The Journal of Academic Librarianship xxx (xxxx) xxx

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

Subject-access metadata on ETD supplied by authors: A case study about keywords, titles and abstracts in a Brazilian academic repository

Ana Lúcia Terra [a,*], Carmen Agustín Lacruz [b], Óscar Bernardes [a], Mariângela Spotti Lopes Fujita [c], Gema Bueno de la Fuente [b]

[a] Porto Accounting and Business School, Polytechnic Institute of Porto, Porto, Portugal
[b] University of Zaragoza, Spain
[c] São Paulo State University - UNESP, Brazil

ARTICLE INFO

Keywords:
Subject-access metadata
ETD
Electronic theses and dissertations
Keywords
Academic repository

ABSTRACT

The keyword lists are rich in terminology and as such, they are also characterized by great semantic ambiguity, which presents the problems of synonymy and polysemy typical of uncontrolled language, which is both an advantage and a drawback. However, assigning keywords has gained particular importance in the current open scientific communication ecosystem in digital environments, mainly in the academic dissertation and repositories of theses.

This article reports a study on the organization of knowledge, using subject-access metadata of master's theses from the digital repository of the University of São Paulo (USP), Brazil, applying analytical techniques to a big amount of data. The objective was to analyze the number of keywords in each record and keywords repeated in the title and abstracts of the dissertations in Portuguese and English. The analysis of the 48,501 metadata records of master's theses submitted to the repository, between 2001 and 2019, presents a total of 223,867 keywords in Portuguese, with an average of 4.62 keywords per record, and a total of 216,521 keywords in English, equivalent to an average of 4.59 per record. Although, the attribution of keywords in Portuguese and English by the author is an economic way to expand access to the content of theses and dissertations in institutional repositories, it is necessary to define rules for authors about the choice of keywords and the preparation of the abstract, as well as its translation into English. The task of controlling keywords requires a partnership between authors and librarians who can enrich the quality of indexing languages.

Introduction

Keywords can be used to represent the topics of the documents. Many different information systems have used keywords systems to enhance access over the years. The origin of the keywords stems from the elaborate thematic indexes that historically aided users in locating the information within the books, in the scriptoria of the medieval European monasteries (Gil Leiva et al., 2013, 231). These tools have received many names over time: bibliographic indexes, subject heading lists, descriptors list, etc. With small but significant semantic differences, all these words designate different instruments that describe, express and inform the topics of the documents and therefore represent the information contained in them.

Keyword lists can be used in a post coordinated setting allowing users to freely coordinate terms at the time of assignment or in the search progress. The design of a key word list does not have to correspond to any category and grammatical form, admit spelling variants and designate concepts with very precise or imprecise meanings. As a result keyword lists are characterized by great semantic ambiguity and therefore present the problems of synonymy and polysemy typical of uncontrolled language. In sum, keyword lists can provide rich in terminology which is both an advantage; but they also present a drawback. Its vocabulary is very expressive because it evolves at the same time as the terminology used in documents and therefore it is updated quickly and easily. On the other hand they allow an important economy of human resources and great agility in the management of technical tasks (Slype, 1991, 22-30) which is why they are used by most information systems

* Corresponding author at: Porto Accounting and Business School, Polytechnic Institute of Porto, Rua Jaime Lopes, Amorim, s/n, 4465-004 S. Mamede de Infesta, Portugal.
E-mail address: anaterra@iscap.ipp.pt (A.L. Terra).

https://doi.org/10.1016/j.acalib.2020.102268
Received 17 July 2020; Received in revised form 25 September 2020; Accepted 27 September 2020

Assigning keywords has gained particular importance in the current open ecosystem of scientific communication in digital environments. Here academic repositories stand out as a kind of system whose primary mission is to keep and disseminate the scientific production of authors linked to the institution they belong. Academic or institutional repositories have become one of the main resources to preserve and maximize the impact of research carried out in higher education institutions. In these systems, it is necessary to organize knowledge with representation activities which are important for access and retrieval and also contribute to the intellectual and social organization of knowledge (Hjorland, 2003).

On the other hand, currently, in the formal structure of academic works, the title, abstract and keywords are academic paratexts (Genette, 2000), which the author of the work must write/register and put together with the text. They constitute privileged elements of the pragmatic dimension of the works, and they have an important influence on the reader. Thus, they have become the essential elements of knowledge representation, at all academic levels (undergraduate works, master's and doctoral theses).

In the process, title, abstracts and keywords have become metadata available in open access within repositories, over which information professionals have little control. Thus, the specialized techniques and tools used in the representation of knowledge have given way to other non-professional approaches that result in content generated by the users/authors of academic works. The same applies to serial publications, in which the authors are currently responsible for the elements of knowledge representation (title, abstract and keywords) of their papers when they fill in the electronic submission metadata in the journals and later in the repositories of the institutions to which they belong (Fujita et al., 2018).

These subject metadata in institutional repositories provide important access points for the search and navigation functions within collections as a whole. These metadata can be subdivided into two distinct types: controlled vocabulary metadata that extract values from formally maintained term lists and free text metadata that rely on natural language and are freely chosen by the authors (Zavalina, 2011).

This paper reports a study on the knowledge organization of master's theses from the digital repository of the University of São Paulo (USP), Brazil, using analytical techniques applied to a big amount of data. Data from the fields of master these year, keywords, title and abstract, in Portuguese and English, from 48.501 records were collected. The collected data served for a quantitative analysis of the relationship between keywords, titles and abstracts, in order to better understand the metadata provided by the author, from the point of view of subject metadata.

Literature review

The literature review on keywords analysis in subject metadata of Electronic Theses and Dissertations (ETD) prioritized the analysis of subject metadata and its standardization, in addition to the combination of keywords with controlled vocabulary descriptors for knowledge representation and information retrieval.

The analysis of metadata regarding the standardization of its format was carried out in some studies, like Park and Richard (2011) and Tarver et al. (2015). Park and Richard (2011) studied the metadata of electronic theses in Canadian academic institutional repositories and verified variations and inconsistencies in the application of metadata in Dublin Core format. The level of inconsistency and variation is particularly significant in certain elements, such as the date or the identification of the course to which the academic work is related to, but does not seem to exist in the object fields of our study (title, abstract and keywords). The analysis of subject metadata, limited to the dc:subject field values, was performed quantitatively in a study by Tarver et al. (2015) with a large dataset of more than 8 million records from the Digital Public Library of America (DPLA), a library with a metadata aggregation system. Using a "big data" approach, the variations in the subject metadata field were analyzed. The study concluded that there is a great variation in the number of instances of subject fields across records, ranging from zero subject terms to over a thousand subject terms. On the other hand, there is a very high percentage of terms with a single use, due to the lack of a common controlled vocabulary. The analysis provides a framework for general discussion on metadata subjects in digital collections in order to verify the integrity and quality of the record regarding subject metadata and its capacity of being a meaningful representation of the content.

Other works have dealt with knowledge representation and information retrieval of ETDs, addressing the advantages and disadvantages of using keywords and descriptors and carrying out comparative analyzes (Ansari, 2005; Davarpanah & Iranshahi, 2005; Sassen, 2017; Voorbij, 1998).

Voorbij (1998) conducted two studies with monographs in the humanities and social sciences, in the online catalog of the National Library of the Netherlands, in order to compare the value of subject descriptors and the keywords of titles as subject search entries. In the first study, twelve librarians made a comparison between the subject descriptors and the keywords of 475 records and the conclusions were that 37% of the records is considerably improved by the subject descriptor and 49% slightly or considerably improved. In the second study, librarians searched for topics using keywords from the title and subject descriptors on the same topic. The relative recall amounted to 48% and 86%, respectively. The results showed that keywords in the title do not always offer enough clues to information search. Therefore, the author points out that the descriptors can improve the record of a publication and can control the vocabulary in order to remove the concern of vocabulary control from the user. On the other hand, the study highlights that truncated keywords in the title retrieve many relevant results, but also many irrelevant ones and that the descriptors can counteract this deficiency. Another study points to similar conclusions. Davarpanah and Iranshahi (2005) studied the effectiveness of the keywords in the title and the descriptors assigned by the authors to represent the theses of different subject areas indexed in the database of Iranian theses. For them, the title of a work is the main element to attract the reader's attention because it provides a general indication of what the document is about, although sometimes, in some disciplines, literary titles are created that do not inform the content. The findings of the study establish that retrieval by subject descriptors offers better results than searching by title keywords. The combination of title keywords and controlled descriptor index terms is a powerful tool to indicate what a publication is about. A title can be useful, but at the same time it can be improved by a descriptor. Ansari (2005) examined the degree of exact and partial coincidence between the keywords of the title and the descriptors attributed to medical theses in Farsi indexed at the Central Library of the University of Medical Sciences in Iran. In the comparison over time, it was observed that the number of exact matches increased, indicating that authors have become more attentive in choosing the title. The author notes that there are keywords in the title of the thesis with a high value of information that were not included in the descriptors and recommends considering these keywords and inserting them as indexing descriptors.

Sassen (2017) carried out a study on practices of cataloging dissertations of academic libraries of the Association of Research Libraries in order to discern how libraries provide access to subjects, as well as to the names of academic departments and advisors. An analysis of catalog records revealed that this information is recorded more often on uncontrolled notes and access points than on authorized access points. In nearly 45% of the catalogs, uncontrolled-subject-access points are used and only 25% of the catalogs use LCSH to represent the subject of ETDs. Undoubtedly, the cataloging of ETDs can be completed more quickly if notes or uncontrolled access points are used to register names and subjects. Although these practices reflect a movement toward cataloging efficiency, they must be considered in the context of ETDs discovery.

Besides improving the retrieval of theses and dissertations by

optimizing access points, it is also important to develop practices that take advantage of the keywords and metadata provided by the authors. Strader (2009), Zavalina (2014), Han et al. (2016) and Maurer and Shakeri (2016) dealt with this topic. These authors consider that libraries now face metadata created by noncataloguers, who often use subject terms not available in established controlled vocabularies. They therefore insist that keywords are better aligned with established discipline-specific controlled vocabularies.

Strader (2009) focuses her study at Ohio State University and the conclusions follow previous works that consider that both keywords and controlled vocabularies complement each other, and both show high concordance with the significant words in the title of the papers. Exact and partial matches were counted, as well as singular and plural differences and other variants that could affect the user's search results. Zavalina's works on the mediations between the access needs of users and the metadata in digital collections are very interesting. Zavalina (2014) deals with the complementarity between free text subject metadata and descriptors of a controlled vocabulary in three large-scale digital libraries that aggregate digital collections of cultural heritage. The results of this study empirically demonstrate that the inclusion of information on subjects in free texts and with controlled vocabulary is a common practice among some of the large-scale digital libraries. More detailed collection-level metadata records, including free-text subject metadata and controlled vocabulary, enable a more complete representation of the intellectual content of information objects and ultimately improve access to subjects.

Han et al. (2016) collected 32,696 keywords from 5365 master's and doctoral theses submitted to the University of Illinois at Urbana-Champaign's institutional repository between 2010 and 2014. The authors suggest ways to improve ETD subject metadata as libraries move toward linked open data and semantic web and metadata reconciliation work is required. Based on the analysis of keywords provided by the author, the study shows that domain-specific controlled vocabularies have unique terms that are not available in LCSH that could be useful in aligning additional keywords if remediation or reconciliation work is considered.

Maurer and Shakeri's (2016) research, carried out in the Kent State University Library catalog, refers to the frequency of the attribution of keywords provided by the author and subject headings provided by the cataloger for records of electronic theses and dissertations (ETD) in different disciplines. The results show that, on average, more keywords assigned by the author and more LCSH subject headings assigned by the cataloger were ascribed to works in the arts and humanities than to works in the social sciences and sciences, technology, engineering and mathematics (STEM). The STEM disciplines, in particular, received a lower amount of topical metadata, in part due to under-assignment of metadata related to names, geographic locations and corporate entities. The authors also comment on the problems that usually occur when accessing keywords and consider that:

> "Today keyword access is the de facto search mechanism for most library catalogs, although often automation vendors provide unique and sometimes proprietary indexing and relevance routines. Regardless, catalogers recognize that their work within the library catalog entails optimizing the bibliographic record for keyword searches by the user." Maurer & Shakeri (2016, p. 217)

Despite the importance of keywords for information access, and for ETDs access, Maurer and Shakeri (2016) stress that there is little research on the number of keywords provided by student-authors and on the differences of author-assigned keywords for ETD in different disciplines.

The combination of keywords and descriptors for knowledge representation and organization is a discussion in the current literature and reveals a trend whose advantages and disadvantages are influential in decisions regarding information retrieval of theses and dissertations. Standardization of metadata is also critical with respect to metadata collection systems. Analysis of the literature reveals that up to now, little research has been conducted to specifically assess subject metadata in digital repositories of theses and dissertations.

Objectives and study context

This paper aims to characterize the organization of knowledge through the analysis of subject metadata contained in keywords, as well as in titles and abstracts of master's theses available in the digital repository of the University of São Paulo (USP), Brazil.

This general objective was achieved through the following specific goals:

a) Analysis of the number of keywords in each record, in Portuguese and in English.
b) Analysis of repeated keywords in the title of master's theses, in Portuguese and English.
c) Analysis of repeated keywords in the abstracts of master's theses, in Portuguese and English.

Based on the study findings, we sought to provide some insights in order to improve the quality of the subject metadata provided by master's theses authors to be included in academic repositories and to better the overall repository.

To this end, a survey was made in some fields of the metadata in the records of all master's dissertation theses submitted to the repository of the University of São Paulo (USP), in the fields of the year of the dissertation, keywords, title and abstract, in Portuguese and English.

The Digital Library of Theses and Dissertations of the University of São Paulo (https://teses.usp.br) was created to make available on the Internet the knowledge produced by academic works submitted to the University of São Paulo, allowing the Brazilian and international communities to have access to complete digital version of theses and dissertations. The Digital Library was launched in 2001 along with the Portal do Conhecimento. For the authors of theses and dissertations, the Digital Library is a unique opportunity to spread out their works, in a quick and easy manner. This will foster professional growth in national and international context. The same opportunity will be given to research advisors and postgraduate courses, which will have a significant increase in the impact of their research, both in Brazil and anywhere in the world with available Internet access. The Digital Library is associated with a global initiative recognized by UNESCO, the Networked Digital Library of Theses and Dissertations (NDLTD), which guarantees greater reliability and coverage, and also associated with the Brazilian Institute of Information on Science and Technology (Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT) of the Ministry of Science and Technology, through the Brazilian Digital Library of Theses and Dissertations. In April 2019, according to data available on the website, it had a total of 87,098 documents including PhD thesis (35,426), habilitation thesis (662) and master's dissertations (51,010).

Methodology

Academic libraries are great at stockpiling knowledge; every year thousands of students, professors and researchers publish new data in institutional repositories. A huge number of documents and specific metadata are made available allowing new analysis approaches. For this paper, we try to contribute to this new analysis approach collecting a large amount of metadata from a Brazilian academic repository.

To this end, a survey was made of some fields of the metadata of the records of all master's theses submitted to USP repository, between 2001 and 2019. Data were collected from 48,501 records, regarding the fields of the year of dissertation, keywords, title and abstract, in Portuguese and English. The temporal distribution of the records is shown in Table 1.

Table 1
Temporal distribution of master's theses records.

Year	Number of records
2001	105
2002	345
2003	395
2004	642
2005	798
2006	1485
2007	2808
2008	2633
2009	3268
2010	3396
2011	3032
2012	3284
2013	3921
2014	3778
2015	4354
2016	4198
2017	4019
2018	4683
2019	1200
Blank	157
Total	48,501

As we process all metadata fields of 48.501 records, representing 95% of all master's dissertations available in the institutional repository of the University of São Paulo in May 2019, over 1 million records were stored, e.g.: dates, keywords, subjects, authors, e-mail, DOI, statistics, etc. However, in this paper we will work only with data from titles, keywords and abstracts fields.

The treatment and analysis of very large amounts of data in order to make it useful is called big data (Lozada et al., 2019). The technologies now offer a plethora of opportunities to leverage and optimize the big data we acquire and collect from libraries (Jantti, 2016).

In order to make it more useful, first we need to crawl the data from the web or internal databases, using web-scraping techniques.

Web-scraping focuses on the transformation of unstructured data into machine-readable indexes or semantic information, making it ready to be used as big data, for auditing, immerse search analysis and better decision making. This technique can be divided into four main processes, according to Slamet et al. (2018): 1st- creating a scrapping template; 2nd- Exploring site navigation; 3rd- Automating navigation and extraction: from the processes number 1 and 2; and 4th- extracting data and package history: the information acquired from process number 3 is saved in tables and database.

For the propose of this research, CG enterprise web-scraping software was used, which was designed by Sequentum for organizations with a critical reliance on structured data and it includes sophisticated features for monitoring success criteria of data extraction, legal compliance and production fail-over that are not available in other solutions. There are other solutions, exposed by Olmedilla et al. (2016) like Python, Java, Ruby and PHP. We chose CG enterprise because it is more dynamic, fast to implement and has strong error handling features. Our scraping cycle went through four stages. First, we designed an agent to track the library data, establishing which information we would like to extract from the USP Digital Library, thus defining the validation rules for each data. Second, we did some debugs to test how the agent (spider) is performing the crawl, and if all data is structured on the correct fields. After little adjustments on XPath and C# code language, we run the spider to scrap automatically the entire library to extract only thesis simulating a human behavior, e.g.: scrolls, mouse click, etc. XPath stands for XML Path Language and uses "path like" syntax to identify and navigate through nodes in an XML document, the method is based on a tree, and provides the ability to navigate around the tree, selecting nodes through a variety of criteria. This is very important and relevant for libraries in order to understand how their data is being treated by other websites, e.g.: information aggregators, repositories, fraud software, etc.

Conclusively, we stored the final data into a MySQL database because it is efficient, ubiquitous and has an open source engine available for all major platforms.

Results

The 48,501 master's theses records available in the USP repository present a total of 223,867 keywords in Portuguese, with an average of 4.62 keywords per record and a standard deviation of 1.77.

Analyzing the distribution of the number of Portuguese keywords per record (Table 2), it appears that more than half have four (24.33%) or five keywords (27.63%). Records with three keywords represent 18.43% of the repository and those with six keywords reach 12.35%. The remaining keywords groupings are insignificant, ranging from 5.47% for registrations with seven keywords and 3.19% for those with only one keyword. Even though it has a very residual value of 0.28%, it is important to refer to the blank records because they correspond to an incorrect deposit, with different practices ranging from the effective omission of keywords, to the inclusion of keywords in Portuguese in the field of English keywords. In addition, when the keywords are in a language other than Portuguese, for example in Spanish, they were not considered in our information data collection because they were in a field with another name. Another residual value is that of records with more than eight keywords, as for example, one record with 48 keywords and another with 32. Although they are unique cases, they show that the repository does not seem to define a maximum number of keywords, which should perhaps be corrected.

In turn, the total number of keywords in English is 216,521, equivalent to an average of 4.59 per record, with a standard deviation of 1.81.

With regard to the distribution of the number of keywords per record, it appears that, as expected, the values are very similar to those found in the Portuguese keywords. Thus, records with five (27.52%) and four keywords (23.91%) predominate, followed by those with three keywords (17.34%) and six (11.77%).

Since the values are not exactly the same as for the keywords in Portuguese, this means that the field of keywords in English is not, in some cases, a literal translation of the first, and there is no strict equivalence between both fields as it would be advisable. This situation is more evident in the case of blank records, which represent 2.87% of the total of English records, while in Portuguese keywords they were only 0.28%. In total, there are only 104 records without keywords in either Portuguese or English, which corresponds to a value without significant expression in the total set of 48,501 records.

Regarding the number of Portuguese keywords repeated in the title, it appears that the redundancy between the two fields is not very significant, with a standard deviation of 1.12. Indeed, in 27.01% of the records there is no repetition of keywords in the title and in 33.01% there is a coincidence in only one keyword and the words in the title. Two keywords in the title occur in 24.93% of the records while the repetition of three keywords in the title corresponds to 10.88% of the Portuguese records. A residual value of 3.18% refers to cases in which four keywords occur in the title. Some records have a blank value (0.28%), which is due to the fact that the titles appear in a language other than Portuguese, for example Spanish, or because even if they are in Portuguese, the words of the title are in italics or bold and in that situation they are not counted.

The English keywords included in the title present values very close to those recorded in Portuguese with an almost equal standard deviation of 1.11. In fact, there are 25.36% of records without repetition of keywords in the title and 33.54% with repetition of one keyword. Moreover, there are 23.51% of the records with two keywords in the title and only 10.16% with three keywords. Titles with four keywords also correspond to a residual value of 3.02%. The most significant difference regarding the repetition of keywords in Portuguese in the title is in the percentage of blank records (3.74%). This difference stems largely from the fact that the records do not have the title in its English version, and they have

Table 2
Number of keywords in each record (PT & EN).

	0	1	2	3	4	5	6	7	8	More than 8
PT	136	1546	1433	8937	11,799	13,403	5989	2655	1306	1297
	0,28%	3,19%	2,95%	18,43%	24,33%	27,63%	12,35%	5,47%	2,69%	2,67%
EN	1393	2156	1012	8411	11,595	13,347	5707	2422	1199	1259
	2,87%	4,45%	2,09%	17,34%	23,91%	27,52%	11,77%	4,99%	2,47%	2,60%

italic or bold words that are not counted in our metadata survey (Table 3).

The number of keywords repeated in the abstract is an indicator of the consistency between the keywords and the ideas conveyed in the text that aims to summarize the most relevant aspects of the document to which it relates. In addition, the abstract serves to develop, frame and specify the meaning of the keywords provided. In this sense, it is expected that keywords will be repeated in the abstract. Thus, in the Portuguese records, it appears that almost a quarter (24.73%) of the abstracts include two keywords and 22.39% three keywords. The relationship between keywords and the abstract does not seem to be very significant, an idea reinforced by the fact that 18.05% of the abstracts only repeat one of the keywords used. Conversely, only 14.17% of abstracts include four of the keywords. Bearing in mind that the average number of keywords per record is 4.62, these results seem to show a slight complementarity between the field of keywords and that of abstracts. Five keywords repetition in the abstract field occurs only in 6.60% of the cases and six or more keywords are included only in 2.50% of the abstracts.

The repetition of English keywords in the English abstract presents values that follow what happens in Portuguese records. In fact, the two highest percentages correspond to two (23.99%) and three (22.05%) keywords repeated in the abstract, as in the case of the Portuguese records. The abstracts in which no keyword appears (11.46%) and in which only one appears (17.07%) correspond to more than a quarter of the records. The inclusion of four keywords in the abstract (13.83%) or five (6.25%) represents only one fifth of the records. In English records, a higher number of blank records appears due to the fact that in some cases there are no keywords in English, there is no abstract in English or because the words are in bold or italics and are not counted (Table 4).

Discussion

Comparing the results obtained with other studies on the analysis of subject metadata provided by the authors of academic works, it appears that the research herein stands out for the amount of records analyzed. In fact, data were collected from 48,501 master's dissertations, representing 223,867 keywords in Portuguese and 216,521 keywords in English. This research also distinguished itself because it focuses on a uniform typology of academic work, the master's thesis, and covers an extended period of time, of almost 19 complete years (from 2001 to mid-2019). Other works have dealt with a much smaller number of data. Focusing on the keywords of the titles and the descriptors assigned to the records of the theses from an Iranian database, Davarpanah and Iranshahi (2005) selected a sample of 600 theses, extracting 5669 keywords from the title field. Strader (2009) collected data from 285 theses and dissertations, from the catalog of Ohio State University (USA) between June and October 2005, and with keywords assigned by the authors, in a total of 1681. Han et al. (2016) used a corpus of 32,696 keywords relating to 5365 doctoral theses (3270) and master's dissertations (2095), submitted to the University of Illinois at UrbannaChampaign (USA) repository, between 2010 and 2014. Maurer and Shakeri (2016) looked at the 1255 records of theses and dissertations existing in the Kent State University (USA) repository, having collected 6595 keywords. Out of the scope of academic work repositories but focusing on the analysis of a large volume of records, specifically in relation to subject metadata, although not limited to keywords produced by users, Tarver et al. (2015) worked with 8,012,390 records from the Digital Public Library of America. This brief review shows that the studies already published mainly focus on the North American context. In this sense, the this work also presents itself as a pioneer with respect to the Brazilian context.

When comparing the number of keywords in Portuguese and in English, the different number of keywords in the two languages stands out. Remembering that 223,867 keywords were collected in Portuguese and 216,521 in English, there are 7346 fewer keywords in English than in Portuguese. This happens because there are records in which not all Portuguese keywords are translated into English or in which there is no keyword translated, with only the indication "not available". It should also be noted that the order in which the keywords are presented in Portuguese and in English is not parallel. There are also situations in which the records have a greater number of keywords in English than in Portuguese. As there are no studies comparing the coexistence of keywords in a language and their translation into English, we cannot draw any comparative analysis. However, in this specific repository, there seems to be a need to improve the coherence between keywords in both languages, in order to create an equivalent representation of the document's content. It also appears that this will be a topic of analysis that can be explored in other academic repositories.

As mentioned, each record has an average of 4.62 keywords in Portuguese and 4.59 keywords in English. When considering the number of keywords allocated to the description of the subject of each dissertation, we should take into account that, theoretically, this kind of academic work addresses new research themes. This can be a challenge in terms of using controlled vocabularies, which have more difficulty in the rapid incorporation of new terms. On the other hand, the inclusion of each additional keyword can mean an improvement in the visibility of the work, which will be more likely to be recovered in searches (Lubas, 2009). However, this can also result in a dispersion problem, as in the case studied by Han et al. (2016, p. 3) in which "most of the keywords have only one associated thesis", which makes it difficult to locate works on the same subject.

The average of keywords per record in the USP repository differs slightly from other similar studies. Indeed, Strader's research (2009) counted 5.9 keywords per record, without distinguishing master's dissertations from doctoral theses. Han et al. (2016) found that, specifically for master's theses, the average number of keywords was five for each record, while for doctoral theses it was six. Maurer and Shakeri (2016)

Table 3
Distribution of the number of keywords repeated in the title (PT & EN).

	0	1	2	3	4	5	6	7	9	Blank
PT	13,099	16,011	12,090	5277	1543	297	39	7	2	136
	27,01%	33,01%	24,93%	10,88%	3,18%	0,61%	0,08%	0,01%	0,00%	0,28%
EN	12,299	16,265	11,402	4927	1466	287	36	3	–	1816
	25,36%	33,54%	23,51%	10,16%	3,02%	0,59%	0,07%	0,01%	–	3,74%

Table 4
Distribution of the number of keywords repeated in the abstract (PT & EN).

	0	1	2	3	4	5	More than 5	Blank
PT	5464	8753	11,996	10,858	6872	3200	1211	147
	11,27%	18,05%	24,73%	22,39%	14,17%	6,60%	2,50%	0,30%
EN	5558	8278	11,635	10,694	6708	3029	1193	1406
	11,46%	17,07%	23,99%	22,05%	13,83%	6,25%	2,46%	2,90%

obtained an average of 5.3 keywords per record, with 9.5% that did not have any keywords.

In our study, records without keywords or with only one have no significant percentages. In fact, there are only 0.28% of the records without keywords in Portuguese and 3.19% with only one. In the case of English records, the values are slightly higher, with 2.87% without keywords and 4.45% with only one, but in the latter case, records without keywords in English are also counted because they present the formula "not available". To better contextualize these results, we can relate them to the data collected by Sassen (2017), who, analyzing the cataloging practices of theses and dissertations of 114 affiliated libraries of the ARL (Association of Research Libraries), found that the identification of subjects occurred in 87% of the records catalog. In the Digital Public Library of America, Tarver et al. (2015) found 22.8% of records without subject identification, which represents almost a quarter of the total. In this sense, we think that, in USP repository, the fact that it is up to the authors to provide the subject metadata significantly increases the number of records with identification of the subject. Another aspect to be highlighted is the great variation in the number of keywords between records, with those that have none and those that have 48 (maximum number in Portuguese records) or 53 (maximum number in English records). Thus, a reality noted by Tarver et al. (2015, p. 37) in the Digital Public Library of America, who found that "one noticeable finding is the high variability of the number of instances of subject fields across records, ranging from no subjects to more than one thousand" which could be explained by several reasons such as "(…) may be due to workflow issues, a lack of tools to discover incomplete records or resources to fix known deficits, or even local practices that do not require or encourage subject representation". In the Brazilian repository under study, the reasons for this disparity also need to be investigated, but these may be some of the topics pointed out by Tarver et al. (2015).

Analyzing the tripartite relationship between keywords, titles and abstracts is the more innovative approach of our research. On the one hand, because most research studies on the subject-related metadata provided by the authors in thesis or dissertation repositories seek to establish a relationship between the keywords provided by the authors of the academic works and controlled vocabularies, such as subject headings, with a focus on the use of LCSH (Han et al., 2016; Maurer & Shakeri, 2016; Schwing et al., 2012; Strader, 2009). On the other hand, the tripartite relationship between keywords, titles and abstracts has been little explored in the available literature, except for the work of Strader (2009), who conceived a six-level correspondence hierarchy between these metadata and LCSH.

In USP repository, we found that in 27.29% of the records there is no repetition of keywords in the title in Portuguese, a value that rises slightly to 29.10% in the case of English records. These values are much lower than those recorded by Strader (2009), where it appears that 43.78% of the keywords have no occurrence in the title. However, as the author pointed out, in Ohio State University repository, where the study was focused, authors were discouraged from using words from the titles as keywords. Schwing et al. (2012) collected very similar values, noting that 43.34% of the keywords provided by the authors of theses and dissertations from Kent State University appeared in the titles of the documents, a value that rises to 52.90% if variants of the same word are considered. According to these authors, "(…) the author-supplied keywords do add uniqueness and therefore increase the discoverability of their respective ETDs when compared to terms in titles" (Schwing et al., 2012, p. 920).

Most published work focuses on the relationship between the words in the title and the descriptors of the controlled vocabulary used to represent the subject of the documents. Thus, we will not be able to make a direct comparison with the results we obtained, since we related the keywords to the words in the title. However, the analysis of this relationship can also provide us with analytical clues for our own approach.

Engelson (2013) found that, in religion, theology and biblical studies, only 42% of the titles of the works in the used sample present keywords that could be accurately used to represent the subject. Wang (2006) used the titles of works in Computer Science to show that the title may give rise to new entries in a thesaurus for the area, stressing, however, that this is only feasible if the titles explicitly reflect the content of the document, which may not be the rule in all scientific areas. For the Iranian context, there are two works that examine the coincidence between the words in the titles and the descriptors used to represent the content of theses and dissertations (Ansari, 2005; Davarpanah & Iranshahi, 2005). In one case, with 45% of exact match between words in the title and the descriptors, this percentage rises to 67% if we consider cases in which there is similarity between the words in the titles and the descriptors. It is also noted that there is an average of seven words per title and that the greater the number of words in the title, the greater the number of descriptors, which also increases the level of coincidence (Davarpanah & Iranshahi, 2005). The results obtained by Ansari (2005) are in line with these, with a 70.3% of coincidence between words in the title and descriptors of doctoral theses in different fields of medicine. However, unlike the previous study, the author did not find an increase in the level of coincidence between the words in the title and the descriptors, in longer titles. Another interesting aspect of this research stems from the fact that there is an increase in the degree of coincidence between words in the title and descriptors over time.

The data collected in USP repository show that there seems to be a higher degree of coincidence than in the referred studies, since we have more than 70% of Portuguese records where the same words occur in the title field and in the keywords field. Even so, the degree of redundancy does not appear to be very high because only 14.44% of the records have three or four words in the title included in the keyword field. Note that we are referring to repetitions of exactly the same words and not variants of the same word. In English records, we found that there are some differences from Portuguese records, showing problems resulting from the lack of uniformity in filling in the fields in the repository and not so much due to aspects inherent specifically to each of the languages.

With regard to the overlapping of words between the keywords and the abstract fields, Strader's work (2009) identifies an exact match of 54.61% of the keywords with words from the abstract, with 10.59% of the keywords not appearing in the abstract. The data from USP repository show a very similar percentage with regard to records where there is no coincidence of words between the two fields with values of 11.27%, for the Portuguese records, and 11.46% for the English records.

The analysis of the occurrence of keywords in the abstract field also seems to be one of the innovative aspects of this research because few works dedicated to this approach were found. However, the complementarity between keywords and abstracts is a significant element for the representation and organization of knowledge, as well as for research and information retrieval. In this sense, there seems to be a need to deepen this approach, in order to optimize the dynamics

between the two fields, providing the user with more substantive clues to know the content of the documents.

Conclusions

The study herein has made it available to know, using a big amount of data and analytical techniques, how the authors of the academic works preserved in the repository of the University of Sao Paulo in Brazil participate in the organization of knowledge.

Assigning the authors of academic works the duty to provide keywords and abstracts in Portuguese and in English represents a more economical way of extending the alternatives of accessing the content of the theses and dissertations in institutional repositories. However, in order to optimize the potential of this solution, it is necessary to define rules for the authors, with regard to the choice of keywords and the preparation of the abstract, as well as its translation into English. Additionally, these products need to be verified and validated by professionals, in order to guarantee the quality of the repository metadata.

It is important that this task of monitoring and controlling keywords is considered a partnership between authors and librarians, in a context in which libraries are redefining their functions within academic information systems, and their professionals have to be flexible and adaptive to new environments. These collaborative dynamics bring multiple benefits: both for authors, whose works gain greater visibility, by improving the quality of their indexing, and for librarians, who find in authors a source of authority to contrast and enrich the quality of indexing languages. Keywords from theses and dissertations present updated vocabulary from the scientific areas, that could be used to enrich the controlled vocabularies.

Given the knowledge of a specific repository resulting from our research, we can make some suggestions to improve the creation of metadata by the authors of academic works, in addition to aspects inherent to the parameterization of the repository itself, with regard to the choice of keywords and the preparation of abstracts:

1) Include the same number of keywords in Portuguese and in English.
2) Follow the same order of presentation of the keywords in Portuguese in their translation into English.
3) Define the minimum and maximum number of words, in the abstract in Portuguese and in its translation into English, emphasizing the need for equivalence in terms of extension of the abstract in both languages.
4) Make controlled vocabulary available as a user consultation tool, in scrowldown mode, for defining keywords without the need for typing.
5) Include a warning about the advantages of consistency between words in the title, abstract and keywords.
6) Include subject analysis procedures for self-archiving in thesis and dissertation repositories.
7) Incorporate within the information systems, resources, lists of frequently asked questions, or tutorials and other materials that provide the basic skills to perform the tasks of semantic description of resources.

With regard to future work, we believe that the analysis of the variations between scientific areas and over time, especially with regard to the coincidence between the keywords, the words of the title and the abstracts constitutes a research line that deserves to be deepened. The use of a large quantity of data collected from institutional repositories can offer new insights to the research on knowledge organization of academic works and thus improve their access and visibility.

Declaration of competing interest

None.

References

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review, 54*(7), 410–414.

Davarpanah, M. R., & Iranshahi, M. (2005). A comparison of assigned descriptors and title keywords of dissertations in the Iranian dissertation database. *Library Review, 54*(6), 375–384.

Engelson, L. (2013). Correlations between title keywords and LCSH terms and their implication for fast-track cataloging. *Cataloging and Classification Quarterly, 51*(6), 697–727.

Fujita, M. S. L., Agustín-Lacruz, M.-D.-C., & Terra, A. L. (2018). Journals' guidelines about title, abstract and keywords: An overview of information science and communication science areas. *European Science Editing, 44*(November), 76–79.

Genette, G. (2000). *Palimpsestes: la littérature au second degré*. Paris: Éditions du Seuil.

Gil Leiva, I., Fujita, M. S. L., & Díaz-Ortuño, P. (2013). Elaboración de índices para libros: perspectivas de actuación y formación profesional en España y Brasil. In F. Ribeiro, & M. E. Cerveira (Eds.), *Informação e/ou Conhecimento: as duas faces de Jano* (pp. 230–244). Faculdade de Letras da Universidade do Porto - CETAC.MEDIA.

Han, M.-J. K., Harrington, P., Black, A., & Kudeki, D. (2016). Aligning authorsupplied keywords for ETDs with domain-specific controlled vocabularies. In *Classification & Indexing Satellite Conference*.

Hjorland, B. (2003). Fundamentals of knowledge organization. *Knowledge Organization, 30*(2), 87–111.

Jantti, M. (2016). Libraries and big data: A new view on impact and affect. In J. Atkinson (Ed.), *Quality and the academic library: Reviewing, assessing and enhancing service provision* (pp. 267–273). Oxford: Chandos Publishing.

Lozada, N., Arias-Pérez, J., & Perdomo-Charry, G. (2019). Big data analytics capability and co-innovation: An empirical study. *Heliyon, 5*(10), Article e02541, 1–7.

Lubas, R. L. (2009). Defining best practices in electronic thesis and dissertation metadata. *Journal of Library Metadata, 9*(3–4), 252–263.

Maurer, M. B., & Shakeri, S. (2016). Disciplinary differences : LCSH and keyword assignment for ETDs from different disciplines. *Cataloging & Classification Quarterly, 54*(4), 213–243.

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2016). Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards and Interfaces, 46*, 79–87.

Park, E. G., & Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *The Eletronic Library, 29*(3), 394–407.

Sassen, C. (2017). Enhancing bibliographic access to dissertations. *Technical Services Quarterly, 34*(1), 40–53.

Schwing, T., McCutcheon, S., & Maurer, M. B. (2012). Uniqueness matters: Authorsupplied keywords and LCSH in the library catalog. *Cataloging and Classification Quarterly, 50*(8), 903–928.

Slamet, C., Andrian, R., Maylawati, D. S., Suhendar Darmalaksana, W., & Ramdhani, M. A. (2018). Web scraping and naïve bayes classification for job search engine. *IOP Conference Series: Materials Science and Engineering, 288*(1), 1–7.

Slype, G.v. (1991). Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. Retrieved from http://softwaredocumental.org/repositorio/Texto-completo/1991-vanSlype,Hipola,MoyaAnegon-Loslenguajesdeindizacionconcepcion,construccionyutilizacionenlossistemasdocumentales.pdf.

Strader, C. R. (2009). Author-assigned keywords versus Library of Congress subject headings: Implications for the cataloging of electronic theses and dissertations. *Library Resources & Technical Services, 53*(4).

Tarver, H., Phillips, M., Zavalina, O., & Kizhakkethil, P. (2015). An exploratory analysis of subject metadata in the digital public library of America. In M. C. Malta, & S. A. B. G. Vidotti (Eds.), *Proceedings of the international conference on Dublin Core and metadata applications* (pp. 30–40). Dublin Core Metadata Initiative.

Voorbij, H. J. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation, 54*(4), 466–476.

Wang, J. (2006). Automatic thesaurus development: Term extraction from title metadata. *Journal of the American Society for Information Science and Technology, 57*(7), 907–920.

Zavalina, O. L. (2011). Free-text collection-level subject metadata in large-scale digital libraries: A comparative content analysis. In *International conference on Dublin Core and metadata applications* (pp. 147–157). Retrieved from https://dcpapers.dublincore.org/pubs/article/view/3630/1856.

Zavalina, O. L. (2014). Complementarity in subject metadata in large-scale digital libraries: A comparative analysis. *Cataloging & Classification Quarterly, 52*(1), 77–89.