Thematic Article

Institutional Repository Keyword Analysis with Web Crawler

Mariângela Spotti Lopes Fujita[1], **Isaque Katahira**[2], **Jéssica Beatriz Tolare**[3]

Recommended citation:

Fujita, M. S. L., Katahira, I., & Tolare, J. B. (2022). Institutional Repository Keyword Analysis with Web Crawler. *Central European Journal of Educational Research, 4*(2), 54–59. https://doi.org/10.37441/cejer/2022/4/2/11395

Abstract

This study aims at investigating procedures of semantic and linguistic extraction of keywords from metadata of documents indexed in the Institutional Repository Unesp. For that purpose, a web crawler was developed, that collected 325.181 keywords from authors, in all fields of knowledge, from February 28th, 2013 to November 10th, 2021. The preparation of the collection, extraction and analysis environment used the Python programming language, composed of three program libraries: library requests, which allows manipulation of hyperlinks of webpages visited through web crawler; BeautifulSoup library, used to extract HTML data through webpage analysis; and Pandas library, which has an open code (free software) and stands for providing tools for high performance data manipulation and analysis. The final listing consisted of 273,485 keywords, which represents 15.9% of the listing initially collected. Results indicated that the most recurring problem was the duplication of keywords, with 51,696 duplicated keywords, representing indicators of inconsistencies in the search for documents. It is concluded that the refinement of keywords assigned by authors eliminates the incorporation of a set of symbols that do not represent the authors' keywords with the same spelling, but with upper/lower case variations or lexical variations indexing different documents.

Keywords: institutional repositories, web crawler, indexing by author

Introduction

Information retrieval system is a mediating environment between a set of documents and their potential requester. Its efficiency depends on the quality of representation of information items (documents) and requests from its users (searches) (Neto & Ferneda, 2016). In this sense, the representation process needs to converge so that the user locates the documents of interest. As for the storage, in possession of a document, the information system must represent it already thinking about expressions that define it and are intuitively recognized by the user. As for the retrieval, when accessing the system, the user types search expressions which represent their search need and to be successful they must correspond to what is stored in the system. Considering these two representations (of the system and the user), the retrieval system will be as efficient as "the search expression is represented in a similar way to the way the documents were represented, so that a comparison between two representations is possible" (Neto & Ferneda, 2016, p. 33).

Among information retrieval systems, the establishment of institutional repositories (IR) is related to the observation of institutional needs for organization and dissemination of services provided, as well as research carried out. According to Rieger (2007), several points need to be observed when evaluating the context of implementing an IR, from content characteristics to user needs and available human, financial and technological resources. The IR play a fundamental role in the dissemination of knowledge, especially academic publications. Running on a wide range of software platforms, with great diversity of installation, settings and support systems, IRs remodeled the ways for storing, organizing, and retrieving materials, making these processes more efficient.

[1] São Paulo State University – Unesp, Marília, São Paulo, Brazil; mariangela.fujita@unesp.br
[2] São Paulo State University – Unesp, Marília, São Paulo, Brazil; isaque.katahira@fatec.sp.gov.br
[3] São Paulo State University – Unesp, Marília, São Paulo, Brazil; jessica.tolare@unesp.br

The expansion of relatively low-cost technologies and the growing need for organization, availability and retrieval of information have advanced considerably in recent decades, since the 1990s, with the Open Archives Initiative (OAI) and Open Access (OA). Such advances were associated with discussions on free access to academic-scientific productions (Sanchez et al., 2019). Associatively, the growth and sharing of data by the organization of repositories were so substantial that contextualizing the point we are at becomes a risky task, given the uninterrupted, and even dizzying, evolution of related studies.

In this scenario, it is understood that the information can be represented in different formats, such as texts, images, sounds, among others. By representing, it is meant to create substitutes for a given object; choosing some elements for summarizing such as abstracts, keywords or metadata. According to Novellino (1996), the representation is characterized, mostly, by a process of substituting and summarizing.

The main feature of the process of information representation is the replacement of a long and complex linguistic entity – the document text – by its summarized description. The use of such summarization is not only a consequence of practical restrictions regarding the volume of material to be stored and retrieved. This summarization is desirable for it aims at demonstrating the essence of the document. Thus, it works as a way to emphasize what is essential in the document considering its retrieval, besides being the ideal solution for organizing and using information (Novellino, 1996, p. 38).

Due to their growing importance, IR, in addition to being an essential source of information in contemporary times, have also become an object of study, since they enable not only the conduction of research, but also the knowledge of the entire search path performed by the user, since, when navigating through the digital system, each user action can be retrieved by the manager.

When thinking about retrieval systems with an exponential number of publications and users willing to retrieve specific materials, strategies for information retrieval must be investigated for recognition and comparisons between controlled and natural vocabularies. Thus, the investigation proposed in this article aims to provide a list of keywords that can be used by the team responsible for indexing the materials. This list would help in the convergence between indexed terms and terms searched by users. Obviously, this work team could also use the list of subjects in the repository (authors' keywords) for the building or updating tasks, if refined beforehand.

The standardization proposal is an attempt to minimize the incompatibilities between the indexing language and the search language, since the server software used to manage the repository provides users with several tools to assist them in their searches, in order to improve the search mechanisms and results. However, in order to successfully achieve this task, it is essential to understand how users Interact with IRs, so that system updates will be able to eliminate or minimize the obstacles users encounter when trying to access the desired material, and they can also improve the operation and efficiency of recovery systems.

In this sense, this study aims at investigating procedures of semantic and linguistic extraction of metadata keywords in documents indexed in repositories, understanding that, for the effective representation of a document, the selection and accurate description of metadata is essential. Reis (2008) points to the keyword metadata as an example, highlighting the functions of organizing, classifying, hierarchizing documents in the repository and facilitating retrieval by subject. For this purpose, web crawler was implemented for collecting keywords from the Institutional Repository at Unesp (IR – Unesp) and creating a separate database for analyzing keyword standardization of keywords through frequency analysis. The database generated in this study can be used to standardize and update the databank of the above-mentioned repository and thereby improve the retrieval of relevant information.

Research design and Methods

The collection and extraction of authors' keywords was carried out at the Institutional Repository (IR) Unesp, 5th largest IR in Brazil and 22nd largest in the world, on November 10, 2021, using a web crawler, a tool that acts as a network crawler capable of browsing through all IR webpages in a methodical and automated way in order to collect data of interest to the researcher.

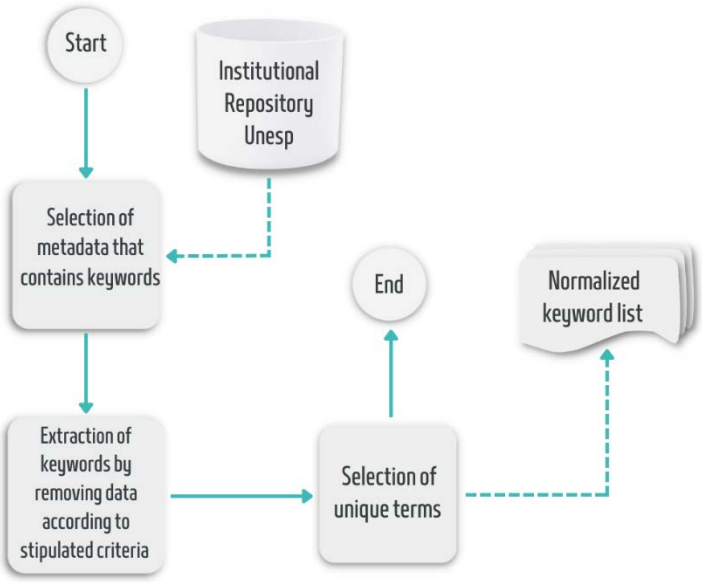
The web crawler collected 325,181 keywords from authors, from all areas of knowledge, allocated in the Institutional Repository at Unesp since its implementation up to the day information was collected.

The environment for collecting and analyzing the authors' keywords was prepared for the use of Python programming language. Python is widely used to address problems involving large volumes of data (Barbosa & Perico, 2019) as it provides fast learning curves (Curty & Serafim, 2016). In addition, this programming

language has many libraries that can be used for data analysis (Chiavegatto, 2015), a fact that justifies its choice for manipulating the dataset discussed in this article.

In the case analyzed, the authors' keywords were retrieved by collecting the metadata that is on the IR-Unesp web pages, located on the left side; followed by extracting the authors' keywords, selecting unique terms, and producing the standardized list of keywords, as shown in Figure 1.

Figure 1. Method used for selecting, extracting and creating standardized keywords



The selection of unique terms was performed through: I) exclusion of regular expressions, (keywords, 3, [r'"', r'^[0-9]+', r'^[0-9]+\-[0-9]+']⁴); II) removal of symbols that do not represent the authors' keywords, such as "#", "*" and "&" , as they cause damage to information retrieval, as they are prejudicial to information retrieval. For example, "#ProstateCancer", which should read "Prostate Cancer". Another example is in "–Learning" active, which should be passed as Learning active. Not processing these symbols characterizes the improvement of information retrieval without causing damage to the user's search; III) removal of leading and trailing spaces from keywords and IV) unification of keywords with uppercase/lowercase characters. To eliminate stopwords, the NLTK platform was used, due to its collaboration for programming in Python to work with human language data <stopwords=nltk.corpus.stopwords.words('portuguese')>. Functions for prefixes and suffixes analysis were not applied in this experimental phase.

The web crawler was able to go through the 32 web pages from A to Z, also including pages with expressions starting with numbers (0-9), each with approximately 10,000 keywords, totaling 325,181 keywords of authors collected, from all areas of knowledge assigned at RI Unesp since its implementation on February 22, 2013 until the date of data collection (November 10, 2021). The address for accessing the authors' keywords is available at: https://repositorio.unesp.br/browse? type=subject.

For the manipulation of the listings, the Pandas Dataframe was added to the tool, which is developed for efficient analysis and manipulation of data. Subsequently, the re (regular expressions) library was added, which provides operations for matching regular expressions often called regex. To homogenize the accents, the unidecode library was added. Finally, the Beautiful Soup library was applied to extract keywords from the authors of RI Unesp pages (HTML and XML).

After this basic step of cleaning the keywords, the API available at: https://detectlanguage.com/ was used to identify the languages, with the highest occurrence of three: English (61.9%); Portuguese (17.5%); Spanish (6.7%) and other languages (13.9%). Language identification is a relevant feature for analyzing multilingual databases.

The treatment of keywords assigned by the authors helps the user to successfully carry out his research. The database cleanup, extraction of authors' keywords constituted by selection of metadata, using web crawler, is an accessible procedure and can be performed by the repository manager. With this procedure, the responsible

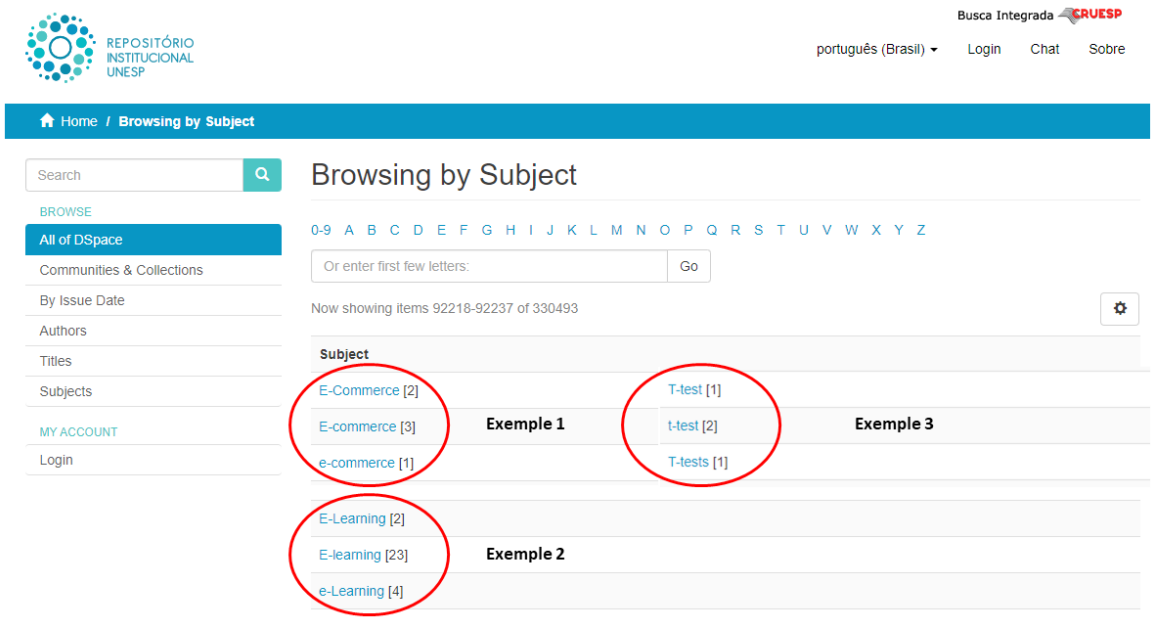
[4] This is a python-based regular expression. It takes keywords that have three or more characters. The letter "r" stands for the unprocessing of keywords that are empty or contain only numbers, which are represented by means of r'^[0-9].

for the information storage system will be able to eliminate, or at least minimize, obstacles users find when trying to access indexed academic publications, besides improving the operation and efficiency of retrieval systems by generating a new list of standardized keywords.

Results and discussion

To create the list of the terms used by the authors, web crawler was implemented. Such web crawler visited 32 pages of the IR Unesp and collected 325,181 keywords from authors registered in the indexed documents until 11/10/2021. Through data collection, it was found that the list of keywords from authors in the IR Unesp does not have control of vocabulary in natural language, as exemplified in Figure 2:

Figure 2. Example of finding errors in the list by keywords



Source: https://repositorio.Unesp.br/browse?type=subject

In examples 1 and 2 (Figure 2), it is possible to identify those identical words, but with uppercase or lowercase letters index different materials. In example 3, we observe that there is no plural/singular treatment either, which can hinder the search for the desired material and prevent its retrieval by the user. In addition to measures to eliminate retrieval losses due to literal non-correspondence of words, to improve retrieval strategies within RI Unesp, it is essential to know better the language used by users, in order to recognize the relevance of this vocabulary. By looking for interrelationships that allow semantic and syntactic aggregations that eliminate redundancies and ambiguities, the retrieval system increases its effectiveness, since the analytical processing is organized to offer the retrieval with the maximum of useful information and the minimum of useless documents, so that the user finds what he needs, but does not waste time with dissociated results (LANCASTER, 2004).

In view of the above, the first criteria for vocabulary treatment were applied as in Table 1:

Table 1. 1st Stet on the treatment of keywords from authors in the IR Unesp

List of keywords from authors in the IR Unesp	
Period of analysis	02-22-2013 up to 11-10-2021
Total of keywords	325,181
Duplicated keywords	51,696
Total of valid keywords	273,485
Average extension of keywords	2,42

Source: Authors

To perform the data analysis the API[5] was used, which is a programming application used to provide a general programming interface. From this analysis it was possible to observe that the list of valid keywords assigned by authors to their documents is mainly composed of three languages: English (50.39%), Portuguese (25.76%), and Spanish (14.94%).

In addition, it was found that the authors' list of keywords composed of 325,181 records did not have any homogenization of natural language, which constitutes an obstacle to the recovery of materials and, therefore, requires investigation. Therefore, we found that the lack of control in the vocabulary used to represent the materials indexed in the IR Unesp can impact the retrieval of these materials by users, making it necessary to consider this factor to improve the retrieval of information by users, from strategies that increase compatibility between the way of indexing and the way of searching for information.

Similarly, Miguéis et al. (2013) addressed the importance of keywords and their attribution in the metadata by authors in the health field, comparing them with the use of MeSH controlled vocabulary in the final published version. Miguéis et al.'s (2013) study is similar to the results obtained in the research herein, even with a difference of almost 10 years in its development. In both works, the refinement of authoritative forms of subject headings attributed by the authors to the documents that will be indexed in Institutional Repositories was considered necessary, as during the process of indexing words, a set of symbols that do not represent the authors' keywords may occur, thus making the indexed documents unretrievable. In order to avoid this lack of document retrieval, the adoption of controlled vocabulary is directly associated with the standardization and consistency of retrieved materials in Miguéis et al.'s (2013) work.

A more recent study developed by Fujita and Tartarotti (2020) confirms the importance of controlled vocabularies for the representation and retrieval of documents. The authors aimed at analyzing the keywords of the scientific production of researchers on a digital platform where their academic curricula are recorded. Results showed that there is a lack of standardization in the keywords recorded in the scientific production of researchers in this digital environment, resulting in low consistency in the retrieval of subjects. In order to overcome the problem, the authors recommended the development and establishment of guidelines for the authors and for the platform itself.

Results presented by Fujita and Tartarotti (2020) are in line with results herein, as both show the need to standardize keywords, which is an attempt to minimize the incompatibilities between the search and retrieval of scientific productions by users.

Conclusions

Given the importance of Institutional Repositories, the standardization of authors' keywords is an attempt to minimize incompatibilities between authors' vocabulary and potential readers (repository users). However, to perform this task, the IR systems need to provide users with tools that help them in their searches and, consequently, enable the improvement of search engines through authors' keywords.

The refinement of authors' keywords in Institutional Repositories is necessary, as at the time of indexing the words, a set of symbols that do not represent the authors' summarization intentions may be incorporated, making the indexed documents non-retrievable. In this sense, the implementation of a tool for collection using a web crawler proved to be efficient for the analysis, management and organization of keywords assigned by authors in documents indexed in institutional repositories and can be used for document retrieval more efficiently.

From the total of 325,181 keywords available in the IR Unesp, 231,308 (71.13%) were used only once, that is, they are related to only one document. The number of co-occurrence of pairs of authors' keywords was 93,873.

Knowing the inconsistencies of keywords assigned by authors in documents indexed in IR-Unesp, knowledge of the current situation of stored materials is essential to propose strategies that improve the representation of materials and, consequently, the recovery of these documents, according to users' interests. Periodic updating of controlled vocabularies is considered complicated and difficult due to losses, distortions of author's words, attributions of erroneous terms, high cost; in addition, there may be difficulties in interoperability information due to incompatibility between controlled languages. Thus, the development of strategies for Natural Language Processing is recommended to improve the formulation of updated and efficient controlled vocabularies.

[5] Available at: https://detectlanguage.com/

Therefore, the list of keywords obtained by the web crawler's extraction can offer possibilities for creating and maintaining vocabulary control tools, from the alphabetical list of keywords and even thesauri, but it needs to undergo strict terminological and linguistic control in order to be used.

Funding: This research was funded by CAPES.

Acknowledgments: We thank Johnathan Dabney for the English language editing.

Author Contributions: All authors actively participated and together in the development of the study. Author 1 developed the web crawler. Author 2 carried out the theoretical deepening using related studies. Author 3 was responsible for the translation and revision of the study.

References

Barbosa, P. F. M., & Perico, D. H. Desenvolvimento de material didático para ensino a distância de python básico. https://fei.edu.br/sites/sicfei/2019/cc/SICFEI_2019_paper_87.pdf.

Chiavegatto Filho, A. D. P. (2015). Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, *24*, 325–332.

Curty, R. G., & da Silva Serafim, J. (2016). A formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Informação*, *21*(2), 307–331.

Fujita, M. S. L., & Tartarotti, R. C. D. E. (2020). Análise de palavras-chave da produção científica de pesquisadores: o autor como indexador. *Informação & Informação*, *25*(3), 332–374.

Jurafsky, D., & Martin, J. H. (2000). Speech & language processing. Pearson Education India. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Miguéis, A., Neves, B., Silva, A. L., & Trindade, Á. (2013, October). A importância das palavras-chave dos artigos científicos da área das Ciências Farmacêuticas, depositados no Estudo Geral: estudo comparativo com os termos atribuídos na MEDLINE. In *4ª Conferência Luso-Brasileira sobre Acesso Aberto*. https://estudogeral.uc.pt/bitstream/10316/24485/1/A%20import%C3%A2ncia%20das% 20palavras-chave.pdf

Neto, J. J., & Ferneda, E. (2016). Ontologia como recurso de padronização terminológica no processo de recuperação de informação. *Informação em Pauta*, *1*(1), 30–45.

Novellino, M. S. F. (1996). Instrumentos e metodologias de representação da informação. *Informação & Informação*, *1*(2), 37–45.

Ochando, M. B. (2014). Nuevos retos de la tecnología web crawler para la recuperación de información. *Métodos de Información,* 4(7), 115–128.

Pandas. Python Data Analysis Library. 2021. https://pandas.pydata.org/.

Python Software Foundation. Documentação de Aplicações para Python. 2021. https://www.python.org/search/?q=+BeautifulSoup&submit=.

Reis, C. M. B. (2008). Otimizando a recuperação da informação em Repositórios Institucionais. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz. https://www.arca.fiocruz.br/handle/icict/2950.

Rieger, O. Y. (2007). Select for success: key principles in assessing repository models. *D-lib Magazine*, 13(7/8), 1–8.

Sanchez, F. A., Vechiato, F. L., & Vidotti, S. A. B. G. (2019). Recomendações de encontrabilidade da informação para repositórios institucionais. In *Seminário em Ciência da Informação – SECIN*, 8., Informação social no contexto da Ciência da Informação. Londrina: Universidade Estadual de Londrina. http://www.uel.br/eventos/cinf/index.php/secin2019/secin2019/paper/viewFile/577/406.