The Canadian Journal of Information and Library Science
La Revue canadienne des sciences de l'information et de bibliothéconomie

Forms and Functions of Author Keywords in Theses and Dissertations at the UNESP Institutional Repository (Brazil)

Mots-clés et descripteurs dans les thèses et mémoires en sciences de l'information

Mariângela Spotti Lopes Fujita

Volume 47, Number 2, 2024

Bobcatsss 2024 Special Issue
Numéro spécial Bobcatsss 2024

URI: https://id.erudit.org/iderudit/1116006ar
DOI: https://doi.org/10.5206/cjils-rcsib.v47i2.17628

See table of contents

Publisher(s)

Canadian Association for Information Science - Association canadienne des sciences de l'information

ISSN

1195-096X (print)
1920-7239 (digital)

Explore this journal

Cite this article

Spotti Lopes Fujita, M. (2024). Forms and Functions of Author Keywords in Theses and Dissertations at the UNESP Institutional Repository (Brazil). *The Canadian Journal of Information and Library Science / La Revue canadienne des sciences de l'information et de bibliothéconomie*, *47*(2), 166–175. https://doi.org/10.5206/cjils-rcsib.v47i2.17628

Article abstract

This research aimed to prepare guidelines for authors by investigating forms and functions of keywords assigned by authors in theses and dissertations defended in 2023 in the Graduate Program in Information Science at Unesp. The exploratory and descriptive study utilized a sample collected in the Unesp Institutional Repository. A corpus of 31 theses and 14 dissertations submitted to the Unesp Institutional Repository comprised a total of 183 keywords in Portuguese without duplicates and an average of 4.7 keywords, considering 213 keywords with duplicates. The analysis results initially identified that the Repository has a tutorial on using the Unesp Thesaurus to control vocabulary and that the authors use natural language to assign keywords. The findings reveal that, out of the 183 keywords, 89 (48%) are exclusive, singular and specific to the area of Information Science, candidates for descriptors in the Unesp Thesaurus. The other 94 keywords (51.3%) have 40 (21.3%) exact descriptors, and the other 54 (29.5%) present forms and functions that serve as examples for inclusion in the tutorial instructions. Based on the results obtained, it is concluded that the percentage of 21% overlap between keywords and descriptors reveals that the Unesp Thesaurus was consulted by the authors when filling out keyword metadata and that the low number of exact descriptors and exclusive keywords indicate that they need to be included as new terms. It is recommended, therefore, to define an Indexing Policy that considers the need for hybrid coexistence between natural language and vocabulary control.

The Canadian journal of information and library science - La Revue canadienne des sciences de l'information et de bibliothéconomie (CJILS-RCSIB)
Vol. 47, No. 2 (2024) - Bobcatsss 2024 conference special issue
DOI: 10.5206/cjils-rcsib.v47i2.17628
Conference paper

Forms and Functions of Author Keywords in Theses and Dissertations at the UNESP Institutional Repository (Brazil)

Mariângela Spotti Lopes Fujita
São Paulo State University, Marília, Brazil

This research aimed to prepare guidelines for authors by investigating forms and functions of keywords assigned by authors in theses and dissertations defended in 2023 in the Graduate Program in Information Science at Unesp. The exploratory and descriptive study utilized a sample collected in the Unesp Institutional Repository. A corpus of 31 theses and 14 dissertations submitted to the Unesp Institutional Repository comprised a total of 183 keywords in Portuguese without duplicates and an average of 4.7 keywords, considering 213 keywords with duplicates. The analysis results initially identified that the Repository has a tutorial on using the Unesp Thesaurus to control vocabulary and that the authors use natural language to assign keywords. The findings reveal that, out of the 183 keywords, 89 (48%) are exclusive, singular and specific to the area of Information Science, candidates for descriptors in the Unesp Thesaurus. The other 94 keywords (51.3%) have 40 (21.3%) exact descriptors, and the other 54 (29.5%) present forms and functions that serve as examples for inclusion in the tutorial instructions. Based on the results obtained, it is concluded that the percentage of 21% overlap between keywords and descriptors reveals that the Unesp Thesaurus was consulted by the authors when filling out keyword metadata and that the low number of exact descriptors and exclusive keywords indicate that they need to be included as new terms. It is recommended, therefore, to define an Indexing Policy that considers the need for hybrid coexistence between natural language and vocabulary control.

Keywords: theses and dissertations, keywords, author-supplied keywords, controlled vocabularies, unesp, institutional repository

Introduction

The combination of keywords and descriptors is a current discussion in the literature, and this reveals a trend whose advantages and disadvantages are influential in decision-making regarding information representation and retrieval of theses and dissertations.

Maurer and Shakeri (2016) observe a lack of information about how the number of keywords provided by researchers/authors of theses or dissertations correlates with their experience in submitting articles. In various library environments, the trend is to improve the retrieval of academic theses and dissertations by optimizing access points and developing practices that take advantage of keywords and metadata provided by the authors.

Han, Harrington, Black and Kudeki (2016) argue that, in recent decades, libraries have undergone an evolution in their retrieval services. They have transitioned from OPACs to web-scale discovery services that enable access to both OPAC resources and articles and chapters available in major database subscriptions, whose resources are described with more specific subject terms.

Libraries can handle increasingly more metadata created by non-cataloguers (e.g., author-supplied metadata) that often use subject terms not available in established controlled vocabularies, and continually update them with the most specific keywords from specialized domains.

Thus, libraries could use subject metadata keywords filled in by the authors of theses and dissertations and continuously update them in controlled vocabularies of specialized domains. However, for theses and dissertations, authors need instructions on assigning keywords, which include using controlled vocabulary, reducing ambiguities, and understanding the role of keywords in accurately representing significant content.

The research aimed to develop guidelines for authors on the importance of representing the content of theses and dissertations through keywords. The objective of the research was to investigate the forms and functions of keywords assigned by authors by conducting exploratory and descriptive research on a sample of theses and dissertations defended in 2023 in the Graduate Program in Information Science, collected from the Unesp Institutional Repository.

Correspondence concerning this article should be addressed to Mariângela Spotti Lopes Fujita: mariangela.fujita@unesp.br

Theoretical Framework

Natural language keywords and controlled vocabulary descriptors, such as thesauri, are part of the subject metadata of journal articles, books, theses, and dissertations in information systems. This hybrid combination of natural language and descriptors is a product of digital interoperability across various information systems that share bibliographic records with full texts. These bibliographic records are composed of metadata filled in by the authors themselves, with authorship data, title, abstract, publication data (year, publisher, place of publication), keywords, descriptors, biographical notes, Uniform Resource Locators (URLs), digital texts, sounds or image files and others, depending on data storage and retrieval needs. One of the best-known metadata standards is Dublin Core. Yi-Fang & Quanzhi (2008) refer to metadata as useful for searching and browsing digital library collections and that subject metadata can help improve document indexing and retrieval. They recommend that the subject metadata be completed by authors using controlled vocabulary in combination with keyword assignment.

When authors complete the metadata for theses, dissertations, or other scientific works, they consequently alleviate the workload of professionals. Additionally, they reduce labor and time costs for the information system, considering the rapid pace and large volume of publications, which is impossible to be monitored (Mathes, 2004; Gonçalves, 2008).

In some institutional repositories metadata of bibliographic records of theses and dissertations contain keywords assigned by the authors and also descriptors assigned by librarians in a hybrid system of subject representation.

In this repository, the librarian adds descriptors from a controlled vocabulary derived from an indexing process, which primarily differentiates a descriptor from a keyword. This task, performed by professional indexers, is a decision by the information system to complement the representation coverage using descriptors common to other information systems. This procedure ensures greater consistency with other documents in the same subject area and enhances exhaustiveness in retrieval. Professional indexers know about the indexing process and the use of controlled vocabularies developed to obtain greater consistency in the representation of content. While authors assign keywords, they have different objectives. They are unaware of the indexing process and are not trained in using controlled vocabularies.

Névéol, Dogan & Zhiyong (2010) argue that there are significant differences in terms of form and perspective between and author's assignment of keywords to an article and an indexer's use of Medical Subject Headings (MeSH) for indexing terms. The descriptor results from querying the index term in a controlled vocabulary. The difference lies in the indexing process, which consists of the conceptual analysis of the document and the translation of the indexing term into the controlled vocabulary descriptor (Lancaster, 2003). Meanwhile, the keyword can be extracted from any part of the document with or without vocabulary control and can be assigned by authors and editors or even be generated automatically (Gonçalves, 2008) without carrying out a standardized conceptual process unknown to the authors.

Lardera and Hjørland (2020) consider authors of theses and dissertations as expert indexers who possess knowledge on their scientific domain. They argue that these authors understand the content and the function of keywords for researchers within that particular area of knowledge, relevant for theses or dissertations. Authors, professional indexers and users are involved in assigning keywords, even if they have different objectives (Mathes, 2004). In a study on convergence and divergence between tags, keywords and descriptors, Kipp (2009) reports that few studies on authors' keywords compared to descriptors have been conducted and that the controlled vocabularies used by professional indexers require training to be used by authors or users. However, the terminology used by the author will always be different from the controlled vocabulary used by the professional because it is the result of knowledge generated by the evolving domain area, by its scientific nature itself, and is more specific rather than standardized to obtain consistency of terms.

Keywords indicate a subject's main concepts and coverage that will be useful for indexing and retrieving information (Ercan & Cicekli, 2007). Nevertheless, Gonçalves (2008) assesses that judging the relevance of keywords assigned by specialist indexers with the content to which they are linked (title, abstract and text) is challenging. The indexing process carried out by professional indexers is different because they are trained to select indexing terms according to a specific protocol (Névéol, Dogan & Zhiyong, 2010). Another difference highlighted by Névéol, Dogan & Zhiyong (2010) and reinforced by Lardera & Hjørland (2020) refers to the difference in the objectives of selecting indexing terms by professional indexers who consider the article in the larger scope of the collection and specialist indexers (authors) whose focus is the keywords they consider important to describe the content of the article to readers who are researchers in that specific area of knowledge.

Conversely, automatic indexing extracts keywords based on linguistic and statistical parameters of word frequency and the importance of location in the textual structure (Gil Leiva, 2017), which results in keywords whose relevance can be judged by a human indexer in the case of the semi-automatic indexing system. Therefore, in automatic indexing, the system is developed to carry out a standardized indexing process and select keywords, including the use of controlled vocabulary for representation by descriptors.

Taking as a reference the study by Holstrom (2019) about the author, considered as a domain expert, being one of the actors in subject indexing in addition to professional indexers, casual indexers and machine algorithms, it is possible to apply his actor-based model proposition for subject indexing as a complement to existing models aimed at the professional indexer based on the definition of key properties:

> 1) actors are the primary drivers of subject indexing work, 2) observing and understanding many types of actors' processes in real-life situations is as valuable as prescribing correct methods for professional subject indexing, and 3) multiple and different types of actors can perform subject analysis work and subject representation work on the same information objects, and these hybrid (multi-actor) approaches to subject indexing are explicitly supported. (Holstrom, 2019, p.125).

In order to observe and understand the processes of domain experts, studies were carried out by Gil Leiva & Alonso Arroyo (2007); Strader (2009); Névéol, Dogan, & Zhiong (2010); Woolverton, Hoover & Fowler (2011); Schwing, McCutcheon & Maurer (2012); Maurer & Shakeri (2016); Han, Harrington, Black & Kudeki (2016); Zhang et al. (2016), Khatir & Ganiefar (2018); Li (2018); Munan (2018); Freitas & Dal´Evedove (2019); Lu, Li, Zhifeng & Cheng (2019); Philips, Tarver & Zavalina (2019); Fujita & Tartarotii (2020); Golub, Tyrkkö, Hansson & Ahlström (2020); Terra, Agustín Lacruz, Bernardes, Fujita & Bueno de La Fuente (2021); and Fujita (2024).

The studies by Munan (2018); Lu, Li, Zhifeng & Cheng (2019); Li (2018) and Fujita (2024) carried out investigations into the functions or categories that the keywords selected by authors take in representing the content of the text which highlights a proposed methodological standard used by authors to ensure the completion of a subject indexing process.

The comparison between keywords assigned by authors and controlled vocabulary descriptors was investigated in studies by Gil Leiva & Alonso Arroyo (2007); Strader (2009); Névéol, Dogan & Zhiong (2010); Zhang, et al. (2016); Schwing, McCutcheon & Maurer (2012); and Golub, Tyrkkö, Hansson & Ahlström (2020). The studies by Gil Leiva & Alonso Arroyo, (2007) and Golub, Tyrkkö, Hansson & Ahlström (2020) used controlled vocabularies from specialized databases for comparison, while Schwing, McCutcheon & Maurer (2012); Maurer & Shakeri (2016); Han, Harrington, Black & Kudeki (2016); and Strader (2009) used the Library of Congress Subject Headings (LCSH), and the Medical Subject Headings (MeSH) was used by Névéol, Dogan, & Zhiong (2010).

Comparisons between keywords and descriptors in the investigations mentioned above obtained lower equivalence means between keywords and descriptors, but the authors were in favour of the combined use of descriptors and keywords, considering the complementation that adds up and increases the access and retrieval vocabulary by the user, in addition to the possibility of enriching the controlled vocabularies with new terms assigned by the authors despite the problems identified due to the use of natural language.

Investigations into keywords assignment by authors of theses and dissertations were the objective of studies by Strader (2009); Woolverton, Hoover & Fowler (2011); Schwing, McCutcheon & Maurer (2012); Han, Harrington, Black & Kudeki (2016); Maurer & Shakeri (2016); Khatir & Ganjefar (2018); Phillips, Tarver & Zavalina (2019); Freitas & Dal´Evedove (2019); and Terra, Agustín Lacruz, Bernardes, Fujita & Bueno de La Fuente (2021).

The study of overlap between author-supplied keywords and Library of Congress Subject Headings (LCSH) in bibliographic records of electronic theses and dissertations was first developed by Strader (2009) in the Ohio State University online catalogue and replicated by Schwing, McCutcheon & Maurer (2012) and Maurer & Shakeri (2016) in the Kent State University online catalogue. The findings support most of Strader's conclusions, including the complementary nature of the keywords and controlled vocabularies. Complementarity is emphasized in all three studies with evidence that there is value in uniqueness because both keywords and LCSH provide unique terms that enhance access. Researchers have also innovated with regard to partial matching, particularly within LCSH. The fact that exclusivity is important has implications for the continued use and maintenance of LCSH and for future research.

The complementary effect between author keywords and controlled terms is observed by Golub, Tyrkkö, Hansson & Ahlström (2020), who highlight the authors' deficiency of training in indexing and guidelines.

Phillips, Tarver & Zavalina (2019, p.66) observed, in the same way that, "[. . .] though there are more total keywords, a slightly larger percentage of LCSH terms used in the records (61%) are unique compared to keywords (59%)." The analysis carried out by the authors considers that subject headings are longer compared to the simple terms used for keywords, which reinforces the argument of uniqueness and complementarity.

Maurer & Shakeri (2016) highlight the results referring to the relatively higher average of keywords assigned to doctoral dissertations (5.4) in relation to Master's theses (5.1), in line with Wolverton & Hoover (2011) in research with 82 institutions that assigned keywords, only eight specified a minimum number, ranging from one to five keywords.

Using analytical techniques applied to metadata for accessing master's theses from the Digital Repository of São Paulo University, Terra, Agustín Lacruz, Bernardes, Fujita & Bueno de La Fuente (2021) obtained an average of 4.62 keywords in Portuguese and 4.59 in English per record. Given the findings, they consider that it is necessary to define rules for authors' keyword assignment and selection and that vocabulary control requires mutual collaboration between authors and librarians. In an analysis of keywords in the scientific production of researchers for the submission of articles in journals indexed in Scopus and published on the *Portal Docentes Unesp*, Fujita & Tartarotti (2020) found a non-standardized practice for keywords and low levels of consistency in terms of indexing assessment. The study recommended that an information organization and representation policy be drawn up, along with guidelines for authors regarding keyword assignment.

The comparison between keywords assigned by authors and indexing terms by professional indexers in theses and dissertations is investigated by Khatir & Ganjefar (2018) and Freitas & Dal´Evedove (2019). The findings obtained by Khatir & Ganjefar (2018) reveal that the similarity between the two indexes is only 8% and that 40% of the author's keywords and 45% of the professional indexer's terms are extracted from the first 20% of the abstract. The findings reached by Freitas & Dal´Evedove (2019) demonstrate agreement between indexings, assessed with more exhaustiveness in the author's indexing and more precision in the librarian's indexing. It is pointed out that the author needs to be guided regarding using controlled vocabulary for assigning indexing terms.

In summary, the studies that investigated keyword assignment by authors of theses and dissertations carried out quantitative and qualitative analyses whose main observations refer to the author's role as an expert indexer, the average number of keywords assigned to theses and dissertations, the complementary effect between natural language and controlled language and recommendations. The author is one of the main actors in subject indexing by filling in the subject metadata of their scientific productions (theses, dissertations, articles, proceeding papers), which minimizes costs and time for information systems and contributes to the terminological evolution of their area of knowledge. However, they do not receive specific training or guidelines on indexing, and they are unaware of the controlled vocabularies specially constructed for professional indexers' activities. On the other hand, the average number of keywords assigned to dissertations is relatively higher than the average assigned to theses, at around 5.4 and 5.1, respectively, and, in another study, 4.62 to Master's theses. Regarding hybrid combinations between keywords and descriptors, complementarity is a beneficial factor for information systems due to the uniqueness offered by natural and controlled languages, as they provide exclusive terms in each case. The main recommendations refer to developing indexing guidelines for authors to assign keywords with the availability of controlled vocabularies.

Research methods and objectives

The exploratory and descriptive research analyzed keywords assigned by authors in a self-archiving system for theses and dissertations from the Unesp Institutional Repository compared with descriptors from the Unesp Thesaurus. The aim was to develop and improve guidelines for authors on the importance of representing the content of theses and dissertations for consistency between theses and dissertations keywords and controlled vocabulary descriptors.

Description of the sample universe

The incorporation of the Unesp Thesaurus into the self-archiving guidelines played a crucial role in selecting theses and dissertations from the Unesp Institutional Repository for sample identification and analysis. This importance is highlighted by the research of Fujita and Panuto (2024), who found that among the 10 Brazilian repositories examined in the study, only the Unesp Institutional Repository offers some guidelines for subject representation through keyword assignment and author-controlled vocabulary.

The Unesp Thesaurus[1] has a vocabulary with specialized terms from the areas of knowledge of teaching, research and extension activities at São Paulo State University - UNESP. It is built with the combination of controlled vocabularies from the Library of Congress Subject Headings (LCSH), Terminology from the National Library of Brazil (BN), Medical Subject Headings (MeSH), and Descriptors in Health Sciences (DeCS) in MARC21 authority records. The TemaTres software provides queries, and access to the Unesp Thesaurus used to represent the most significant information content of books, theses, dissertations, monographs, final papers, journal articles, documents, legislation, etc. The Permanent Committee of the Unesp Thesaurus made up of catalogers from the libraries of São Paulo State University - UNESP and researchers in the area of Knowledge Organization, has maintained the Unesp Thesaurus since 2013. Maintenance work is continuous to update the correctness of keywords and the consistency of hierarchical relationships between terms. Currently, it contains a total of 209,062 terms, among these, 133,129 (63.68%) are preferred terms and 75,933 (36.32%) are non-preferred terms. There are 25,753 (22.9%) hierarchical relationships, 86,728 (77.1%) associative relationships, and 4,140 Explanatory Notes[2] among the preferred terms.

It was developed to be used in the self-archiving of theses and dissertations in the Unesp Institutional Repository[3] and in the integrated search interfaces of library databases to retrieve information about any document indexed in any Unesp database. To do this, the user may simply access the Unesp thesaurus link below the search interface, type a word, part of a word or phrase in the search box and choose the keyword most representative of the information needed. During the self-archiving of theses and dissertations, the Unesp Institutional Repository provides a form that contains specific metadata for identifying data (authorship, title, Graduate

[1] https://www.biblioteca.unesp.br/tesauro/vocab/index.php

[2] https://www.biblioteca.unesp.br/tesauro/vocab/sobre.php

[3] https://repositorio.unesp.br/home

Program, etc.) and in the case of keywords, the authors are guided to go to the thesaurus and the "Tutorial for using the Unesp Thesaurus" ("Tutorial para o uso do Tesauro Unesp") (UNESP, S.d).

The "Tutorial for using the Unesp thesaurus" (UNESP, S.d) informs the author that the use of the Unesp Thesaurus aims to provide visibility and retrieval in databases and repositories and that its vocabulary adopts three terminological sources widely used by other institutions: Library of Congress Subject Headings, Terminology of the National Library of Brazil and Medical Subject Headings.

In "Research tips for Unesp Thesaurus" ("*Dicas de pesquisa no Tesauro Unesp*"), the tutorial recommends describing subjects using descriptors that best represent the content of the thesis or dissertation and suggests a minimum of 3 descriptors, one of them representing the area of knowledge. Next, it guides the author in typing the term in the thesaurus search interface with correct accentuation and spelling to check the list of occurrences referring to the typed term. If the desired term exists, the tutorial clarifies that it is possible to "browse" more general or specific terms than the searched term, as the thesaurus provides a semantic network with terms related to the entered term. In addition to this instruction, the author is alerted to synonymous terms adopted as preferred by the thesaurus to replace the typed term. In this case, the author must use the preferred term from the thesaurus to ensure representation that aligns with the specialized terminology of their domain.

When the typed term does not exist in the thesaurus, the tutorial's instructions recommend that the author request its inclusion in the institution's library where he/she completed the Graduate Program. The policy of the Unesp Thesaurus Permanent Committee is to maintain constant updating with the inclusion of terms originating from scientific productions that reflect the scientific and technological development of different areas of knowledge.

Finally, authors are advised to avoid acronyms, slang and jargon, phrases with many words, complex concepts, commercial names and generic words (e.g. study, analysis). They are also informed that the Unesp Thesaurus does not contain proper names, names of institutions or geographical locations.

Sample selection criteria

Aiming at investigating the forms and functions of keywords assigned by authors of theses and dissertations while self-archiving into the Unesp Institutional Repository, the corpus of analysis comprised 45 records, 31 records of theses and 14 of dissertations from the Postgraduate Program in Information Science at Unesp submitted by the authors in 2023 to the Unesp Institutional Repository by self-archiving.

The limited sample size and focus on a single knowledge domain aimed to develop and apply an analysis method to observe form and function aspects. This method is based on the manual annotation approach by Lu, Li, Zhifeng, and Cheng (2019), which is applied to each thesis and dissertation record. It also involved assessing the authors' command of terminology and their indexing expertise in the Information Science field during subject analysis. The manual annotation record, which consists of data extracted from the metadata, includes the title, authorship, abstract, and a list of keywords in Portuguese and English provided by the authors. Each keyword from the records list was examined for its form and function.

Each record, composed of data extracted from the metadata of theses and dissertations from the Unesp Institutional Repository, contains data identifying title, authorship, abstract and the list of keywords assigned by the authors in Portuguese and English. In the records of theses and dissertations in Dublin Core metadata format, no Unesp Thesaurus descriptors are assigned by professional indexers, and only keywords are assigned by the authors. For this research, the keywords were extracted in Portuguese to use the Unesp Thesaurus, whose vocabulary primarily comprises terms in Portuguese.

Following this procedure, all keywords assigned by the authors of theses and dissertations in the Portuguese language were extracted, and a single alphabetical list was drawn up with a total of 213 keywords in Portuguese from the lists of 152 keywords extracted from the dissertations and 61 keywords extracted from the theses.

The analysis procedures were divided into instructions for authors of theses and dissertations on keyword assignment during self-archiving in the Unesp Institutional Repository and analysis of data relating to keywords assigned to theses and dissertations aiming at investigating forms and functions of keywords assigned by authors.

Findings and Discussion

The analysis results initially identified that the Unesp Institutional Repository provides, alongside the self-archiving guidelines for theses and dissertations, access to the Unesp Thesaurus controlled vocabulary accompanied by a tutorial on how to use vocabulary control. This tutorial could be linked to the Indexing Policy registered in the Unesp Institutional Repository. In this case, the indexing policy could establish a hybrid indexing process that considers the author as both cataloger and indexer. Additionally, it would emphasize the significance of keyword assignment carried out as an integral part of updating the specialized vocabulary of Unesp Thesaurus, as recommended by Terra, Agustín Lacruz, Bernardes, Fujita & Bueno de La Fuente (2021) and Fujita & Tartarotti (2020).

This tutorial offers objective instructions on using descriptors and the correctness and consistency of terms to be assigned in the keyword field, but it lacks instructions on the importance of vocabulary control for keyword forms

Table 1

Single terms

keyword
Análise de domínio (Domain Analysis)
Análise de domínio (Domain Analysis)
Análise do discurso (Discourse analysis)
Aprendizado de máquina (Machine Learning)
Arquitetura da informação (Information Architecture)
Arquivologia (Archival science)
Bibliometria (Bibliometrics)
Biblioteca universitária (University library)
Biblioteca universitária (University library)
Ciência da Informação (Information Science)
Ciência da Informação (Information Science)
Ciência da Informação (Information Science
Ciência da Informação (Information Science)
Competência digital (Digital literacy)
Competência em informação (Information literacy)
Competência em informação (Information literacy)
Competência em informação (Information literacy)

assigned in 4 theses or dissertations and is the most recurrent keyword in the registration corpus. Information Science is the name of an area of knowledge, hence the reason for its duplicity, however, others such as "domain analysis" and "discourse analysis" are names for research methods used by more than one thesis or dissertation that employed them in the investigation.

Each keyword extracted from a thesis or dissertation was compared to the Unesp Thesaurus descriptors to verify the degree of accuracy or use of non-existing vocabulary. Table 2 reveals that the first 12 keywords in the alphabetical list were not found in the Unesp Thesaurus, nor did keywords have an exact match (Content Analysis), non-preferred terms (Machine Learning), or terms in the plural form (School library).

When comparing each keyword with the Unesp Thesaurus's descriptors, it was identified that, out of 183 keywords, 89 (48%) did not exist as exact or partial descriptors in the Thesaurus, and 94 (51.3%) were listed as descriptors. The group of 89 keywords is representative of the exclusive terms in natural language. In the group of 94 keywords, 40 (21.3%) were exact or identical, and the remaining 54 (29.5%), despite having been identified, had other forms and functions.

Regarding the 54 keywords in the Unesp Thesaurus, the results demonstrate that six (11.11%) terms in singular and plural forms (see examples in Table 3), 31 (57.4%) were single and non-compound terms (see examples in Table 4 and Table 5), and 17 (31.48%) were non-preferred terms, whose synonyms could have been adopted (see examples in Table 6)

Regarding the 89 keywords that do not exist in the Unesp Thesaurus, the results demonstrate that they are:

- 58 (65.16%) new terms, specific to the area of Information Science not yet included (e.g., Data access; Altmetrics; Citation analysis; Cocitation analysis; Domain analysis; Data architecture; Digital competence).
- 12 (13.48%) phrases (e.g., Teaching, research and extension; Social function of knowledge organization systems; Supply chain management; Homeless people; Thesaurus of the United Nations Bibliographic Information System)
- 4 (4.5%) acronyms (DICT; CARE Principles; FAIR principles; LMS).
- 2 (2.24%) geographic names (Angola; São Paulo)
- 4 (4.5%) broad words (Critical perspectives; Interrelationship; Sentiment analysis; Lexical adherence)
- 9 (10.11%) Terms without adherence to the area of Information Science (Police reports; Public sphere; State of the art; Innovation generation; Institutional Identity; Intelligence Process; Psychosocial Risk; Television; Social vulnerability)

The authors could have resolved the forms and functions of the keywords presented by the identified and unidentified terms when using the Unesp Thesaurus and the Tutorial, which could apply these results in their instructions. Conversely, the addition of the 89 new keywords, particularly the 54 keywords that are the new terms specific to the area of Information Science that are not yet in the Unesp Thesaurus, highlights a unique aspect offered by the keyword analysis of the sample as they may be candidates for new descriptors.

Conclusion

The analysis of keywords assigned by Information Science authors in the Unesp Institutional Repository's self-archiving system for theses and dissertations compared with descriptors from the Unesp Thesaurus aimed to investigate keyword forms and functions and prepare guidelines for authors on the importance of content representation and retrieval.

From the findings, the investigation concluded that the 21% overlap of keywords and descriptors reveals that the Unesp Thesaurus was consulted by the authors when filling out the keyword metadata. The overlap suggests that expertise in Information Science aids in selecting preferred descriptors from the Unesp Thesaurus. However, it also highlights limitations due to inadequate indexing skills for identifying concepts that represent significant content and for applying keyword functions that enhance precision in selection by Method, Technique, Topic, etc., as well as standardized

Table 2

Keywords from theses and dissertations and descriptors from the Unesp Thesaurus

Keyword	Unesp Thesaurus Descriptors
Acesso a dados (Data access)	
Acoplamento bibliográfico (Bibliographic coupling)	
Acoplamento bibliográfico de período (Period bibliographic coupling)	
Aderência Lexical (Lexical Adhesion)	
Agência de checagem de fatos (Fact-checking agency)	
Altmetria (Altmetrics)	
Ambiguação da informação (Information ambiguation)	
Análise de acoplamento (Coupling Analysis)	
Análise de citação (Citation analysis)	
Análise de citação relational (Relational citation analysis)	
Análise de citação uni-variada (Uni-variate citation analysis)	
Análise de cocitação (Co-citation analysis)	
Análise de conteúdo (Content analysis)	*Análise de conteúdo* (Content analysis) (Communication)
Análise de domínio (Domain Analysis)	
Análise de sentimento (Sentiment Analysis)	
Análise do discurso (Discourse analysis)	*Análise do discurso* (Discourse analysis) BT Semântica
Angola	
Aprendizado de máquina (Machine Learning)	Use *Aprendizado do computador* (Use computer learning)
Arquitetura da informação (Information Architecture)	*Arquitetura da informação* (Information Architecture) BT *Ciência da informação* (Information Science)
Arquitetura de dados (Data architecture)	
Arquivo audiovisual (Audiovisual archive)	
Arquivologia (Archival Science)	*Arquivologia* (Archival Science) BT *Ciência da Informação* (Information Science) BT *Documentação* (Documentation)
Arquivo de mulheres (Women’s archives)	Arquivos (Archives)
Arquivos pessoais (Personal archives)	Arquivos pessoais (Personal archives) BT Arquivos (Archives)
Arquivos sonoros naturais (Natural sound archives)	*Arquivos sonoros* (Sound archives) BT *Arquivos* (Archives)
Autocitação (Self-citation)	
Aves (Birds)	*Aves* (Birds) BT *Vertebrados* (Vertebrates)
Bibliometria (Bibliometrics)	*Bibliometria* (Bibliometrics)
Biblioteca escolar (School library)	*Bibliotecas escolares* (School libraries)
Biblioteca universitária (University library)	*Bibliotecas universitárias* (University libraries)
Bibliotecário (Librarian)	*Bibliotecários* (Librarians) BT *Bibliotecas* (Libraries)

Table 3

Plural terms

Keyword	Unesp Thesaurus's Descriptor
Biblioteca universitária (University library)	*Bibliotecas universitárias* (University libraries)
Bibliotecário (Librarian)	*Bibliotecários* (Librarians)
Currículo (Curriculum)	*Currículos* (Curricula)
Estudo de caso (Case study)	*Estudo de casos* (Case studies)
Método (Method)	*Métodos* (Methods)

Table 4

Single terms

Keyword	Unesp Thesaurus's Descriptor
Genealogia acadêmica (Academic genealogy)	*Genealogia* (Genealogy)
Ética da informação (Information ethics)	*Ética* (Ethics)
Investigação científica (Scientific investigation)	*Investigação* (Investigation)
Mediação cultural indígena (indigenous cultural mediation)	*Mediação cultural* (Cultural mediation)
Museologia social (Social museology)	*Museologia* (Museolgy)

Table 5

Compound terms

Keyword	Unesp Thesaurus's Descriptor
Estrutura semântica de dados (Semantic data structure)	*Estrutura de dados* (*Computação*) (Data structure (Computing))
Estrutura sintática de dados (Syntatic data structure)	*Estrutura de dados* (*Computação*) (Data structure (Computing))
Recuperação e visualização da informação (Information retrieval and visualization)	*Recuperação da informação* (Information retrieval) *Visualização da informação* (Information visualization) BT *Ciência da Informação* (Information Science)

keyword forms, such as plural descriptors, simple words, or preferred descriptors aimed at improving retrieval and indexing in information systems. The findings indicate that despite the authors' expertise in Information Science, standardizing keyword attribution remains challenging due to insufficient knowledge of the indexing process during the analysis and representation phases. This challenge underscores the need for detailed guidance, which should be incorporated into a more comprehensive "Indexing tutorial using the Unesp Thesaurus" than the current version, emphasizing the importance of subject analysis in the indexing process as a determinant of successful outcomes in the representation stage with controlled vocabulary.

On the other hand, the low number of exact descriptors and the exclusive and specific keywords in the area of Information Science can be considered indicative that the thesaurus has not

Table 6

Non-preferred terms

Keyword	Unesp Thesaurus's Descriptor
Aprendizado de máquina (Machine learning)	USE *Aprendizado do computador* (Computer learning)
Competências profissionais (Professional skills)	*Competência profissional* (Professional skill) USE *Profissionalismo* (Professionalism)
Comunicação científica (Scientific communication)	USE *Comunicação na ciência* (communication in Science)
e-Saúde (e-Health)	USE *Telecomunicações na medicina* (telecommunications in medicine)
Feminicídio (Feminicide)	USE *Crimes contra as mulheres* (Crimes Against women)

followed the terminological update of the area of Information Science in the inclusion of new terms, whether as preferred or non-preferred. In addition to the continuous updating of the controlled vocabulary with new terms, it is recommended that the system provide more explanatory tutorial instructions to authors on the use of preferred descriptors and keywords.

Analysis of the results obtained highlights two improvement proposals to be carried out by repository managers:

- Inclusion of the subject analysis stage of the indexing process and guidance on assigning keywords and descriptors in an indexing tutorial with the use of controlled vocabulary to expand the scope of a tutorial only using controlled vocabulary that is more restricted to the representation of subjects; and,

- Definition of an Indexing Policy that considers the need for hybrid coexistence between natural language and vocabulary control to enable the inclusion of new terms specific to domain areas in vocabularies controlled based on the attribution of keywords by the authors.

References

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing and Management*, 43, 1705-1714. https://doi.org/10.1016/j.ipm.2007.01.015

Freitas, M. P., & Dal´Evedove, P. R. (2019). Consistência na indexação por atribuição no repositório institucional da UFSCAR. In XX Encontro Nacional de Pesquisa Em Ciência da Informação. Florianópolis: Universidade Federal de Santa Catarina. https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1203/811

Fujita, M. S. L. (2024). Analysis of the functions of Keywords assigned by authors in scientific publications in events and journals. *Digital Journal of Library and Information Science*, Campinas, SP. v.22, e024020, 2024. https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8676208/en

Fujita, M. S. L. & Panuto, J. C. (2024) Guidelines on assigning the subjects of theses and dissertations in repositories. *IFLA Journal, 50*(1), 160-9, 2024. https://doi.org/10.1177/03400352231217275

Fujita, M. S. L., & Tartarotti, R. Dal´E. (2020). Análise de palavras-chave da produção científica de pesquisadores: o autor como indexador. *Informação & Informação, 25*(3), 332 – 374. http://www.uel.br/revistas/informacao

Gil Leiva, I. (2017) SISA—automatic indexing system for scientific articles: experiments with location heuristics rules versus TF-IDF rules. *Knowledge Organization,* 44(3) 139-162. http://dx.doi.org/10.5771/0943-7444-2017-3-139

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science & Technology, 58*(8), 1175–1187. https://doi.org/10.1002/asi.20595

Golub, K.,Tyrkkö, J., Hansson, J., & Ahlström, I. (2020). Subject indexing in humanities: a comparison between a local university repository and an international bibliographic service. https://doi.org/10.1108/JD-12-2019-0231

Gonçalves, A. L. (2008). Uso de resumos e palavras-chave em Ciências Sociais: uma avaliação.*Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, *13*(26), 78-93. https://doi.org/10.5007/1518-2924.2008v13n26p78

Han, M-J. K., Harrington, P., Black, A., & Kudeki, D. (2016). Aligning author-supplied keywords for ETDS with domain-specific controlled vocabularies. In: *Classification & Indexing Satellite Conference* (pp. 1-10). http://hdl.handle.net/2142/97879

Holstrom, C. (2019). Moving Towards an Actor-Based Model for Subject Indexing. *NASKO: North American*

Symposium on Knowledge Organization 7(1), 120-128. https://doi.org/10.7152/nasko.v7i1.15631

Khatir, A., & Ganjefar, S. (2018). The analysis of the distribution and focus of keywords in theses and dissertations and compliance with descriptors, title, and abstract. Iranian Journal of Information Processing and Management, 34(1) pp.411-428. https://www.academia.edu/106472352/The_Analysis_of_the_Distribution_and_Focus_of_Keywords_in_Theses_and_Dissertations_and_Compliance_with_Descriptors_Title_and_Abstract

Kipp, M. (2009). User, author and professional indexing in context: an exploration of tagging practices on CiteULike. *Canadian Journal of Information and Library Science*, *35*(1), 1-41.

Lancaster, F. W. (2003) *Indexing and abstracting in theory and practice*. 3rd ed. Facet Publishing.

Lardera, M., & Hjørland, B. (2020). Keyword. In: Hjørland, B. & Gnolli, C. (2020) *Encyclopedia of knowledge organization*. https://www.isko.org/cyclo/keyword

Li, M. (2018). Classifying and ranking topic terms based on a novel approach: role differentiation of author keywords. *Scientometrics*, 116, 77–100. https://doi.org/10.1007/s11192-018-2741-7

Lu, W., Li, X., Zhifeng, L. & Cheng, Q. (2019). How do author-selected keywords function semantically in scientific manuscripts? *Knowledge Organization, 46*(6), 403-18 https://doi.org/10.5771/0943-7444-2019-6-402

Mathes, A. (2004). Folksonomies – cooperative classification and communication through shared metadata [Online Report]. *Journal of Computer-Mediated Communication*, 47. https://adammathes.com/academic/computer-mediated-communication/folksonomies.pdf

Maurer, M.B. & Shakeri, S. (2016). Disciplinary differences: LCSH and keyword assignment for ETDs from different disciplines. *Cataloging & Classification Quarterly, 54*(4), 213-243. https://doi.org/10.1080/01639374.2016.1141133

Névéol, A., Doğan, R. I., & Zhiyong, L. (2010). Author keywords in biomedical journal articles. AMIA 2010 Symposium Proceedings. p.537-541. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041277/

Phillips, M. E., Tarver, H., & Zavalina, O. L. (2019). Using metadata record graphs to understand controlled vocabulary and keyword usage for subject representation in the UNT theses and dissertations collection. *Cadernos bad (portugual)*, (1).

Schwing, T., McCutcheon, S., & Maurer, M. B. (2012). Uniqueness Matters: LCSH and Keywords in the Library Catalog's ETD Records. *Cataloging and Classification Quarterly, 50*(8), 903-928. https://doi.org/10.1080/01639374.2012.703164

Strader, C. R. (2009) Author-assigned keywords versus Library of Congress Subject Headings: implications for the cataloguing of electronic theses and dissertations. *Library Resources & Technical Services*, *53*(4), 243-50. https://doi.org/10.5860/lrts.53n4.243

Terra, A. L., Agustín Lacruz, C., Bernardes, Ó., Fujita, M. S. L. & Bueno De La Fuente, G. (2021). Subject-access metadata on ETD supplied by authors: A case study about keywords, titles and abstracts in a Brazilian academic repository. *Journal of Academic Librarianship*, 47, 102268. https://doi.org/10.1016/j.acalib.2020.102268

UNESP. (n.d.) Rede de Bibliotecas da Unesp. Comissão Permanente do Tesauro Unesp. *Tutorial para uso do tesauro Unesp*. Unesp.

Wolverton, R. E., Hoover, L., & Fowler, R. (2011). Subject Analysis of Theses and Dissertations: A Survey. *Technical Services Quarterly, 28*(2), 208–209. https://doi.org/10.1080/07317131.2011.546276

Yi-Fang, B.W., & Quanzhi, L. (2008). Document keyphrases as subject metadata: incorporating document key concepts in search results. *Information Retrieval*, *11*, 229–49, 2008 https://doi.org/10.1007/s10791-008-9044-1

Zhang, J., Yu, Q. Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author. keywords: a case study of patient adherence research**.** *Journal of the Information Science and Technology*, 67(4) 967–972. https://doi.org/10.1002/asi.23437