How Do Authors Choose Keywords for Their Theses and Dissertations in Repositories of University Libraries? An Introspection-Based Enquiry

Mariângela Spotti Lopes Fujita, Roberta Cristina Dal'Evedove Tartarotti, Paula Regina Dal´Evedove, and Maria Carolina Andrade e Cruz

Considering the importance of subject retrieval for scientific visibility, and the need to guide authors in self-archiving their papers in institutional repositories of university libraries, this study observed the patterns and strategies used by authors while indexing for keyword assignment. The study examined four categories of analysis: criteria for keyword assignment; use of controlled vocabulary for keyword assignment; understanding of the importance of keywords; and ordering criteria and function of assigned keywords. The study found that, while assigning keywords, authors: consider fundamental concepts for representing significant content of the text; act as domain expert indexers; and are unaware that keyword assignment is an indexing process that requires controlled vocabularies. The research suggests that institutional repositories implement a hybrid information representation and retrieval system to allow for both the representation of more specific subjects of knowledge domains, as well as controlled vocabulary indexing terms.

Introduction

Scientific communication is shaped by the characteristics of a scientific communication system, with several interdependent component actors whose objectives are related and interconnected.[1]

---

*Mariângela Spotti Lopes Fujita is Professor in the Pos Graduate Program in Information Science at São Paulo State University (UNESP), São Paulo, Brazil, and Researcher from CNPq, Level 1B, e-mail: mariangela.fujita@unesp.br; Roberta Cristina Dal´Evedove Tartarotti is PhD in Information Science and librarian of the State University of Campinas (UNICAMP), Brazil, e-mail: roberta_tartarotti@yahoo.com.br; Paula Regina Dal´Evedove is Professor in the Pos Graduate Program in Information Science at Federal University of São Carlos (UFSCar), and Professor in the Pos Graduate Program in Information Science at São Paulo State University (UNESP), São Paulo, Brazil, e-mail: dalevedove@ufscar.br; Maria Carolina Andrade e Cruz is a PhD in Information Science and is a Librarian at São Paulo State University (UNESP), São Paulo, Brazil, e-mail: maria.andrade@unesp.br.

Institutional repositories of university libraries store collections of digital objects and provide basic deposit and retrieval methods, and in many cases provide additional features such as security, and a protocol for remote and distributed access. Institutional repositories are intended for assembling and storing all the intellectual output of a given institution, or consortium of institutions, such as universities for long-term preservation, access, and distribution.[2] Within these context, the librarian acts as an information specialist who focuses on the primary, interdependent elements of the scientific-academic communication system in two main aspects: the encouragement of research production by the university faculty, and the means of communication available for sharing these results.

This information professional specialist, by developing information representation activities with different discourse domains, uses knowledge organization processes and systems with the specific goal of producing descriptive or thematic metadata. This metadata has precise and specific value in a future and likely retrieval in information systems' search interfaces (web, databases, institutional repositories, online catalogs, search engine websites, etc.). In this way, information resources are identified, described, organized, and communicated to serve specific purposes.

Traditionally, in university libraries, the professional information specialist librarian performs this representation; however, scientific publications in born-digital format—such as journal articles, proceeding articles, theses and dissertations—are increasingly published through interactive software which enables the submission and representation of the original papers by the author. In the submission process of institutional repositories of university libraries, considered self-archiving, the author of the original paper produces the descriptive and thematic metadata that represent the material and content description. However, when submitting to the institutional repository, and while filling out the descriptive and thematic metadata, authors are unaware of the processes and knowledge organization systems. In addition, authors are not guided to efficiently complete this task, nor to understand the implicit objectives of the results to be obtained in information retrieval search systems to facilitate the publication citation.

When filling in keyword metadata within the institutional repository—to represent the content of their original papers—authors become indexers. The keyword assignment for representation, even in natural language, carries the meaning of the content. The authors' process of keyword assignment is loaded with subjectivity as it depends on their individual cognition or knowledge which, without guidance on the purpose of the representation process, will not perform the integration between individual and social levels.

Authors are experts on their thesis or dissertation topic, and knowledgeable in the discourse domain, whose main goal is to communicate and disseminate their knowledge; however, they are not information professionals. Thus, the main question guiding this paper is: how do thesis and dissertation authors choose keywords in institutional repositories of university libraries?

Efficient retrieval, by Topic/Subject, in an academic library's institutional repository facilitates the visibility of the that university's research. Thus, it would be wise to provide guidance for authors self-archiving their work in such information retrieval systems. The objective of this investigation was to observe strategies used by researchers, particularly authors of theses and dissertations, while indexing for keyword assignment in self-archiving institutional repositories at Brazilian universities. To this end, a theoretical and methodological study was carried out on the

introspective and retrospective observation of patterns and strategies in keyword assignment by authors as indexers. The observation of introspection and retrospection of thesis and dissertation authors as domain expert indexers was performed using the Individual Verbal Protocol (IVP) technique to analyze the mental process of keyword assignment and, through qualitative analysis of these authors' cognition, report criteria for assignment, use of controlled vocabulary, understanding about the importance of the keyword and criteria for keyword ordering and function.

Literature Review

The concept of a keyword is linked to natural language (i.e. without vocabulary control) and to the publication's author, who freely assigns it. Natural language is how the keyword is configured, according to the ANSI/NISO standard (American National Standards Institute; National Information Standards Organization, 2005).[3]

In keyword assignment, the author provides the main ideas of their work and chooses keywords without consideration for controlled vocabulary.[4,5,6] Authors select the keywords they believe best represent the content of their writing,[7] and often select with care.[8]

Keywords are not exclusively assigned by authors; publishers and/or machine algorithms may also assign keywords. Zhang et al. conducted research on comparative assessment between Author Keywords and Keywords Plus, extracted from the titles of references cited by Thomson Reuters, whose results revealed similar search trends.[9]

The free choice mode of keywords refers to the lack of a more systematized process. With free choice, keywords are subject to each author's individual judgement. This is different from the subject indexing process practiced by professional indexers, in which the identification of indexing terms is performed by subject analysis of the textual content and then represented by terms from a controlled vocabulary. Essentially, keywords and indexing terms result from different processes: keywords can be extracted from any part of the document without having a vocabulary control to be applied, whereas with indexing terms, "the term is the result of complex mental activities, which involve the processes of conceptual analysis (identification of document subjects) and translation (conversion of the conceptual analysis into a given set of terms)."[10] The main factor is that authors select keywords that represent what they consider important to describe the content of their own article, and indexers, in contrast, consider the article in the larger scope of the collection.[11] Névéol et al. agree that the significant differences between keyword assignment and subject indexing are due to the fact that, "authors are asked to choose a small number of keywords, without reference to a controlled vocabulary; whereas indexers are trained to select indexing terms according to a specific protocol."[12]

More than one actor—the author and the indexing librarian—performs the indexing of scientific publications, journal articles, theses, and dissertations undertaken in digital libraries. Holstrom et al. consider that subject indexing can be done by four types of actors: professional indexers, domain experts, casual indexers, and machine algorithms.[13,14] Their article investigates and discusses the feasibility of what he termed a "hybrid approach" in which several different actors perform subject indexing of the same information object, result in benefits for subject searches.[15]

Although keywords lack standardization, Gonçalves identified keywords with relevant functions and characteristics, such as type of research (e.g. exploratory, theoretical, etc.) and context of the study.[16] For example, proper names perform the function of keywords as exemplified by Gonçalves in the citation of proper names such as "Gramsci" or "Kant."[17]

Considering the different functionalities that keywords assume, Lu et al. investigated how selected keywords function semantically in scientific publications.[18] To do this, they performed manual processing of articles from the Journal of Informetrics and performed a manual annotation scheme of keyword functions such as "research topic," "research method," "research object," "research area," "data," and "other" based on content analysis of the texts. The results showed that the diversity of keyword functions decreases, but irregularity increases with the number of keywords assigned by the author. The conclusions indicated that research should take into account the different types of keywords selected by the author.

In research on applying bibliometric analysis in automatic keyword extraction, Li used function differentiation for keywords, combined with topic term classification of texts, and concluded that the proposed function differentiation for keywords partially improves the selection of high-frequency words, particularly for text topic queries.[19]

The current landscape of digital environments endowed with digital tools and objects supports different actors performing indexing, and it is both possible and necessary for information professionals, domain experts, and authors of academic papers to improve subject indexing experiences on these information objects.

However, according to Fujita, et al., submission rules of scientific journals or self-archiving systems, in general, offer no guidelines for authors on subject indexing procedures for keyword assignment.[20] Pereira de Oliveira et al. analyzed the submission guidelines of various Brazilian journals of Information Science for advice on assigning keywords to articles. They found that journal guidelines largely only addressed the number of keywords, and did not provide guidance regarding how to select keywords, nor the use of controlled vocabulary. The authors recommended developing an indexing policy that provides clear advice to the authors at the time of keyword assignment.[21] Similarly, while self-archiving their academic papers in institutional repositories of university libraries, authors are usually not given information about keyword assignment, use of controlled vocabularies, or vocabulary expansion;[22] when they are guided, the submission guidelines are not publicly accessible in the system.

Very specialized areas require controlled vocabularies, or natural language keywords, which reflect the innovative and unique vocabulary of a particular groups of researchers.

The lack of an indexing method that performs the representation of indexed content with controlled vocabulary is the main disadvantage of keyword assignment. Using the Verbal Protocol to observe the mental strategies used by theses and dissertation authors during keyword assignment is an innovative approach, and this study contributes to indexing research by considering the perspective of ordering and functions assigned to keywords in the context of the authors' area of scientific expertise.

Methodology

The Individual Verbal Protocol, also known as "Thinking Aloud," has its origins in studies in Cognitive Psychology in the precursor studies by Ericsson and Simon.[23, 24] It is based on the use of introspection for "observing, obtaining and describing structures of the content of subjects' conscious experiences, with a focus on discovering the similarities of human behavior."[25] The methodology consists of recording the verbal externalization of thoughts during the reading activity, that is, the individuals read and interpret at the same time, verbally externalizing everything that "crosses their mind" while reading.

This introspective technique has been applied in many areas of knowledge, including eighty-one papers published in the area of Information Science from 1989 to 2013, according to Alonso-Arroyo et al.[26] In Information Science, the "Thinking Aloud" technique has been used—by Fujita et al.,[27] Fujita and Rubi,[28] and Redigolo, et al.,[29] among others—to research information search, usability, image search, relevance judgments, terminological understanding, visual information processing, abstract preparation and reading during indexing documents. This study used Individual Verbal Protocol—introspection and retrospection observation of thesis and dissertation authors as expert indexers of their domain—to analyze not only authors' mental process of keyword assignment but also, through a qualitative analysis of these authors' cognition, to report criteria for assignment, use of controlled vocabulary, understanding keyword importance, and criteria for keyword ordering and function.

The qualitative validity of a reduced sample of Individual Verbal Protocols represents the results of observing individual cognitive processes regarding the performance of a given task by individuals qualified for the task. Research applying the verbal protocol for observation and analysis of the thought process, behavior, and strategies used during task execution contribute to a better understanding of the phenomena under study.[30] This study applies the verbal protocol as a technique for collecting and analyzing information to obtain introspective verbal reports, which reveal the strategies employed by theses and dissertation authors in keyword assignment. Thinking aloud represents an additional task for participants to perform; therefore, five protocols are analyzed. This technique provides more accurate information about keyword assignment right at the moment it is being performed, that is, when particular cognitive behaviors occur.[31] This provides more direct access to an author's actual mental process during keyword assignment, whereas other techniques occur after the keyword assignment process.

Although this study's sample size is small, it benefits from analyzing the authors' choices while they performed the task of assigning keywords. Using a qualitative approach to observing how participants carried out their task, analysis and results were based on participants' cognitive expression.

By observing the UNESP institutional repository, we have verified that chosen keywords were, in fact, the authors' natural language. The natural language keywords assigned by the authors is available in the collected scientific production records, in the on-demand archiving, and in the self-archiving by thesis and dissertation authors. No vocabulary control is performed with the keywords collected from the metadata, which are available in an alphabetical list for consultation by users during the search. This alphabetical list presents variations of the same word with the use of singular or plural, capital letters or lowercase, use of quotation marks and other signs such as hyphen, and so on. No treatment is carried out to reduce these variations.

Despite having a search interface and several "filters" for refining searches, the authors use natural language with no vocabulary control tools in the formulation of the search for retrieval. The expanded query feature during the search is not available.

The definition of participating authors was based on the category of graduate students who self-archive their thesis or dissertation in the UNESP Institutional Repository upon completion of their master's or doctoral degree, in Information Science or Education on the Marília/SP campus. This study's sample size was five authors.

For the application of the online semi-structured interview using the Individual Verbal Protocol (IVP) qualitative methodology on the self-archiving of theses and dissertations in

the UNESP Institutional Repository, the following procedures were outlined: prior to self-archiving; while self-archiving; and post self-archiving.

Procedures Prior to Self-archiving

Definition of the Research Universe

The organizational context of the UNESP Institutional Repository was chosen for the observation of the self-archiving process of theses and dissertations through semi-structured interviews. Implemented in 2013, the Universidade Estadual Paulista (UNESP) Institutional Repository aims to, "store, preserve, disseminate, and enable open access, as a global public good, to the university's scientific, academic, artistic, technical, and administrative production," guided by the Internal Regulations of the UNESP Institutional Repository (UNESP, 2019). The UNESP Institutional Repository performs, in addition to automatic collection, archiving on demand, and self-archiving. In the self-archiving modality, it provides researchers and authors of theses, dissertations, and final papers with an interface for filling out the metadata of material and thematic description and submission of the original paper.

Selection of Authors/Participants

The study's five participants were graduate students in the master's and doctoral courses at the Marilia campus. An informal conversation was held with each of the authors through social media and/or e-mail, resulting in the acceptance and scheduling of the interview dates. For the analysis of the Verbal Protocol transcripts, the authors' identities were anonymized through specific initials, according to the level and graduate course to which they belong (table 1):

TABLE 1
Authors Participating in the Research and their Initials for Identification and Analysis of the Verbal Protocols

Level	Course	Identification
Master's Degree	Graduate Program in Information Science	M-CI
Doctorate	Graduate Program in Information Science	D-CI1
	Graduate Program in Information Science	D-CI2
	Graduate Program in Education	D-E1
	Graduate Program in Education	D-E2

Informal Conversation with the Authors

Informal conversations with the authors went as follows:

- The research objectives were explained to the participating authors and each was given the Informed Consent Form (ICF) signed by the researchers. The authors were asked to sign the document and formally accept participation in the research; an original document was sent to them. The researchers highlighted that their identity would remain anonymous. The purpose was to make the authors as comfortable as possible not compromise the data during the interview and the data collection recording.
- The authors were asked to briefly explain their research.
- The authors were introduced to the Individual Verbal Protocol (IVP) methodology and its basic guidelines (Appendix A).

Procedures While Self-archiving

Self-archiving in the UNESP Institutional Repository

The researchers started videoconference sessions with the participating authors. Participants were instructed to start the self-archiving process in the system and share their screen with the researchers for later analysis. Both the self-archiving process performance and the authors' verbalizations were recorded. At this moment, the researchers turned off video and audio not to interfere with the self-archiving process or verbalizations.

Keyword Assignment

The researchers verified the process the participants used to assign keywords to their theses and dissertations, writing down the keywords selection both in Portuguese and in English, as well as the cognitive and metacognitive strategies the participants used.

Retrospective Interview

Retrospective interviews were conducted with the authors so that they could complement their opinions about the performed activity. The questions were developed from the initial research question: "How do authors of theses and dissertations choose keywords in the institutional repositories of university libraries?" and were as follows:

- What criteria did you use to assign your thesis/dissertation keywords in the Repository?
- Did you feel the need to use a vocabulary control for choosing keywords?
- How did you decide on the order of the keywords and what function does each one have?
- How important are keywords to you?

At the end of each semi-structured interview, the recordings were automatically saved in the Google Drive tool.

Procedures Post Subject Searches

Literal Transcription of the Authors' Verbalization Recordings

The transcription of the authors' verbalizations was carried out during the self-archiving at UNESP Institutional Repository. For this, specific IVP notation was used (Appendix B).

Analysis of the Authors' Verbalization Recording

A detailed reading of the recording transcriptions was carried out, to search for significant phenomena for the elaboration of categories of analysis

Definition of the Categories of Analysis

These were based both on the retrospective interview questions and on the data collected from the Verbal Protocol application to the authors during the self-archiving of theses and dissertations in the UNESP Institutional Repository (table 2), considering the initial research question: "How do authors of theses and dissertations choose keywords in the institutional repositories of university libraries?"

Rereading of the Data to Extract Excerpts that Exemplified Each Category of Analysis

The semi-structured interview transcripts were reread to extract excerpts from the discussion that best exemplified each category of analysis by synthesizing the main observed aspects.

TABLE 2
Retrospective Interview Questions and Categories of Analysis

Question—Retrospective Interview	Category of Analysis
What criteria did you use to define your thesis/ dissertation keywords in the Repository?	Criteria for assigning keywords
Did you feel the need to use a vocabulary control for assigning keywords?	Use of controlled vocabulary for keyword assignment
How important are keywords to you?	Conception about the importance of keywords
How did you decide on the order of the keywords and what function does each one have?	Criteria for ordering and function of assigned keywords

Results

The presentation of the results was carried out in a qualitative way to allow the study of the strategies used at the moment of self-archiving in the UNESP Institutional Repository, and to analyze the cognitive and metacognitive aspects of the participants in view of the questions and the investigation objective. The small number of participants allowed the researchers to use the Individual Verbal Protocol (IVP) to analyze the methods participants used to select keywords. The categories of analysis were created according to the authors' answers and the questions of the retrospective interview as guidelines for their elaboration (see table 2).

The presentation of the results shows the categories of analysis and their rationale, the perceptions obtained during the procedures adopted in the self-archiving in the Repository, and the excerpts taken from the interview with the authors. Different acronyms were used to anonymize the participants, as shown in table 2.

Criteria for Assigning Keywords

Category Description

Criteria for assigning keywords refers to the parameters adopted by the authors to choose the terms that best represent the research developed in the thesis/ dissertation.

Author D-CI1 expressed difficulties in locating representative terms in the controlled vocabularies, as many did not contemplate the level of specificity of the terms worked in his/her research. For example, the author cited the term "social media," which is not included in the information representation instruments; "social networks" is offered as an equivalent, however, this term is conceptually different. Thus, D-CI1 stated, "we are tied to the instruments, but the instruments do not represent what we research." Author M-CI justified the use of their keywords from the object to be investigated, while for D-CI2, the keyword identification follows the employed scientific methods. When D-CI1 said they were, "always thinking about the one who is going to find my thesis," and "that those words had a direct identity with my work," they expressed the direct relationship between the developed research and the users' search process in the system as criteria for assigning keywords, in addition to retrieval and the visibility of their research.

To decide on the keyword order, author D-E2 considered the objective of their work, following a logical sequence that contemplates the object of the study, the type of the object, the subject acting on the object, and the aspect through which the subject is observed. Based on their research, the participant explained why they chosen certain keywords:

> First the object of all investigation that was the writing (…) After the definition of this first one, I had to define the writing of what? Writing of argumentative statement, then I defined the genre OPINION ARTICLE. Writing of the opinion article. But opinion article writing, second keyword, by whom? By students, by subjects, so I used the word, the term, SUBJECT. (…) I defined SOCIAL AWARENESS because awareness is a very broad term, very broad. So social awareness is in the sense of human values and meanings.

Author D-E1 stated that, "[keywords] are the words that really give the general idea of our work," and that the adopted criteria are related to the role each keyword plays in the research, with the first assigned keyword, "Afro-Antillean woman," as the main object. Next, the place or geographical limit contemplated in the research, "'Rondônia', and the area of knowledge, 'Education,' as "RONDÔNIA is because of our region, which … is the limit … is … spatial, geographical and … EDUCATION because the thesis is on EDUCATION." The fourth keyword, "bibliographical analysis," is linked to the method used in the research. The participant explained, "because we start from the historical text, but we also use documentary source, confronting with the documentary source and … and BIBLIOGRAPHICAL." Finally, the keyword phrase "cultural studies," represents the theoretical approach of the research, that is, the "theoretical line."Regarding keyword translation, author D-CI1 had doubts about the insertion of terms in Spanish, asking, "should I also assign in Spanish, knowing these terms but not being in my thesis?…~~ If I think I explored the environment in Spanish … is … from Spain and that I want people to retrieve that thesis, maybe I put them, but what if that is … block … my submission and delay my certificate issuance? …~~ Do I call the chat … to help me? … A librarian? … I'll … ah, no." Although the participant recognized the possibility of enhancing retrieval of their dissertation by assigning keywords in an additional language, they decided not to include them. Similarly, author D-E1 only assigned keywords in Portuguese and English.

Use of Controlled Vocabulary for Keyword Assignment

The use of controlled vocabulary for assigning keywords refers to the use of controlled language for choosing keywords. It is divided into two sub-items to distinguish when controlled vocabularies are used: when assigning the keywords in the paper and when self-archiving it in the UNESP Institutional Repository.

Use of Controlled Vocabulary for Assigning Keywords in the Thesis or Dissertation

When inserting the dissertation keywords in self-archiving the paper in the UNESP Institutional Repository, author D-CI1 was asked about the guidelines to consult the UNESP Thesaurus to assign the keywords; they replied, "here I have a doubt … if I have already assigned in my text the keyword … which I thought … the words that represent my subject … is it an obligation …. does UNESP make me put in my dissertation the terms that are in the UNESP thesaurus? … And now, if this is the case … I will have to consult the terms … that I chose, and if I don't have them here, will I put them in or not?" This guideline is available only at the time of author self-archiving (i.e., there is no formal guidelines for authors prior to the submission process to adopt the UNESP Thesaurus while the thesis/dissertation is still under development) causing confusion to the authors.

Author D-E1 stated that they did not consult the UNESP Thesaurus to assign the keywords for their thesis, due to the level of specificity reached by the first assigned keyword, as well as the object of study of the research: "Afro-Antillean woman." The participant stated:

> yes … so I was sure that it doesn't have. (…) But I am sure it has … RONDONIA, EDUCATION, BIBLIOGRAPHICAL AND DOCUMENTAL ANALYSIS, and CULTURAL STUDIES ((RI)). […] The only keyword that I was sure I did not have and I said, ah, I will not consult.

This quote highlights the importance of updating the controlled language, in this case, the UNESP Thesaurus, in order to follow the dynamics of the scientific development of the respective areas of knowledge it covers. As for other keywords, participants did not consult the UNESP Thesaurus because they were sure the keywords would be located in this controlled language; however, there was no need to perform such consultation/validation.

Author D-E2 did not feel the need to consult a controlled vocabulary for assigning keywords for their dissertation, pointing out that they had not used a controlled vocabulary for their selection; instead they said that, "the choice of keywords and their order followed the philosophy of language theory." It can be observed that the author was probably unaware of the function of a controlled vocabulary. Author M-CI consulted the UNESP Thesaurus, a controlled language used by UNESP Network to standardize the subjects assigned to the documents inserted in the Athena catalog, but did not find the desired term used in their research. This participant found a conceptually similar term, but chose not to follow the vocabulary, saying, "I ended up not using [the thesaurus], I was supposed to use DOCUMENTARY LANGUAGE, because we defined that the term to refer to it would be INDEXING LANGUAGE."

Use of Controlled Vocabulary During Self-archiving of Thesis or Dissertation

Three participating authors did consult the UNESP Thesaurus and understood the importance of having a control over the terminological issue. However, not all terms were found in the vocabulary and the authors were in doubt about how to proceed, opting to maintain the terms they had already defined as keywords. For example, author D-CI1 stated, "I will risk it, I will add my keywords, I hope that with this I can signal to them that this term is important, is represented in the research, that they can incorporate them, sometimes I am helping … and giving suggestions for terms." Author D-CI2 consulted the controlled vocabulary to better represent their work, saying, "when assigning keywords, we try to verify … if those keywords … were large areas, if they were being placed in the right way, if they really represented the work… So, in our case, we tried to do a control, yes." Author D-CI1 tried to consult the UNESP Institutional Repository tutorial to seek guidelines to help their decision at this point but didn't find the guidelines.

When choosing keywords for their thesis, author M-CI had no concern about not finding the desired term; however, M-CI became concerned when self-archiving in the UNESP Institutional Repository because they only found two relevant terms in the institution's vocabulary. Therefore, in selecting other keywords, M-CI decided to follow the keywords already predefined in their research, saying, "And now I don't know what I to do (…) As a related term it gives me DOCUMENTARY LANGUAGE (…) should I put DOCUMENTARY LANGUAGE or INDEXING LANGUAGE? So, I will look at the thesaurus manual (?) It shows how to use

it, but in my case, I will have to make a choice. I will add INDEXING LANGUAGE." Author D-CI1 suggested training for authors using the UNESP Institutional Repository, on how to select keywords, and what to do when the terms are not found in the used vocabulary, sharing, "I feel that it lacks … a … instruction, a capacity building, a training on whether it is mandatory, not mandatory, (…) I understand that these instruments … they are not thought up for nothing, is to find the information, but … we need to train the user in advance." The other two authors did not consult any type of vocabulary. One author explained that they did not consult it because they knew they would not get the desired term, because their research uses very specific terms, saying, "I did not consult the thesaurus, the UNESP thesaurus … because I knew I wouldn't have … I was sure there would be no AFRO-ANTILLIAN WOMAN… So … I kept typing and including my keywords."

Understanding the Importance of Keywords

The study checked the authors' opinion and understanding about keywords.

The participating authors understood the importance of keywords and two main themes emerged. First, participants understood that keywords function as a way to represent the main points of their research in a condensed way. As author D-E1 expressed, "the keywords go back to your work, it is the map of the work, it is the face of the work. The keywords, in five words, describe … the most factual, the most … is … the identity of the work, the identity of the work." Author D-E2 agreed with this perspective stating that keywords guide "both for the one who writes and for the future reader, for all potential readers." In short, as author M-CI stated, "keywords are the basic summaries of even the essentials of the document. If not well defined, they will not represent the essence of the work."

The second function of keywords frequently noted by the participants was for information retrieval. Participants understood that their papers would be retrieved through the keywords. As explained by author D-CI2 "I think it is extremely important not only for the identification issue, but also for the retrieval of the work itself, because if you correctly assign the keywords you end up retrieving exactly what you are… you want to make available… for your reader, when you don't assign the keywords correctly, normally he will not retrieve what he would like to retrieve." This situation causes retrieval problems in the Institutional Repository, as "many times you do research, a bibliographical survey, and end up getting papers that are not coherent with what is written in the keywords."

Criteria for Ordering and Function of Assigned Keywords

The criteria for ordering and function of the assigned keywords identifies the principles for deciding the sequence of keywords assigned by the authors.

In choosing the order of keywords, author D-CI1 was guided by the research objective, and used a broader concept—"Digital Media"—after contemplating other terms, such as "Resources" and "Sharing Networks." D-CI1 explained, "SOCIAL MEDIA is an umbrella concept, it is the object of research and it is the term we defend; So, because it is a more generic term and represents the most, it was the first one to be chosen." D-CI1 chose the term "Web 2.0 Technologies" "because Web 2.0 is the concept where it emerged, without it, SOCIAL NETWORKS and SOCIAL MEDIA would not exist. So, in a conceptual matter of supporting this object today that underlies and sustains it, it is WEB 2.0 TECHNOLOGIES." Next, D-CI1 chose "Social media," "because the term is widely used by the discursive community."

According to the D-CI1, "MEDIA, TECHNOLOGY and NETWORKS are three terms from the same family used … to represent these information environments, it is … online, from the web, so we used them together." The two last terms were defined by the research setting: "Libraries;" "and by the methodology used: 'Domain analysis.'"Defining the keywords based on the scope of the term, as well as the form used by the discourse community, suggests that participants were concerned with making their selected keywords representative enough to achieve a broader understanding of the research. Starting from the main research object, author D-E1 chose the term "Afro-Antillean woman" as the first keyword, explaining, "the woman is the main reference, the woman in EDUCATION, the AFRO-ANTILLIAN WOMAN in EDUCATION and as the text, the title says, so we chose the question of the woman, because it is the basis, … of our work." The second and third terms were selected based on the research setting, that is, the region of the state of "Rondônia," since "*RONDONIA is the space, the region, the Amazon region,*" whereas "Education" is the research theme centered on the education of Afro-Antillean women. Author D-E1 explained, "EDUCATION in third place… because we work on the issue of women in EDUCATION, black teachers in EDUCATION in the Amazon." Finally, D-E1 chose "Bibliographic and documental analysis," and "Cultural studies" as terms, stating that these elements form "the basis of [their] research." In this instance, keyword ordering was guided by the main concepts of the research; keywords were seen as fundamentally representing the work. The two authors used similar criteria to choose the two final keywords, based on the research setting and the methodology used.

At first, author D-CI2 reported that keyword ordering was based on the research methodology. D-CI2 chose "domain analysis," explaining that this keyword "came first because it was the method I used to be able to develop the rest of my work." However, D-CI2 did not select their other keywords based on importance, sharing, "the others … I cannot say that there was a… how do you say? A degree of importance." From the questions asked in the retrospective interview, the author reports reflecting on the issues raised about the keyword ordering, considering a new sequence from this reflection.

On the other hand, author D-E2 reported following an specific order beginning with the research object, "writing," followed by a logical sequence for the definition of the other keywords: "opinion article genre," "subject," and "social awareness." D-E2 shared, "we imagined a sequence, THE WRITTEN language, argumentative genre, OPINION ARTICLE. Maybe it wouldn't be illogical to think first about the subject and then choose a genre, would it? First, we think of the language, the WRITTEN language and then we define the genre and then we define who is going to construct that genre, the subject."

Analysis of Results

In response to the research question, "How do thesis and dissertation authors choose keywords in institutional repositories of university libraries?" the results indicate that the way theses and dissertations authors choose keywords is directly related to the level of specificity of the research conducted in the domain. Therefore, many new terms are used and could be aggregated as related to the controlled vocabulary authorizers, as Peset also observed.[32] The selected terms directly relate to the topic, the object and objective of the research, and according to scientific methods used; however, keywords were also selected with a view to the research retrieval and visibility. In the keyword ordering and function, ordered sequences of functions are indicated by the authors. Participating authors contemplated the object of study, the type

of the object, the subject acting on the object, and the aspect by which the subject is observed; this aligns with Lu et al.'s findings.[33] In another sequence, authors assigned keywords that represent the topic of the subject, the geographical location, the method, and the theoretical approach, as Gonçalves also noted.[34] Therefore, an ordering principle for assigning keywords is verified, guided by semantic functions from the main concepts disseminated in the research and representative of the specialized domain in which the thesis or dissertation was generated.

Keyword assignment is performed at two moments, after the abstract in the pre-textual part of the thesis or dissertation, and during self-archiving for filling the subject metadata in the Repository. The guidance for using controlled vocabulary is available during self-archiving, but not during the formal preparation of the thesis and dissertation text. The use of controlled vocabulary was not necessary for the authors who were probably unaware of it. However, when they self-archived participants understood the importance of vocabulary control. If participants were unable to find a precise match for their choice of keywords with the controlled vocabulary terms, they decided to maintain their previously assigned keywords which were conceptually compatible with the meaningful content of the text, as Li et al.,[35] Zhang et al.,[36] and Peset also found.[37] The use of the controlled vocabulary was not understood by the authors who did not find a tutorial available with guidelines to assist them in their decisions, and not all keywords were compatible with the controlled vocabulary terms already observed by Oliveira et al.[38] and Freitas and Dal'Evedove.[39]

Conclusions

This investigation provides important results about how authors assign keywords, how they choose keywords that are more specific and pertinent to the object and purpose of the research, and how keyword ordering is guided by the conceptual function the keywords represent. Moreover, this study found that authors were aware that keywords are important for visibility and retrieval, and that they kept this in mind while selecting keywords. In general, participants were unaware of controlled vocabulary and its vocabulary control function prior to self-archiving, and they had no guidance on controlled vocabulary use and function. These results provide further research directions in investigating content representation in hybrid information systems, aimed at making high-quality open scholarly resources available.

The Individual Verbal Protocol provided personalized results by revealing the cognitive manifestations of each participant while selecting keywords. The immersive and introspective analysis of the procedures employed by the authors in choosing keywords for their theses and dissertations reveals a scenario of discussion and guidelines that help institutions and professionals to improve the quality of subject metadata and support the author in assigning keywords.

This study concludes that author keyword assignment was guided by concepts fundamental to the representation of the meaningful content of the text, and that keywords were ordered based the main theme of the research, as well as an awareness of the need for visibility and retrieval.

Finally, this study's results show that authors act as domain expert indexers, but are unaware that keyword assignment is an indexing process that requires representation by controlled vocabularies. To this end, the study recommends that self-archiving systems include tutorials on keyword assignment with vocabulary control without requiring authors to exclusively use controlled terms. Keywords tend to represent more specific subjects within

the sciences while the indexing terms of a controlled vocabulary tend to be more stable and connect to broader subjects. Keywords and controlled vocabulary indexing terms are complementary and neither should be used exclusively. The better option is a hybrid information representation and retrieval system which allows keywords and controlled vocabulary indexing terms to coexist.

Acknowledgements

We would like to thank the postgraduate students in Education and in Information Science at São Paulo State University, Marília Campus, Brazil, for their participation in our survey. We are grateful to National Council for Scientific and Technological Development (CNPq), Brazilian government agency, for support to scientific research.

Appendix A. Introduction to the Verbal Protocol Technique

Procedures: Through the Verbal Protocol, we will observe the patterns and strategies used by the authors while performing subject indexing of keyword assignment to their final graduate papers in the São Paulo State University (UNESP) Institutional Repository. The research volunteers are students from the Graduate Programs in Human and Social Sciences at the Faculty of Philosophy and Sciences (FFC) of UNESP, Marília, who defended their theses or dissertations and are in the process of self-archiving their paper in the UNESP Institutional Repository.

Instructions to authors: The Verbal Protocol (VP) is the data collection technique used in our research. We will be recording this moment. The VP technique consists of the author "thinking aloud" the procedures being carried out, that is, verbalizing the procedures being carried out, in our case, at the moment of self-archiving the thesis/ dissertation in the UNESP Institutional Repository. We highlight that your personal data will be kept confidential, and only the data concerning the research will be used. You will be sent the Informed Consent Form (ICF) of the research, with the researchers' signature, informing about the research, for your authorization.

Appendix B. Specific Notes for Transcribing the Interviews*

italics: author's vocalization

( ): researcher's questions or comments

….: short pauses

…~~: long pauses

(…): omission of a passage that is not relevant in the transcription of the interview

((RI)): author's or researcher's laughter

((RM)): author's or researcher's tone of irony

"…": author's or researcher's paraphrase

{…}: excerpt from the base text† verbalized by the author

[ ]: inclusion in the transcriptions, of description of the author's significant gestures or the researcher's analytical comments

MAYBE: keywords discussed or assigned by the authors to the theses and dissertations

Underlined: specific passage that demonstrates the studied phenomenon

* Adapted from original notes for verbal protocols transcription by Marilda do Couto Cavalcanti, "*Reader-text interaction: aspects of pragmatic interaction*" (Campinas: UNICAMP Press, 1989).

† The base text refers to the information contained in the form used by the authors to self-archive the thesis/ dissertation/ in UNESP Institutional Repository.

Notes

1. John Budd, *The academic library: its context, its purpose, and its operation* (Englewood: Libraries Unlimited, 1998).

2. Marcos Gonçalves. "Digital libraries," in *Modern information retrieval: the concepts and technology behind Search*, by Ricardo Baeza-Yates and Bertier Ribeiro-Neto (Harlow: Pearson, 2011).

3. American National Standards Institute - ANSI; "National Information Standards Organization - NISO. Z39:19-2005: guidelines for the construction, format, and management of monolingual controlled vocabularies." Bethesda: NISO Press, 2005.

4. Karen Spark Jones, *Automatic keyword classification for information retrieval*, (London, UK: Butterworths, 1971).

5. Ana Miguéis and Bruno Neves, "Uma abordagem à linguagem de indexação dos artigos científicos depositados no repositório científico da universidade de Coimbra," *Ponto de Acesso* 7, no. 1 (May 2013):116-131.

6. Fernanda Peset et al., "Survival analysis of author keywords: an application to the library and information sciences area," *Journal of the Association for Information Science and Technology* 71, no. 4 (May 2020): 462-473. doi.org/10.1002/asi.24248

7. Ling-Li Li et al., "Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006," *Scientometrics* 80, no 1 (March 2009): 39-58. doi.org/10.1007/s11192-008-1939-5

8. Juan Zhang et al., "Comparing keywords plus of WOS and author keywords: a case study of patient adherence research," *Journal of the American Society for Information Science and Technology* 67, no. 4 (April 2016): 967-972. doi.org/10.1002/asi.23437

9. Zhang et al. "Comparing keywords," 972.

10. Aline Lima Gonçalves, "Uso de resumos e palavras-chave em Ciências Sociais: uma avaliação", *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação* 13, no. 26 (October 2008): 78.

11 Aurélie Névéol, Rezarta Dogan and Zhiyong Lu, "Author keywords in biomedical journal articles," AMIA…Annual Symposium Proceedings 2010 (November 2010): 537-41.

12. Névéol, Dogan and Lu, "Author keywords," 541.

13. Chris Holstrom, "Moving towards an actor-based model for subject indexing," *North American Symposium on Knowledge Organization* 7, no. 1 (2019): 120-128. https://doi.org/10.7152/nasko.v7i1.15631

14. Marco Lardera and Biger Hjorland, "Keywords," *Knowledge Organization* 48, no.6 (November 2020): 430, https://doi.org/10.5771/0943-7444-2021-6-430

15. Holstrom, "Moving towards an actor-based," 128.

16. Gonçalves, "Uso de resumos e palavras-chave," 93.

17. Gonçalves, "Uso de resumos e palavras-chave," 93.

18. Lu et al., "How do author-selected keywords function semantically in scientific manuscripts?" *Knowledge Organization* 46, no. 6 (October 2019): 403, https://doi.org/10.5771/0943-7444-2019-6

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

20. Mariangela S. L. Fujita, María-del-Carmen Agustin-Lacruz and Ana Lúcia Terra, "Journals' guidelines about title, abstract and keywords: an overview of information science and communication science areas," *European Science Editing* 44, no.4 (November 2018): 76-79.

21. Lais Pereira de Oliveira et al., "Política de indexação em periódicos da Ciência da Informação: um estudo das diretrizes para atribuição de palavras-chave aos artigos," *Perspectivas em Ciência da Informação* 25, no. 4 (March 2020): 140-169.

22. Marina Penteado Freitas and Paula Regina Dal'evedove, "Consistência na indexação por atribuição no repositório institucional da UFSCAR," *Encontro Nacional de Pesquisa em Ciência da Informação*, October 2019. https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1203

23. Anders Ericsson and Herbert Simon, "Verbal reports as data," *Psychological Review* 87, no. 3 (1980): 215-251, https://doi.org/10.1037/0033-295X.87.3.215.

24. Anders Ericsson and Herbert Simon, "Verbal reports on thinking," in *Introspection in second language research*, ed. Claus Faerch and Gabriele Kasper (Clevedon: Multilingual Matters, 1987), 24.

25. Roberta C. D. Tartarotti, Paula Regina Dal'Evedove and Mariângela S. L. Fujita, "Protocolo Verbal em grupo e a pesquisa brasileira em organização e representação do conhecimento", *Encontros Bibli: Revista eletrônica de Biblioteconomia e Ciência da informação* 22, no. 48 (January 2017): 41-58, https://doi.org/10.5007/1518-2924.2017v22n48p41.

26. Adolfo Alonso Arroyo et al., "Protocolo verbal: análisis de la producción científica, 1941-2013," *Informação & Sociedade: Estudos* 26, no. 2 (September 2016): 76.

27. Mariângela S. L. Fujita, Maria Isabel Nardi and Silvana Aparecida Fagundes, "Observing documentary

reading by verbal protocol," *Information Research* 8, no. 4 (July 2003): 61.

28. Mariângela S. L. Fujita and Milena P. Rubi, "Modelo de lectura profesional para la indización," *Scire* 12, no. 1 (June 2006): 47.

29. Franciele M. Redigolo, Mariângela S. L Fujita and Isidoro Gil-Leiva, "Guidelines for Subject Analysis in Subject Cataloging," *Cataloging & Classification Quarterly* 60, no. 5 (August 2022): 424, https://doi.org/10.1080/01639374.2022.2093300

30. Anders Ericsson and Herbert Simon, *Protocol analysis: verbal reports as data*. (Cambridge, UK: MIT Press, 1993).

31. Andrew Cohen, "Using verbal reports in research on language learning," in *Introspection in second language research*, ed. Claus Faerch and Gabriele Kasper (Cleverdon: Multilingual Matters, 1987), 82.

32. Peset et al., "Survival analysis," 473.

33. Lu et al., "How do author-selected keywords function semantically in scientific manuscripts?" 418.

34. Gonçalves, "Uso de resumos e palavras-chave em Ciências Sociais," 93.

35. Li et al., "Global stem cell research trend," 58.

36. Zhang et al., "Comparing keywords plus of WOS and author keywords," 972.

37. Peset et al., "Survival analysis of author keywords," 473.

38. Oliveira et al., "Política de indexação," 169.

39. Freitas and Dal'evedove, "Consistência na indexação."