ISSN 1678-765X

doi 10.20396/rdbci.v20i00.8668751

Research Article

RDBCI
Revista Digital de Biblioteconomia e Ciência da Informação
Digital Journal of Library and Information Science

Systematization of the evaluation model of vocabulary control in repositories: research report with the Unesp Institutional Repository

Mariângela Spotti Lopes Fujita [1]

ABSTRACT

Introduction: Vocabulary control in information resource storage, treatment and retrieval systems is necessary to obtain consistency between indexing and retrieval in order to avoid informational dispersion. Digital repositories in universities are currently fundamental in the organization and management of knowledge generated by scientific, technological, artistic, and administrative production, however, it is necessary to verify the availability of controlled vocabulary and how vocabulary control is carried out. Objective: With the objective of systematizing a proposal for vocabulary control and the use of controlled vocabularies in university repositories managed by libraries, an evaluation model was developed that proposes to systematize methods, procedures, resources, and techniques. Methodology: For this, the development of the investigation carried out exploratory research with bibliographic and documentary research and applied research in the Unesp Repository. Results: The results obtained constituted an Action Plan, discussed and elaborated by the Study Group, composed of six actions and nine studies to evaluate and control vocabulary in university repositories about: vocabulary control in indexing by professionals and non-professionals; vocabulary control in retrieval; use of subject metadata from academic papers; keyword matching; analysis of terminological variations at semantic, syntactic and pragmatic levels; analysis of transaction logs for searches by subject. Conclusion: It is concluded that the systematization of actions in an evaluation model is relevant for university repositories to incorporate the advances offered by vocabulary control in their routines and, mainly, for the contribution of new terms arising from scientific and technological evolution.

KEYWORDS

Vocabulary control. Documentary languages. Institutional repositories. Evaluation.

Author's correspondence

[1]Universidade Estadual Paulista
Marília, SP, Brazil /
e-mail: mariangela.fujita@unesp.br

Sistematização de modelo de avaliação do controle de vocabulários em repositórios: relato de pesquisa com o Repositório Institucional Unesp

RESUMO

Introdução: O controle de vocabulário em sistemas de armazenamento, tratamento e recuperação de recursos de informação é necessário para se obter consistência entre a indexação e a recuperação de modo a evitar a dispersão informacional. Repositórios digitais em universidades são, atualmente, fundamentais na organização e gestão do conhecimento gerado pela produção científica, tecnológica, artística e administrativa, entretanto, é preciso verificar a disponibilização de vocabulário controlado e como se realiza o controle de vocabulário. Objetivo: Com o objetivo de sistematizar proposta para controle de vocabulário e uso de vocabulários controlados em repositórios universitários administrados por bibliotecas foi elaborado modelo de avaliação que se propõe a sistematizar métodos, procedimentos, recursos e técnicas. Metodologia: Para isso, o desenvolvimento da investigação realizou pesquisa exploratória com pesquisa bibliográfica e documental e pesquisa aplicada no Repositório Institucional Unesp. Resultados: Os resultados obtidos constituíram-se em um Plano de Ação, discutido e elaborado por Grupo de Estudos, composto de seis ações e nove estudos para avaliação e controle de vocabulário em repositórios universitários acerca de: controle de vocabulário na indexação por profissionais e não profissionais; controle de vocabulário na recuperação; uso de metadados de assuntos de trabalhos acadêmicos; compatibilização de palavras-chave; análise de variações terminológicas em nível semântico, sintático e pragmático; análise de logs de transação de buscas por assuntos. Conclusões: Conclui-se que a sistematização das ações em modelo de avaliação é relevante para que repositórios universitários incorporem os avanços oferecidos pelo controle de vocabulário em suas rotinas e, principalmente, pela contribuição de novos termos oriundos da evolução científica e tecnológica.

PALAVRAS-CHAVE

Controle de vocabulário. Linguagens documentárias. Repositórios institucionais. Avaliação.

CRediT

- Recognitions: Thanks to the General Coordination of Libraries of Unesp.
- Funding: This study was funded by the Brazilian agencies National Council for Scientific and Technological Development (CNPq) for the grant granted. Process No. 304760/2017-0
- Conflicts of interest: Authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
- Ethical approval: Yes.
- Availability of data and material: The data sets generated and/or analyzed during the present study are available in the https://dmptool.org/plans/60792

Authors' contributions: Authorship is the only one that performed conceptualization, formal analysis, research, methodology, supervision, validation, original draft writing: FUJITA, M. S. L.

JITA: HS. Repositories.

Article submitted to the similarity system

Submitted: 25/03/2022 - **Accepted**: 26/05/2022 – **Published**: 31/05/2022

1 INTRODUCTION

With the accelerated production of born-digital documents (created in digital environments) and the concern with preservation, repositories appear as an alternative for the safe deposit of digital objects. The great advantage is that they can be used in public and private institutions to disseminate all the produced research, in addition to providing self-archiving by the authors, which provides better research dissemination. Therefore, many repositories are registered worldwide, and the Ranking Web of Repositories[1] counts 3885 repositories around the world in its latest edition of February 2022, among which 3751 are institutional repositories.

It constitutes a scientific information service (in a digital and interoperable environment) that manages the intellectual production of an educational and research institution. It gathers, stores, organizes, preserves, retrieves and, mainly, disseminates the scientific information produced in an institution. According to Lynch (2003, p. 2) it is "[...] a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members."

Furthermore, a repository contains a system for retrieving information through document and metadata access points, which allows access to the digital document contained in it. In the case of the subject, as in other databases, it is possible to verify whether vocabulary control was used for representation in indexing or in search.

With the possibility of self-archiving performed by the author, the repository becomes a more dynamic and friendly information system as it allows interactivity with the users who start to build it socially and, with that, want to be visible in addition to the need to ensure digital preservation in an institutional system that will provide probative reliability to the funding agency, other institutions and the scientific community. This interaction provided by self-archiving requires a commitment to descriptive and thematic standards from the Repository that are continually assessed and applied to ensure visibility.

On the other hand, the repository's author and user interaction provides benefits of its applied knowledge domain terminology for keyword assignment in subject metadata during self-archiving and during the search strategy. This terminology is specialized in knowledge domains used among peers in the scientific community that follows the evolution and innovation whose document content comes from scientific research to generate new knowledge. On the other hand, this scientific terminology has terminological variations, mainly at the syntactic and semantic level, which need vocabulary control with the use of controlled vocabularies to ensure the desired visibility.

Considering that the repository practices the combined use of natural and controlled languages in a hybrid way, this situation can benefit the repository if there is an indexing policy for authors and librarians that provides the necessary guidelines for vocabulary control. For this, it is necessary to study vocabulary control assessment in repositories to develop an adequate proposal for the use of controlled vocabulary during subject assignment as well as the use of natural language for continuous updating of the controlled vocabulary.

Aiming at elaborating a proposal for vocabulary control and use of controlled vocabularies in university repositories managed by libraries, an assessment model was developed, which proposes to systematize methods, procedures, resources and techniques for the elaboration of an indexing policy for repositories.

[1] TRANSPARENT RANKING: All Repositories (February 2022).
Available at: https://repositories.webometrics.info/en/transparent

2 THEORETICAL FRAMEWORK

Vocabulary control is exercised with the aid of controlled vocabulary such as thesauri, authorized alphabetical lists of terms, subject heading lists, among others.

The ISO 25964-2 Standard (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011, p.16) on thesauri for information retrieval considers vocabulary control essential because in common speech a term can have more than one meaning and the choice of a preferred term to represent a specific concept is never straightforward because concepts can be expressed in many ways. Therefore, the thesaurus plays an important role in mediating between the terms used in speech and those that work effectively for information retrieval, which implies that the user accepts a degree of artificiality in the controlled vocabulary to achieve benefits in retrieval.

The vocabulary control concept for Standard Z39.19-2005 "Guidelines for the construction, format, and management of monolingual controlled vocabularies" (AMERICAN NATIONAL STANDARDS INSTITUTE/ NATIONAL INFORMATION STANDARDS ORGANIZATION, 2005, p.10) means that it is

> [...] the process of organizing a list of terms; (a) to indicate which of two or more synonymous terms is authorized for use; (b) to distinguish between homographs; and (c) to indicate hierarchical and associative relationships among terms in the context of a controlled vocabulary or subject heading list.

The norm, therefore, considers the organization of the controlled vocabulary as the vocabulary control itself, however, at the same time, highlights the functions that the controlled vocabulary performs, such as the indication of the authorized synonym term and what hierarchical and associative relationships exist between terms. Therefore, vocabulary control is present in a controlled vocabulary which in turn is used to perform vocabulary control.

Vocabulary control, therefore, is linked to the use of a controlled vocabulary that "[...] is essentially a list of authorized terms." (LANCASTER, 2004, p.19, our translation), however, it goes beyond a list as the authorized terms are organized in a semantic structure that controls synonyms, homographs and related terms, either by hierarchical relationship or associative relationship.

Controlled vocabularies characteristically have a dual function for the purpose of reciprocity in vocabulary control, because they are used during the representation and search processes. Lancaster (2002, p.22, our translation) demonstrates the dual role for vocabulary control purposes:

> 1. Facilitate the consistent representation of subjects by indexers and users who retrieve, avoiding the dispersion of related elements. This is achieved with the control (grouping) of synonyms and quasi-synonyms and the distinction of homographs;
> 2. Facilitate a broad search on a subject by linking terms with paradigmatic and syntagmatic relationships

Hjorland considers that the principle of controlled vocabulary follows Cutter's rule that "[...] it is always the most specific, most appropriate expressions that should be looked up in the vocabulary of notations and assigned to documents." And that, "In this way, the expressions for the topics to be made retrievable are rendered most predictable." (HJORLAND, 2008, p.89)

The advances of indexing in terms of its assessment are valid measures to ensure the advantage of the use of controlled vocabularies in providing consistency, both in representation and in search, so that "[...] a concept or theme always appears expressed in the same way" (MOREIRO GONZALEZ, 2004, p. 51, our translation).

Another advantage is that the evolution of controlled vocabularies, probably influenced by social indexing on the internet, has offered increasingly intuitive visualization modes designed for non-professional users who need terminological support to achieve specificity or exhaustiveness in their searches. In the evolution of controlled vocabularies, the thesaurus is undoubtedly used by national and international institutions as a form of presentation that facilitates the understanding of concepts and their contextualization in a certain area of knowledge. In this regard, the ISO 25964 Standard (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011, p.vi) considers "[...] in the past thesauri were designed for information professionals trained in indexing and searching, today there is a demand for vocabularies that untrained users will find to be intuitive [...]".

Based on this premise the ISO 25964 Standard supports the application of the thesaurus as a controlled vocabulary, also in situations where computers make choices, that is, "If both the indexer and the searcher are guided to choose the same term for the same concept, the relevant documents will be retrieved" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011, p.vi).

Controlled vocabularies are critical when indexing. It is a natural language translation tool aimed at reducing the problems of terminological variation in addition to inconsistencies in the written form. Controlled vocabulary is a list of authorized terms that helps indexing, improving aspects related to retrieval with the use of terms consistent with the content of the documents; control synonyms, opting for a single standardized form, with cross-references in all other forms; differentiate homographs; bringing together or linking terms whose meanings are more closely related to each other. However, it is more than a list, since the authorized terms are organized in a semantic structure that allows the control of synonyms, homographs and related terms, either by hierarchical relationship or associative relationship (LANCASTER, 2004, p.14 -19).

However, using vocabulary control in institutional repositories is not an easy task due to a series of factors and peculiar aspects not present in other information systems, such as the modalities of self-archiving by the author and automatic populating. With the results obtained from the analysis of the criterion for the use of controlled vocabularies or terminological tools in a sample of 35 Spanish university repositories, Barrionuevo Almuzara, Alvite Díez, Rodríguez Bravo (2012, p.98, our translation) observed that "[...] the main function was handled by lists of subject headings and keywords, and in lesser degree, classifications, thesauri, and descriptor lists." And that, with the option of self-archiving, authors can determine their own keywords without consulting the subjects the system offers. They conclude that "[...] the volume of uncontrolled terms that can be included in repositories is not limited, a circumstance that seems to require some form of standardization."

In an investigation on controlled vocabularies, Fujita and Tolare (2019) performed an analysis on interface resources of 86 Brazilian repositories to identify types of controlled vocabularies. The results identify that 81% use lists of terms in alphabetical order without vocabulary control and 65% of the repositories include natural language keywords and terms from controlled vocabularies in their metadata. They consider that the list of terms in alphabetical order, derived from natural language keywords, are less complex controlled vocabularies compared to thesauri and that they could be improved with the application of vocabulary control. Regarding the integration of terms from controlled vocabularies with natural language keywords assigned by the authors, Fujita and Tolare (2019) consider it necessary to update vocabularies for two reasons: the term lists incorporate terms and keywords and the number of terms and keywords in metadata increase the visibility of scientific production archived in the repository.

3 METHODOLOGY

The methodology adopted to carry out this investigation on vocabulary control assessment and the use of controlled vocabularies in institutional repositories has an exploratory descriptive character of an ethnographic nature due to the need to extract data and information directly from reality.

Ethnographic research aims to discover new relationships and new ways of understanding reality based on the participants' observation and perspective of the meanings of the results obtained in their daily practice. For this, participant observation techniques, interview and interview analysis were used by the researcher based on the observed aspects and results obtained from theoretical studies, as well as analysis of the documentation, which were organized in three stages: 1) exploration, which consists of the selection of problems, location and the first contacts with the field of study; 2) decision, or data search to understand and interpret the phenomenon; and 3) explanation of the reality, through the analysis of the entire process experienced by the researcher, through the professionals' reports about the developed activity (MAIA, 2007).

The development of the ethnographic investigation was carried out with two methodological guidelines: the first one discussed, observed, interacted on the use and vocabulary control assessment with a Study Group formed with researchers on the subject, university libraries' catalogers, repositories' manager and support professionals; and, the second one that carried out and discussed viable proposals for vocabulary control assessment in a university repository with group monitoring.

The first guideline, of ethnographic nature, constituted the Study Group to study the vocabulary control assessment and the use of controlled vocabularies in institutional repositories through biweekly meetings for two years, in order to promote critical reflection to identify problems or assess changes during seminars on professional experiences or sharing of experiences. In these seminars, the theoretical and methodological systematization of vocabulary control in information representation and retrieval and vocabulary control assessment in institutional repositories managed by libraries was presented from the analytical and comparative considerations of research development, as well as the proposal of a methodological model for vocabulary control assessment in university repositories.

The researcher and the group interacted with the research object and analyzed the entire process that allowed the improvement of the methodological model with the proposal of an Action Plan for implementation, maintenance and assessment whose systematization resulted in the vocabulary control assessment model in institutional repositories, object of analysis of the second methodological guideline of ethnographic research. The second guideline performs the systematization of the methodological model for vocabulary control assessment in a university repository with group monitoring. The university repository used for the plan execution was the Unesp Institutional Repository, coordinated by the Management Group and developed by the Executive Coordination that monitors the activities of the Technical Team. The second guideline unfolded into two phases: the phase of knowledge about the case study with the Unesp Institutional Repository and the discussion phase of the Action Plan for the elaboration of the methodological model of vocabulary control assessment with Study Group monitoring.

For the development of the first phase, named analytical-descriptive study of vocabulary control at the Unesp Institutional Repository, joint meetings of the Study Group with the Executive Coordination and Technical Team were held as a way of obtaining a degree of interaction with the professional reality. We conducted interviews with members of the Executive Coordination and Technical Team to obtain insights into the procedures regarding the activities of the Repository, as well as the need and importance of vocabulary control and the use of controlled vocabularies. Documentation analysis was performed to contextualize and complement the information collected in the interviews.

In the analytical-descriptive study of vocabulary control in the Unesp Institutional Repository, we sought to analyze the relevance of vocabulary control and its use in the Unesp Institutional Repository from meetings with the Study Group as well as through an interview with the Executive Coordination and Technical Team. For the interview, the documentation was analyzed to contextualize the questions and complement the information collected in the interviews.

The interview with the Executive Coordination and Technical Team aimed to obtain an overview on the procedures regarding the activities of the Repository, as well as on the need and importance of vocabulary control and the use of controlled vocabularies. The questions were prepared based on the analysis of: a) the literature and documentation on the Unesp Institutional Repository (RI-Unesp); and from the answers to a questionnaire on indexing policy[2] completed by the RI-Unesp Executive Team. The questionnaire prepared by the Southeast Network Work Subgroup had the theme of indexing policy in repositories. The purpose was to obtain an overview of the procedures regarding the activities of the Repository, as well as the need and importance of vocabulary control and the use of controlled vocabularies. We sought to understand, for further analysis, the institutional context, the reality of the work of professionals regarding the elements, variables, processes and instruments that involve the management of vocabulary control to elaborate a diagnosis of the indexing policy in the repositories.

In the second phase of discussing the feasibility of applying the Action Plan in the Unesp Institutional Repository, the researcher and the Study Group held meetings to analyze and discuss the Action Plan to systematize the methodological model of vocabulary control assessment with the participation of the Executive Coordination and Technical Team. In this way, the results of the analytical-descriptive study of vocabulary control in the Unesp Institutional Repository of the first phase and the discussion of the Action Plan for the elaboration of the methodological model of vocabulary control assessment with monitoring of the Study Group of the second phase, will be presented, respectively, in the following two sections.

4 ANALYSIS OF THE UNESP INSTITUTIONAL REPOSITORY FROM THE PERSPECTIVE OF VOCABULARY CONTROL

In this research, the study of the Unesp Institutional Repository had a special focus and was the environment for the development of the investigation. In order to prepare an Action Plan with the Study Group to systematize a methodological model for vocabulary control assessment in university repositories, an analysis of the Unesp Institutional Repository was initially prepared from the perspective of vocabulary control through the literature on its creation and operation through the analysis of a questionnaire and an interview with the Executive Team.

In order to understand the history of the repository creation and implementation, a review of legal frameworks and publications from the implementation team itself was used for contextualization. The Unesp Institutional Repository, created in October 2013 together with the repositories of the São Paulo University (USP) and the State University of Campinas (UNICAMP), are added to the CRUESP Repository of Scientific Production (Conselho de Reitores das Universidades Estaduais Paulistas).

The repository was created from Unesp Ordinance number 88, February 28, 2013, which established the Unesp Institutional Repository Policy Management Group (GRI-Unesp), responsible for the development, implementation and maintenance of the university repository with the objective of "store, preserve, disseminate and enable open access, as a global public good, to the scientific, academic, artistic, technical and administrative production of the University." (UNIVERSIDADE ESTADUAL PAULISTA, 2013, p. 47, our translation). The General Coordination of Unesp Libraries (CGB) is part of the Management Group and is

[2] Questionnaire provided by the Southeast Network of Institutional Repositories

responsible for the executive coordination of the project. It is responsible for ensuring the inclusion of the production in the Repository with the Technical Team, formed by librarians and professional systems analysts.

For the implementation of the Unesp Institutional Repository, four goals were defined reflecting its objective (ASSUMPÇÃO; SILVA; FERREIRA; BASTOS, 2014, p. 4, our translation):

> 1. inclusion of institutional scientific production published from 2008 to 2012 and indexed on Web of Science;
> 2. inclusion of institutional scientific production published on SciELO journals;
> 3. inclusion of institutional scientific production published from 1976 to 2007 and indexed on Web of Science;
> 4. inclusion of institutional scientific production indexed on Scopus.

The opening of the CRUESP Repository made Unesp aim to include all the university's scientific production in the repository. Thus, the initial goal was to include the production of university researchers indexed on Web of Science and Scopus databases and published on Scientific Electronic Library Online (SciELO) journals. To achieve this goal efficiently, even with the deadline for the opening of the CRUESP Repository and the Unesp Institutional Repository, processes of collection, conversion and automatic import of records referring to this scientific production were used.

The Repository is organized into communities that represent Unesp’s university units and divided into subcommunities that represent the departments and Graduate Programs, where collections are of different types of documents authored by professors and/or students associated with the department or the Graduate program. Likewise, scientific articles and other materials are entered on the Repository only by the Technical Team responsible for this activity. It is not possible to publish on the Unesp Institutional Repository, as it does not edit and does not publish any documents, only those already published in scientific journals, proceedings, etc.

The Repository is part of the open access movement of scientific production and anyone can access and download the Repository documents. There is only one restriction on some dissertations and theses regarding access to the full text when it is restricted during a period chosen by the author (embargo period). Anyone can register in the Repository; the author can subscribe to receive notifications about new documents added to the repository. To do so, one needs to register on the repository's website with their e-mail and follow the instructions on the page.

The activities involving the implementation, maintenance and improvement of the Repository are developed by the General Coordination of Libraries (CGB) through its partnerships with the Distance Learning Center (NEaD), with the Chancellor of Graduate, Undergraduate, Research, Extension and Administration, with Fundação Editora Unesp and Unesp Innovation Agency (AUIN).

Since its creation, the Unesp Institutional Repository has achieved satisfactory results that fully serve the academic community and, in July 2016, it obtained the sixth position in the ranking of *Web of Institutional Repositories*, with national repositories, and the 233rd position in the world ranking of repositories. In 2022, it is one of the five largest repositories in Brazil, according to the February 2022 edition of the Web Ranking of Institutional Repositories, and is among the 24th largest institutional repositories in the world (WEBOMETRICS, 2022).

In March 2022, the repository has 173848 records of which 55% are articles (95420), 17% are master's theses (30248), 9% are doctoral dissertations (16278), 6% are final term papers (10383) and other materials such as conference papers (8902), abstracts (6579), reviews (1984), editorials (754), letters (643), book chapters (550), books (487), patents (408) , podcasts (275), errata (254), notes (209), professorship theses (167), bulletins (78), magazines (74), research data (56), newspapers (34), reports (26) , data papers (13), data management plans (9), biographies (7), educational objects (6), musical score (2), regulation (1) and video (1). In short, the Unesp Institutional Repository satisfactorily serves its community, allowing the dissemination and access to the production developed at the university.

To include the production of researchers from the university indexed on the databases, resources were used to collect, convert and automatically import the production records. For this, automatic collections helped to include records from data sources such as Scopus, Web of Science, SciELO, PubMed, Lattes Curriculum and the Athena catalog that are integrated to the Open Research and Contributors Identification (ORCID) profile of Unesp's professors and researchers. The repository uses ORCID to integrate the scientific production of researchers on the database. Unesp was the first university to use it in Brazil, so the entire academic production of researchers is part of the researchers' record and avoids the rework of filling in and updating their data on other sites.

The Repository is organized into communities that represent Unesp's university units and is divided into subcommunities that represent the departments and Graduate Programs, it has collections of all types of documents from the university community. The general information about the Unesp Institutional Repository contained on the website explains that:

- The used software is DSpace;
- In the left sidebar of the repository there is a search menu with the type of production, document date, author, title, keyword. In keywords, it presents an alphabetical list of 329,103 terms in Portuguese, English, French and Italian, which are present in the documents' metadata.
- The Metadata Format: Dublin Core.
- Type of production: academic and scientific production, administrative production, artistic production, commemorative production - Unesp 40 years, cultural production, technical production.
- Types of materials: it has 28 different types of materials that add up to 173821; the three types with the highest numbers are: articles (95420), master's theses (30237), doctoral dissertations (16270) and final term papers (10375).[3]

To understand the current context of the repository and how professionals develop the functions and operations necessary for its operation, an interview was carried out with the manager and two librarians from the Executive Team of the Unesp Institutional Repository in October 2021. They answered the questions during the recorded interview. According to the explanations of the methodology, 12 questions were formulated, according to Appendix 1. The first part with five questions was formulated from published literature on the Unesp Institutional Repository and its documentation, and the second part, with seven more questions, were based on the answers to the questions in the "Questionnaire on Indexing Policy in Repositories" (analyzed above by the author).

According to the answers to the **Questionnaire on Indexing Policy in Repositories**, we analyzed that:

- The **formation and development of the Unesp Institutional Repository digital collection** involves populating (collection, capture or harvesting carried out automatically or semi-automatically), deposits (document submission) and self-archiving;
- The **repository team** is composed of nine university employees, two analysts, three scholarship students from the Librarianship course, an assistant and three librarians, among them the coordinator who also makes up the Management Committee of the Repository. It currently has 36 librarians from the Unesp library network dedicated to indexing and cataloging tasks.

[3] Dada from February 24, 2022 available at: Repositório Institucional UNESP

- The Repository uses **DSpace platform** and auxiliary software for processing information: Duplicate Checker; Oxygen; Adobe; MarkEdit; LibreOffice. The used metadata standard is Dublin Core;
- The repository has **metadata standardization** for defining mandatory, repetitive and description fields in order to improve the quality of document storage.
- The **profile of the repository's users** is its academic community and also encompasses the external community in general;
- In terms of **indexing practice**, the repository does not carry out an authority control for records migrated from sources external to the repository, nor does it have a written and formalized indexing policy. However, it presents instructions for practical procedures of the subject indexing process and, during the process, the Unesp thesaurus is used as an automatic aid to facilitate the operation;
- Regarding the **quality of indexing**, the level of specificity of the indexing terms is not established in determining the documents' subjects, the repository has an indication of at least three terms or subjects per document;
- Regarding **indexing tools**, the Unesp Institutional Repository does not use automatic validation/correction tools for terms/subjects to ensure correction and consistency of subjects and names (geographic, names of people, identifiers, series and titles), it uses terms/subjects without vocabulary control, in natural language (keywords) combined with the indexing language for thematic representation;
- The repository uses more than one **indexing language for thematic representation** such as list of subject headings, thesauri, subject headings and institution authorities, institution thesaurus, Library of Congress and National Library authorities. Likewise, it does not enrich and maintain the indexing language that can cover the interoperability/semantic compatibility of controlled vocabularies and does not offer a tagging system for indexing texts by users;
- With regard to **assessment of indexing**, currently, the repository does not carry out tests or trials for the periodic assessment of the practice of indexing by retrieval and does not have published reports on this assessment.
- The interviewee reports that the university has a group of librarians who study the Unesp language and indexing who, among their activities, carry out studies on these procedures, however, the group is not linked to the repository team.

Continuing, on the analysis of the answers to the questions of the first part of the interview with the Executive Team of the repository, the manager of the Unesp Institutional Repository (Manager) clarified that:

a) **Indexing policy (for subject validation):**

The indexing policy already established for the Athena catalog can also be used in the future for the institutional repository. Currently, the repository does not share the descriptors assigned in subject validation based on the thesaurus, as there is no catalog interoperability with the repository. The library network has several databases, one of which is the repository and the other is the Unesp catalog.

> *We believe that the validation of subjects from the bibliographic catalog could be used, with the proper updating/inclusion of new terms. But we are not sure about the technical issues, as we need a specific study to carry out tests for this reuse.* (Manager)

On the other hand, the catalog indexing policy needs to reuse and review the keywords chosen by the author, as it is the author's indexing product during self-archiving in the Unesp Institutional Repository and should not be eliminated from the Athena online catalog, but rather go through a validation of new terms for the thesaurus, especially because the catalog does not allow the author to self-archive and determine keywords.

> *So, [...] I believe that now ... the policy is established, I understand that this indexing policy, it is alive, active [...] and then I believe it is about ... activating the group and do, bring these reflections to this group, so that they can also understand this importance of the language used by the user and to actually enrich, I believe, this I think always leads to an enrichment and an improvement in information retrieval, so I actually see it this way, that it would be something in the sense of activating the group and so that we could bring this reflection and update of the policy itself.* (Manager)

b) **Goals of the Unesp Institutional Repository for 2014:**

Libraries and the CGB will not be able to achieve the goals without the political and institutional will of the university. For the repository to really develop well, the university manager's engagement in this project is paramount, as it is a benefit not only for libraries, but for the entire university and the external community.

> *[...] the institutional repository has an ordinance that regulates it, recently, we had an open access policy approved by all, [...], I realize today that we are sought to, for example, to carry out a study of the "impact of the pandemic " so the university recognizes that the repository environment has data and these data speak, recently [...] people are beginning to understand mainly because, because of the fact that it came from the funding agency, for example, which is here, it launches an open access policy and makes it mandatory to have the production stored in the repository, not only the production but the research data, to deliver a management plan, so I also understand that [...] Unesp, before 2013, it had already gone through three repository initiatives and none of them worked, the fourth attempt, which is precisely when Fapesp says to the three state universities "you must implement the institutional repository because we are going to release an ordinance informing the researchers that if they do not deposit, that if the production is not in the repository, they will not receive the funding, the financial resource [...]"* (Manager)

> *[...] in the repository environment we are, because of the infrastructure. This is a little bigger, it goes beyond the management team in the execution part and goes a little beyond into the information technology coordinator, which has now bought space in the cloud so that we can take advantage of and store this content, which is bigger. So, the videos are usually stored and deposited on Youtube and what we do, in fact, is to give the link, describe the video and give the link to where this object is hosted.* (Manager)

c) **Scientific production of Unesp researchers on the Unesp Institutional Repository:**

The example of the Arts Institute was mentioned, which was not usually contemplated by the Chancellor's Office due to its different type of production. However, currently, more repositories comprise artistic and museum works and, in view of the diversity of scientific production that the university has and considering the relevance of the humanities area, it is important to understand, according to the manager, that there will be a growing diversity of document types, even greater than what there is today.

> *[...] so in relation to the Arts Institute [...] we even started conversations because a lot of the materials they produce, we will also need to check the copyright part and they say [...] that regulate their production because they usually produce these musical scores, anyway, and they already sell them, so we need to verify all this part there, right in the case of the scores, that we have here in the repositories [...] they are, if I'm not mistaken, within a community that celebrates Unesp's 40th anniversary where the Institute, the Arts Institute... developed and created the scores for Unesp's hymn, so very likely it is linked to this community, that's it. Videos, for example, we do not store them...* (Manager)

d) Theses and Dissertations

Theses and dissertations represent a significant amount of Unesp's scientific production in the Repository and we are aware that this type of material is also on the online catalog, whose software has been updated and we have been working on technical issues to ensure data migration between both bases of data.

> *At this moment, because we haven't .... changed the Aleph software to the Alma platform, so before, with the Aleph software, we still uploaded ... from one to the other, so if the theses and dissertations are entered in the repository ... this is migrated to the catalog. Now, with the ALMA platform, we have been working on this [...] relationship, so to have in the repository what we also have in Athena [...] we believe that the subject assessment of what there is in the bibliographic catalog, of course ... with the proper updating and inclusion of new terms, we do believe that we can take advantage of it for the repository ... but still and I think that [...] we, in fact, we are still not sure about these technical issues, of what would it involve if we would be able to, how would this use, this reuse would be, so I understand that we need to carry out a specific study, to be able to have this reuse of subject validation that already exists in the bibliographic catalog.* (Manager)

e) **The Unesp Tesauro in self-archiving tutorials:**

Librarians directly act in the process of reviewing records of different types of documents that enter the repository. To carry out this activity, they follow the recommendations of the Verification Tutorial available on the repository's website regarding the use of Unesp Tesauro.

> *[...] so, as we have this tutorial, it has this recommendation to use the following typologies: articles, book chapters, research data, management plan, final term paper (TCC), thesis and dissertation which are currently undergoing review because we will include a part of the justification ... it was asked us to review the ordinance that regulates self-archiving so the theses and dissertations documents are undergoing review and will certainly have this type recommendation [...] recently we opened the TCC self-archiving for example [...] the repository is gaining an increasing dimension of insertion within our works [...] it is not possible to have only one professional to do this validation, so they want to put more people from the team to help in the validation of these records [...] they have a recommendation to use the Unesp thesaurus, that's what they have in our tutorial of verification and validation of records within the repository [...]* (Manager)

In the **second part of the interview**, the Manager answered the questions from the responses to the "Questionnaire on Indexing Policy in Repositories" in order to clarify the following aspects:

a) **Number of professional librarians dedicated to the task of indexing and cataloging:**

There are 36 professional librarians from the Unesp library network who review the records that enter the repository as a result of self-archiving, however, they do not index these records.

> *[...] **these 36 professionals**, in fact, are inserted in the process of reviewing the record that enters the repository and there they have, in the review manual, the instructions for reviewing records, they do have a recommendation that they should consult the… Tesauro Unesp.* (Manager)

> *The [...] professionals are responsible for validating records within the institutional repository and for this activity they use the recommendations described in the tutorial for verification (articles, book chapter, research data, management plan and TCC).* (Manager)

b) **Function of auxiliary software used for information treatment**

Each software has a specific and important function. It is possible, for example, to check for duplication in other procedures that are not only in the repository and import the metadata records from an e-books base into the bibliographic catalog itself:

> ***Duplicate Checker:*** *it is one of the tools of the metadata quality module [...] it shows possible duplicate records within the repository. It is currently used for other projects.*
>
> ***Oxygen:*** *tool for developing and applying style sheets for transforming database records into the format adopted by DSpace.*
>
> ***Adobe Acrobat:*** *software for editing PDF files.*
>
> ***MarcEdit:*** *it transforms metadata from MARC21 to MARC XML. We haven't been using this tool for the repository anymore.*
>
> ***Libreoffice:*** *is an Office package. We use Libreoffice Calc because it can [...] in more effective ways than Microsoft Excel.* (Manager)

c) Metadata standardization

The repository uses Dublin Core for metadata standardization and ISO 639 for language definition.

d) Indexing and indexing policy for the Unesp Institutional Repository

No indexing activity is performed for any type of document and the professionals follow the tutorial for checking self-archiving records, which recommends the use of the Unesp thesaurus to select the descriptors. There is no defined indexing policy for the Repository and the Unesp Tesauro is adopted for vocabulary control for reviewing the records of some document typologies by professionals and during self-archiving by authors. It is not yet possible to include the Unesp Tesauro in automatic or semi-automatic support incorporated into the Unesp Institutional Repository.

> *[...], but we don't have indexing within the repository, we don't have anyone working there [...] the only thing we have [...] they have a tutorial for verification and validation of records, in this tutorial there is a recommendation to use the Unesp thesaurus.* (Manager)

The analysis of the Unesp Institutional Repository from the perspective of vocabulary control obtained through literature, documentation, questionnaire and interview results reveals that vocabulary control is carried out with the recommendation of using the Unesp Tesauro specifically for the tasks of reviewing records by the librarians and during self-archiving by the authors of some document typologies for which specific tutorials are available. The modality with highest number of archived documents on the Unesp Institutional Repository is the automatic populating, without the possibility of indexing by the professional; and the second modality is the self-archiving by the authors whose records are validated by the professionals. In addition, the mediated archiving modality in which the professional prepares the record and indexes using a controlled vocabulary does not apply. The Unesp Institutional Repository does not have an indexing policy described in a manual, but, recently, the Executive Team has invested a significant effort in the elaboration of tutorials especially towards records review for professionals and tutorials towards self-archiving of different document typologies in which the use of the Unesp Tesauro for vocabulary control.

It is noteworthy that the involvement with the Executive Team during the development of this investigation was inspiring for the development of the tutorials. However, guidelines will be necessary for the thematic treatment and standardization of subject metadata that will certainly help in the formalization of an indexing policy towards the specificities of management and operation of the Unesp Institutional Repository. Another point to be observed is the absence of tests or trials to assess vocabulary control, whose results could serve as a parameter to calibrate the valid indicators in the definition of the indexing policy. With this view, we detail below the systematization of the assessment model for vocabulary control in Institutional Repositories based on studies developed by the Study Group with the collaboration and participation of the Executive Team of the Unesp Institutional Repository.

5 SYSTEMATIZATION OF A VOCABULARY CONTROL ASSESSMENT MODEL FOR INSTITUTIONAL REPOSITORIES (IR)

The methodological model for vocabulary control assessment for Institutional Repositories was developed based on the contributions of studies carried out in the Action Plan by the researchers of the Study Group according to the first and second ethnographic methodological guidelines reported in the methodology section. The systematization of the study results proposes methodologies for assessing vocabulary control to be applied with defined objectives in information organization and representation processes carried out in institutional repositories. Therefore, it was necessary to indicate the methodologies proposed in such studies and to identify compatible processes and information organization and representation systems.

The Action Plan, composed of six actions and nine studies, was initially discussed with the Executive Team after a seminar to present the research project, followed by meetings to discuss the main problems and demands of the Institutional Repository. The actions were defined based on the result of these meetings, whose interaction between researchers, catalogers, IT professionals, librarians and the manager required nine study proposals included in six actions, as shown in Chart 1 below:

Chart 1. Action Plan: corresponding actions and studies

ACTIONS	STUDIES
ACTION 1: Assessment of vocabulary control in indexing	STUDY 1: Assessment of subject analysis by authors in self-archiving of theses and dissertations in repositories: observational study with Verbal Protocol;
	STUDY 2: Proposal for standardization of keywords assigned by researchers in the submission of scientific production in different information systems that perform management and dissemination: a proposal for an indexing policy for researchers and authors
	STUDY 3: Indexing policy in institutional repositories in Brazil: diagnostic study in the perception of managers and indexers;
ACTION 2: Assessment of vocabulary control in retrieval	STUDY 4: Assessment of the subject indexing process by retrieval: an analysis of CRUESP's institutional repositories;
ACTION 3: Metadata analysis of theses, dissertations and TCCs in the Repository;	STUDY 5: Metadata of subjects of theses and dissertations in the Unesp library catalog and in the Unesp Institutional Repository: exploratory study on vocabulary control
ACTION 4: Study of the matching of keywords (natural language) and Unesp thesaurus according to Santos (2020):	STUDY 6: Vocabulary control in Unesp Institutional Repository: mapping proposal and matching with the Unesp Tesauro
ACTION 5: Analysis of terminological variations at the semantic, syntactic and pragmatic level using natural language processors.	STUDY 7: Analysis of terminological variations at a syntactic, semantic and pragmatic level in the vocabulary of the Unesp Institutional Repository
ACTION 6: Study of log analysis	STUDY 8: Updated bibliographic review on log analysis and methodology and verification of the influence of log analysis studies to assess information retrieval in repositories

	STUDY 9: Experimental analysis of the database of search logs by subjects of the Unesp Institutional Repository

Source: By the author (2022)

Each study was carried out by groups of researchers from the Study Group with the collaboration of the Executive Team that followed and participated in the development during the years 2020 and 2021. The Unesp Institutional Repository served as a research universe for all studies according to methodological orientation of ethnographic nature.

Actions 1 and 2 are dedicated to vocabulary control assessment whose studies develop specific methodologies in view of indexing and retrieval processes as an object of research. In action 3, the research object is the subject metadata that provides an assessment of vocabulary control standardization. Actions 4, 5 and 6 focus on natural language as an object of research with a view to standardization and matching for continuous updating of controlled vocabularies. The methodologies for vocabulary control assessment in repositories developed by each study are identified by their respective objectives and functions according to Chart 2 below:

Chart 2. Methodologies for vocabulary control assessment in repositories: objectives and functions

Methodologies	Objective	Function
Study 1: Observation of standards and strategies used by authors of theses and dissertations during indexing to assign keywords in self-archiving in institutional repository	Vocabulary control assessment in indexing	Vocabulary Control in indexing
Study 2: Analysis of keywords assigned by researchers for the submission of articles from journals indexed on the Scopus database and on the Unesp Faculty Portal, regarding the standardization and vocabulary control for different functions in information storage and retrieval systems	Vocabulary control assessment in indexing	Vocabulary Control in indexing
Study 3: Diagnostic study of indexing policies in Brazilian institutional repositories through the perception of cataloging-indexing managers and librarians	Indexing policy Assessment	Indexing policy Development
Study 4: Assessment of subject indexing through the retrieval approach with users in the Institutional Repository through searches for subjects in the Unesp Institutional Repository using natural language (keywords of the users' ongoing research) and in a second moment using controlled language (Unesp Tesauro)	Vocabulary control assessment in retrieval	Vocabulary control in retrieval
Study 5: Comparative analysis of the procedures for treating existing theses and dissertations in the catalog and in the Repository to assess vocabulary control	Metadata analysis	Standardization of subject indexing for vocabulary control
Study 6: Mapping and matching of keywords from Unesp Institutional Repository (RIU) records with the Unesp Tesauro, through syntactic matching, considering the processes of equality and similarity between the terms.	Mapping and matching of keywords with Tesauro Unesp	Updating controlled vocabularies
Study 7: Identify and describe the concept and types of terminological variations at a syntactic, semantic and pragmatic level that occur or are likely to occur in the controlled vocabulary of the Unesp Institutional	Analysis of terminological variations	Updating controlled vocabularies

Repository and develop strategies for their terminological treatment.		
Studies 8 and 9: Analysis of user search logs to update controlled vocabularies	log analysis	Updating controlled vocabularies

Source: By the author

In order to establish a sequence of methodology application for vocabulary control assessment in institutional repositories, the contributions of each study were analyzed aimed at their applicability in the Unesp Institutional Repository as transcribed below:

Study 1: To recommend that self-archiving systems include tutorials on keyword assignment with vocabulary control without imposing that they are required to use only the controlled terms. The most current characteristic of the keyword tends to represent more specific subjects within the sciences and, in comparison, the indexing terms of a controlled vocabulary tend to be more stable and to connect to broader subjects, which determines a complementarity between them and does not allow exclusion, but coexistence in a hybrid system of information representation and retrieval.

Study 2: To elaborate an indexing policy proposal for standardizing keywords assigned by authors and researchers in the submission of scientific production in different information systems that perform scientific management and dissemination. To discuss the elaboration of a policy for information organization and representation to be followed, which can be continuously assessed and updated and which provides guidelines to professors/researchers in the standardized attribution of keywords in their bibliographic productions. It is recommended to inform professors of the results of this research, so that they can correct their article keywords, in the light of guidelines for standardization and consistency.

Study 3: To elaborate indexing policy manual and recommend indexing policy elements and variables in Unesp Institutional Repository to improve the thematic treatment with standardization of conduct regarding metadata;

Study 4: To adapt and apply a methodology for assessing indexing by retrieving information with users in institutional repositories; to compare the results obtained between institutional repositories in order to develop an overview of the retrieval situation by subject; to recommend elements of indexing policy in Unesp Institutional Repository;

Study 5: Development of guidelines for indexing theses and dissertations for authors and catalogers using the Unesp Tesauro to carry out vocabulary control;

Study 6: To generate the mapping of keywords with Unesp Tesauro through syntactic matching, considering the processes of equality and similarity between the terms;

Study 7: To identify a set of elements that allow guiding or reorienting the indexing policies used in the vocabulary of the Unesp Institutional Repository and, as indirect benefits, to be verified in the medium term, the increase in the representativeness of such vocabulary from the perspective of its user community.

Studies 8 and 9: To extract by Transaction Log Analysis (TLA) of the natural language used by users to perform searches in three sequential phases: data collection, preparation and analysis.

The analysis of the contributions of each study, carried out in 3 joint meetings of the Study Group with the manager and the Executive Team, was decisive for the preparation of a proposal for systematization and application of the methodologies in three axes described below:

AXIS 1 - Diagnostic study of the indexing policy in the repository: it will begin with the diagnostic study of the indexing policy in the repository (Study 3) with the purpose of obtaining the necessary information on organizational requirements, users and financial resources, as well as elements and variables of information organization and representation. To complete the diagnostic study, the analysis of subject metadata (Study 5) of the documents that enter the repository through the self-archiving modality is indicated in order to ensure a standardization of indexing and define guidelines for the elaboration of guidelines for authors. Still in the diagnostic study, the study of indexing evaluation by the retrieval approach by subjects with users (Study 4) is recommended for comparative analysis of the natural language with the controlled language of the Unesp Tesauro that allows to obtain results regarding the correction, specificity and exhaustiveness in the search system and in the Unesp Tesauro.

AXIS 2 - Development of the repository indexing policy: The three Axis 1 studies will provide results for the elaboration of the Repository Indexing Policy, essential for authors' self-archiving and metadata validation by librarians. The Indexing Policy, therefore, will need to consider two indexing stages: from subject analysis to keyword attribution and from the representation of keywords by controlled vocabulary. For the first stage, the application of assessment methodologies of *analysis of keyword attribution studies* (Study 2) and *observation of standards and strategies used by authors during indexing to assign keywords in self-archiving* (Study 1) to adapt the guidelines in the indexing policy.

AXIS 3 - Elaboration and updating of controlled vocabulary: Tesauro Unesp is the controlled vocabulary to be used for the stage of keyword representation by authors and librarians and, therefore, must be continually updated so that it meets the quality criteria of the indexing, correction, specificity and exhaustiveness. In this sense, studies aimed at updating controlled vocabularies are vital to maintain adherence to the use of vocabulary control during indexing and validation of indexing in the self-archiving modality. The study of mapping and matching of keywords present in the metadata (Study 6) with Unesp Tesauro is carried out through processes of equality and similarity of syntactic matching that can increase the number of relations with authorized terms, in particular the relation of equivalence. The analysis of terminological variations at a syntactic, semantic and pragmatic level (Study 7) that occur in Unesp Tesauro is a study that may help in the elaboration of terminological treatment strategies to include continuous updates. On the other hand, the natural language used by Repository users during the search strategy can be studied by analyzing user search logs (Studies 8 and 9) for continuous updating of Unesp Tesauro with the insertion of new terms that will be selected from the results of a specific log analysis procedure.

6 FINAL CONSIDERATIONS

The proposal for the systematization and application of methodologies for vocabulary control assessment in repositories fundamentally consists of developing an indexing policy whose elaboration is carried out based on 3 axes: Diagnosis, Development of the Indexing Policy and Update of the Controlled Vocabulary.

With this proposal, the manager and the Executive Team assessed the systematization of the methodologies and, depending on the work routines with the Unesp Institutional Repository and the engagement of the Study Group, they defined that initially, in the short term, the study of mapping and matching of keywords present in the metadata (Study 6) the execution is viable and, mainly, that the activity of self-archiving publications and academic works by university researchers has demanded an increasingly specialized vocabulary in each specialty domain. Such demand is increasing, which leads to the generation of new keywords without vocabulary control. In addition, the list of keywords available in the Unesp Institutional Repository presents vocabulary control problems that overburden its use by repository users. From the mapping study and matching of keywords present in the metadata, new terms can be assessed by the Permanent Commission of Unesp Tesauro. The urgency of this study is supported by the fact that there is a controlled vocabulary for use by librarians that is available to users and authors, whose use has been recommended in self-archiving tutorials, which increases the interest in assigning controlled terms both in indexing and in retrieval.

This decision, although important and justified for the absence of a repository controlled vocabulary, it does not exclude the need to carry out the diagnosis and indexing policy to be carried out afterwards. The executive team is small and has only two librarians and a professional from the area of Computer Science and, therefore, the Study Group decided to continue to develop the studies until the final elaboration of the indexing policy with the support of the Permanent Commission of the Unesp Tesauro.

The proposal for vocabulary control and the use of controlled vocabularies in university repositories managed by libraries requires that assessment studies be carried out to systematize methods, procedures, resources and techniques suitable and feasible for the repository context. The repository structure, operation and management present significant differences in relation to other information retrieval systems by allowing the interaction of authors in the self-archiving, the assignment of subjects by the authors, the storage of different types of documents and information resources and, most notably, the possibility of maintaining a hybrid information representation system with natural language keywords and controlled vocabularies. This context requires the elaboration of a suitable indexing policy for repositories that considers all the actors and factors in favor of vocabulary control and terminological richness of natural language specialty.

REFERENCES

AMERICAN NATIONAL STANDARDS INSTITUTE/NATIONAL INFORMATION STANDARDS ORGANIZATION. Z39.19-2005. **Guidelines for the construction, format, and management of monolingual controlled vocabularies**. Bethesda, Maryland: NISO Press, 2005. Available at: http://www.niso.org/standards/resources/Z39-19-2005.pdf . Access on: 6 Mar. 2022.

ASSUMPÇÃO, F. S. *et al*. A conversão de registros na implantação de repositórios institucionais: o caso do Repositório Institucional UNESP. *In*: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 18, 2014, Belo Horizonte. **Anais** [...] Belo Horizonte: UFMG, 2014. p. 1-16. Available at: http://repositorio.Unesp.br/handle/11449/123645 Access on: 6 Mar. 2022.

BARRUONUEVO ALMUZARA, L. *et al*. A study of authority control in Spanish university repositories. **Knowledge Organization**, v.39, n.2, p. 95-103, 2012. Available at: http://www.ergon-verlag.de/isko_ko/downloads/ko_39_2012_2_e.pdf. Access on: 6 Mar. 2022.

HJØRLAND, B. What is knowledge organization (KO)? **Knowledge Organization,** v.35, n.2/3, p.86-101, 2008. Available at: https://bit.ly/3a8YgCm. Access on: 30 May 2022.

FUJITA, M. S. L.; TOLARE, J. B. Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros. **Informação & Informação** (Online), v.24, p. 93 - 125, 2019. Available at: http://www.uel.br/revistas/uel/index.php/informacao/article/view/37985. Access on: 30 May 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 25964-1:2011 **Information and documentation --** Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization, 2011.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 25964-1:2011 **Information and documentation --** Thesauri and interoperability with other vocabularies --

Part 2: Interoperability with other vocabularies. Geneva: International Organization for Standardization, 2013.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2.ed.rev.atual. Trad. de Antonio Agenor de Briquet de Lemos. Brasília: Briquet de Lemos/Livros, 2004. 452p. (Original title: Indexing and abstracting in theory and practice)

LANCASTER, F. W. **El control del vocabulario en la recuperación de información.** 2.ed. rev. Trad. de Alejandro de la Cueva Martín. València: Universitat de València, 2002. (Original title: Vocabulary control for information retrieval; Educació. Materials, 12)

LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. **Association of Research Libraries,** Washington, DC., n.226, p. 1-7, fev. 2003. Available at: https://bit.ly/3NKVBx9. Access on: 6 Mar. 2022.

MAIA, G. Z. A. Pesquisa etnográfica e estudo de caso. *In*: MACHADO, L. M. M. *et al*. **Pesquisa em educação**: passo a passo. Marília: Edições M3T, 2007. p.83-94.

MOREIRO GONZÁLEZ, J. A. **El contenido de los documentos textuales**: su análysis y representación mediante el lenguage natural. Gijón: Trea, 2004.

PANUTO, J. C. **A abordagem do controle de vocabulário nos repositórios institucionais e sua importância para a Arquivologia**. Marília: Faculdade de Filosofia e Ciências, 2021. 63p. (Relatório final de pesquisa IC-CNPq)
SAYÃO, Luís Fernando. Repositórios digitais confiáveis: conceitos, tecnologias e padrões. *In*: FÓRUM DE CIÊNCIA E TECNOLOGIA: REPOSITÓRIOS CONFIÁVEIS DE DOCUMENTOS ARQUIVÍSTICOS DIGITAIS, 2011.**Tópico temático** [...]. Campinas: Unicamp, 2011.

TARTAROTTI, R. C. D. **Avaliação do processo de indexação de assuntos em repositórios institucionais pela abordagem da recuperação da informação**. 2019. 370p. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Marília, 2019. Available at: https://bit.ly/3wYVirU. Access on: 6 Mar. 2022.

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO". **Portaria Unesp nº 88, de 28 de fevereiro de 2013**. Dispõe sobre a criação do Grupo Gestor da Política do Repositório Institucional UNESP (GRI-UNESP). São Paulo: Unesp, 2006. Available at: https://bit.ly/38EAfCQ. Access on: 22 Mar. 2022.

WEBOMETRICS. **Ranking web of repositories.** Available at: https://docs.python.org/pt-br/3/library/json.html. Access on: 6 Mar. 2022.

Appendix 1 - Preparation of interview questions

Objective: to obtain an insight into the procedures regarding the activities of the Repository, as well as the need and importance of vocabulary control and the use of controlled vocabularies

Elaboration method:
- Analysis of the literature and documentation on the Unesp Institutional Repository (RI-Unesp);
- Analysis of the responses to a questionnaire on indexing policy completed by the RI-Unesp Executive Team;

A- Questions on literature analysis and documentation on the Unesp Institutional Repository:
1- "Regarding the indexing policy already established for the Athena online catalog, the Unesp repository manager believes it can also be used in the future for the institutional repository." (TARTAROTTI, 2019, p.162)

Comment on the use of subject validation performed in dissertations and theses by librarians for the Athena catalog;
2- "For its implementation, four goals were defined that reflected the objective of the Repository (ASSUMPÇÃO; SILVA; FERREIRA, BASTOS 2014, p. 4):
1. inclusion of institutional scientific production published from 2008 to 2012 and indexed on Web of Science;
2. inclusion of institutional scientific production published on SciELO journals;
3. inclusion of institutional scientific production published from 1976 to 2007 and indexed on Web of Science;
4. inclusion of institutional scientific production indexed on Scopus."

2.1 After 2014, were other goals set?

2.2 How many goals have been achieved to date?

3 - "In August 2021, the repository has 171337 records, of which 56% are articles (95392), 17% are master's thesis (29389), 9% are doctoral theses (15738), 5% are final term papers (9350) and other materials such as papers presented at an conferences (8902), abstracts (6579), reviews (1984), editorials (754), letters (643), book chapters (549), books (483), patents (393), podcasts (274), errata (254), notes (209), professorship theses (167), bulletins (78), magazines (74), newspapers (34), reports (26), data papers (13), biography (7), educational object (6), data management plan (5), musical score (2), regulation (1) and video (1)." (PANUTO, 2021, p.53)

3.1 Do these figures correspond to the scientific production of professors or part of it?

3.2 Are theses and dissertations all in the repository?

3.3 Why is the Unesp Tesauro not mentioned in the keyword assignment tutorials?

B - Questions elaborated from the answers to the indexing policy questionnaire
The "Questionnaire of the Southeast Network Work Subgroup: indexing policy in repositories" was answered by Flávia Maria Bastos, manager of the Unesp Institutional Repository.

4- The answer that 36 professionals from Unesp are dedicated to indexing and cataloging tasks means that RI-Unesp performs indexing of which types of documents?

5- Explain the function of each of the software used in the treatment of information:
Duplicate Checker:
Oxygen:
Adobe:
MarcEdit:
Libreoffice:

6- What type of metadata standardization is used to define mandatory, repetitive and description fields?

7- Considering that the practical procedures of the subject indexing process are covered by some institution's manual, does it mean that they carry out indexing? For which types of documents? Describe the indexing process carried out and cite the manual that contains it.

8- Unesp Tesauro was mentioned as an automatic or semi-automatic aid to facilitate the indexing process, however, it does not have any semi-automatic or automatic aid incorporated into RI-Unesp. It is possible?

9- Several answers indicate that they follow an indexing policy manual or a service manual, but it is not mentioned, why?

10- Is it allowed to use more than one indexing language, indicated in the answer about the description of indexing languages: do they use 4 languages? For all document types?

11- Comment on the last answer: "The university has a group of librarians who study the Unesp language and indexing that, among their activities, carry out studies on these procedures, but this group and these activities are not linked to the Institutional Repository team"

12- Is there any planning for this to happen in the future?