQUEST['url'];
<thead>
 <tr>
 <th>Select</th>
            <th>No</th>
            <th>Link</th>
            <th>Image</th>
            <th>Size</th>
            <th>Time Processing</th>
</tr>
</thead>
            foreach ($value as $key => $val) {
 //echo $key . '=>' . $val['src']. '<br/>';
 //var_dump($val);
 $img_url = $val['src'];
```

Figure 3: Image extraction using JSON coding

V. RESULT AND DISCUSSION

A data extraction engine need to be able to extract all the data that are required from web page. We need to define the unified resource locator (*url*) of the web page where the objective data is located. This is initial process to extract data from a specific web page. Figure 4 shows Setiu Wetlands web page namely WWF- Malaysia. WWF stands for World Wide Fund for Nature. It was formerly known as the World Wildlife Fund but adopted its current name to show that it also works on other environmental issues, and not just wildlife.



Our Creative Multimedia team captures stunning still and video footage of our conservation work. These are available for sale, with all proceeds being channeled to our environmental protection efforts.

Nick, a baby orangutan, in Bukit Piton Forest Reserve.

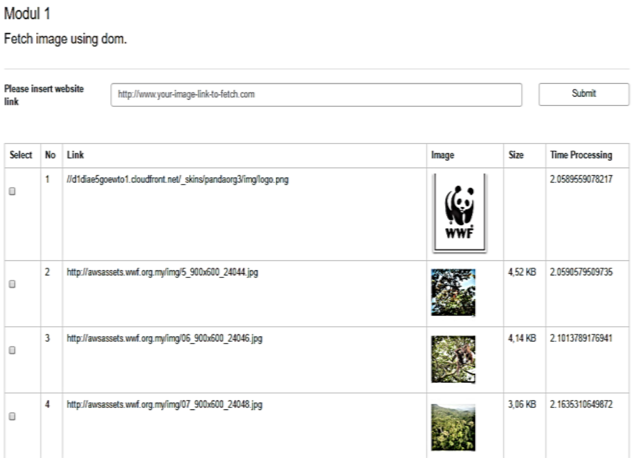
© WWF-Malaysia

FOR PHOTO PURCHASE QUERIES, PLEASE CONTACT

Rahana Husin
Senior Executive, Creative Multimedia
WWF-Malaysia,
Peninsular Malaysia
+603 7450 3773

Figure 4: Setiu Wetlands web page

A sample image extraction from WWF web page is shown in Figure 5. The extraction process done using DOM approach. The extracting process will mine images and their information and transform them into tabular format.



Modul 1

Fetch image using dom.

Please insert website link: http://www.your-image-link-to-fetch.com [Submit]

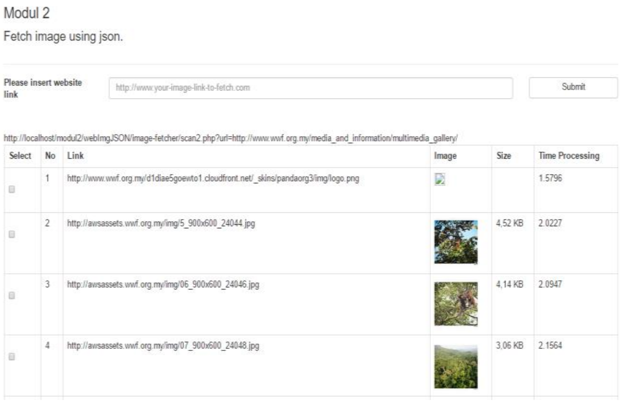
Select	No	Link	Image	Size	Time Processing
□	1	//d1diae5goewto1.cloudfront.net/_skins/pandaorg3/img/logo.png			2.0589559078217
□	2	http://awsassets.wwf.org.my/img/5_900x600_24044.jpg		4,52 KB	2.0590579509735
□	3	http://awsassets.wwf.org.my/img/06_900x600_24046.jpg		4,14 KB	2.1013789176941
□	4	http://awsassets.wwf.org.my/img/07_900x600_24048.jpg		3,06 KB	2.1635310649872

Figure 5: Image extraction using DOM approach for Setiu Wetlands homepage.

Figure 6 shows image extraction using JSON approach. The concept of extraction is same to extract images information into structure format but JSON is speedy than DOM.

Figure 7 shows the comparison of time extraction processing between DOM and JSON approach in extracting images from Setiu wetland web page.

An example using biodiversity site: http://www.biodiversityexplorer.org/ as shown in Figure 8. This web page focuses on the links between biodiversity, conservation and local people's livelihoods. It is organized by International Institute for Environment and Development (IEED).



Modul 2

Fetch image using json.

Please insert website link: http://www.your-image-link-to-fetch.com

Submit

http://localhost/modul2/webImgJSON/image-fetcher/scan2.php?url=http://www.wwf.org.my/media_and_information/multimedia_gallery/

Select	No	Link	Image	Size	Time Processing
□	1	http://www.wwf.org.my/d1diae5goewto1.cloudfront.net/_skins/pandaorg3/img/logo.png			1.5796
□	2	http://awsassets.wwf.org.my/img/5_900x600_24044.jpg		4,52 KB	2.0227
□	3	http://awsassets.wwf.org.my/img/06_900x600_24046.jpg		4,14 KB	2.0947
□	4	http://awsassets.wwf.org.my/img/07_900x600_24048.jpg		3,06 KB	2.1564

Figure 6: Image extraction using JSON approach for Setiu Wetlands homepage.

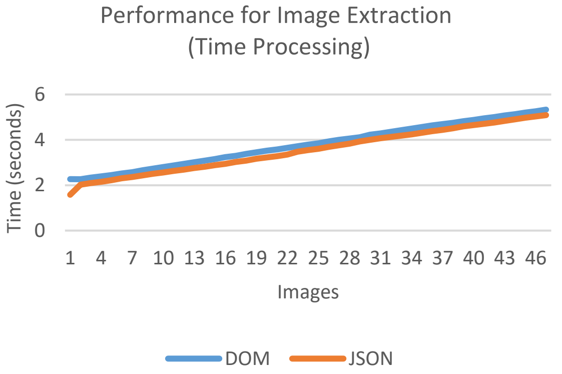
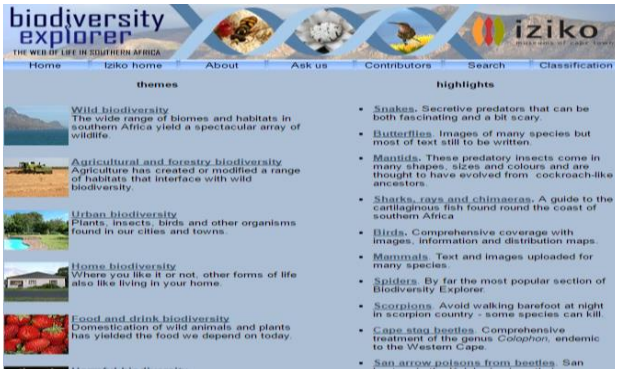


Figure 7: Time performance for image extraction using JSON and DOM



themes

Wild biodiversity
The wide range of biomes and habitats in southern Africa yield a spectacular array of wildlife.

Agricultural and forestry biodiversity
Agriculture has created or modified a range of habitats that interface with wild biodiversity.

Urban biodiversity
Plants, insects, birds and other organisms found in our cities and towns.

Home biodiversity
Where you like it or not, other forms of life also like living in your home.

Food and drink biodiversity
Domestication of wild animals and plants has yielded the food we depend on today.

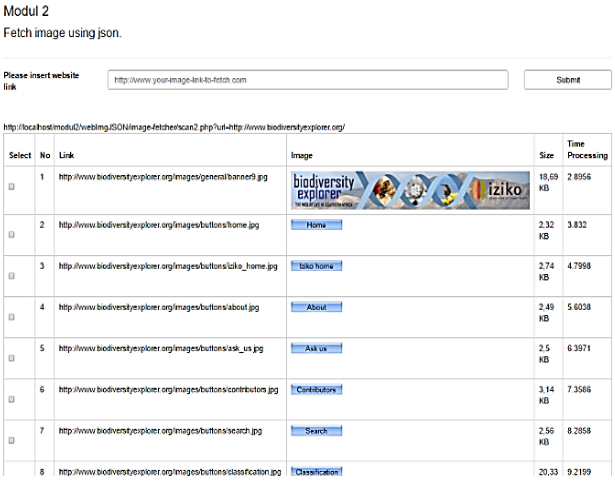
highlights

- Snakes. Secretive predators that can be both fascinating and a bit scary.
- Butterflies. Images of many species but most of text still to be written.
- Mantids. These predatory insects come in many shapes, sizes and colours and are thought to have evolved from cockroach-like ancestors.
- Sharks, rays and chimaeras. A guide to the cartilaginous fish found round the coast of southern Africa
- Birds. Comprehensive coverage with images, information and distribution maps.
- Mammals. Text and images uploaded for many species.
- Spiders. By far the most popular section of Biodiversity Explorer.
- Scorpions. Avoid walking barefoot at night in scorpion country - some species can kill.
- Cape stag beetles. Comprehensive treatment of the genus *Colophon*, endemic to the Western Cape.
- San arrow poisons from beetles. San

Figure 8: Biodiversity explorer web page

This web is used as a platform for organizer to provide various information such as promoting sustainable development to improve livelihoods and protect the environment on which these livelihoods are built. Many valuable information such as text and images can be found in this web page. The main motivation for this paper is to extract images and mine image details such as images, links of images, size of images and store selected images in single multimedia database. In ideal scenario, if people want to save image, it can be extracted manually. People can extract them manually by saving each image as many as possible. But how to extract and mine images manually if there are large of volume images? Therefore, another solution must be developed to extract images automatically to reduce time consuming. The important part of extraction system is database of records. This is because the records that have been extracted and saved can be used for beneficial purpose such as documentation, analyze reports and so on.

An experimental has been done to extract image using different approach. This experimental is important to identify the characteristics of DOM and JSON such as the ability to extract images and time taken for extraction process. Figure 9 shows the experimental of image extraction using JSON approach. The extraction of semi-structured data, images will be displayed in tabular format as shown in Figure 8. There are six fields in the table; column to select images, the number of images, image, size of image and time for extraction processing. After extraction process, user can select images that want to be stored in database.



Modul 2

Fetch image using json.

Please insert website link: http://www.your-image-link-to-fetch.com

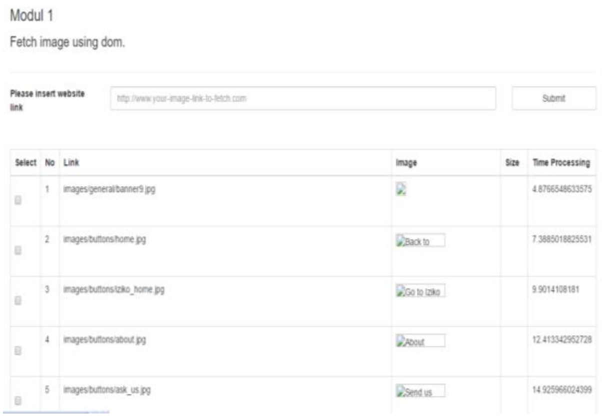
Submit

http://localhost/modul2/webImgJSON/image-fetcher/scan2.php?url=http://www.biodiversityexplorer.org/

Select	No	Link	Image	Size	Time Processing
□	1	http://www.biodiversityexplorer.org/images/general/banner9.jpg		18,69 KB	2.8956
□	2	http://www.biodiversityexplorer.org/images/buttons/home.jpg	Home	2,32 KB	3.832
□	3	http://www.biodiversityexplorer.org/images/buttons/iziko_home.jpg	Iziko home	2,74 KB	4.7998
□	4	http://www.biodiversityexplorer.org/images/buttons/about.jpg	About	2,49 KB	5.6038
□	5	http://www.biodiversityexplorer.org/images/buttons/ask_us.jpg	Ask us	2,5 KB	6.3971
□	6	http://www.biodiversityexplorer.org/images/buttons/contributors.jpg	Contributors	3,14 KB	7.3586
□	7	http://www.biodiversityexplorer.org/images/buttons/search.jpg	Search	2,56 KB	8.2858
	8	http://www.biodiversityexplorer.org/images/buttons/classification.jpg	Classification	20,33	9.2199

Figure 9: Image extraction using JSON

Figure 10 shows image extraction using DOM approach. In Figure 10, image cannot be retrieved as detail as Figure 8. This is the limitation of extraction using simple DOM. There are certain images that cannot be extracted.



Modul 1

Fetch image using dom.

Please insert website link: http://www.your-image-link-to-fetch.com

Submit

Select	No	Link	Image	Size	Time Processing
□	1	images/general/banner9.jpg			4.8766548633575
□	2	images/buttons/home.jpg	Back to		7.3885018825531
□	3	images/buttons/iziko_home.jpg	Go to Iziko		9.9014108181
□	4	images/buttons/about.jpg	About		12.413342952728
□	5	images/buttons/ask_us.jpg	Send us		14.925966024399

Figure 10: Image extraction using DOM

As a result, Figure 11 shows comparison based on time performance for image extraction using JSON and DOM. This graph shows JSON can extract images in speedy than DOM. The result of image extraction consists of 51 images includes of noisy images. The objective data was extracted by defining *url* in the configuration rules for both techniques; JSON and DOM. Some specific noisy images are contained in web pages such as logo and buttons are also extracted because of the structure of html documents. A mediator tool call Wrapper for Extraction Images using Document Object Model (DOM) and JavaScript Object Notation Data (JSON) approach WEIDJ will be developed and discussed in future work to overcome those limitations.

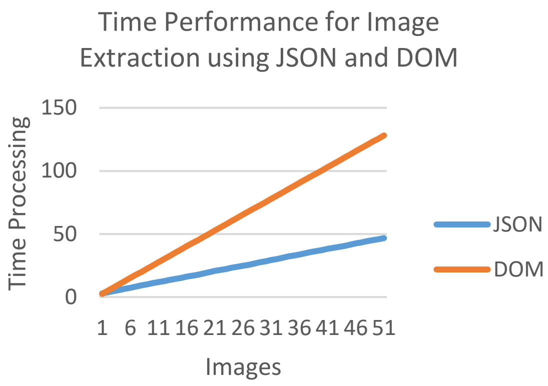


Figure 11: Time performance for image extraction using JSON and DOM

VI. Conclusion

The summary from the findings, we can conclude that extensive research need to be focused on how to integrate and extract semi-structured data focusing on images from multi-sources. It also focusses on arranging the extracted data in a tabular format then how to store them automatically into database from multi-web pages. The combination of both techniques should ensure data extraction process can be done fast and effectively.

This paper proposes an alternative approach to extract images from semi-structured web pages automatically. The propose technique is to hybrid between JSON and DOM in extracting data. In this paper, users need to specify the target of unified resource locator (url). This research aims the extraction target of structured data images from web pages details in tabular format. The advantages of having specific extraction target are avoidance of unnecessary process in extraction and improvise of the quality of extraction.

This paper validates the result through the experimental of images extraction process using two techniques; JSON and DOM. We test the input for different domain as discussed in section result and discussion. The process for extraction is time consuming, this research working on the enhancement of WEID algorithm which will extract images of web pages in efficient way.

Future work, by leveraging the input description using mediator tool called as WEIDJ, our tool harvest more real-world situation without failure. The prototype tool, WEIDJ could be further enhanced by integrating heterogeneous sources for input descriptions. Then the data extraction that stored in multimedia database could be manipulate into any manageable forms such as reports, statistics, documentation and others.

Acknowledgment

This research is supported by Scholarship of Universiti Malaysia Terengganu (BUMT) and Niche Research Grant Scheme (vot53131).

References

[1] S. López, J. Silva, and D. Insa, “Using the DOM tree for content extraction,” in *Proceedings 8th International Workshop on Automated Specification and Verification of Web Systems*, 2012, pp. 46-59.

[2] S. M. Narawade, N. M. Prabhakar, N. S. Maruti, S. M. Bhagwat, and B. Burghate, “A web based data extraction using hierarchical (DOM) tree approach,” *International Journal for Innovative Research in Science and Technology*, vol. 2, no. 11, pp. 255-257, 2016.

[3] T. Weninger, W. H. Hsu, and J. Han, “CETR: content extraction via tag ratios,” in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 971-980.

[4] A. Manjaramkar and R. L. Lokhande, “DEPTA: An efficient technique for web data extraction and alignment,” in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2307-2310.

[5] V. B. Kadam and G. K. Pakle, “DEUDS: Data extraction using DOM tree and selectors,” *International Journal of Computer Science and Information Technologies*, vol. 5, no.2, pp. 1403-1410, 2014.

[6] B. Mehta and M. Narvekar, “DOM tree based approach for web content extraction,” in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, 2015, pp. 1-6.

[7] D. Gibson, K. Punera, and A. Tomkins, “The volume and evolution of web page templates,” in *Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005, pp. 830-839.

[8] K. S. K. Niasi, E. Kannan, and M. M. Suhail, “Page-level data extraction approach for web pages using data mining techniques,” *International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1091-1096, 2016.

[9] P. Rawat, S. Sayyad, S. Surinder, and S. Shelke, “Application for web data extraction and analysis,” *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 7, pp. 148-152, 2016.

[10] K. Kanaoka and M. Toyama, “Effective web data extraction with ducky,” in *Proceedings of the 19th International Database Engineering & Applications Symposium*, 2015, pp. 212-213.

[11] D. Peng, L.-D. Cao, and W.-J. Xu, “Using JSON for data exchanging in web service applications,” *Journal of Computational Information Systems*, vol. 7, no. 16, pp. 5883-5890, 2011.

[12] D. Buttler, L. Liu, and C. Pu, “A fully automated object extraction system for the World Wide Web,” in *21st International Conference on Distributed Computing Systems*, 2001, pp. 361-370.

[13] C. Hong-ping, F. Wei, Y. Zhou, Z. Lin, and C. Zhi-Ming, “Automatic data records extraction from list page in deep web sources,” in *Asia-Pacific Conference on Information Processing 2009 (APCIP 2009)*, 2009, pp. 370-373.

[14] L. Liu, C. Pu, and W. Han, “XWRAP: An XML-enabled wrapper construction system for web information sources,” in *16th International Conference on Data Engineering*, 2000, pp. 611-621.

[15] V. Crescenzi, G. Mecca, and P. Merialdo, “Roadrunner: towards automatic data extraction from large web sites,” in *VLDB*, 2001, pp. 109-118.

[16] C.-N. Hsu and M.-T. Dung, “Generating finite-state transducers for semi-structured data extraction from the web,” *Information Systems*, vol. 23, no. 9, pp. 521-538, 1998.

[17] C.-H. Chang and S.-C. Lui, “IEPAD: information extraction based on pattern discovery,” in *Proceedings of the 10th International Conference on World Wide Web*, 2001, pp. 681-688.

[18] M. K. Yusof and M. Man, “Efficiency of JSON approach for Data Extraction and Query Retrieval,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 4, no. 1, pp. 203-214, 2016.

[19] M. Vagač, M. Melicherčík, M. Marko, P. Trhan, A. Michalíková, R. Kliment, *et al.*, “Crawling images with web browser support,” in *2015 IEEE 13th International Scientific Conference on Informatics*, 2015, pp. 286-289.

[20] K. Kanaoka, Y. Fujii, and M. Toyama, “Ducky: a data extraction system for various structured web documents,” in *Proceedings of the 18th International Database Engineering & Applications Symposium*, 2014, pp. 342-347.

[21] S. Z. Abidin, N. M. Idris, and A. H. Husain, “Extraction and classification of unstructured data in WebPages for structured multimedia database via XML,” in *2010 International Conference on Information Retrieval & Knowledge Management,(CAMP)*, 2010, pp. 44-49.

[22] N. Derouiche, B. Cautis, and T. Abdessalem, “Automatic extraction of structured web data with domain knowledge,” in *2012 IEEE 28th International Conference on Data Engineering*, 2012, pp. 726-737.

[23] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, “Web data extraction, applications and techniques: a survey,” *Knowledge-Based Systems*, vol. 70, pp. 301-323, 2014.

[24] M. Man, I. A. A. Sabri, M. M. A. Jalil, N. H. Ali, and S. Muhamad, “Information integration architecture system for empowering rural woman in Setiu wetlands,” presented at the *Seminar Ekosistem Setiu 2016: Sains Marin & Sumber Akuatik Untuk Kelangsungan Hidup*, Universiti Malaysia Terengganu, 2016.

[25] I. A. A. Sabri and M. Man, “Multiple types of semi-structured data extraction using WEID,” presented at the *Regional Conference on Sciences, Technology and Social Sciences (RCSTSS)*, Copthorne Hotel Cameron Highlands, 2016.

[26] J. Creech. (2012, 31 May 2017). *Biodiversity Web Resources*. Available at http://www.istl.org/12-fall/internet.html

[27] I. A. A. Sabri and M. Man, "WEIDJ : An improvised algorithm for image extraction from web pages," presented at the T*he 8th International Conference on Information Technology*, Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan, 2017.