

INFORMATION INTEGRATION ARCHITECTURE SYSTEM FOR EMPOWERING RURAL WOMAN IN SETIU WETLANDS, TERENGGANU, MALAYSIA

MUSTAFA MAN¹, ILY AMALINA AHMAD SABRI^{1*}, NORAIDA ALII AND SURIYANI MUHAMAD²

¹*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia*

²*School of Social and Economic Development, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia*

*Corresponding author: ilylina@yahoo.com

Abstract: Nowadays, there is a shift in the dynamics of technology, economy and society. Empowering rural women by promoting women's entrepreneurship using e-business is viewed as an important approach to improve living standards and to further sustain development. There is a large volume of information available from the empowerment of women's activities via their web pages. The information on the web is available in the form of structured, semi-structured and unstructured data. Those data can be transformed into meaningful information records. Such information records such as demographic profile, economic activities in each villages are important to be managed properly and stored in a central database. It is necessary to extract such information records to provide relevant information needed by decision makers for developing new policy about women's activities. There are many techniques and algorithms used in data mining and machine learning. Certain information needs to be combined or integrated to gain a more comprehensive and meaningful data. The integrated information also can be used for querying and reporting a comparative business activity. This paper proposed information integration architecture for handling all the empowering women information activities in several villages at Setiu Wetlands, Terengganu, Malaysia in semi-structured data (images) by using Document Object Model (DOM) and JavaScript Object Notation (JSON).

KEYWORDS: Economic empowerment, semi-structured data, data extraction, data mining, information integration

Introduction *Climate Change*

Around the world, resilient and resourceful rural women contribute in a multitude of ways through different livelihood strategies to lift their families and communities out of poverty. They work as unpaid and are self-employed as on-farm and non-farm labourers, as on and non-farm wage labourers for others in agriculture and agro-industry, as entrepreneurs, traders, and providers of services, as leaders, as technology researchers and developers, and as caretakers of children and the elderly. They work in permanent and temporary employment and work along a rural-urban continuum cross-border context, with increasing numbers of rural women migrating for daily, seasonal, or permanent work in urban areas (Hill, 2011).

Rural women work long hours and many of their activities are not defined as "economically active employment" in national accounts but are essential to the well-being of their households. They also constitute a significant proportion of the labour on their family farms, whether producing for household consumption or for enterprise or both. Their potential is limited by multiple and diverse constraints by persistent structural gender disparities that prevent them from enjoying their economic and other rights. Rural women are constrained by unequal access to productive resources and services and inadequate or inaccessible infrastructure. The limitations rural women face in turn impose huge social, economic, and environmental costs on society as a whole and rural development in particular including lags in agricultural productivity (Hill, 2011).

A huge amount of data is available on the empowerment of women activities via web pages and it continues to grow rapidly. While there is a vast amount of data available, the importance issue should be concern are tools and methods to manage the semi-structured data to usable information.

In order to access information from a variety of heterogeneous information sources, web data extraction as a tool has to be able to translate queries and data from one data model into another. User commonly retrieve data from web by browsing and keyword searching but browsing is not suitable for locating particular item of data (Embley *et al.*, 1999). This is because the following links are tedious and it is easy to be off track. Browsing is not cost effective as users have to read the documents to find desired data. Keyword searching is sometimes more efficient than browsing.

Web data extraction system is a software applications that can extract data from web sources Laender *et al.* (2002). Ferrara *et al.* (2014) stated that this application usually interact with a web source and extract data stored in it. The extracted content could consist of elements in the HTML Web page. Finally, the extracted data might be post-processed, converted in the most appropriate structured format and stored for further usage.

Multimedia data such as images, video clips, animations, graphics and audio have increased rapidly over the past several years. Users have begun to expect that multimedia contents should be easily accessed. Relevant photo images that appear in webpages, video clips related to text articles are accessible. It is important to provide integrated access to diverse types of multimedia semi-structured data stored in disparate data sources. Many web data extractors today deal with multimedia data.

In this day and age, many systems for data extraction from web pages are developed (Ferrara *et al.*, 2014). A traditional approach is to write specific programs called as “extractor” or

“wrappers” is developed to extract the contents of the web pages based on certain criteria. A survey that offers a rigorous taxonomy to classify web data extraction systems has been presented by Laender *et al.* (2002). Chang *et al.* (2006) introduced a tridimensional categorization of web data extraction systems. The criteria are based on task difficulties, technique used and degree of automation.

In 2008, a relevant survey on Information Extraction was discussed by Sarawagi (2008). This paper posits that the automatic extraction of information from unstructured sources has opened up new avenues for querying, organizing, and analysing data by drawing upon the clean semantics of structured databases and the abundance of unstructured data. Flesca *et al.* (2004) surveyed approaches, techniques and tools for web wrappers. Baumgartner *et al.* (2009) surveyed about web data extraction.

Sangeeta (2016) proposes a tool that can process RDF Based Search and DOM Based Search to extract the relevancy data. Resource Description Framework is used along with DOM to provide precise answers to users’ queries. The DOM segment fusing algorithm is used to analyse and fuse the extracted information from web. Alarte *et al.* (2015) propose a method that can remove irrelevant information from web template. DOM tree is used to analyse the similarity between collections of webpages that are detected using a hyperlink analysis. The extraction of context to improve similarity data records and to extract the relevant result records based on stored URL list and Run Time Generated has been proposed by Mehta (2015). These papers are recommended to those who intend to approach these disciplines.

This study has been proposed to develop web data extractor that focus on extracting data based on different page levels. The different page levels may consists of extracting data from surface of web (Man & Sabri, 2017; Sabri & Man, 2017; 2018b), multi-source of web page and deep web.

Material and Method

In this section, we present the proposed architecture for information extraction in various web pages using DOM and JSON approach. Figure 1 illustrates the basic research framework for the overall process involved in information extraction. The prototype tool built

using PHP is designed and developed. The framework consists of User, Interface, Web page, XML and Multimedia Database layers. These layers communicate with each other in order to retrieve and construe data from user to multimedia database. The research framework is summarized in Table 1.

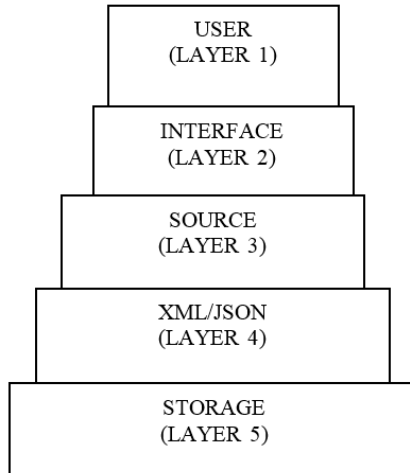


Figure 1: Research framework

Table 1: Explanation for research framework

Layer		Description
1	USER	<ul style="list-style-type: none"> Represent the user who will be using the implemented system
2	INTERFACE	<ul style="list-style-type: none"> Interaction medium between user and the source location that allows user to manipulate data. User will identify the useful data to be extracted to the source.
3	SOURCE	<ul style="list-style-type: none"> This layer consists of structured, semi-structured and unstructured data of web page. User need to identify the useful data to be extracted from the source. Extraction data later will be stored in a storage location that can handle various types of the multimedia elements. Data will be classified on the type of data such as text, image, audio, or video before it can be allocated in the storage.
4	XML/JSON	<ul style="list-style-type: none"> After classification process, the result will be placed into XML or JSON document. The structured data is then transmitted to the next layer.
5	STORAGE	<ul style="list-style-type: none"> The storage is a multimedia database. This database is used to organize huge amount of multimedia types of data such as text, audio, images and videos format.

Extraction and Classification

In this section, the proposed architecture is presented. The prototype tool built using PHP was designed and developed. Extraction and classification is a process to extract multimedia data from webpages. Classification is a main process before extraction. The multimedia data is classified as either in text, image, video and audio format. In Setiu wetland, a lot of information related to women activities conducted by World Wildlife Fund Organization (WWF) could be accessed via webpages in image format.

Figure 2 illustrates the architecture of the research architecture. This architecture is important in implementing the prototype system. It consists of four main components; web, adapter, metadata repository and multimedia database. It also consists of three

layers; Client Tier, Application Tier, and Data Tier. These layers communicate with each other in order to retrieve and analyse data from user to multimedia database.

Web is a collection of information in World Wide Web (WWW). There are structured, semi-structured and unstructured data in websites. These data need to be extracted for many purposes in different field. Data classification is a process that involves in adapter. Metadata repository is a central storage area for this architecture. It provides information about data sources for users. Due to its semi-structured and self-describing characteristics, for this implementation, XML and JSON are employed for testing algorithm. Metadata repository is a temporary storage location that stores the results of data extraction from the web.

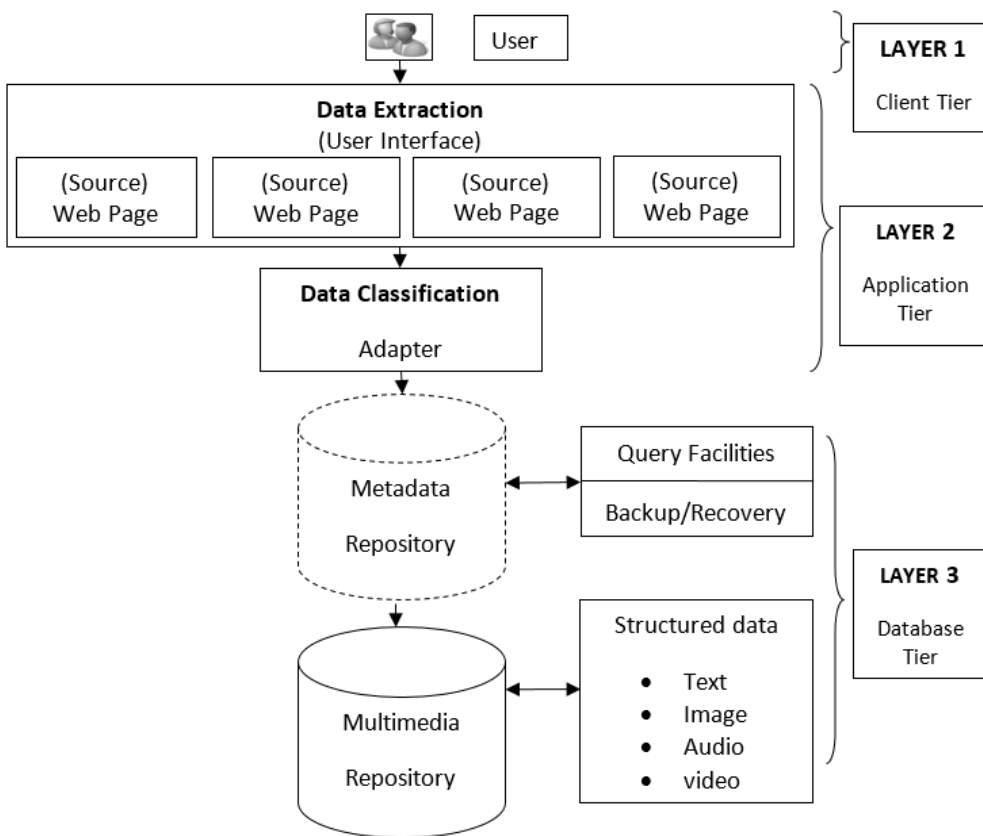


Figure 2: Multiple types of semi-structured data integration architecture

Web page comprises in the form of structured and unstructured objects known as data records. Most of the web page are in Hypertext Markup Language (HTML) format (Das & Kumar, 2014). Table 2 shows HTML tag descriptions for web page contents. By using HTML description, information of web page contents can be transformed into DOM tree structure. Document Object Model (DOM) is

a technique that is employed to find the correct data in the HTML document.

This technique allows DOM events to be processed simultaneously in data extraction as shown in Figure 3. It shows the body node which is in HTML tag. DOM functions to filter unnecessary nodes such as script, style or other items to be eliminated.

Table 2: HTML tag descriptions

No	HTML tags	Descriptions
1	<body></body>	The content of the document
2	<div></div>	Defines a division or a section in an HTML document
3	<a>	Defines a hyperlink, which is used to link from one page to another
4		Provides a way to add a hook to a part of a text or a part of a document
5		defines an image in an HTML page

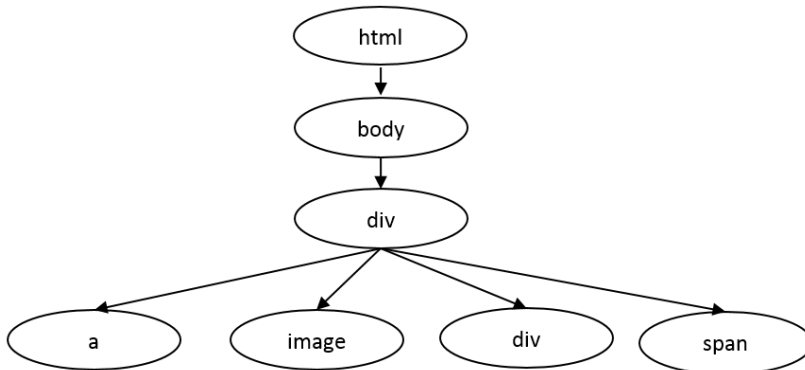


Figure 3: DOM tree structure

Multiple Types of Semi-Structured Data Extraction Algorithm

Algorithm relies on the DOM tree representation of a web page (Sabri & Man, 2018a). The algorithm discussed in this research is for extracting semi-structured data focusing on image that works on web page of Setiu Wetlands automatically. This algorithm will delete unnecessary HTML tags to reduce the processing time. This algorithm also indicated the block-level tags contain a significant amount of useful information and display. The algorithm below describes the aforementioned technique as shown in Figure 4:

1. Retrieve all data from web page
 - Compute the page pattern and identify

the list on each page. Compute a set of features (tags and html attributes)

2. Identify classification of data
 - Identify the list of image
3. Display the extracted data

Algorithm 1 for image extraction from web page

Input : P : a web page

K : keywords used to annotate the leaf nodes with the role r

Output : V : a list of image retrieve from web page

Begin

1. Open the web page internally.
2. Parse the web page in equivalent HTML code
3. cleans up the bad HTML tags and syntactical errors in P , and turns the body of P into a DOM tree, T .
4. discard HTML attributes and representation tags, such as b , i , and font, from T
5. **for** each leaf node i in T do search for data
6. **if** the content of i matches any keyword in K^r then
7. annotate i with the unidentified role and
8. retrieve the image from web page.
9. **if** the content of i does not match any keyword then
10. annotate i with the unidentified role

11. and move to next records.
12. Display all the relevant extracted data
13. return V

End

Figure 4: Algorithm 1 for image extraction from web page

Results and Discussion

This paper studies the problem of automatically extracting structured data encoded in a given collection of pages by inserting URL of web page. Extracting structured data from web pages is clearly very useful, since it enables to store the structured data from semi-structured data. Figure 5 shows WWF web pages (<http://www.wwf.org.my/>) related to the activities in Setiu Wetlands that has been tested for image extraction.



Figure 5: WWF web pages (Setiu wetland activities)

The useful multimedia data is classified using the regular expression, and DOM tree path learning

algorithm. Figure 6 depicts an example of HTML file for image source that has been used to extract image.

```








<div class="row">
  <div class="main-column" style="">
    <div class="container col6">
      <h3>Image Gallery </h3>
      <div id="galleria3852">
        <a href="http://awsassets.wwf.org/my/img/5_900x600_24043.jpg" id="img1_900x600"> | 1  | //d1diae5goewto1.cloudfront.net/_skins/pandaorg3/img/logo.png |  |         | 1.7837441158295 |
| <input type="checkbox"/> | 2  | http://awsassets.wwf.org.my/img/05_900x600_24044.jpg          |  | 4,52 KB | 1.7840991210938 |
| <input type="checkbox"/> | 3  | http://awsassets.wwf.org.my/img/06_900x600_24046.jpg          |  | 4,14 KB | 2.6468181610107 |
| <input type="checkbox"/> | 4  | http://awsassets.wwf.org.my/img/07_900x600_24048.jpg          |  | 3,08 KB | 3.2783249874115 |
| <input type="checkbox"/> | 5  | http://awsassets.wwf.org.my/img/08_900x600_24050.jpg          |  | 2,42 KB | 3.9007329940798 |
| <input type="checkbox"/> | 6  | http://awsassets.wwf.org.my/img/09_900x600_24052.jpg          |  | 2,5 KB  | 4.5711231231689 |
| <input type="checkbox"/> | 7  | http://awsassets.wwf.org.my/img/10_900x600_24054.jpg          |  | 3,46 KB | 5.2091779708882 |

Figure 8: Output of the testing algorithm using DOM

Figure 10 shows performance for data extraction using DOM and JSON, which implements the above discussed algorithms. The evaluation consists of image extraction. In this step, time consumed

for image extraction is verified. Experiments are performed on a similar web page using two different techniques; DOM and JSON. As a conclusion, JSON is more efficient for extracting data in time processing.

### Search Image



Figure 9: Output of the testing algorithm using JSON



## Conclusion

The World Wide Web contains large unstructured and semi-structured data. Researchers are welcomed to develop and implement various techniques to extract data from web sources due to the need for structured information. A wide range of web data extraction in several fields has been developed and continues to be proliferated. In the first part of this paper, the applications of web data extraction systems to real world scenario were reviewed. The focus was on how the application can work in practises and classify two techniques; DOM and JSON that have been applied. The second part of this paper is about the proposed framework and architecture. A simple implementation was provided about the proposed architecture in extracting multimedia data focusing on image using Setiu Wetlands web pages for empowering rural womens' activities. In future work, we plan to extend our approach to extract data from multi-web page. The performance of image extraction will influence the time for extraction process. Proposed technique gives good result in time processing in extracting data. The impact of the study for the nation building is the extraction of image that can be used for other purposes.

## Acknowledgements

The authors are greatly acknowledged to the research grants of NRGs Grant (Vot 53131).

## References

- Alarte, J., Insa, D., Silva, J., & Tamarit, S. (2015). Analysis of hyperlinks and DOM comparison for site-level web template extraction\*.
- Baumgartner, R., Gatterbauer, W., & Gottlob, G. (2009). Web data extraction system *Encyclopedia of Database Systems* pp. 3465-3471, Springer.
- Chang, C.-H., Kayed, M., Girgis, M. R., & Shaala, K. F. (2006). A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10), 1411-1428.
- Das, N. N., & Kumar, E. (2014). Automatic extraction of data from deep web page. *International Journal of Computer & Mathematical Sciences (IJCMS)*, 3(1), 86-91.
- Embley, D. W., Campbell, D. M., Jiang, Y. S., Liddle, S. W., Lonsdale, D. W., Ng, Y.-K., & Smith, R. D. (1999). Conceptual-model-based data extraction from multiple-record Web pages. *Data & Knowledge Engineering*, 31(3), 227-251.
- Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70, 301-323.
- Flesca, S., Manco, G., Masciari, E., Rende, E., & Tagarelli, A. (2004). Web wrapper induction: a brief survey. *AI Communications*, 17(2), 57-61.
- Hill, C. (2011). Enabling rural women's economic empowerment: Institutions, opportunities, and participation. Paper presented at the Background paper: UN women expert group meeting Accra, Ghana.
- Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2), 84-93.
- Man, M., & Sabri, I. A. A. (2017). The proposed algorithm for semi-structured data integration: Case study of Setiu Wetland data set. *Journal of Telecommunication Electronic and Computer Engineering*, 9(3-3), 79-84.
- Mehta, B., & Narvekar, M. (2015). DOM tree based approach for web content extraction. Paper presented at the Communication, Information & Computing Technology (ICCICT), 2015 International Conference on.
- Sabri, I. A. A., & Man, M. (2017) WEIDJ : An improvised algorithm for image extraction from web pages. Paper presented at the The 8th international conference on information

- technology, Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan.
- Sabri, I. A. A., & Man, M. (2018a). Improving performance of DOM in semi-structured data extraction using WEIDJ model. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(3), 752-763.
- Sabri, I. A. A., & Man, M. (2018b). Multiple Types of Semi-structured Data Extraction using Wrapper for Extraction of Image using DOM (WEID). In N. A. Yacob, Mohd Noor, N.A., Mohd Yunus, N.Y., Lob Yussof, R., Zakaria, S.A.K.Y. (Eds.) (Ed.), *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)* (pp. 67-76): Springer.
- Sangeetha, M. K. (2016). Component based information retrieval using DOM. *International Journal of Software Engineering and Its Applications*, 10(2), 117-126.
- Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3), 261-377.