UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Marília

OFICINA UNIVERSITÁRIA
CULTURA ACADÊMICA Editora

Analysis of indexing language in Covid-19 subject representation

Maria Carolina Andrade e Cruz
Jessica Beatriz Tolare
Mariângela Spotti Lopes Fujita

Como citar: CRUZ, Maria Carolina Andrade e; TOLARE, Jessica Beatriz; FUJITA, Mariângela Spotti Lopes. Analysis of indexing language in Covid-19 subject representation. *In:* TERRA, Ana Lúcia; FUJITA, Mariângela Spotti Lopes (org.). **Integrating Information Science for Sustainable Development:** Topics and Trends. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2025. p. 545-572. DOI: https://doi.org/10.36311/2025.978-65-5954-624-4.p545-572

Analysis of indexing language in Covid-19 subject representation

Maria Carolina Andrade e Cruz [1]
Jessica Beatriz Tolare [2]
Mariângela Spotti Lopes Fujita [3]

Abstract: The study aims to investigate how indexing languages and databases represent the subject Covid-19. The proposal involves analyzing the representative terms available in different languages and comparing them with the subject categories in databases, based on retrieval of the topic. As a methodology, the research is characterized as qualitative and exploratory with a comparative method. The first part of the research involved conducting searches for the topic Covid-19 in databases and the second one in indexing languages. Databases that had subject categories were used as criteria: Global Index Medicus (6,895 results), Unesp (1,138 results), Library of Congress (13,256 results), Publications office of the European Union (542,515 results) and LILACS (15,333 results), in the period between 2020-2023. The selected languages were: Health Sciences Descriptors/Medical Subject Headings, the European Union Controlled Vocabulary and the Library of Congress Subject Headings. The study found that the terms defined by indexing languages have a specificity that may not be suitable for databases; they differ in the sense that subject categories do not present a systematic hierarchical order based on the central subject, they are generic and retrieve several papers with themes that are not exclusive to the health area, whereas indexing languages present a logical sequence and greater specificity of the researched topic.

Keywords: Indexing language. Covid-19. Subject representation.

1 São Paulo State University (Unesp), Brazil.
E-mail: maria.andrade@unesp.br | ORCID ID: https://orcid.org/0000-0001-8307-3448

2 São Paulo State University (Unesp), Brazil.
E-mail: jessica.tolare@unesp.br | ORCID ID: https://orcid.org/0000-0002-8637-7989

3 São Paulo State University (Unesp), Brazil.
E-mail: mariangela.fujita@unesp.br | ORCID ID: https://orcid.org/0000-0002-8239-7114

https://doi.org/10.36311/2025.978-65-5954-624-4.p545-572

Introduction

In mid-2019, a disease caused by a previously unidentified virus emerged. The Coronavirus, or Covid-19, as it was denominated, had its first major outbreak reported in the city of Wuhan, China, from where it spread to several countries. The proliferation of the virus is believed to have occurred due to the sale of wild animals and seafood in a densely populated market where some people began exhibiting severe respiratory symptoms (Wu et al., 2020).

In addition to the pandemic, inequality in access to information, to health and to the possibility of prevention increased. In parallel, some governments discredited the seriousness of the disease, influencing thousands of people not to take precautions and to doubt Science. These attitudes were specifically aimed at not combating the Coronavirus. Given this scenario, the search for a medical solution was essential for people's survival; and the role of information professionals was also intensified in the mission to treat newly discovered information and make it available in an efficient and secure way (Fernandes et al., 2021).

Thus, many renowned scientific content providers have mobilized to make research results available on their platforms free of charge to combat Coronavirus, such as Elsevier, Oxford, Wiley, BMJ, Nature, Sage, Emerald, Cambridge and others (Ali & Bhatti, 2020). The databases of these scientific content providers are scientific dissemination tools, which use quality criteria in the assessment of their indexed journals and were used as a complement in the fight against disinformation, in this pandemic scenario.

In the same way, indexing languages contribute to information representation and retrieval. To this end, the tools need to be in line with the informational resources, with the indexing policy of the environment, and with the needs of the community. Indexing languages need constant maintenance and updating so that subject indexing can be specific enough to represent the features of current topics, such as Covid-19.

Considering the fundamental role of scientific databases and indexing languages in organizing information representation, especially with regard to the proposed theme Covid-19, we sought to investigate how

the subject "Covid-19" is treated in the subject categories of the databases and in indexing languages. This term was chosen because it is common among the terms authorized in indexing languages when referring to the virus and the pandemic, as will be verified in the course of this research. The proposal consists of analyzing the representative terms available in different languages and comparing them with the subject categories assigned in databases, based on the search and retrieval of the topic in order to investigate deeper into the importance and influence they have to be accessible to the population.

Theoretical Framework

One of the main objectives of researchers and the scientific community is to make their research results visible by publishing in scientific journals. Databases, especially the most competitive ones, use quality criteria to disseminate these papers, serving as an important factor in the visibility and dissemination of researchers and scientific journals. Furthermore, they present bibliometric data to evaluate various aspects of scientific production, depending on their scope.

To achieve these goals, a preliminary step should be taken: the organization and representation of these scientifically relevant publications using tools specifically developed to support and assist in this demand. According to Hjørland and Gnoli (2016), the process of knowledge organization involves classification, indexing, and Knowledge Organization Systems (KOS), which are used in the selection of concepts based on the subject analysis of informational resources, with indications of the semantic relationships among these concepts. These systems can include classification systems, subject heading lists, thesauri, and others. They may also be understood as indexing languages, a term used in this study synonymous with KOS (Fujita et al., 2018).

As observed in the studies by Fujita et al. (2018) and Barité (2011), other terms are used in the literature to describe these tools for organizing and representing information, which result from terminological variations and theoretical currents. This paper does not intend to compile these

variations present in each theoretical current, but rather to highlight the term used in this study, "indexing language," which stems from the Anglo-Saxon tradition and whose concept aligns with what we propose, following the definition presented by the researchers below.

According to Mazzochi (2018), indexing languages are responsible for the storage and retrieval of documents. More specifically, Gollub (2011) defines indexing languages as "[...] a specific kind of controlled vocabularies representing formalized languages designed and used to describe the subject content of documents for the information retrieval purposes."

Therefore, in Information Organization and Representation, in particular the indexing process, artificially constructed specialized languages are used to represent, in a standardized way, the concepts contained in documents. Dahlberg (1978) introduces that efforts should be made so that this type of language competently defines the concepts of documents and seek precision in the process. Specialized languages, also referred to here as indexing languages, promote vocabulary control and information organization and representation.

Indexing languages are tools to be used in two moments: during the indexing process and when users search in library catalogs, institutional repositories, digital collections, databases and any other system from which they wish to find information. In this regard, Sarkhel (2017) adds that language should meet three purposes: (i) represent the thematic content of documents; (ii) organize a searchable collection; (iii) and represent the thematic content of user queries during searches. The author believes that an indexing language is effective when there is a match between the way the indexer represents the content and the terms users use in searches. This effectiveness depends on several factors, including the way the collection is organized, the language used, the indexing policy and user awareness.

It is noteworthy that indexing languages are not considered inflexible tools, as they should follow technological, scientific and social advances, and, therefore, require the maintenance and updating of the terms within their composition. This is a considerable fundamental point for document representation, with current topics and their retrieval on the agenda.

International standards establish criteria for the construction and maintenance of indexing languages, such as thesaurus. Two of the most current standards stand out here: American National Standards Institute/ National Information Standards (ANSI/NISO) Z39.19 (R2010) and International Standardization Organization (ISO) 25694 (2011, 2013), the latter is the most recent standard published on this subject.

The two guidelines are introduced in this paper as they address important and complementary definitions of the presented concepts, without being mutually exclusive. The ANSI/NISO Z39.19 (R2010) focuses on vocabulary control, while ISO (25964) focus on the management of the construction and maintenance of thesauri in its first part, and in the second, on the issue of interoperability with other types of vocabulary.

ANSI/NISO Z39.19 (2010) defines indexing language as:

> A controlled vocabulary or classification system and the rules for its application. An indexing language is used for the representation of concepts dealt with in documents [content objects] and for the retrieval of such documents [content objects] from an information storage and retrieval systems. (p.6).

Indexing terms are used to describe the content of information objects based on the intellectual process involved in indexing. The assignment of these indexing terms is derived from an indexing language (ISO 25964-1, 2011).

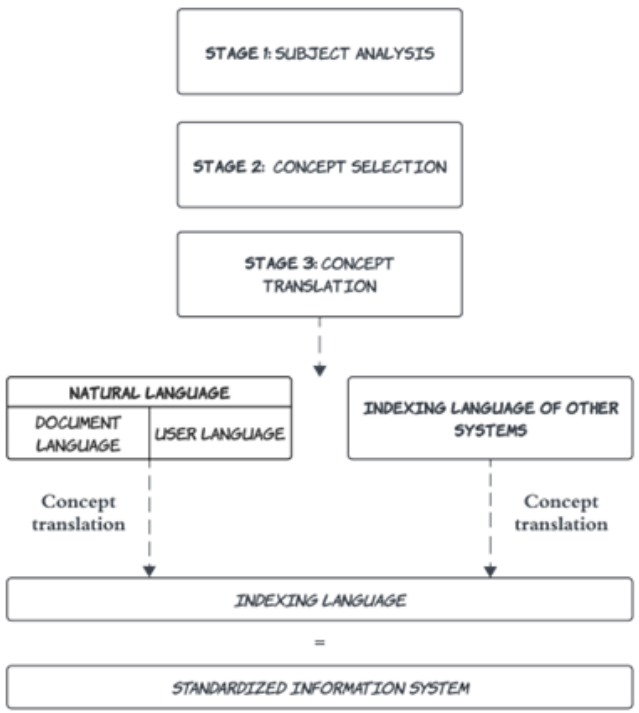
While Sarkhel (2017) points out that the indexing language is an artificially constructed language, composed of expressions which connect several representative terms, its function is to organize semantic content and provide access points for those seeking information, unlike natural language, which can cause obstacles in conceptual representation. The concepts absorb the standard description established by the indexing language and the indexer is responsible for expressing them through terms representative of the extracted essence.

Determining a term to represent a specific concept means establishing vocabulary control, since a single concept can be described in different ways

using different terms (ISO 25964, 2011). Indexing languages aim for this control to avoid discrepancies and misunderstandings in the description of concepts.

Cruz (2017) presents a diagram in Figure 1 to show the use of indexing language within an information system.

Figure 1: Use of indexing language in an information system



Source: Cruz (2017, p.25).

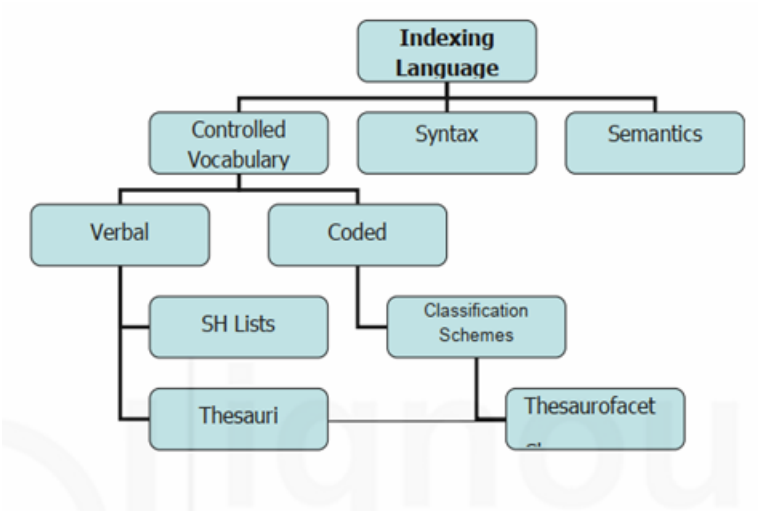
According to Figure 1, the indexing process is divided into its three main stages: subject analysis, concept selection and translation. The indexing language is characterized in the third stage of concept translation. According to the author, the librarian needs to deal with language variables such as: users' language and documents' language, both expressed in

natural language; and other systems' language. The author points out that indexing consistency will depend on the decision of which indexing language will be used, one built by the institution itself or one already published, consolidated and available for consultation.

During the selection or development of an indexing language for integration into a system, the institution should prioritize and scrutinize its quality. Therefore, the institution should observe the composition and development of the structure, semantic and syntactic relationships of the selected language.

Indexing languages, according to Pinto (1985), are made up of vocabulary and syntax. Vocabulary includes the relationship of terms to identify the thematic content of documents, while syntax refers to the rules designed to combine terms, aiming at representing the thematic content of documents. In this sense, Sarkhel (2017) developed a small diagram to show the indexing language structural composition (Figure 2).

Figure 2: Structure of the indexing language



Source: Sarkhel (2017).

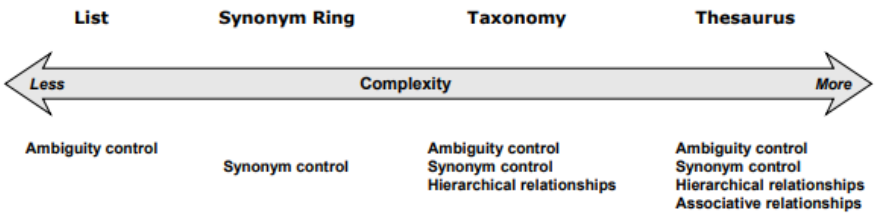
Figure 2 shows the structure of the indexing language. For Sarkhel (2017), it consists of three elements: (i) controlled vocabulary; (ii) syntax; and (iii) semantics. Controlled Vocabulary aims to ensure the relationship between systematically structured terms, which can be verbal (subject heading lists and thesauri) or codified (classification schemes). Syntax refers to a set of rules or grammar that will govern the sequence of words, whether in subject headings or classification notations. It will provide a pattern of relationships, which is recognized between the terms used in the system. And semantics addresses the systematic study of how meaning is structured, expressed and understood in the use of indexing language. Various types of semantic relationships occur, such as equivalence, hierarchical and associative relationships (Sarkhel, 2017).

The indexing language has attributes that play an important role in the effective organization of the inserted collection. For Sarkhel (2017), they are: vocabulary control used to standardize representative terms; coordination of concepts, in which the indexer finds the essence of the document; multiple access; syndetic devices, which is an organizational structure to which related subjects are linked in an underlying classificatory structure; manifestation of relationships to interconnect the terms that make up the language and show the degree of relationship; and structural presentation, so that the structure of the language visibly illustrates the semantic network of concepts and their relationship.

The indexing language is composed as an array of specialized tools for information treatment (Lara, 2004). Pedraza-Jiménez et al. (2009) and Vállez et al. (2015) explain that this allows the use of various tools, including thesauri, taxonomies, ontologies, and lists of authorities, among others.

In Figure 3, ANSI/NISO Z39.19 standard (2010, p.101) presents some types of indexing languages, which are characterized by their degree of complexity in vocabulary control processes.

Figure 3: Structural complexity of controlled vocabularies



Source: ANSI/NISO Z39.19 (2010, p.101).

Figure 3 shows the complexity presented in the structure of controlled vocabularies, according to the ANSI/NISO Z39.19 (2010) standard, from the lowest to the highest degree. Starting from the left to the right, with List that has Ambiguity control, and with Synonym Ring, both have the lowest degree of complexity to control vocabulary. Followed by Taxonomy, with Ambiguity and Synonym Control and Hierarchical Relationships. Thesaurus had one of the greatest complexities, as it addresses Ambiguity and Synonym Control and Hierarchical and Associative relationships.

Regarding semantic relationships, ISO 25964 (2011) also establishes the existence of three types of relationships: (i) hierarchical, in which concepts are presented according to the classificatory order. In this type, there is a correlation between subjects originating from a main class or category, followed by specific and subordinate terms. Hierarchical relationships can be characterized by the acronyms Broader Term (BT) and Narrower Term (NT); (ii) associative, which refers to the association between terms and may or may not have the same meaning. This category aims to relate the terms defined by the language with the term used in the search by the user, and the following acronym can be used to represent Related Terms (RT); and (iii) equivalence, relate to terms with equivalent meanings, though represented by different expressions. In this category, cross-references, non-preferred terms are also established. Their function is to direct the user to preferred terms, based on the term used in the search. The acronyms used for this type of relationship are USE/UF+.

These technical decisions need to be defined according to the planning of the tool and the physical and/or digital collection to be represented. The relationships in indexing languages operate like gears that connect concepts, allowing a logical and systematic understanding by those responsible for maintaining them, as well as users and indexers. This ensures that the indexing language is coherent and effective, facilitating information retrieval and navigation through the contents of the collection.

During the indexing process, indexers can choose the concepts they believe best represent the subject of the document, its essence. Then, they perform a query in the established language to check whether there is a match between the concept chosen by the indexer and the selection of the term that will represent it in the catalog. In this context, some scenarios may arise, such as: (i) the chosen concept matches the defined term; (ii) the chosen concept does not match the defined term, therefore, the term from the indexing language should be used to standardize; (iii) introduction of a new and significant concept; this process ensures that the indexing language remains relevant and comprehensive, according to reality (Tolare, 2021). Fujita et. al. (2018, our translation) adds by explaining that the indexing language serves as a "Commutative code between different linguistic perspectives involved in the documentary system: user, indexer and system, being the main components to obtain the representation and retrieval of adequate information." (p.224).

In addition to well-developed indexing languages, effective systems, and highly qualified professionals, users also play a crucial role in conducting searches. Therefore, there has been discussion about users' information literacy when seeking information from high-quality and reliable sources. Rocha (2022) explains that information literacy is an ability to search, assess and locate the desired information. However, the author emphasizes that if information representation is not solid, even more experienced users will not be able to retrieve information, as the search terms will not correspond to the indexed terms. That said, the fundamental political role played by information professionals is highlighted in promoting tools capable of keeping up with the advances and needs of humanity. This role also corresponds to that of scientific databases. Dantas, in 2002, reported that

in "current" society "information plays a vital role and has great political and economic value. Now, information is considered an economic good or commodity," considering that databases often serve large corporations or institutions focused on an expertise.

According to Franciscatto (2019), databases are essential for the dissemination of consistent and relevant work. When indexed in a database, especially in an established one, these papers are endowed with credibility and recognition of the quality of that scientific research. Therefore, databases are another element that favors scientific production and knowledge dissemination.

Research Methods and objectives

This research is characterized as qualitative and exploratory. To achieve its objectives, the study was divided into: (i) search and retrieval of the theme Covid-19 in databases and indexing languages; and (ii) comparative analysis of the subject categories of the databases and terms of retrieved indexing languages.

Search and retrieval of the theme "Covid-19" in databases and indexing languages

Before carrying out the searches, the following databases were selected: Global Index Medicus (GIM), as it encompasses publications from all continents; Library of Congress (LC), as it is a database of significant value, which has the power to influence other systems; Publications Office of the European Union, as it encompasses publications from the entire European continent; Latin American and Caribbean Literature in Health Sciences (LILACS), as it is specific to the health area; and the service management systems of São Paulo State University (Unesp), as it has an integrated search with publication from CAPES Periodical Portal, and the university's Institutional Repository and Digital Library.

At first, the databases were chosen as they are specifically of the health field and have a significant global influence. The lesser-known

databases were considered for being outside the expected standards as they offer perspectives that help contribute to the analysis.

The aim was for the analysis to encompass all continents, so search systems and databases representing them were chosen: one from Europe, one from North America, one from South and Central America, primarily because the authors are Brazilian, and another more general database with publications from Africa and Asia. Therefore, we could achieve more comprehensive coverage, observing the focus of the publications retrieved in each result.

The criteria used for their selection consisted of: (i) being focused on the health area, encompassing public health and biomedicine, as they are subareas focused on studying the protection of people's health and the study of microorganisms, respectively, and have the skills to develop research on Covid-19 and its impact on the population (GIM, LILACS); (ii) presenting publications with global coverage (GIM, LC, Publications Office of the European Union, LILACS); and (iii) making publications available to users free of charge, making them more accessible. Unesp database was selected as it presents publications from the Brazilian territory, with scientific publications developed by professors, students and researchers belonging to São Paulo State University, and it shows an overview of the reality of the impact of Covid-19 in the country. The links for accessing them are available in chart 1.

Chart 1: Links to access the databases

Selected databases	**Access links**
Global Index Medicus (GIM)	https://www.globalindexmedicus.net/
Library of Congress (LC)	https://www.loc.gov/
Publications Office of the European Union	https://op.europa.eu/en/home
Latin American and Caribbean Health Sciences Literature (LILACS)	https://lilacs.bvsalud.org/
São Paulo State University (Unesp)	https://unesp.primo.exlibrisgroup.com/discovery/search?vid=55UNESP_INST:UNESP

Source: by the authors

GIM database provides access to world literature, focusing on subareas such as public health and biomedicine, which were produced by low-middle income countries. The Library of Congress is considered one of the largest existing libraries, which has a vast collection and provides access to information on a global scale, while the Publications Office of the European Union has publications from countries that are part of the European Union members. LILACS database is specialized in the health area, with scientific and technical literature from more than 26 countries in Latin America and the Caribbean and with free access. Unesp database has an integrated service of a network of the institution's libraries, which provides access to the physical collection, the bases subscribed by the university, the content available on the CAPES Periodicals Portal, the repository and the institution's digital library, with a scientific research production developed by the entire body of students, teachers and researchers.

Indexing languages were also selected for the study as they are constructed artificially, through procedures and aiming at controlling vocabulary, enabling access and ensuring information retrieval. Among them, the following languages were selected for the research: Health Sciences Descriptors and Medical Headings (DeCS/MeSH); European Union controlled vocabulary; Library of Congress Subject Headings (LCSH); and the Macrostructure of the Global Index Medicus itself.

Initially, the criteria for selecting these languages were based on their specialization in the area of health and presenting specific language (Health Sciences Descriptors and Medical Headings - DeCS/MeSH and Global Index Medicus). However, LCSH and the European Union Controlled Vocabulary were also selected due to their importance for global coverage and their ability to influence the development of other systems in other countries. The indexing languages were accessed through Unesp, with free access for all users.

Chart 2: Links to access indexing languages

Indexing Languages	Access Links
DeCS/MeSH	https://decs.bvsalud.org/
European Union Controlled Vocabulary	https://op.europa.eu/pt/web/eu-vocabularies/thesauri
LCSH	https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html
GIM Macrostructure	https://pesquisa.bvsalud.org/gim/decsocator/?output=site&lang=en&from=0&sort=&format=summary&count= 20&fb=&page=1&index=tw&q=%22Covid-19%22

Source: by the authors

DeCS/MeSH is a thesaurus intended to serve as a single language for indexing different materials (articles, books, technical reports, etc.) and to be used in research and retrieval of scientific literature subjects in different information sources (Virtual Health Library, LILACS and MEDLINE). The European Union Controlled Vocabulary is made up of alignments (interoperability projects and tasks in databases); ATTO tables (used to manage different types of tags such as metadata, editorial and graphical user interface content); authority tables and code lists (both were defined to harmonize and standardize codes and labels used by Publication Services and in interinstitutional data exchange); taxonomies (have a unique hierarchical structure and show different types of relationships such as parent/child); and thesaurus (composed of Digital Europa Thesaurus – DET, ECLAS and EuroVoc, intended to represent concepts through representative terms and present their relationships). LCSH is a subject authority database, which encompasses personal names, company titles, meeting or conference titles, uniform titles, general subject titles and specific subjects, such as geographic.

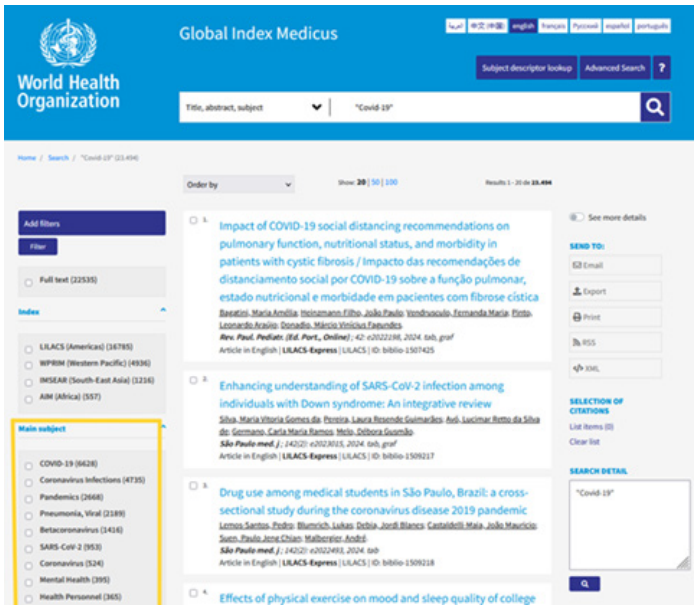
Analysis of database subject categories and indexing language terms

To carry out searches in the selected databases and indexing languages, the term "Covid-19" was used, with a delimited period from

2020 to 2023. This term was chosen as it is the most commonly used among its variations.

From the obtained results, the existence of information that made up "Main subject" or the equivalent in the search refinement area was observed. They aimed at filtering publications by their subjects assigned by the database, as observed in Figure 4. Therefore, to analyze the research, we chose to incorporate this refinement in the sampling, as it allows us to investigate these subjects, the similarities and differences with the term Covid-19 and with the representative terms present in indexing languages. To use in the analysis, we named this sampling as "Subject Categories". These categories can generally be found on the left or right side of the database interface.

Figure 4: Example of subject categories



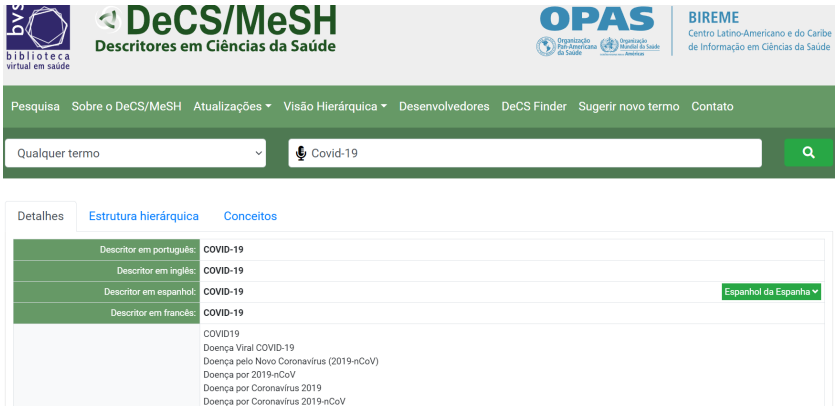
Source: by the authors

These subject categories show the publications assigned to each subject. There is no explicit information about how this process occurs within the

selected databases. However, through it, refining the search and finding even more specific research, as required by the user's needs, is possible. From these, the study established that the first 50 subject categories would be selected, as they occur in descending order, starting from the largest to the smallest number of studies assigned to a subject category.

While indexing languages present different dynamics on their platforms, for the analysis, the chosen terms were those that made up the field "Entry term(s)", as they address specific terms, and users can start the search by them or are redirected to them. An example of the search performed by the selected term in the indexing language is shown in Figure 5.

Figure 5: Example of searching the term covid-19 in indexing language



Source: https://decs.bvsalud.org/ths/resource/?id=59585&filter=ths_termall&q=covid-19

The example shown in Figure 5 presents the results of the search using the term Covid-19 in the DeCS/MeSH indexing language. It can be observed in its details, in the form of a hierarchical tree structure and through its concepts. This type of structured language enables the analysis of the relationships between different types of representative terms, which originate from a central theme. Its construction took place through

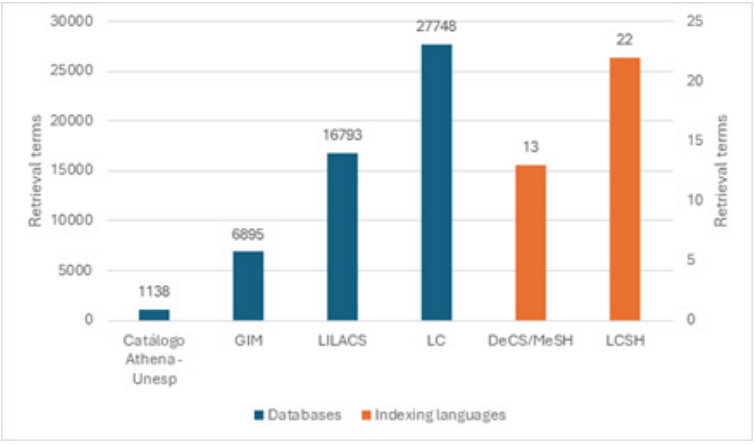
specialist professionals to ensure that the information has an adequate representation and effective retrieval during searches.

Findings and Discussion

The findings are initially presented by the subject categories found in the databases and subsequently, the results found in the analyzed indexing languages.

Some of the analyzed systems present "subject categories" in their search results, which are also described as "keywords", "subjects" or "main subject". These categories are generated from the works that make up the base, where they are grouped by similar themes. The results were organized from the first 50 terms in order of assigned results, in other words, from the categories with the most assigned research to the categories with the fewest studies, carried out manually in a ranking table, created in word.

Graph 1: Terms retrieved from databases and indexing languages



Source: by the authors

The graph shows the number of terms retrieved from databases and indexing languages. In total, 52,574 terms were retrieved in the databases,

of which 27,748 were from LC, followed by LILACS with 16,793. The smallest number is from GIM with 6,895 and Catalog Athena - Unesp with 1,138. In indexing languages, a total of 35 terms were retrieved, divided into DeCS/MeSH (13) and LCSH (22).

Next, the data found in the two defined areas will be presented and discussed: the databases and the indexing languages.

Databases

Global Index Medicus database retrieved 6,895 results with the term Covid-19, including publications of papers presented at conferences, articles, theses and dissertations. Publications Office of the European Union database includes documents retrieved from laws, rights, publications, summaries of legislation, web pages and the official EU contact list. The Latin American and Caribbean Literature in Health Sciences database, retrieved 16,793 papers. All available formats were considered at the Library of Congress (LC), both online and printed. The search for the term "Covid-19" retrieved 27,748 results. Initially, all the searches were generally conducted in all fields. The results were then filtered by subject.

The analyzed databases presented different results in relation to the most incident subject categories. It was noted that in the Publications Office of the European Union and in the Global Index Medicus, the most common category of results was not the term properly defined for the search, Covid-19, but rather "EU Member State" and "Coronavirus", respectively.

Data from the Publications Office of the European Union refer to documents published by European Union entities. Therefore, they present decisions, reports and communications carried out by the European Communities and the European Union; these results were publications on the Coronavirus. In the Global Index, the most frequently found results stand out for the categories that involve: health consequences of the virus (pneumonia, coronavirus infection, Severe Acute Respiratory Syndrome, respiratory issues, mortality, etc.), public health decisions (Public Policy,

Unified Health System, Quarantine, Public Health, Social Isolation, Personal Protective Equipment), human behavior in the face of the pandemic (Mental Health, Occupational Health, Social Behavior, Stress, Psychological, Social Work), studies on the virus (Betacoronavirus, SARS-CoV-2, Coronavirus) and medical procedures (Nursing, Primary Health Care, Pediatrics, Thorax, Aged, Patient-to Professional, hospitalization).

Unesp is a public university and its collection covers all areas of knowledge. Regarding the theme Covid-19, the category presenting the highest number of publications was entitled Covid-19, with papers published by professors and students at the University. However, we found that the subject categories also identified a high incidence of topics focused on technology (Artificial Intelligence, Application Software, Computer Networks, Computer Communication Networks, Computers, Software Engineering, etc.). We suggest that this result showed the efforts of areas other than health to contribute to the migration of the teaching environment that needed to happen, from physical to digital. Universities, as well as several private companies, schools and other institutions needed to invest in technologies to continue operating remotely. Adetayo (2023) emphasizes that before the COVID-19 pandemic, emerging countries widely used traditional research teaching methods in their university libraries. Subsequently, they were forced to quickly find new ways to continue operating and serving both the scientific community and the general population through online education. This work of training and continuous guidance to researchers is understood as crucial for enabling researchers to access the desired information. We highlight here the use of databases and indexing languages, which are elements that information professionals should be familiar with to assist in the process of organizing and retrieving information.

Lilacs presents categories consistent with the profile of its database, which is specifically focused on health, presents subject categories for topics such as: health consequences resulting from the virus (Coronavirus Infections, Viral Pneumonia, Severe Acute Respiratory Syndrome , Neoplasms), studies on the virus (Betacoronavirus, SARS-CoV-2, Coronavirus), public health policies (Social Isolation, Contingency

Plans, Personal Protective Equipment, Immunization Programs, Physical Distancing), hospital action measures (Epidemiological Monitoring, Intensive Care Units, Primary Health Care, Nursing Care, Biological Risk Containment), however, one of the subject categories that stood out in relation to other databases was "Telemedicine", which became a reality and allowed to serve people regardless of their location. Lilacs is considered to be similar to the Global Index Medicus, despite GIM database is not exclusive to a single area. It is noteworthy that Lilacs and Global Index Medicus use DeCs/MeSh as their indexing language, which may contribute to their similarity in subject categories, even if the indexing language and subject categories are distinct elements. According to Lancaster (2004), satisfactory search results in databases can be considered "useful" and "relevant" when they meet the user's needs. This can be linked to the categories of subject with exhaustive indexing, as pointed out by the author. Similarly, GIM uses a criterion of comprehensiveness in its subject categories to encompass more papers on a specific topic, while Lilacs uses them more specifically to meet the needs of health researchers more pertinently. Issues such as exhaustiveness and specificity in indexing should be defined in an indexing policy when belonging to an information system (Lancaster, 2004; Fujita et al., 2016), comprehensiveness in the coverage of concepts laid out in the structure of indexing language (ISO 25694-1, 2011). Here, we can associate these same aspects with the criterion of comprehensiveness used by each database.

Library of Congress retrieved a large number of works with the proposed central theme and the presence of categories that address American cities and the country itself was perceived as a factor that distinguishes it from other databases (United States, New York (State), Bronx, Queens, Brooklyn, Manhattan), the way in which this period of human history was portrayed (Digital Photographs, Time Lapse: Covid-19, Blogs, Fashions, Newspapers, Portraits), in addition to the government measures and laws taken to combat the virus (Face Masks, Law, Law Library, Government, Legal Notice).

Indexing languages

The selected indexing languages were: Health Sciences Descriptors / Medical Subject Headings (DeCs/Mesh); the European Union Controlled Vocabulary - EU VOCABULARIES; and the Library of Congress Subject Headings (LCSH). In DeCs/MeSH, the search for the term Covid-19 retrieved 13 results. However, the results showed that the preferred term has 60 cross-references with terms considered to be alternative or non-preferred which have no vocabulary control.

In the indexing language of the Library of Congress Subject Headings, 22 results were found and the term closest to the one searched was Covid-19 (Disease) with nine cross references. Research carried out in the European Union Controlled Vocabulary - EU VOCABULARIES demonstrates that its structure only presents two cross-referenced terms.

In the analysis of indexing languages, the study found that DeCs/MeSh retrieved 13 results that addressed Covid-19, with LCSH bringing 22 results in the search. As expected, DeCs/MeSh brought specific descriptors from the health area (Serological Test for COVID-19, Acute Post-COVID-19 Syndrome, Serotherapy for COVID-19, ChAdOx1 nCoV 19) and in the term "Covid-19", it presented 60 cross-references. This is to standardize all the terms that have emerged to address the same subject. The importance of a tool such as an indexing language that directs both the indexer and the user when choosing the ideal term for their representation and/or search is thereby observed. Cruz and Fujita (2021) highlight that making the indexing language available to the user can contribute to information retrieval by allowing for the development of a search strategy that is more coherent with each database used. This approach ensures that the terms used in indexing and searching are consistent.

LCSH addressed more social and comprehensive themes in its structure (COVID-19 Pandemic, 2020- , in art; COVID-19 (Disease) (Islamic law); COVID-19 Pandemic, 2020- , in motion pictures; COVID-19 Pandemic, 2020- , in mass media, COVID-19 Pandemic, 2020- , in popular culture; COVID-19 Pandemic, 2020- , in literature). This is due to the fact that LCSH is not a specialized indexing language

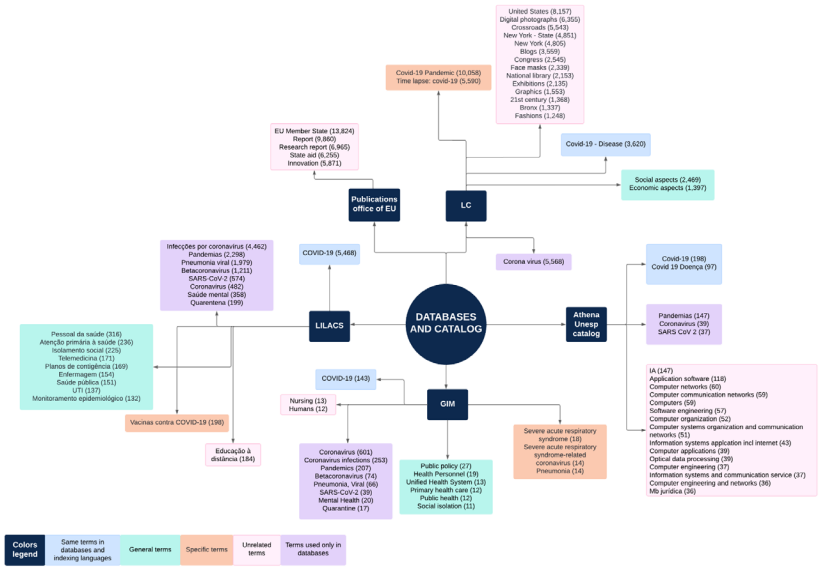
in one area of knowledge, but seeks to serve other types of information resources. The proposed descriptor "Covid 19" had 9 cross-references that presented related terms written differently, considered synonymous (Coronavirus disease-19, SARS coronavirus 2 disease, SARS CoV-2 disease).

In EU VOCABULARIES, the term Covid-19 is in a structure with three general terms attached to it "infectious disease; illness; health policy" and with only two cross references: "NextGenerationEU" and "recovery from the coronavirus pandemic", demonstrating that it is not an indexing language that brings specificity to areas related to health, but rather focused on the social aspect.

Therefore, subject categories differ from indexing languages, especially if we conduct a thematic approach as in this research. When analyzing the theme Covid-19, we noticed that subject categories do not present a systematic hierarchical order based on the central subject. Several papers with related themes are retrieved, but not necessarily in the health area, such as in indexing languages that present specificity of the researched topic and relate the variations of the term, represented by the cross-references and standardizing them in a single descriptor.

Chart 3 is a diagram that presents a list of broader and narrower terms that are unrelated or that were only assigned by databases or indexing languages.

Chart 3: Results in databases and indexing languages



Source: by authors.

The scheme in Chart 3 shows how the terms are related in databases and indexing languages. Each color corresponds to a different category, such as broader or narrower terms; unrelated terms; terms that were assigned only by databases or indexing languages; and the terms that were assigned by databases and indexing languages.

The term "Covid-19" was the only one assigned by both indexing languages and databases. Furthermore, the broader terms are related to the health area, such as "Social isolation" or "Public policy", while the narrower terms, in orange, are directly linked to the term "Covid-19". Most of the terms assigned by the databases are unrelated terms, such as "Software engineering", "Computers" or "blogs", which demonstrates a certain concern regarding the quality of the assignment of subject categories and keywords in the studies. Furthermore, the terms defined by the indexing languages have a specificity that may not be suitable for databases specific to the health area due to the lack of coverage corresponding to what the specificity of the area requires. In parallel, it is relevant to reflect on the

coverage of terms assigned by these languages, such as "Coronavirus-19 Disease". Such reflection can extend to how these terms are reused by information units such as libraries to thematically represent their collection.

The indexing language used by Global Index Medicus is DeCs/MeSh, which is available for user query in the Institution's online catalog.

Conclusion

As highlighted in the introduction, indexing languages require periodic maintenance to remain updated so that subject indexing can accurately represent specific topics, such as "Covid-19" in this case. The research problem consisted of investigating how the term is treated within the subject categories and indexing languages analyzed. We concluded that the subject categories presented in the database interfaces have distinct characteristics: they are comprehensive, where there is no intention to specify the subject; they are linked to subjects not evidently related to Covid-19 and the health field; and there are more specific categories that directly address issues related to the researched subject. These subject categories function as a resource for categorizing databases to gather documents. However, they are not tools that adhere to a specific policy or are used in a subject analysis process, such as indexing. They are developed and made available to act merely as search filters.

Indexing languages adhere to the purpose defined by their responsible institution, and the databases that utilize them are aware of the area in which each language operates, as well as its degree of specificity and coverage of topics. This can be observed in the results from the LCSH when searching for the term "Covid-19", which yields general terms compared to DeCS/MeSH, a language specialized in Health Sciences, and which includes numerous cross-references (60) to describe the concept of Covid-19. EU VOCABULARIES, despite presenting the chosen term, brought a more political and social bias. Again, this reflects the purpose of each indexing language and its objectives in representation

We can consider that the research proposal, to analyze the representative terms available in different languages, such as in English and Portuguese, and to compare them with the subject categories assigned in databases, was specific within the scope of studies on indexing languages, subject representation, and information retrieval. This is because the investigation aimed not only at one information organization tool but at two: indexing languages, developed with a solid, predefined purpose used for indexing and information retrieval, and subject categories, used as a resource for grouping common documents. Additionally, we noticed that both have different roles within information organization. Indexing languages seek to standardize the terms used both in the representation and in the search strategy for valuable information for users. Subject categories comprehensively bring together how a subject is related to others through a qualitative perspective of the works indexed in the database. The topic Covid-19 was chosen precisely because of its importance for the health area and for public and governmental interests.

With that, Libraries and information systems play a fundamental role in building scientific knowledge through information processing. The role of librarians as indexers is politically necessary to combat misinformation through adequate representation of informational content. Indexing languages and databases are part of this gear in the precise representation of informational resources. Therefore, it is evident that databases are rich information environments that require the functionality of their interfaces, considering a facilitative aspect in website navigation, with the organization of information present within them and the use of consistent indexing languages to coherently represent the subject addressed in the types of documents provided by each one.

Reflecting on the emergence of new threatening diseases and the technological evolution we have been witnessing, we consider that although we are currently transitioning to new technologies, such as the progressive establishment of artificial intelligence (AI) which can significantly contribute to the information organization and representation in information systems, such as databases, this does not eliminate the need for the improvement of indexing languages through their constant

updating. This is especially true as we are at the beginning of a new technology that is not yet widely known or available to everyone.

During the Covid-19 pandemic, it became evident that the tools managing and providing information, along with scientific databases and information systems, are key elements in combating misinformation and can be crucial for the success of research that significantly impacts human life. The topic of Covid-19 was chosen due to its importance to the health area, as well as public and governmental interests. Therefore, this study aims to convey the scope of the tools that establish vocabulary control and their usability in databases. Through the use of subject cross-references, it is possible to group the most commonly used terms to describe a single concept, thus avoiding discrepancies in search results and facilitating the researcher's work.

On this matter, it is important to emphasize the role of information professionals in empowering their users about the use of available resources for navigating and searching their databases, whether through tutorials, asynchronous training, or other methods. Information representation, as an area of interest and dedication in the organization of information, should solidify knowledge and successful practices to evolve and enhance information systems.

References

Adetayo, A. J. (2023). Post Covid-19 pandemic and library users' education: Impact on examination and survey. *The Journal of Academic Librarianship, 49*, 102695. https://doi.org/10.1016/j.acalib.2023.102695

Ali, M. Y., & Bhatti, R. (2020). COVID-19 (Coronavirus) pandemic: information sources channels for the public health awareness. *Asia Pacific Journal of Public Health*, *32*(4), 168-169. https://journals.sagepub.com/doi/full/10.1177/1010539520927261.

American National Standards Institute/National Information Standards Organization – z39.19-2005. (2010). *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. NISO Press.

Barite, M. (2011). Sistemas de organización del conocimiento: una tipología actualizada. *Informação & Informação*, *16*(3), 122-139.

Cruz, M. C. A. (2017). *Linguagens de indexação no contexto da política de indexação:* estudo em bibliotecas universitárias. [Trabalho de conclusão de curso]. Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, Brasil.

Cruz, M. C. A., & Fujita, M. S. L. (2021). O uso de linguagem de indexação por bibliotecas universitárias brasileiras. *Informação & Informação*, *26*(1), 2021.

Dahlberg, I. (1978). Teoria do conceito. *Ciência da Informação*, *7*(2), 101-107.

Dantas, M. (2002). *A lógica do capital-informação: a fragmentação dos monopólios e a monopolização dos fragmentos num mundo de comunicações globais* (2a ed.). Contraponto.

Fernandes M. R., Freire Junior, A. M., & Souza, A. D. (2021). Atuação do bibliotecário clínico em tempos de pandemia da covid-19. *Revista Brasileira de Biblioteconomia e Documentação*, *17*, 1-20.

Franciscatto, R. (2019). *Base de dados científica*. Prof. Dr. Roberto Franciscatto. https://www.franciscatto.com.br/bases-de-dadoscientificas/#:~:text=Os%20 trabalhos%20presentes%20em%20uma,sua%20pr%C3 %A1tica%20a%20 ser%20desenvolvida.

Fujita, M. S. L, & Gil Leiva, I. (2016). Avaliação da indexação por meio da recuperação da informação. *Ciência da informação*, *41*(1), 50-66.

Fujita, M. S. L., Moreira, W., Santos, L. B. P., Cruz, M. C. A., & Ribas, R. R. de B. (2018). Construction and evaluation of hierarchical structures of indexing languages for online catalogs of libraries: an experience of the São Paulo State University (Unesp). *Knowledge Organization*, *45*(3), 220-231.

Fujita, M. S. L., Santos, L. B. P., & Alves, R. V. (2018) ¿Son los lenguajes de indización y documentales sistemas de organización del conocimiento?: Un análisis bardiano de la variación terminológica. *Scire, 24*(2), 23-33.

Golub, K. (2011). *Knowledge Organisation Systems. Technical Foundations, UK*(6). http://technicalfoundations.ukoln.ac.uk/subject/knowledge-organisation-systems.html

Hjørland, B., & Gnoli, C. (2016). Knowledge organization. *Knowledge Organization, 43*(6), 475-484. Também disponível em Hjørland, Birger, ed. *ISKO Encyclopedia of Knowledge Organization*, https://www.isko.org/cyclo/knowledge_organization

International Organization for Standardization. (2011). *ISO 25964-1: Information and documentation - Thesauri and interoperability with other vocabularies - part 1: Thesauri for information retrieval*. Genebra.

International Organization for Standardization. (2013). *ISO 25964-2: Information and documentation: Thesauri and interoperability with other vocabularies: Part 2: Interoperability with other vocabularies*. Genebra.

Lancaster, F. W. (2004). *Indexação e resumos: Teoria e prática* (2a ed.). Briquet de Lemos.

Lara, M. L. G. de. (2004). Linguagem documentária e terminologia. *Transinformação*, *16*(3), 231-240.

Mazzocchi, F. (2018). Knowledge organization system (KOS). *Knowledge Organization 45*(1), 54-78. https://www.isko.org/cyclo/kos

Pedraza-Jiménez, R., Codina, L., & Rovira, C. (2009). Metadatos en la web semántica: lenguajes de marcado para la organización de sistemas de información. In L. Codina, M. C. Marcos, & R. Pedraza-Jiménez (Ed.), *Web semántica y sistemas de información documental* (pp. 13-42). Trea.

Pinto, M. C. M. F. (1985). Análise e representação de assunto em sistemas de recuperação da informação: linguagens de indexação. *Revista da Escola de Biblioteconomia da UFMG*, *14*(2), 169-186.

Rocha, F. M. S. (2022). Análise da produção científica sobre competência informacional no contexto da Ciência da Informação no Brasil. *Revista Ibero Americana da Ciência da Informação*, *15*(1), 52-75.

Sarkhel, J. (2017). *Indexing languages.* Indira Gandhi National Open University.

Tolare, J. B. (2021). *O uso de linguagem de indexação na representação temática de livros em bibliotecas universitárias: Observação com Protocolo Verbal Individual* [Dissertação de Mestrado]. Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília.

Vállez, M., Pedraza-Jiménez, R., Condia, L., Blanco, S., & Rovira, C. (2015). Updating controlled vocabularies by analysing query logs. *Online Information Review*, *39*(7), 870-884.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhaozhong, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Wenjie, T., Cheng, G., & Jiang, T. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe*, *27*(3), 325-328.