The Fourteenth International Conference on Advanced Geographic Information Systems, Applications, and Services

GEOProcessing 2022

June 26, 2022 to June 30, 2022 - Porto, Portugal

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

Clóvis Santos Jr (UFR) - presenter
clovis@ufr.edu.br

Carina F. Dorneles (UFSC)
carina.dorneles@ufsc.br

# Short Bio - presenter

Cóvis is a Professor at the Federal University of Rondonópolis. He holds a degree in Computer Science from the University of Alfenas (1996), a Master's in Knowledge Management and Information Technology from the Catholic University of Brasília (2003), and a Ph.D. in Computer Engineering from the University of São Paulo. He finished his Post Doctorate at the Federal University of Santa Catarina in March 2022. He has experience in Computer Science, with an emphasis on interfaces and data quality, working mainly on the following topics: information systems, informatics applied to agribusiness and agriculture, and traceability. Currently, his research focus is on the development of a facilitating tool for generically managing documents, using as validation the existing scenario in the social observatory of the municipality of Rondonópolis-MT. The social observatory is a space that brings together the most significant number of entities representing civil society to contribute to improving public management. The tool will include the storage of metadata associated with existing files and routinely used in the activities of the social observatory.

# Schedule

- Introduction
  - Environmental mapping;
  - Delineation areas;
  - Georeferenced data on the Web;
  - eXtensible Markup Language and KML format.
- Related Work
  - Data Extraction;
  - Data Quality;
  - Data Enrichment.
- Proposal
  - Overview.
  - Main Context.
  - Extraction.
- Results
  - Main Context;
  - Data Enrichment;
  - Information;
  - Information from Enrichment Data.
- Conclusions
- Related Work References

# Introduction

Environmental mapping of degraded areas is usually difficulty;

The delineation of areas can be identified in web repositories, providing alternatives to enriching geo-referenced databases for environmental scenarios;

The problem in this context is to find georeferenced data on the Web related to the degraded areas;

Usually, data sources are made publicly available on the Internet as tables of geographic data for a variety of areas, including urban sustainability, transportation networks, policy studies, and health;

Another essential aspect refers to eXtensible Markup Language, XML has been widely used as the standard language for data exchange. The data extraction used in the paper inspects content with KML format, derived from XML.

# Related Work

1/2

| Contribution | Citation |
|---|---|
| Extraction is not a recent topic, currently, the proposed solutions are essentially used with wrappers or programs to extract information from the web. | LloretGazo (2020) |
| Poor data quality has various causes and is a challenge that should not be underestimated. In some instances, the data source is in the wrong format or range of values. | Otmane and Meena (2021) |
| High data quality is possible when the data are suitable for use and can meet the objectives set by data users. | Gong and Wang (2017) |
| Data integration for information generation can also be created by integrating queries without extracting data. | Kamdar and Musen (2021) |
| In our approach, numerical values represent the domain of the data, so post-processing is unnecessary. However, the mentioned research contributed to the construction of the relevant micro parse for the creation of georeferenced coordinates. | Wenzek et al (2020) |
| Usually, the data is semi-structured or generic, our proposal uses structured data in a specific area for the environment. The authors also developed research focused on enriching data to support agricultural policy. | Rousi et al (2021) |

# Related Work

2/2

| Contribution | Citation |
|---|---|
| The authors present an API with a graphical interface as a facilitator to demonstrate data extraction. Similarly, we developed an extractor for Keyhole Markup Language or KML files. We emphasize that the format used is not restrictive, but a delimiter for the investigated domain. | Zhang (2020) |
| Land degradation as a process in which the biophysical state of the environment is affected by a combination of natural or human-induced processes. The enriched data can identify the potential environmental impacts of nearby threats. | Boer and Hannam (2020) |
| Traditional wrappers rely heavily on structures such as HTML as sources. In this work, syntactic rules identify the content of interest. | Lloret-Gazo (2020) |
| Other environmental monitoring initiatives are used in an open-source software called ”Free and Open Source Software for land degradation vulnerability assessment (FOSS)”. | Imbrenda and Calamita (2013) |
| Another initiative with ”Open Foris”, 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. | Openforis (2021) |

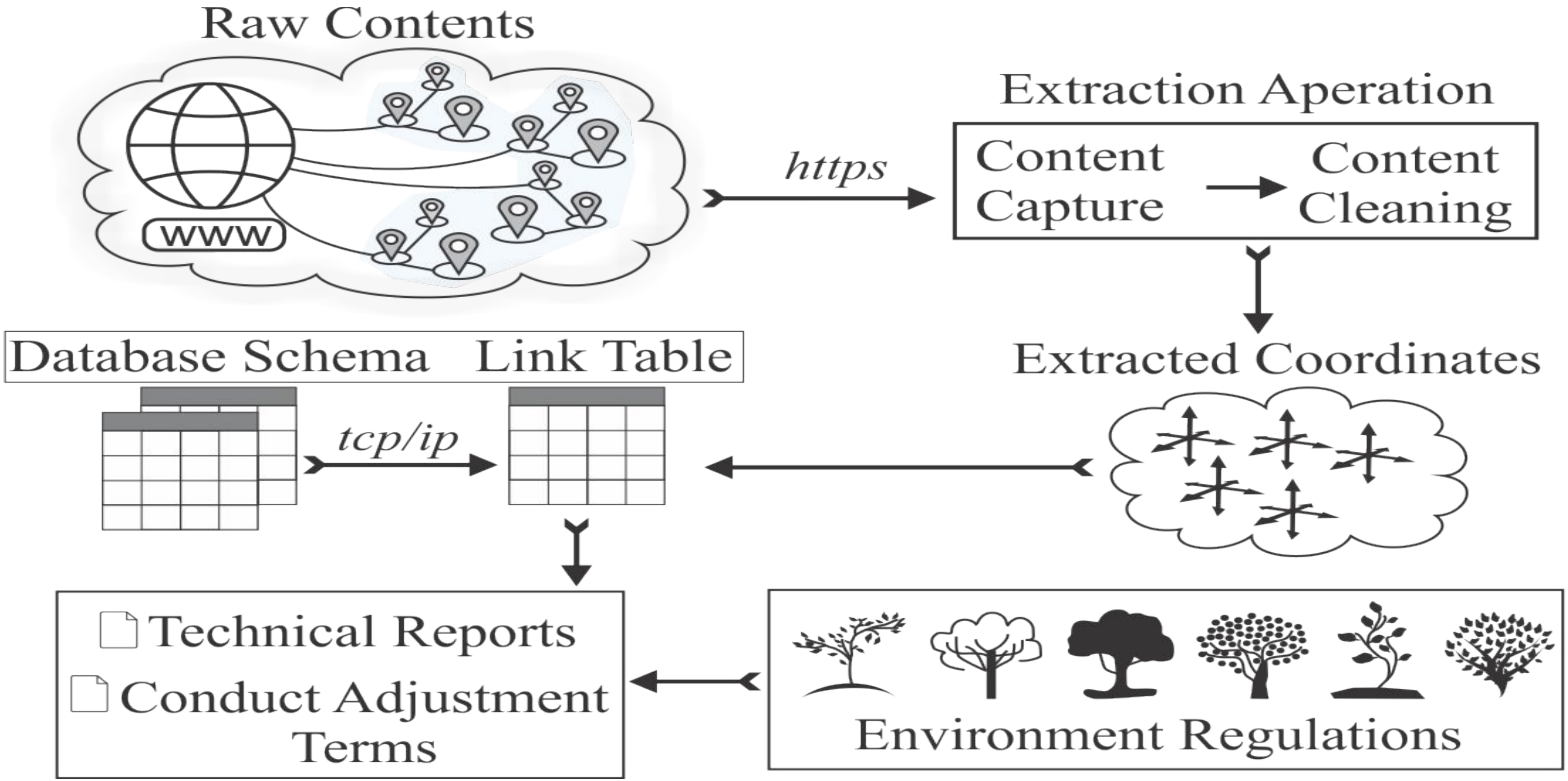
# Proposal - Main Context

The proposal provides a viable, cost-effective solution for aggregating georeferenced information from external sources into entities in databases that improve data enrichment and favor the creation of information using environmental georeferencing data. Another important point is using georeferenced information to Contribute to regulatory requirements, such as delineating permanently protected areas, river sources, rural headquarters, and property boundaries.

Main subjects approach

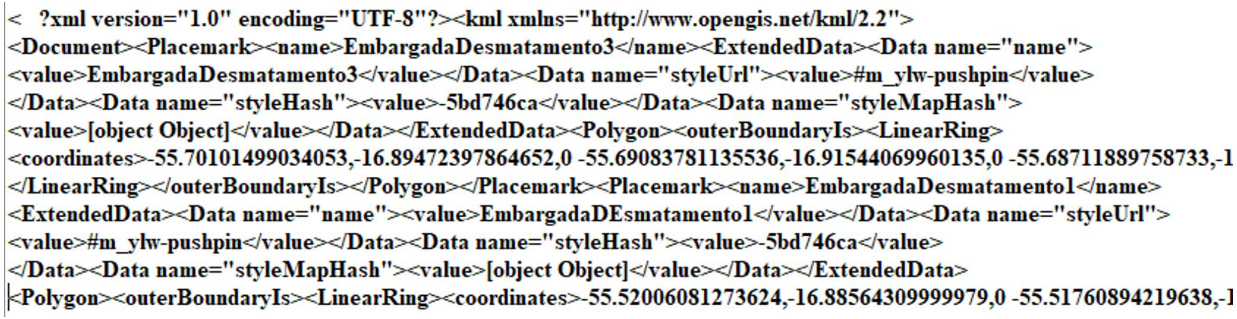
- Identifying the type of data source appropriate for the specific domain;
- Web sources related to georeferencing for the environment using standard keyhole markup language or kml files based on the eXtensible Markup Language or xml structure were chosen;
- Development of a prototype for extracting data about selected sources, allowing integration with environmental databases;
- Enrichment the data with environmental databases to produce information about secondary impacts in urban areas or at sites of ecological importance near degraded areas.

# Proposal - Overview



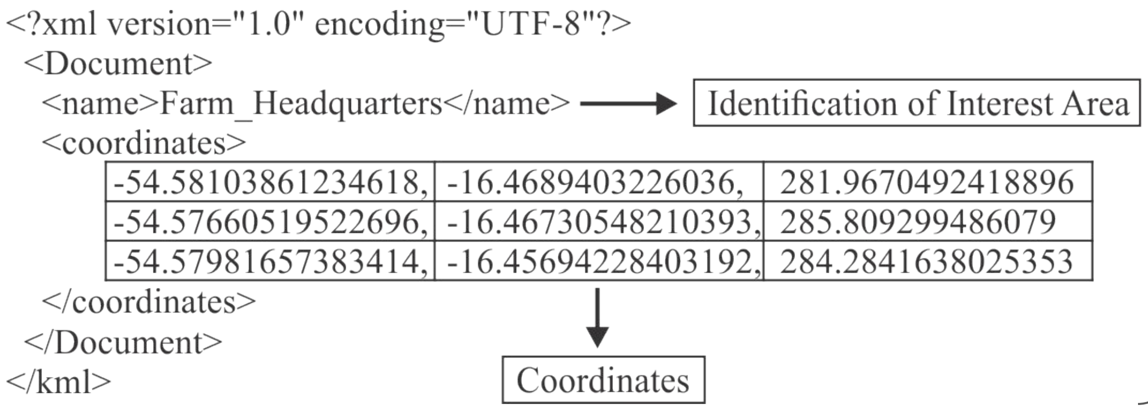
# Proposal - Extraction

Raw Content



```
<   ?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,-16.88564309999979,0 -55.51760894219638,-1
```

Coordinates



```
<?xml version="1.0" encoding="UTF-8"?>
 <Document>
  <name>Farm_Headquarters</name>
  <coordinates>
        -54.58103861234618, -16.4689403226036,  281.9670492418896
        -54.57660519522696, -16.46730548210393, 285.809299486079
        -54.57981657383414, -16.45694228403192, 284.2841638025353
  </coordinates>
 </Document>
</kml>
```

Identification of Interest Area

Coordinates

# Results - Main Context

- The most significant value for enriched data in the paper domain is georeferencing for the environment;

- Environmental georeferencing is essential for water resource identification, rural access, and mapping of degraded areas;

- Web data extraction refers to georeferenced data. This type of data was chosen for its relevance to the environment;

- The data used came from keyhole markup language or kml files after extraction, the tests were performed locally, the datasets used are also made available on the web using the same infrastructure;

- It is important to point out 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, were used the name of the georeferenced area or point and the coordinates to delineate it, essential for enrichment operations performed with the profit and loss database.

# Results - Data Enrichment

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 two algorithms.

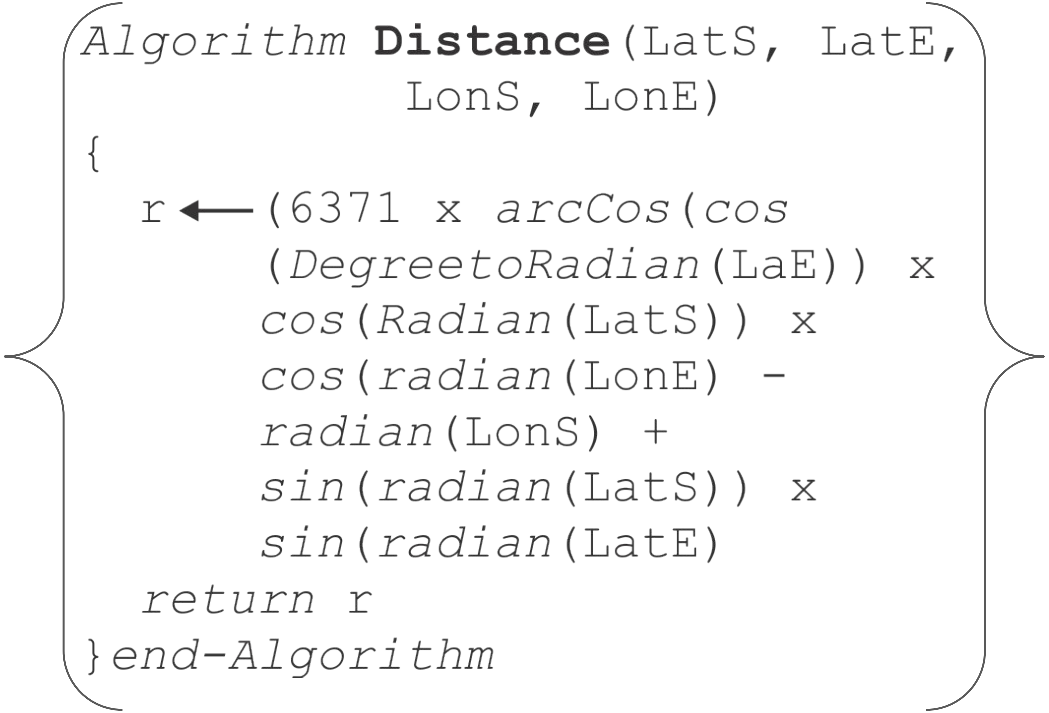
The first identifies the center of the degraded area:

```
Algorithm MidPoint(LatS, LatE,
           LonS, LonE,
           Out_La, Out_Lo)
{
  dLon ← radian(LonE) - radian(LonS)
  bX ← cos(radian(LatS))x cos (dLon)
  bY ← cos(radian(LatS))x sin (dLon)
  Lat ← radian(arcTan(sin(LatS) +
        sin(radian(LatS)),
        sqrt((cos(radian(LatS)) + bX) x
        (cos(radian(LatS)) + bX) + bY x
         bY)))
  Lon ← LonS + degree(LatS)) + bX))
  Out_La ← Lat
  Out_Lo ← Lon
}end-Algorithm
```

In addition, was used an equation for calculating the midpoint for geographic areas of polygons. 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.

# Results - Information

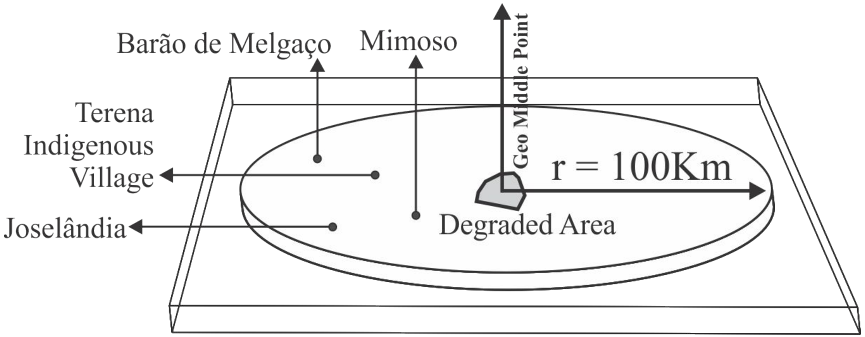
The second calculates the distance between the center of the degraded area and the center of the urban areas according to algorithm based on the Haversine equation:



```
Algorithm Distance(LatS, LatE,
                LonS, LonE)
{
   r ← (6371 x arcCos(cos
         (DegreetoRadian(LaE)) x
         cos(Radian(LatS)) x
         cos(radian(LonE) -
         radian(LonS) +
         sin(radian(LatS)) x
         sin(radian(LatE)
   return r
}end-Algorithm
```

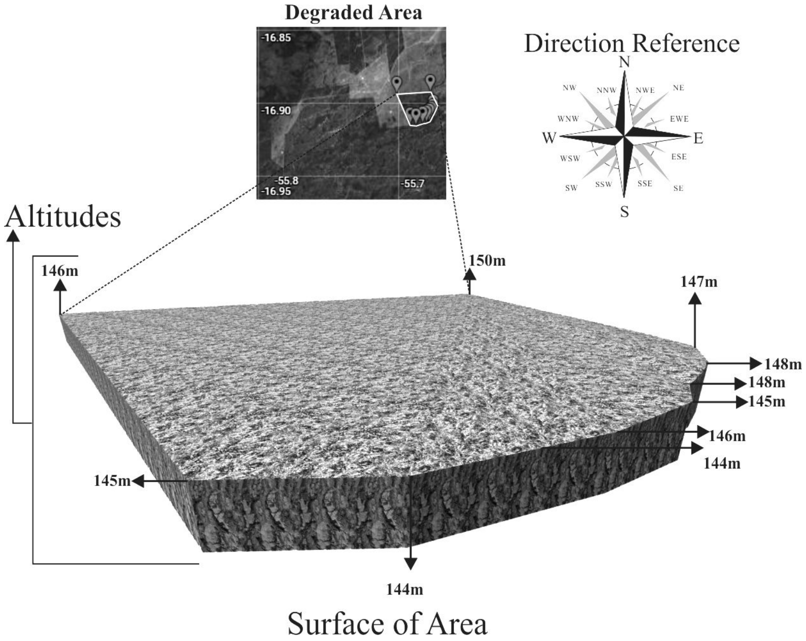
The information generated by calculating distances between damaged areas and points of geographic interest, such as cities and river sources, can help develop preventive measures to restore these areas and prevent further damage nearby. The approach adopted for storing and visualizing the information uses a temporary storage table for display in a graphical user interface.

# Results - Information from Enrichment Data



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.

Another important point related to the analysis of the extracted data refers to the terrain's surface. These analyses include data on heights for each geographic coordinate, which is generally not obtained from GPS devices or even websites that provide open geographic datasets.



# Conclusions

- The paper presents 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 for 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.

- Two central points are essential in research as contributions.
    - The first refers to the enrichment of environmental data through the extraction of 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 important 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.

# Related Work References

Lloret-Gazo, Jorge, A Browserless Architecture for Extracting Web Prices, Association for Computing Machinery, Proceedings of the 35th Annual ACM Symposium on Applied Computing. isbn 9781450368667, New York, NY, USA.

Otmane Azeroual and Meena Jha, Without Data Quality, There Is No Data Migration. MDPI AG. 2021.

Kamdar, Maulik R. and Mark A. Musen. "An empirical meta-analysis of the life sciences linked open data on the web." Scientific Data 8 (2021).

Wenzek, Guillaume et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." LREC (2020).

Rousi, Maria et al. "Semantically Enriched Crop Type Classification and Linked Earth Observation Data to Support the Common Agricultural Policy Monitoring." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021): 529-552.

Zhang, Shuo and Balog, Krisztian, Web Table Extraction, Retrieval, and Augmentation: A Survey, Association for Computing Machinery, New York, NY, USA, issn 2157-6904, ACM Trans. Intell. Syst. Technol.2020.

Boer, Ben and Hannam, Ian, Land Degradation. 2020.

Gong, Dihong and Wang, Daisy Zhe and Peng, Yang, Multimodal Learning for Web Information Extraction, New York, NY, USA, Proceedings of the 25th ACM International Conference on Multimedia, 2017.

Imbrenda, Vito and Calamita, G. and Coluzzi, Rosa and D'Emilio, Mariagrazia and Lanfredi, Maria and Perrone, Angela and Ragosta, Maria and Simoniello, Tiziana,Free and Open Source Software for land degradation vulnerability assessment. 2013.

openforis, Open Foris. 2021. http://openforis.org/, (Accessed on 10/19/2021)