Enriching Georeferenced Environmental Data Using Web Data Extraction to Contribute to Degraded Area Impact Analysis

Clovis Santos
ICEN/UFR
Federal University of Rondonópolis-UFR
Rondonópolis, Brazil
email: clovis@ufr.edu.br

Carina Dorneles
INE/CTC/UFSC
Federal University of Santa Catarina-UFSC
Florianopópolis, Brazil
email: carina.dorneles@ufsc.br

***Abstract*—Data enrichment uses resources to fill gaps in customer data sets, enterprise systems, marketing, product sales, and related applications. Environmental applications have the potential to be enriched by aggregating georeferenced data from external sources. The data available on the Web might be a viable alternative to support data enrichment. Usually, this process is done manually, at a high cost of time and human resources. In this context, georeferenced data enrichment using external datasets is a viable and available resource that can be used to reduce financial costs and improve the geographic localization process, which directly depends on appropriate hardware, such as GPS devices and human availability for data collection. This paper presents two main contributions: (i) a data extraction process, which enriches georeferences in a specific application; and (ii) a data enrichment process, which indicates potential risks in environmental areas with potential soil degradation problems. We validate the extracted data from the Web, using these data in an application to verify the distance between areas classified as degraded and possible points of interest in or near urban areas. Finally, it is essential to point out that the research has an interdisciplinary essence involving information systems and the environment, collaborating with both domains.**

***Keywords*—** Georeferencing, data enrichment, environment, extraction web data, environmental retrieve information

I. INTRODUCTION

Environmental mapping of degraded areas is usually created through *in loco* visits. Identifying degraded areas consists of delimiting the geographic coordinates of their respective boundaries. The delineation of areas can be identified in Web repositories, providing alternatives to enriching geo-referenced databases for environmental scenarios. The first step in enriching data is identifying repositories of interest for the area under study. The paper describes the use of geographic coordinates and site identification to identify potential side effects in nearby regions, as described in Section IV.

The main problem in this context is finding georeferenced data on the Web related to delineating degraded areas. Usually, these are data with limited access, but free or non-commercial tools are essential for delineation with good geographic accuracy. Another critical issue is acquiring georeferenced data, which has a high cost and is a barrier for many rural properties. This difficulty leads to obligatory irregularities for small plots associated with family farms. Data sources are made publicly available on the Web as tables of geographic data for a variety of areas, including urban sustainability, transportation networks, policy studies, and health. However, environmental fields are particular, making it challenging to use generic data common in other fields.

Another essential aspect discussed in this paper is eXtensible Mark-up Language, XML, which has been widely used as the standard language for data exchange [18]. The data extraction used in the paper inspects content with Keyhole Markup Language (KML) format, which uses XML as a framework. The content of this type of file stores various data, but the research explores only a subset referring to coordinates for delimiting polygons or locating specific points in geographic areas.

The paper presents an alternative to the enrichment of georeferenced data that contributes to identifying and monitoring environmental impacts in degraded areas and their consequences when detected near urban areas and essential reference points such as river sources. In this context, the georeferenced data were obtained through web extraction in government portals with public access without commercial licenses. It is important to mention that the proposed approach is not restricted to degraded areas. The delimitation in this subarea of the environmental domain has been used to specialize enrichment. However, the enrichment-related approach can be applied to subsets of data from the same domain, such as agriculture, rainfall, and animal control.

The data enrichment process can be classified into six approaches [6]:

- Data fusion: a procedure for unifying data from multiple sources representing the same entity with consistency and valuable presentation.
- Data entity recognition: a process of identifying words in texts, finding and recognizing names of people, companies, organizations, cities, and other predefined entities.
- Data Disambiguation: a method for disambiguation or elimination of ambiguity is the process of identifying the correct data entity in the presence of inconsistent and ambiguous variations in entity names.
- Data Segmentation: a process of grouping data into a set of predefined attributes.
- Data Input: an approach to estimate values for missing or conflicting data items.
- Data Categorization: a procedure for identifying data into different categories based on topics, events, or other features.

The proposal described in this paper uses data entity recognition to identify valuable data for the proposed enrichment. Although researchers have divided web data extraction into different problems based on the modality of the data, they have faced similar problems, such as identifying relevant data for the domain under study using prior knowledge related to the web environment [8]. The extractor used in this paper is a complementary tool to adequately prepare the data for mapping to the selected entity from the target database. This data is essential to generate reports with georeferencing related to environmental information.

Data extraction in web environments does not have a specific domain for data enrichment. The focus is generally on enrichment without a vertical target for the data. In this context, our proposal contributes with an environmental domain approach, which focuses

 ISBN: 978-1-61208-983-6

on a practical application for data enrichment in a real scenario related to monitoring degraded environmental areas.

We aim to use a simple application architecture as a tool to identify side effects near degraded areas using georeferenced data obtained through data enrichment from web sources. The metric used to evaluate the results was defined with geographic identification of sites within a predetermined radius in geographic areas. It is essential to see that degradation like erosion scan leads to siltation of springs that cause irreversible environmental impacts depending on the degree of degradation. Therefore, the result of the enriched data identify potentially damaged sites. After the site's indications, a comparison between the resulting information and the regional maps is possible.

Our proposal suggests a viable alternative for enriching georeferenced data. Section III, provides an overview of a computational architecture for extracting, cleaning, and linking external data sources in a web environment to tables in databases.

This paper presents four contributions regarding georeferenced data extraction for enrichment related to the natural environment:

- Web data extraction: objectively and practically shows a viable alternative for web data extraction as an alternative to enrich georeferenced databases.
- Georeferencing: presents how to use georeferenced data as a resource for environmental monitoring of degraded areas.
- Data enrichment: presents web application architecture for data extraction with goals to enrich a georeferenced database.
- Environment: describes a computational solution to assist environmental consultants and technicians in monitoring degraded areas identifying possible impacts in nearby regions.

This paper is organized as follows. In Section II, we describe works related to web data extraction and georeferenced data enrichment. In Section III, we present the proposed architecture for data extraction, cleaning, and association. In Section IV, we present a case study. In Section V, we discuss some conclusions and possibilities for future work.

II. Related works

In the context of research, the enrichment of georeferenced data depends on the extraction result. According to [7], data extraction from web pages started in the 2000s. Currently, the proposed solutions are essentially used with wrappers or programs to extract information from the web. Although there are several papers on this topic, this section will cover some works related to the main topics of this paper. This section provides an overview of approaches to extracting data from web content based on the XML structure, such as the KML. We point out the difference we propose related to each work discussed.

According to [2], poor data quality has various causes and is a challenge that should not be underestimated. Sometimes, the data source is in the wrong format or range of values, and data must be cleaned or validated. High data quality is desirable and the main criterion to determine if the enrichment was successful and the statements are correct. Our paper presents an alternative to data enrichment from web extraction. High data quality is possible when the data are suitable for use and meet the objectives set by data users. This definition clarifies that data quality highly depends on context, synergy, needs, uses, and access [1].

The research described by [19] addresses data integration in the biomedical domain. Despite being a different domain from the one addressed in our paper, some similarities, for example, data integration for information generation, can be observed. However, the authors' proposal presents query integration without data extraction. In our research, extraction is one of the steps for using data in georeferencing applications.

The research presented in [20] gives a general analysis regarding techniques for extracting data from the web. Extraction is one of the essential points of the present research, with the differential in data extraction as part of the solution for data enrichment in the environmental domain. This paper referred to some items, such as constructing the micro parse to analyze raw content and structuring it for database storage.

In [21], the authors address data enrichment using web documents as sources, which is the approach we also used in our research. Similar to what was described in [20], the research has the main objective to evaluate improvements in querying the enriched database. Our proposal has no perspective of directly analyzing the enriched data since verifying the data quality is certified by visualizing the data in specific geographical scenarios. Plotting the polygons in environmental preservation areas indicates the accuracy and quality of the enriched data.

The research presented in [22] investigates the domain associated with dictionaries for natural language. In our approach, numerical values represent the data domain, so post-processing for validating the data associated with a specific vocabulary is unnecessary. However, the mentioned research also contributed to constructing the micro parse used to select the relevant numerical values to create georeferenced coordinates.

The paper presented by [17] takes a similar approach. However, it uses generic semi-structured data, while the approach of our proposal is to use structured data in a specific area of the environment.

Analogous to our approach, the authors in [24] also developed research focusing on data enrichment for agricultural policy support. However, the domain is slightly different as we addressed data enrichment for the environmental domain, but they are related or complementary. Other topics that are similar to our needs also relate to the storage and query of the extracted data. Both papers use relational databases as a resource for storing the data arranged in tables for the agility of the queries. We have used the proposal presented in [24] as reference rules for the parse developed in our research. A similar approach identifies georeferenced coordinates storing databases after cleaning them.

Data extraction using the web as a source usually uses applications written specifically for this purpose. Creating a generic application becomes almost unfeasible in the face of the diversity of structures used to build pages that are the public repositories usually used by extractors. In our proposal, we have created a tool with a graphical interface to visualize the extracted data in all stages. A similar approach has been presented in [23], composed of an API and a graphical interface that presents the implementation extraction use. In our approach, the extractor implementation was developed specifically for a dataset with characteristics previously defined for KML files. We highlight that the format used is not restrictive but used as a delimiter for demonstrations presented in the results section.

Finally, the environment targets enriched data, such as the extraction and enrichment of georeferenced data to delineate degraded areas. The paper [5] defines land degradation as a process in which the biophysical state of the environment is affected by a combination of natural or human-induced processes acting on the land. This process can lead to an acceleration of degradation. Plants and animals that generally play a role in restoring the land may not survive. The enriched data can identify the potential environmental impacts of nearby threats.

The information presented at the end of Section IV refers to regional reviews of sites based on data enriched with web extraction and data imported from state datasets. In this scenario, the proposed application is limited to verifying distances between geographic points. The work [7] presents a solution independent of structural changes in the data source. Traditional wrappers rely heavily on structures such as HTML as sources. In this work, syntactic rules identify the content of interest. Other environmental monitoring initiatives are described in [4], where an open-source software called "Free and Open Source Software (FOSS) for land degradation vulnerability assessment" is presented. The idea is to identify different levels of vulnerability using models of Environmentally Sensitive Areas

(ESAs). [3] presents another initiative with "Open Foris", which is a set of free, open-source software tools that facilitate collecting, analyzing, and reporting forest inventory data. It is important to note that the application proposed in the paper covers the extraction, enrichment, and operations with georeferencing of regions in environmental areas. The other works referenced focus on analyzing soils and data related to inventories, so the paper's approach can contribute to different situations.

The biggest challenge is having high-quality georeferenced data, which are usually restricted since they are obtained with expensive equipment or provided through equally expensive private services. Delimiting an area for monitoring with generic datasets, with precision above 30 meters, for example, can compromise results associated with mapping and applications requiring accuracy. In this sense, the article shows a viable and reasonable alternative to aggregate georeferenced data to the integrated database systems for geographic identification and monitoring in the environmental scenario.

III. PROPOSAL

The proposal provides a viable, cost-effective solution for aggregating georeferenced information from external sources into entities in databases. This solution supports applications using environmental georeferencing data. Furthermore, using georeferenced information contributes to regulatory requirements, such as delineating permanently protected areas, river sources, rural headquarters, and property boundaries.

The proposal's development began with identifying the type of data source appropriate for the specific domain. Web sources related to georeferencing for the environment using the standard KML. Next, was developed a prototype for extracting data about selected sources, allowing integration with environmental databases. The final step is to enrich the data with entities from the environmental databases to produce information about secondary impacts in urban areas or at sites of ecological importance near degraded areas.

A. Infrastructure Detail

This section presents an overview of the proposed architecture for extracting, cleaning, and associating georeferenced data in three steps. The first obtains data from external sources from an address (*url*) specified by transferring the content to a local structure. The next step refers to data cleaning. After this step, the captured content is analyzed, and redundant data is removed using a micro-parse to identify relevant content such as territory names and geographic coordinates. The last step refers to the association of the georeferenced data with the data of the target schema. The association is done under user supervision using an intermediate entity. This oversight must show the correspondences between the entity's attribute value and the captured georeferenced coordinates. Section IV presents the use of coordinates for data enrichment in environmental scenarios. Figure 1 shows the structure of the data after extraction without cleaning it.

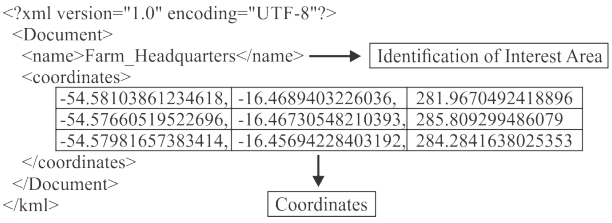

Fig. 1. Remote Data Source Structure.

IV. RESULTS

The most significant value for enriched data is the georeferencing of the natural environment. Environmental georeferencing is essential for water resource identification, rural access, and mapping of degraded areas. In situations where environmental regulations are not respected, such as deforestation in restricted areas, court proceedings verify the development of the regeneration process for the specified area. In this sense, georeferencing is essential for temporal comparisons and tracking. Web data extraction refers to georeferenced data. This data type was chosen for its relevance to the environment, as shown in Figure 2. Figure 2 also shows the data collected at an initial stage without any transformation. The data is not yet usable at this stage, as it is still in the original structure without any modification to make it accessible to external applications and users.

<?xml version="1.0" encoding="UTF-8"?><kml xmlns="http://www.opengis.net/kml/2.2">
<Document><Placemark><name>EmbargadaDesmatamento3</name><ExtendedData><Data name="name">
<value>EmbargadaDesmatamento3</value></Data><Data name="styleUrl"><value>#m_ylw-pushpin</value>
</Data><Data name="styleHash"><value>-5bd746ca</value></Data><Data name="styleMapHash">
<value>[object Object]</value></Data></ExtendedData><Polygon><outerBoundaryIs><LinearRing>
<coordinates>-55.70101499034053,-16.89472397864652,0 -55.69083781135536,-16.91544069960135,0 -55.68711889758733,-1
</LinearRing></outerBoundaryIs></Polygon></Placemark><Placemark><name>EmbargadaDesmatamento1</name>
<ExtendedData><Data name="name"><value>EmbargadaDEsmatamento1</value></Data><Data name="styleUrl">
<value>#m_ylw-pushpin</value></Data><Data name="styleHash"><value>-5bd746ca</value>
</Data><Data name="styleMapHash"><value>[object Object]</value></Data></ExtendedData>
<Polygon><outerBoundaryIs><LinearRing><coordinates>-55.52006081273624,-16.885643099999979,0 -55.51760894219638,-1

Fig. 2. Raw Content.

As shown in Section III, the data used came from KML files after extraction. All the tests were performed locally, and the datasets used are also available on the web using the same infrastructure. It is essential to note that this file format is used and shared by many geoprocessing applications. Despite a large amount of data available in this file format, only two data were used. In this case, the name of the georeferenced area or point and the coordinates were used to delineate it. The coordinates are essential for enrichment operations performed with the profit and loss database.

After the data is extracted and cleaned, it is temporarily stored in a local database. The main objective of this temporary database is to have the possibility of structured data storage, during the cleaning step. The use of a temporary database simplifies the association between the database entities for the enrichment and the local data. Both are data organized in entities and attributes, as shown in Figure 3. In a further step of the enrichment, the temporary data is clear for new operations.

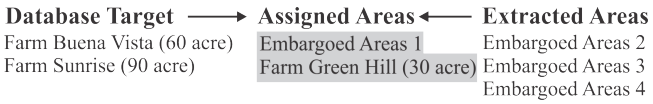

Fig. 3. Area Association.

The contribution of enrichment is the aggregation of georeferenced data on rural properties, which are data related to degraded areas. Erosion and large deforested areas near river sources or urban areas result in significant environmental and urban problems, especially with vertical erosion close to cities.

An important point raised by [12] refers to the classification of the data, which allows comparative analyzes in different regions such as protected areas, national parks, and others. The proposed enrichment process incorporates complementary data, adding resources later used to improve information generation. The application allows manual intervention in the prototype to classify the geographic data extracted and linked to the target database. The areas were classified into six groups according to the specification of an environmental specialist.

A. Information from Enriched Data

An example using the government database is shown to illustrate data enrichment. This database contains the central geographic location of all counties and districts in Brazil. This reference was used to calculate the distance between the center of urban areas or

districts and the center of degraded areas, using the equation 2. For illustration, we have used a radius of 100 km from the center of the degraded areas. Figure 4 shows the pseudocode representation of the algorithm used for checking possible affected areas. Figure 5 illustrates a geographic area with localities in the Amazon border zone identified with the criteria data enriched from the extraction. Two equations were used for this calculation. The first identifies the center of the degraded area, and the second calculates the distance between the center of the degraded area and the center of the urban areas according to the Haversine Equation 1. The Haversine formula is a mathematical equation for calculating the distance between two coordinate points. The distance result obtained by Haversine is considered the smallest circular distance on Earth. The formula avoids a significant rounding error compared to other algorithms such as Pythagoras Equirectangular or Euclidean distance, which is also used to calculate the distance between two coordinates [10].

$$hav(\Theta) = hav(\varphi 2 - \varphi 1) + cos(\varphi 2)hav(\lambda 2 - \lambda 1) \quad (1)$$

In addition, was used an equation for calculating the midpoint for geographic areas of polygons, given in Equation 2. The result defines the first point between the area to be verified and the second point, which is in the middle of the nearby locations.

$$\Delta\sigma = (sin\varnothing_1 sin\varnothing_2 + cos\varnothing_1 cos\varnothing_2 cos(\Delta\lambda)) \quad (2)$$

As seen in Figure 4, the information generated by calculating distances between damaged areas and points of geographic interest, such as cities and river sources, can help develop preventive measures to restore these areas and prevent further damage nearby. The approach adopted for storing and visualizing the information uses a temporary storage table for display in a graphical user interface. As described earlier, the criterion for determining the distance between the central points of the damaged areas and the geographic points of interest was 100 km. This value was the reference used to search the geographic locations within this radius. The relevant areas included in the specified region are temporarily stored in memory with the data: Description of the location, distances, and State, as shown in Table I.

```
Algorithm ImpactAnalysis(param_Dist)
{  MidPoint(CoordMaxMin.Min_Lat,
            CoordMaxMin.Max_Lat,
            CoordMaxMin.Min_Lon,
            CoordMaxMin.Max_Lon,
            Ret_La, Ret_Lo)
  dist←Distance(Ret_La, Dataset_Lat,
                Ret_Lo, Dataset_Lo)
  while Dataset≠ end_records
  if (dist<param_Dist)
  {
    TableDist←Dataset_Distance
    TableDist←Dataset_Location
    TableDist←Dataset_Type
    TableDist←Dataset_State
  }
}end-Algorithm
```

Fig. 4. Impact analysis algorithm.

Confirmation of the degree of degradation and the consequences for nearby sites depends on technical criteria established by technicians and environmental consultants. In this sense, technicians can confidently set parameters for the analysis, such as the type of degradation and the appropriate distance between degraded areas and areas of interest. The scenario proposed in this paper is illustrative because any technical criterion was used to determine that the distance of 100 km is technically valid, but the intent is to present a realistic example of the use of enriched data from data extraction. Although the definition of distance-related metrics is a parameter defined by technicians in the environmental field, there is no commitment to the paper's proposal. The test scenario was created flexibly for adjustments in calculating distances that meet environmental requirements.

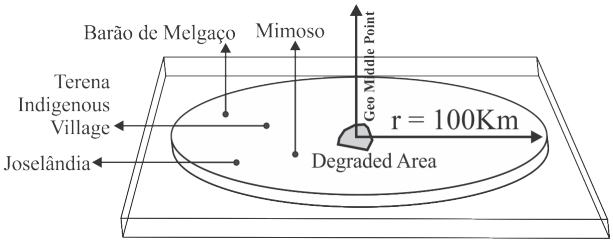

Fig. 5. Search area.

The Table I shows in tabular form the found locations for radius=100km. Although the defined radius is large, it is only illustrative, and values can suit different requirements.

TABLE I
QUERY RESULTS.

Country	Location	Distance	State
Barão de Melgaço	Barão de Melgaço	84 km	MT
Barão de Melgaço	Joselândia	67 km	MT
Santo Antonio	Mimoso	80 km	MT
Santo Antonio	Terena Indigenous Village	61 km	MT

Another important point related to the extracted data analysis is the terrain's surface. These analyses include data on heights for each geographic coordinate, which is generally not obtained from GPS devices or websites that provide open geographic datasets. The Open Elevation API (https:openelevation.com) was used for this type of enrichment, which provides elevation data for geographic coordinates (latitude and longitude).

Figure 6 shows the representation of a degraded area and the elevations for each coordinate. Irregular heights can indicate potential soil erosion. The surface of the degraded region leads to possible rainfall shifts leading to erosion of soil and substrate.

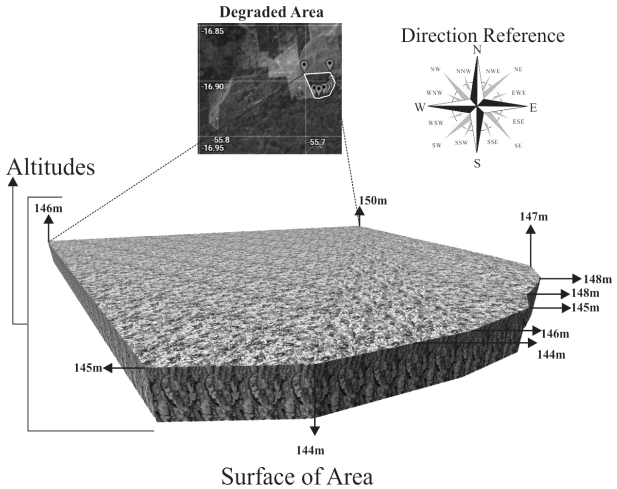

Fig. 6. Map area.

The review of the degraded area contributes directly to the impact that any soil shifts may have on nearby relevant points. With this in mind, an analysis was performed from the lowest elevations of the degraded area to nearby points of interest, according to the potential

direction of possible ground shifts. It is important because degraded areas with sloping surfaces to points such as river sources or highways may have more significant impacts depending on the distance between them.

B. Results Considerations

The results obtained were satisfactory and illustrated extension possibilities for other applications. We have found difficulties creating a micro-parse to analyze the captured content due to its heterogeneity. Although the pages follow the same structure, some may have a different structure depending on the tool used to create them.

Data enrichment was performed satisfactorily with only one limitation: creating an intermediate table to link the georeferenced data to the target table. In data enrichment, changes to the data structure are essential in some cases, as it involves adding new data to a structure that was generally not prepared to accommodate it.

The content presented in this paper related to georeferenced data can be used as references for analyzing the distance between different environmental areas. The availability of this data in the web environment is large and promotes essential data sources. Another critical point is the availability of georeferenced data by government agencies, which also serves as an alternative for data enrichment.

Monitoring degraded areas is an ongoing task that presents numerous difficulties. The paper does not deal with this subject in depth but presents a satisfactory solution that contributes to consultants and environmental engineers of governmental regulatory agencies.

V. CONCLUSIONS AND FUTURE WORK

This paper describes an alternative for data extraction in a web environment in the environmental domain, specifically with georeferenced data. We have presented a consistent way to enrich databases with georeferencing to the environment. The proposal does not aim to exhaust the topic but to present an alternative to the growing demand for environmental data. The cost of obtaining this type of data is relatively high, and as shown in the paper, public access tools are feasible. Government interest in monitoring the conservation or restoration of protected areas is constant, and cost-effective alternatives must be analyzed to meet this demand. Two central points are essential in research as contributions. The first refers to the enrichment of environmental data by extracting data available in both government portals and APIs. The second contribution relates to developing tools to assist environmental engineers in monitoring degraded areas and potential impacts on nearby sites. An essential contribution to this research is to extend the import capabilities to other web data sources besides the file structure presented in this paper. This approach will help add more resources to rural properties in an environmental context, leading to essential allies for inspections in various government administrative areas.

REFERENCES

[1] I. Jaya and F. Jaya and I. Ishak and Lilly Affendey and M. Jabar, “A review of data quality research in achieving high data quality within organization, ” Journal of Theoretical and Applied Information Technology, vol. 95, No 12, pp.2647–2657, June 2017.

[2] O. Azeroual and M. Jha, “Without Data Quality, There Is No Data Migration, ” Big Data Cogn. Comput., Special Issue Educational Data Mining and Technology, vol. 5(2), 2021.

[3] Open Foris. http://openforis.org/, (retrieved: October/2021)

[4] V. Imbrenda and G. Calamita. and R. Coluzzi and M. D’Emilio and M. Lanfredi and A. Perrone and M. Ragosta and T. Simoniello, “Free and Open Source Software for land degradation vulnerability assessment, ” EGU General Assembly Conference Abstracts, vol. 15, pp.11153–, 2013.

[5] B. Boer and I. Hannam, “Chapter 21: Land Degradation Law, ” in Jorge Viñuales and Emma Lees (eds), Oxford Handbook on Comparative Environmental Law, 2019.

[6] S. Azad and S. Wasimi and A. Ali, “Business Data Enrichment: Issues and Challenges, ” 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2018, pp. 98-102, 2018.

[7] J. Lloret-Gazo, “A Browserless Architecture for Extracting Web Prices”, SAC’20: Proceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 2193–2200, March 2020.

[8] X. Dong and H. Hajishirzi and C. Lockard and P. Shiralkar, “Multi-Modal Information Extraction from Text, Semi-Structured, and Tabular Data on the Web, ” KDD’20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3543–3544, August 2020.

[9] S. Zhang and K. Balog, “Web Table Extraction, Retrieval, and Augmentation: A Survey,” ACM Transactions on Intelligent Systems and Technology, vol. 11, pp 1–35, April 2020.

[10] L. Theoson and R. Anthony and J. Purnama, “Distance-Measurement-Decision-Making Backend System Using NodeJS, ” ICONETSI: Proceedings of the International Conference on Engineering and Information Technology for Sustainable Industry, pp. 1–6, September 2020.

[11] O. Emmanuel, “Effects of Deforestation on Land Degradation”, Adebayo Williams (Editor), isbn 978-3-330-34486-0. 2017.

[12] SCITEPRESS - Science and and Technology Publications, Database Design of a Geo-environmental Information System, Proceedings of the 16th International Conference on Enterprise Information Systems. 2014.

[13] , A. Kanukov and P. Ivanov, “IOP Publishing, Geological information database integration into a geographic information modeling system,” IOP Conference Series: Materials Science and Engineering, pp.1–6, 2021.

[14] R. Nicholas and J. Weinman and J. Gouwar and A. Shamji, “Deformable Part Models for Automatically Georeferencing Historical Map Images”, SIGSPATIAL ’19: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 540–-543, November 2019.

[15] A. Agarwal and M. Genesereth, “Extraction and Integration of Web Data by End-Users,” International Conference on Information and Knowledge Management, pp. 2405–2410, 2013.

[16] , R. Fayzrakhmanov and E. Sallinger and B. Spencer and T. Furche and G. Gottlob, “Browserless Web Data Extraction: Challenges and Opportunities, ” WWW ’18: Proceedings of the 2018 World Wide Web Conference, pp. 1095-–1104, April 2018.

[17] D. Gong and D. Wang and Y. Peng, “Multimodal Learning for Web Information Extraction,” MM ’17: Proceedings of the 25th ACM international conference on Multimedia, pp. 288—296, October 2017.

[18] S. Haw and E. Soong, “ Performance evaluation on structural mapping choices for data-centric XML documents, ” Indonesian Journal of Electrical Engineering and Computer Science, pp. 1539–1550, 2020.

[19] M. Kamdar and M. Musen, “An empirical meta-analysis of the life sciences linked open data on the web.” Scientific Data, vol. 8, pp.1–21, January 2021.

[20] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra and K. R. Bodke, “Analysis Of Different Web Data Extraction Techniques,” 2018 International Conference on Smart City and Emerging Technology (ICSCET), 2018, pp. 1–7.

[21] T. Breuer and M. Pest and P. Schaer, “Evaluating Elements of Web-Based Data Enrichment for Pseudo-relevance Feedback Retrieval. ” Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, vol 12880, pp 53–64.

[22] G. Wenzek et al. “CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data.”Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp 4003-–4012, May 2020.

[23] T. Alrashed and A. Zhang. “ScrAPIr: Making Web Data APIs Accessible to End Users.” CHI ’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1—12, April 2020.

[24] M. Rousi et al., “Semantically Enriched Crop Type Classification and Linked Earth Observation Data to Support the Common Agricultural Policy Monitoring, ” in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 529–552, 2021.

 ISBN: 978-1-61208-983-6