INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Metadata, Semantics and Ontologies

ISSN online: 1744-263X - ISSN print: 1744-2621
https://www.inderscience.com/ijmso

Enrichment of data in digital documents with metadata extraction

Clovis Dos Santos Júnior, Carina Friedrich Dorneles

DOI: 10.1504/IJMSO.2023.10060810

Article History:

Received:	30 July 2022
Last revised:	17 March 2023
Accepted:	22 March 2023
Published online:	05 December 2023

Copyright © 2023 Inderscience Enterprises Ltd.

Enrichment of data in digital documents with metadata extraction

Clovis Dos Santos Júnior*

Institute of Exact and Natural Sciences,
Federal University of Rondonópolis (UFR),
Rondonópolis, Mato Grosso, Brazil
Email: clovis@ufr.edu.br
*Corresponding author

Carina Friedrich Dorneles

Department of Informatics and Statistics,
Federal University of Santa Catarina (UFSC),
Florianópolis, Santa Catarina, Brazil
Email: carina.dorneles@ufsc.br

Abstract: Companies have migrated their operational activities from paper documents to automated processes with fully digital storage. This management trend is positive, but printed documents, in most cases, cannot be discarded for administrative or legal reasons. This research used data extraction to enrich the database of a Non-Governmental Organisation (NGO) that monitors the use of public financial resources in counties. The implementation analysed the digital files containing official documents and identified the words with the highest occurrence according to algorithms presented in the research results. The solution created in the research added metadata to improve the search for documents in the database and improve the procedural follow-up of administrative and judicial actions. The results were positive with success in the extraction of the keywords in each document and presented with examples in the results section, showing the steps used to add metadata in the documents.

Keywords: electronic document; text mining; data extraction; NGO.

Reference to this paper should be made as follows: Dos Santos Júnior, C. and Dorneles, C.F. (2023) 'Enrichment of data in digital documents with metadata extraction', *Int. J. Metadata Semantics and Ontologies*, Vol. 16, No. 2, pp.187–193.

Biographical notes: Clovis Dos Santos Júnior holds his degree in Computer Science from the University of Alfenas (1996), Master's degree in Knowledge Management and Information Technology from the Catholic University of Brasília (2003) and PhD in Computer Engineering from the University of São Paulo (2014). He has experience in Computer Science, with emphasis on interfaces and data quality, working mainly on the following topics: information system, informatics applied to agribusiness, agriculture and traceability and data enrichment.

Carina Friedrich Dorneles is Professor at the Department of Informatics and Statistics (INE) of UFSC. Member of the Database Steering Committee of SBC (Brazilian Computer Society). She works in Research, Teaching and Advising Scientific Initiation, Undergraduate, Master and Doctoral Levels. She has been Coordinator of the Graduate Program in Computer Science at UFSC from October 2015 to May 2017. During the PhD degree, she held a sandwich internship at the University of Washington, Seattle, USA, in the Database and Artificial Intelligence research group, under the supervision of Prof. Alon Halevy. Her research interests include data management, information retrieval, mining data with an emphasis on the web, discovery of knowledge and data extraction and matching. She coordinates and participates in research projects in the area, with scientific publications in periodicals and conference proceedings of good quality.

Copyright © 2023 Inderscience Enterprises Ltd.

1 Introduction

The amount of digital data generated worldwide daily is massive. Between 2015 and 2017, the amount of data was more significant than in the entire historical timeline, and the entire digital universe is estimated to reach 44 zettabytes between 2020 and 2025, and 463 exabytes of data will be created daily worldwide. Data to be used must be understandable to the easy search and be made locatable, accessible, interoperable and reusable to improve management. Similarly, it must initially be available in digital form on some online platform for data to be considered valid. Also, in this scenario, data is more accessible with rich metadata.

A highlight of the research is metadata as supplementary data for document retrieval. The use of metadata has enabled data summarisation to assist in digital document search engines. Conceptually metadata can be defined as data about data. A metadata item determines the meaning of a specific piece of data, usually representing information intelligible to people and computer systems.

An essential step towards data standardisation and interoperability is in the creation of metadata, some advantages are: standardisation of the description of data sets, quality of the represented data sets, and interoperability between different processes and physical devices (Löbe, 2021).

The concept of metadata is a resource for creating interfaces that facilitate data retrieval and exchange. The following areas are examples of this concept, highlighting that any area of human knowledge with data needs can benefit from the use of metadata in both data retrieval and data management and information generation.

- Metadata in Health: Health metadata represents the use of structured information to document the creation, management, and use of records over time across all functional domains of medicine and related fields. Systems for the health domain must automatically extract metadata elements from records. They must also allow management of values from metadata in data retrieval by other software (Chekabab et al., 2020).
- Geospatial Metadata: Metadata schemes also contribute to managing data from areas such as Social Sciences and can be shared in an accessible and intelligent way. An example is present in the elements developed by the Federal Geographic Data Committee (FGDC). The FGDC standard is helpful for the use of resources created with a grant from the US Government (Moen, 2000).
- Multimedia: The Moving Picture Experts Group ISO/IEC (MPEG) has developed a set of standards for the representation and encoding of digital video and audio. Two metadata standards: MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938) and MPEG-21 Multimedia Framework (ISO/IEC 21000). MPEG-7 defines the metadata elements, structures and relationships used to describe audiovisual objects, including still images, graphics, 3D models, music, audio, voice, video or multimedia collections (Kim and Chung, 2020).
- Data Warehousing: metadata can also be a resource to standardisation interface for data warehouse applications. They control data flow and describe features such as rules, transformations, aggregations and mappings. Metadata enables the control and management of data warehousing to develop and control intelligent systems without writing code in specific programming languages (Oukhouya et al., 2021).
- Web Metadata: metadata are a resource to the distribution of descriptive elements for web resources, requirements such as markup systems for web pages, conventions and standards to process them and context creation in libraries are some examples. The practice of creating tags or labels for this purpose has taken on a wide variety of applications on the Internet regardless of whether they belong to proprietary or open libraries (Müller and Gleim, 2021).

The research contributed to managing digital documents that make up the collection of the Non-Governmental Organisation called Observatório Social – OSR, Rondonópolis-MT/Brazil.[1] The research domain covers spreadsheets, invoices, memos, bids and other kinds of documents. The collection of NGO documents is public and used for monitoring the management and finances of the municipal executive power. In a transversal way, the article also presents the development of a tool to help the consultations of the documents in the collection, providing improvements in the inspection activities developed by the NGO. The procedural monitoring of the actions of the municipal executive branch requires constant checking of documents in different types of media. The documents used in the research are digital but present difficulties in consultations due to many printed documents for analysis without resources to facilitate the location as metadata.

More specifically, the research shows resources to facilitate the location and identification of documents, also contributing to retrieving data from metadata extracted from the current database.

The approach proposed in the research uses the NGO's database for queries on thematic issues related to each investigation to monitor actions at the municipal level. In this sense, the main problem with managing digital documents lies in the lack of a computational solution capable of retrieving multiple documents from keywords. The metadata added to the database aims to improve the results of these searches.

The management of the documents was done with a specific tool to keep the documents available in their original formats, avoiding storage in databases and avoiding duplications.

The resources for searching and finding information in digital libraries are not new, the availability of full-text access in electronic form can contribute to locating, navigating and displaying information that transcends traditional libraries. The article by Skluzacek et al. (2018) presents an important reference for managing digital files, the research performs an approach regarding storage formats by classifying them as follows:

- *Unstructured*: files containing no specific format or content. Often readable natural language and encoded in ASCII or Unicode. Unstructured files usually are stored in text format (txt).
- *Structured*: are containers of data encoded or organised in a predefined format. It often includes self-descriptive metadata that specific tools and interfaces can extract. The most common formats include: JSON, XML, HDF and NetCDF.
- *Tabular items*: are formatted in rows and columns, usually include a header and column labels. The metadata can be the header, rows or columns. They have metadata for columns such as averages, minima and maxima. Tabular files are often stored as .csv and .tsv files.
- *Images*: Files containing graphical images, such as plots, maps and photographs. Common formats include .png, .jpg and .tif.
- *Compressed*: This type covers any manipulable file with compression / decompression software to produce one or multiple files. Examples of extensions are .zip and .gz.
- *Hybrid*: files that contain multiple types of data, such as tabular data with free-text descriptions and images with text labels.

2 Related works

The research presented by Jeong et al. (2021) addresses data standardisation as a method of data quality management. Also, in the research, the authors address metadata standardisation in databases as a resource for improving data quality, managing the terms used in the metadata, and providing reliability and ease of use. Extracting the metadata also results in a summary that reduces the amount of text to search by removing less necessary content. Although topic-related searches have identification capabilities for relevant texts, there is less emphasis on minimising the diversity of content (Saeed et al., 2020). In this sense, it results in texts summarised in a form poorly suited to the original text's context.

Another fundamental approach refers to big data, and most systems have large amounts and variety of data and lack veracity, which together exceeds the capacity of traditional processing systems. An approach presented by Sawadogo et al. (2019) presents the data lake concept. In practice, it is an enormous repository of raw and heterogeneous data, composed of external sources, allowing users to explore, sample and analyse the data. This research does not directly use the data lake concept, but it is relevant to mention it as conceptual support because, in a way, historical data is stored in files and manipulated in this structure.

Automated metadata extraction from binary documents such as PDFs is key to building a scalable document processing system for libraries and digital collection searches. Metadata extraction has usually used machine learning-based methods. However, relatively small digital documents work fine. In general, digitisation is not satisfactory when the amount of documents is too large, this applies to digitising entire books or thousands of theses and dissertations. The extraction of metadata is a difficult task that requires much manual intervention.

Portable Document Format or PDF is frequently present in research involving text storage and processing because of the prevalence of documents for this file type, as presented in Figure 2. Another important PDF feature has been used as a file delivery format on many platforms for almost all devices. Han and Wan (2018) discussed the shortcomings of using formats such as TIFF and JPEG2000 for file formats for data preservation, showing the PDF format as an essential format for scanning textual documents.

Hashmi et al. (2020) described takes a rules-based approach to metadata extraction, first converting the documents to XML format and then the metadata is extracted. Converting the data set from PDF to XML provides many different features, and when incorporated with textual features, they provide consistent rules and good results. In our proposal, the approach is similar; however, there is no intermediate conversion to XML, we use UTF-8 format text and then the metadata extraction is done. Another essential point is rules; algorithms and parses for textual content analysis provide the pre-processing for extraction in both kinds of research.

Another important point related to the investigated collection is the characteristics of the documents stored in the databases: edicts, memos, norms and contracts. The research presented by Ha et al. (2021) addresses systems for extracting metadata from contracts with error checking and analysis. The present research uses the same principle across the board in selecting and identifying the keywords for each document.

The research presented by Blinston and Blondelle (2017) brings contributions related to digital media formats and storage structure. According to the authors, electronically stored data are in two categories:

- data stored in relational or object databases that are highly structured;
- data located in documents in various formats (TIFF, JPG, PDF, XLS), usually gathered in folders in semi-structured or unstructured content. Typically, this data is divided into 20% structured data versus 80% semi- or unstructured data.

These characteristics even affect the ability to make decisions since software, and analytical systems generally, operate with structured data. Current practices to extract data and metadata from unstructured documents mainly involve manual processes with high financial and human costs (Blinston and Blondelle, 2017).

3 Methodology

The research adopts qualitative and descriptive methods, in this context, a tool converted the textual content into more

readable content in the utf standard, an application used by the NGO realised the conversions to tokens with the bag-of-word concept for subsequent classification of the terms with greater quantitative relevance in each document. The classification results of the elements with the highest occurrences in each text update the database are used to index the physical documents. Figure 1 shows the methodological procedures used to develop the research:

- Original documents: the first step consists of mapping the texts to the original format for content conversion;
- Textual documents: conversion of binary documents to the utf-8 standard;
- Bag-of-words: the creation of lists with tokens and respective quantities of occurrence;
- Database: updating of the database with the relevant keywords.

Figure 2 shows the distribution of files as to the quantity for each type, the identification of this distribution was used as a reference to determine the algorithms used for the classification of the data. Another important point refers to the preparation of the data, elaborated from the file formats used by the algorithms.

Similarly, the amount of data used to store each set of documents according to types, as shown in Figure 3, shows the distribution similar to that shown in Figure 2.

As discussed in Section 2, this research does not address a specific architecture for data management. However, it didactically adopts the concept of data lakes as a reference architecture for the relevant implementations. It is important to note that the data lake concept emerged as an alternative to data warehouses for storing, exploring and analysing large volumes of data. In this sense, data is stored in a raw state and has no explicit schema.

3.1 Data preparation

The preparation of data for specific tools or with specialised algorithms is a fundamental step for the results expected from extraction and classification. Preparing data means organising it in a layout or format required for the demand in question, enabling the necessary operations to generate the expected results. Preparation can range from converting file types to reorganising the arrangement of data sources in specific orders. In this research, data preparation consisted of format conversion. The files are stored initially in binary format with high storage size and structural complexity. It is essential to point out that these characteristics did not hinder data mining. The mining tools have resources for processing various types of files, however, this can compromise the extraction performance in terms of time and feasibility, even for small amounts of data.

3.2 Procedures

The collection of the NGO (OSR) has about 60 GB in digital documents, the strategy adopted was not to store the documents in a database because the collections and contents are in folders, and then a specific computational solution was used to perform the identification and retrieval of digital documents. Figure 4 shows the interface developed for the NGO's demand, which consists of storing only the logical location of the documents and their respective metadata, enabling thematic searches by subject.

Figure 1 Methodological procedures

Figure 2 Distribution of files by quantity

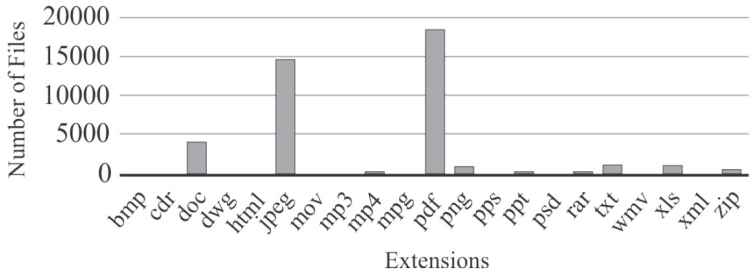


Figure 3 Storage distribution (GB)

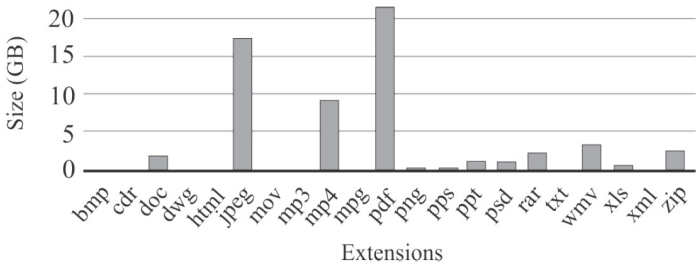
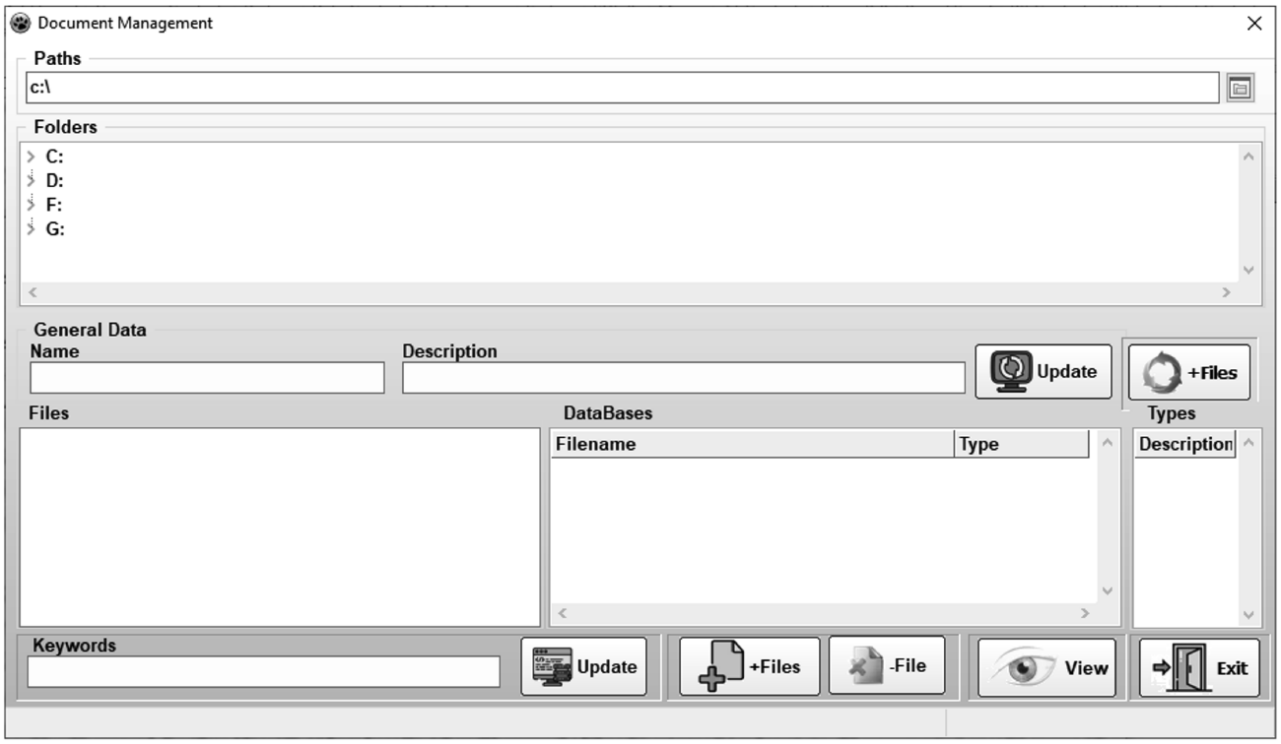


Figure 4 Digital document management system

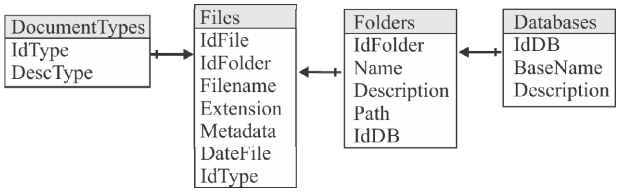


Associating metadata to files requires manual supervision, which is extensive and time-consuming, this justifies using an application to automate the extraction of metadata. In this sense, a tool contributed to speeding up metadata collection to enrich the corporate database. The extraction contributes directly to the location and searches of the digital document database. Binary documents such as PDFs or DOCXs usually do not provide explicit metadata to improve their location. In this sense, complementary data associated with each document significantly improves the search for correlated documents, especially when they are in the same research context, such as biddings, minutes, memos, edicts and cost sheets.

As explained in Section 1, the strategy used for the development of the file manager considered the existing storage structure without discarding it. With this, it was possible to create an application for simple indexing of documents, Figure 5 shows the manager schema. In this representation, it is possible to verify that the application does not store files in the database, the stored data are only references for locating and identifying the documents. The table structure shows the attributes for metadata storage, adding resources for document retrieval, particularly the table files.

In this context, data classification uses machine learning algorithms to summarise quantitatively, create groupings and verify predetermined characteristics or relationships in a specific scenario. Data classification is generally the last step in data set processing, followed by presenting results. The research uses data classification as an intermediate step, performed after extracting data from files and updating the database with metadata for data retrieval.

Figure 5 Database schema for the digital document management system

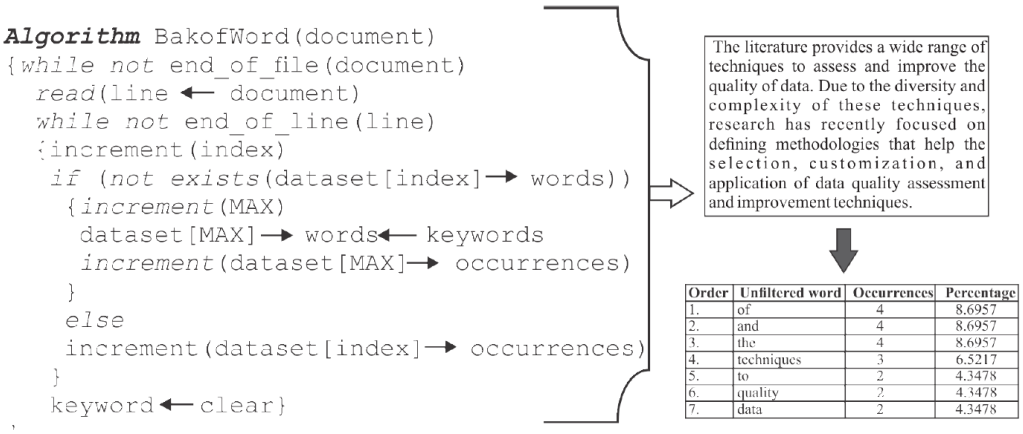


4 Results

The data extraction started with identifying the tokens for each text file, this step used bag-of-words, allowing to individualise the text items without restrictions on words, symbols and punctuations. The model checks word count and disregards grammatical details and word order. The nomenclature comes from the fact that any information about word order or structure in the document is not essential. The model considers only the occurrence of the words, the position in the text is not relevant.

Punctuation marks and other dispensable symbols were removed later in the classification step. The text fragment shown in Figure 6 illustrates how the algorithm for creating the bag-of-words algorithm works. In this illustration, there is no content selection yet, because it is an illustration in which only the seven tokens with the most occurrences were selected, as illustrated in Figure 6.

Figure 6 Simple example of bag-of-word and text fragment extracted from Batini et al. (2009)



The extraction of the keywords uses a ranking algorithm for classifying the number of occurrences of the tokens. Figure 7 shows the illustration of the algorithm used to classify and extract the six elements with the highest occurrence in the text.

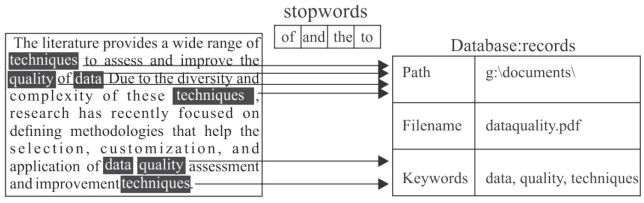
Figure 7 Simple classification an extract keys algorithm

```
Algorithm ExtractKeys(dataset)
{ i←2
  while (i<=MAX)
  { k← 1  j← i-1  ch← c[i]  ed← dataset[i]
    While (j>1) and (k = 1)
    if ch<c[i]
    {c[i+1]←c[i]  e[j+1]← dataset[i] j←j-1
    }else
    k← j+1  c[k]←ch  e[k]←ed
  }
    keyword ← clear    i ←MAX
    while i<=(MAX-6)
      keywords ← keywords+dataset[i]→ words
  return keywords
 }
}
```

The identification of the keywords by checking the number of occurrences for each token in the respective files disregarding characters such as punctuation. After creating the lists containing the words and their respective occurrences, the list is sorted in ascending order, representing the tokens with the highest number of occurrences for the content of each file.

Figure 8 shows the preprocessing of the texts using the stop-word list for exclusion of unnecessary tokens for classification of elective items to keywords.

Figure 8 StopWords



The data extraction represents the most significant difficulty considering the binary format of the files, thus, the initial phase related to the conversion of binary content to text was essential to use the files with the algorithms for classification. As discussed in Section 3, binary contents are not an impeditive factor for the processing used in the research. However, textual contents facilitate the use of algorithms for text mining. In some cases, it is mandatory to use utf-8 (Unicode Transformation-8-bit) standard textual content because they do not work directly with binary files such as PDFs and DOCXs. The retrieval of files with keywords can be available with some strategies according to the demands, such as simple queries using structured query language, specific algorithms for distance calculation and phonetic search algorithms like Soundex.

The conversion of the text into tokens uses an Object Pascal application with a Lazarus GUI. This application uses a simple method of sorting or selection, grouping the tokens according to the number of occurrences of each. The six tokens with the highest occurrence as metadata or keywords for the document are selected and are disregarded words with less than three symbols or letters.

5 Conclusion

The research contributed with a computational solution to enrich an institution's database with social action (NGO). Update metadata to the existing database contributes directly to data enrichment. It was also essential to show the possibility of using the existing data, allowing the entire collection of historical archives without duplicating documents. Data extraction improves the retrieval of unstructured documents, also helping in the creation of information related to thematic groups of documents. Some difficulties faced refer to the formats of the files available in the NGO's collection, as shown in the Figures 2 and 3, this disfavoured the direct extraction of content, it was necessary to perform covers before the extractions. In summary, the research questions were answered satisfactorily with practical demonstrations of data extraction from documents, data enrichment in textual content, and metadata as a resource for data retrieval in archives. Contents such as videos and audio were not addressed in this research, although they represent a relatively small portion of the content, they are relevant and

should also be classified. In this sense, work related to textual data extraction is opportune for this demand. Content stored in video files can be used by extracting the audio layer and then converting the audio into texts, from this conversion, the process for identifying and classifying the keywords follows the same methodology proposed in this research. In the research was adopted that data management uses a local application, this includes data retrieval in queries with pre-established criteria. A relevant contribution is constructing a web or mobile interface for remote queries to the social observatory database. It is essential to highlight that the database is not restricted to NGO use and is available for public consultation. It can collaborate with academic research and public monitoring of documents collected by the NGO in the municipal government's legislative processes and executive actions. Currently, the storage and use of the data are local. The next step of the project concerns the migration of both the data and the search engine to a cloud platform, and this will make the data accessible in locations outside the NGO headquarters so that the managers will be able to perform their work with more mobility, this is an essential point because part of the activities of the NGO represents the supervision of works in their respective execution sites.

References

Batini, C. et al. (2009) 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys*, Vol. 41, pp.16:1–16:52.

Blinston, K. and Blondelle, H. (2017) 'Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents', *Geophysics*, Vol. 36, pp.257–261. Doi: 10.1190/tle36030064.1.

Chekabab, S.M. et al. (2020) 'A health metadata-based management approach for comparative analysis of high-throughput genetic sequences for quantifying antimicrobial resistance reduction in Canadian hog barns', *Computational and Structural Biotechnology Journal*, Vol. 18, pp.2629–2638.

Ha, H.T. et al. (2021) 'Contract metadata identification in Czech Scanned documents', *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, Vol. 2, pp.795–802.

Han, Y. and Wan, X. (2018) 'Digitization of text documents using PDF/A', *Information Technology and Libraries*, Vol. 37, pp.52–64.

Hashmi, A.M. et al. (2020) 'Rule based approach to extract metadata from scientific PDF documents', *Proceedings of the 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, pp.1–4. Doi: 10.1109/CITISIA50690.2020.9371784.

Jeong, H-O. et al. (2021) 'Database metadata standardization processing model using web dictionary crawling', *Journal of Digital Convergence*, Vol. 19, pp.209–215.

Kim, J-C. and Chung, K-Y. (2020) 'Knowledge expansion of metadata using script mining analysis in multimedia recommendation', *Multimedia Tools and Applications*, pp.1–17. Doi: 10.1007/s11042-020-08774-0.

Löbe, M. (2021) 'What metadata? Defining different types of digital assets as application targets of metadata in clinical research informatics', *Studies in Health Technology and Informatics*, Vol. 287, pp.124–125.

Moen, W.E. (2000) *National Information Standards Organization (NISO) Standards Committee AV*, UNT Digital Library.

Müller, L. and Gleim, L.C. (2021) 'Managing versioned web resources in the file system', *Proceedings of the International Conference on Web Engineering*, pp.513–516. Doi: 10.1007/978-3-030-74296-6_41.

Oukhouya, L. et al. (2021) 'A generic metadata management model for heterogeneous sources in a data warehouse', *Proceedings of the 4th International Conference of Computer Science and Renewable Energies*. Doi: 10.1051/e3sconf/202129701069.

Saeed, M.Y. et al. (2020) 'Unstructured text documents summarization with multi-stage clustering', *IEEE Access*, Vol. 8, pp.212838–212854. Doi: 10.1109/ACCESS.2020.3040506.

Sawadogo, P. et al. (2019) 'Metadata management for textual documents in data lakes', *Proceedings of the 21st International Conference on Enterprise Information Systems*, pp.72–83.

Skluzacek, T.J. et al. (2018) 'Skluma: an extensible metadata extraction pipeline for disorganized data', *Proceedings of the IEEE 14th International Conference on e-Science (e-Science)*, pp.256–266. Doi: 10.1109/eScience.2018.00040.

Website

1 http://observatoriosocialroo.org.br/