Thematic article

Evaluation of Indexation Consistency in Publisher Subject Metadata

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

Recommended citation:

Tolare, J. B., Cruz, M. C. A., & Fujita, M. S. L. (2022). Evaluation of Indexation Consistency in Publisher Subject Metadata. *Central European Journal of Educational Research, 4*(2), 28–34. https://doi.org/10.37441/cejer/2022/4/2/11394

Abstract

This work aims at evaluating the indexing of subject metadata published on the MercadoEditorial.org platform. The goal: to verify if the indexing is consistent. First, analysis was done of the tools available on the platform for assigning keywords, and, afterwards, publishers on the platform were documented and verified by running intrinsic evaluation for interconsistency. This was all done to compare the indexing of one work by four authors. The authors were Ciranda Cultural, IBEP, Excelsior Editora, and Via Leitura. The chosen book was “The Alienist” by Machado de Assis, a classic in Brazilian literature. The result of the indexing analysis was that publishers gave keywords such as title, author’s name, characters’ names, names of other books, and excessively used and repeated words. The last category was further broken down being with or without accentuation, being in the singular or plural form. Other terms were assigned that were related to university entrance exams. Thus, it can be concluded that an absence of vocabulary control can make retrieval of a work difficult, simply by assigning terms that inadequately define the subject of the book, and by lacking semantic, syntactic, and morphological standardization among the terms.

Keywords: subject metadata, publishing market, indexing evaluation

Introduction

The publishing market is characterized by its competitive nature in the face of the commercialization of its products, as it strives to gain profits and relevance. Libraries too are competitive; although, rather than seeking monetary success, they aim for the successful dissemination of knowledge. This dissemination is made possible through databases, catalogs, institutional repositories, and any other information system, organized through pre-established standards or procedures.

For this purpose, information organization is a hot-button issue in both the theoretical and practical discussions on Library Science. The ultimate goal of these discussions is to find the best ways to help users access and retrieve information. In this way, processes such as cataloguing and indexing have been developed over the years to more efficiently organize information, which can occur both in libraries and in the publishing market itself.

When running these processes in the publishing market, metadata are used to bibliographically record its publications, to manage, control, and share data, to organize and retrieve information, and to smoothly interoperate between systems; thus, characterization is done in the process of organizing information.

Alves (2018) explains that using metadata is important to the development and use of standards that meet the demands of representation. Metadata and metadata standards aimed at the fields of publishing and book sales can be considered part of the information community, as they appear in the organization of information, and the attribution of terms in representation (Riley, 2009; Alves, 2018). Zeng and Qin (2016), categorize these

[1] Faculty of Philosophy and Science, São Paulo State University (Unesp), São Paulo, Brazil, jessica.tolare@unesp.br

[2] Faculty of Philosophy and Science, São Paulo State University (Unesp), São Paulo, Brazil, maria.andrade@unesp.br

[3] Faculty of Philosophy and Science, São Paulo State University (Unesp), São Paulo, Brazil, mariangela.fujita@unesp.br

more generally, that metadata and their standards are meant for print or digital publication in filling in information. According to Alves (2018, p. 11), the standards of publisher, book merchant and bibliographic metadata have similarities and differences, each one coming from a specific domain, whose objective is to "meet the needs for representation in each domain". The function of editorial and book merchant metadata standards is to help describe the resource, making it easier to identify and find on the web. This is makes it easy to access, to register publishing cost information, and to detail its subject matter. In contrast, bibliographic metadata describe and store information about a resource, while simultaneously allowing access, identification and acquisition of resources, physically or digitally. Therefore, publishing, merchant, and bibliographic metadata standards are considered rich, well-structured and standardized by external coding schemes (Alves, 2018).

Zeng and Qin (2016) pointed out relationships that can be established in these domains (publisher, book merchant and bibliographic domains) such as: benefits available to libraries and cataloguing agencies due to collections of previously described basic metadata; easy price comparison when making purchases, streamlined procedures for assigning identifiers (ISBN, ISSN, DOI) standardization of subjects, literary genres, and authors.

In a study developed by the American Library Association (ALA), it was found that the same metadata developed to help increase bookstore sales were also important to libraries and their users (Baca, 2016). This research enabled the integration of and interoperability among publishing, merchant, and bibliographical standards used by the library. This was significant, since publishing standards offered additional metadata sets that would enrich the bibliographical record for users (Alves, 2018). Baca (2016) points out that metadata records prepared by publishers and merchants can improve the bibliographic records used by libraries.

According to Publishnews (2016) and Alves (2018, p. 3), publishers and book merchants face similar challenges in relation to metadata and the diversity of schemas that represent publications. In many cases, publishers provide the metadata of their publications in their own standards or in standards specific to each merchant. In this process, distribution errors can occur in the metadata, causing problems from delays in publishing to delays in sales and marketing. At the same time, book merchants need to manage a large amount and variety of metadata from different publishers, which often lacks standardization. As a result, inconsistencies in the categorization of materials arise in the catalog. In the case mentioned in our study, both authors cited considered the representation process in the attribution of metadata from publishers and book merchants to be challenging.

It is clear to see that libraries and the publishing market work directly with information, metadata and users who share common processes, systems and objectives, even if the end goal of each might differ. From this perspective, according to Alves (2010), metadata are interconnected with cataloging and indexing, since descriptive information is presented for cataloging (title, author, date, abstract and keywords, etc.) while for indexing, subject terms are presented (descriptors and thematic keywords assigned by the professional responsible for the process).

The relationship between indexing and metadata is in metadata standards, such as MetaTags, whose purpose is to establish the location and retrieval of information resources on the Web. MetaTags helps the indexing process by utilizing search and information retrieval tools, selecting the terms that best represent the material. Though this prevents the material from being indexed in its entirety, the process still moves forward successfully (Alves, 2010, p. 91). The author highlights the fact that vocabulary control patterns can be: thesauri, authority lists, controlled lists, lexicons, classification schemes, and a list of authorized terms among others. This is because these controlled vocabularies will determine the data values, helping to establish relationships between attributes and entities, generating and building more consistent information representations.

In Brazil, an initiative was launched with the creation of the startup MercadoEditorial.org, which aims to develop activity in the field of metadata and seeks to contribute to libraries, publishers, book merchants and users in general. This would be accomplished by generating the sharing and dissemination of free book-related metadata. This study was entirely dedicated to MercadoEditorial.org, which we chose for analysis due to the large number of important Brazilian publishers using the platform. The following were listed as publishers on their platform: Federal University of Rio de Janeiro, The State University of Campinas, Federal University of Paraná, Federal University of Visçosa, São Paulo State University, Publisher of São Paulo University (university publishing houses), Editora 34, Darkside Books, Autêntica and Gutenberg Group (independent users), among others. In addition, the platform is free, making access to and sharing of data very simple.

From this observation on the impact that metadata can have on the representation and retrieval of information in other merchant environments like libraries and others, this work aims to 1) investigate the attribution of subjects in publisher metadata, that the indexing of subject metadata at MercadoEditorial.org 2) might be evaluated. These two steps are necessary to verify the consistency of indexing regarding the use of metadata of shared subjects.

Research design and Methods

The methodological procedures of this research are defined according to the proposed indexing evaluation.

Despite being studied first in the 1950s and being a current hot-button issue, topic evaluation in indexing is not so common or readily available on the publishing market. Therefore, in light of this new universe of research to be investigated, this research is characterized as exploratory.

This item defines the concepts assigned to the evaluation of indexing and the procedures that involve the application of this method, especially in the intrinsic evaluation aimed at inter-consistency in indexing.

The evaluation of subject indexing is a methodology that aims at verifying the quality of the indexing process and its final result. From the evaluation of indexing, it is possible to work with levels of exhaustiveness, specificity, correction and consistency. In this work, the focus of the discussion is on the consistency of indexing by the publishing platform that makes subject metadata available.

According to Gil Leiva et al. (2008) the consistency in indexing is an element that characterizes the degree of similarity in the representation of information in a document through terms defined in the thematic treatment by one or several indexers and, from that, a consistency index is defined.

For the preparation of his research, Gil Leiva (2008) adapted Hooper's formula (1965). Over the years, the author has delved into producing studies that focus on the use of technologies in the indexing process, which can be characterized as automated indexing. Gil Leiva published many works (1999; 2001; 2002; 2008), but the real breakthrough came in a study focused on the development of instruments and methods for the evaluation of indexing using a mathematical formula.

Hooper's (1965) study is considered a precursor in methods to check the consistency level of indexing. The author developed a mathematical formula with the objective of defining the consistency of indexing among indexers. That is, he sought to measure the consistency of indexing performed by two or more indexers in a group. However, for our study, we will use the formula adapted by Gil Leiva (2008), as seen below.

According to Tartarotti (2019) there are two types of intrinsic evaluation: Quantitative intrinsic evaluation through interconsistency – also known as interindexing consistency, when two or more professionals index the same informational resource seeking to compare results. Quantitative intrinsic evaluation through intraconsistency is the comparison of indexing the same informational resource carried out by the same indexer at different times.

It is noteworthy that this study uses intrinsic evaluation through interconsistency because it is performed on the same metadata distribution platform: MercadoEditorial.org, to evaluate the subject metadata of the chosen work, indexed by different publishers. This intrinsic evaluation seeks to analyze the central tasks in indexing like descriptors, headers, sub-headers, and identifiers. It can be quantitative, via evaluations and consensus among experienced professionals, or quantitative, using mathematical formulas. In order to perform, the intrinsic evaluation through interconsistency, the mathematical formula adapted by Gil Leiva (2008) from the study developed by Hooper (1965) was used:

Figure 1. Intrinsic evaluation formula

$$C_i = \frac{T_{co}}{(A + B) - T_{co}}$$

In which

C_i – MercadoEditorial.org consistency index

A – Number of terms assigned in indexing A

B – Number of terms assigned in indexing B

T_{co} – Number of terms in common in both indexing

Using this formula, it is possible to arrive at two consistency indexes: the "flexible" index – which can fluctuate from a value of 0.5 when only part of the descriptors correspond, to a value of 1.0 when descriptors fully match (for example: when indexer A uses "Religion and Philosophy" and indexer B uses "Religion"), and the "rigid" index – when terms completely coincide. For both indexes, when there is no correspondence among subjects, the value is 0 (Gil Leiva et al., 2008; Tartarotti, 2019).

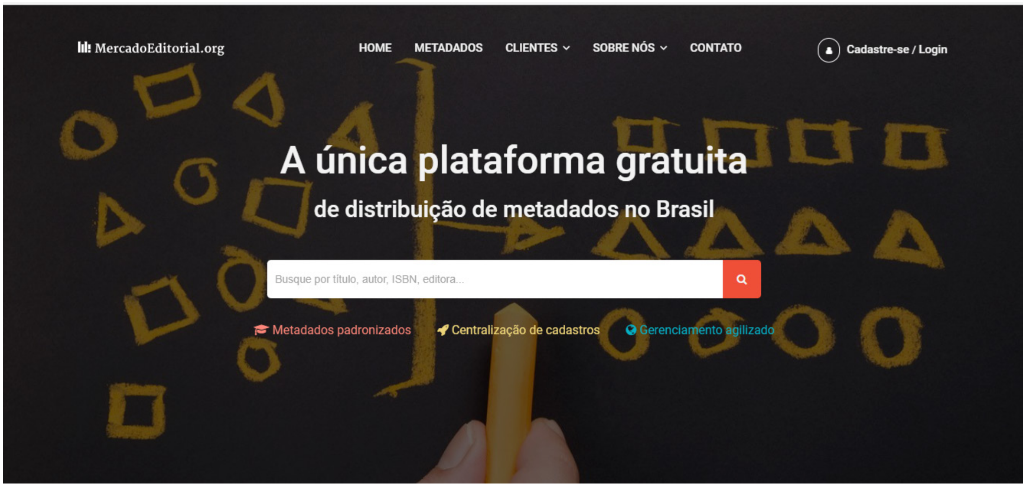
For this study, the work of the renowned Brazilian writer Machado de Assis, "The Alienist" was chosen, and three publishers were selected to apply the intrinsic evaluation method of indexing interconsistency:

- Ciranda Cultural
- Excelsior Editora
- Via Leitura

Universe of study: MercadoEditorial.org

According to the information on the web site, MercadoEditorial.org was founded in 2015, in São Paulo, and it aims to streamline or make efficient the process of releasing metadata in the field.

Figure 2. Homepage of MercadoEditorial.org



A única plataforma gratuita

de distribuição de metadados no Brasil

The first tab "HOME", redirects to the homepage. The second tab, "METADADOS" (METADATA), presents a catalog with several titles and the option of searching by field, on the left-hand side of the page.

The tab "SOBRE NÓS" (ABOUT US), shows a category about metadata intelligence that is a tool to optimize search mechanisms, providing clear information about the books during search, so that clients like publishers can present products more efficiently to the appropriate target group. There are also analytical reports that about information the company selected regarding the products on the publishing market. In addition, it also shows the clients on the platform, namely: publishers, bookstores, distributors, second-hand bookstores, libraries, commercial representatives and integrators. In the same tab, other initiatives are announced, such as "Livraria.ME", created to assist the customers' sales on the platform, in addition to training, frequently asked questions (F.A.Q.), and partners. Finally, contact and registration information are included on the platform.

In the central search box of the homepage, it is possible to search by title, author, ISBN, publisher, among others. By scrolling down the page, a visitor to the page can access "API", which shows how to fill in metadata. In these fields, you can access books by date, publisher, and imprint, along with searching for new books and updating old metadata.

Among the other metadata fields is the "subject" field, in which there are recommendations on how the terms should be placed, with at least three keywords or expressions that describe the theme of the book.

After a brief explanation of the platform, the following section will present the research results with analyzes and discussions.

Results and discussion

To apply indexing evaluation, specifically intrinsic indexing evaluation by interconsistency, the work "The Alienist" from Machado de Assis was chosen.

In order to demonstrate step by step how to apply intrinsic indexing evaluation and to analyze data, tables were formed to compare the publishers, and verify and visualize the respective metadata of subjects they assigned. Flexible consistency is characterized by the subject header or sub-header of a document coinciding with the subject of another document; such consistency can appear as total coincidence, partial coincidence, or no coincidence at all; while rigid consistency is when there is complete and total coincidence of a given subject.

The following table, Table 1, presents the results of the application of Gil Leiva's formula (2008) that was adapted from Hooper' formula (1965). Data from Table 1 were removed and the flexible and rigid consistency indexes were calculated and transformed into percentages.

Table 1. Publishers indexing consistency index

Publishers	Flexible	Rigid
Ciranda Cultural and Excelsior Editora	1.6%	0
Ciranda Cultural and Via Leitura	8%	8%
Excelsior Editora and Via Leitura	7.2%	6%

In Table 1, the publishers Ciranda Cultural, Excelsior Editora and Via Leitura are displayed along with the flexible and rigid consistency indexes. It can be observed that the publishers Ciranda Cultural and Via Leitura have the highest percentage (8%) of flexible and rigid indexing. Excelsior and Via Leitura comes in 2nd place, with 7.2% flexible consistency, and 6% rigid consistency. The lowest percentage registered is for Ciranda Cultura and Excelsior, with 1.6% flexible consistency. In their case, no percentage of rigid consistency was found. The percentages obtained from the consistency evaluation showed compatibility in their results between the publishers Ciranda Cultural and Via leitura. The terms removed to perform the analysis make up each publisher's profile's subject metadata found on MercadoEditorial.org.

Ciranda Cultural and Excelsior Editora

Table 2. Comparison between Ciranda Cultural and Excelsior Publisher

PUBLISHERS	
CIRANDA CULTURAL	**EXCELSIOR PUBLISHER**
Subject Metadata	
Romance, Brazilian Literature, Vestibular (university entrance exam)	Deluxe edition, Posthumous Memoirs, best author, hard cover, Dom Casmurro, alienist, Simão Bacamante, literature, capitu, book, Esau and Jacob, satire, madness, Counselor Ayres' Memorial, The Posthumous Memoirs of Bras Cubas, classic, machado, Braz Cubas, brazilian author, mental hospital,national literature, Quincas Borba, classics, bentinho, national, Bras Cubas, classics, Bruxo do Cosme Velho, The alienist, Machado de Assis

There is a huge difference in the two forms of indexing carried out by these two publishers, both in terms of the amount of subject metadata and in the choices themselves. Ciranda Cultural chose three broader terms, and Excelsior defined terms that go beyond the work under consideration ("The Alienist"), using other titles of the author in the indexing, the type of material edition, characters from other books and variations of the author's name. It is noteworthy that there is an excess of terms that do not correspond to the indexed informational resource. This brought about different results that do not correspond to the desired item, the results in turn appeared in the search engine on the platform, making searching for the real results more difficult. Spelling errors were also identified, along with other mistakes like lowercase beginnings to first names or incorrect accent mark usage.

It was verified through the search on the platform that when terms are used that do not specifically correspond to the informational record in question, the database retrieves multitudes of other possible matches, making it neigh impossible to find the desired item. Exhaustive indexing is an option, but still very little specific, information is used, again causing dispersion of results. In the case of "The Alienist", the target book in question, When put into the search engine as "O alienista", the system found 39 results. Of the 39, books that did not match the search were also brought up, such as other works by the author ("Pai contra mãe" and "Papeis avulsos") and books by other authors and literary genres (essays organized by Arturo Gouveia and thematic tales about the psychiatric hospital).

Results in Table 1 show a low compatibility between indexing and, mainly, the lack of consistency.

Ciranda Cultural and Via Leitura

Table 3. Comparison between Ciranda Cultural and Via Leitura

PUBLISHERS	
CIRANDA CULTURAL	**VIA LEITURA**
Subject Metadata	
Romance, Brazilian Literature, Vestibular (university entrance exam)	vestibular, fuvest 2020, fuvest, vestibular and enem, enem, vestibular 2020, machado de assis the alienist, machado de assis dom casmurro, machado de assis alienist, fuvest 2022, Dom Casmurro, machado de assis, the alienist, dom casmurro, world classic literature, machado de assis yales, Brazilian literature, contemporary Brazilian literature, enem 2020, fuvest 2021, books vestibular, enem books, enem and vestibular, Brazilian and Portuguese literature

When comparing both indexing processes, it is possible to notice the consistency index a little higher in relation to the previous analysis. However, it is still a low index compared to its totality. Via Leitura performed the indexing of informational resources using 24 terms, apparently without any criteria, as it contains mistakes such as the union of two distinct concepts in a single heading (e.g.- "machado de assis tales" or "books vestibular") instead of separating them. In addition, other inconsistencies were observed, such as repetition of terms, names starting with lowercase letters and a conceptually wrong term in: "Brazilian and Portuguese literature", as it is only Brazilian literature. Both flexible and rigid indexes achieved the same result of 8 % indexing consistency.

Excelsior Editora and Via Leitura

Table 4. Comparison between Ciranda Cultural and Via Leitura

PUBLISHERS	
EXCELSIOR PUBLISHER	**VIA LEITURA**
Subject Metadata	
deluxe edition, Posthumous Memoirs, best author, hard cover, Dom Casmurro, alienist, Simão Bacamante, literature, capitu, book, Esau and Jacob, satire, madness, Counselor Ayres' Memorial, The Posthumous Memoirs of Bras Cubas, classic, machado, Braz Cubas, brazilian author, mental hospital, national literature, Quincas Borba, classics, bentinho, national, Bras Cubas, classics, Bruxo do Cosme Velho, The alienist, Machado de Assis	vestibular, fuvest 2020, fuvest, vestibular and enem, enem, vestibular 2020, machado de assis the alienist, machado de assis dom casmurro, machado de assis alienist, fuvest 2022, Dom Casmurro, machado de assis, the alienist, dom casmurro, world classic literature, machado de assis yales, Brazilian literature, contemporary Brazilian literature, enem 2020, fuvest 2021, books vestibular, enem books, enem and vestibular, Brazilian and Portuguese literature

The flexible index was 7.2% and the rigid 6%. It is observed that even using a large number of terms, both publishers achieved a low consistency index. Thus, it is clear that a large number of descriptors is useless when there are no fundamental criteria for their establishment. The characteristics and shortcomings of these descriptors have already been discussed above.

Summary of results

Based on the results achieved, it is clear that there is a lack of guidance and trained personnel for the indexing of information resources on MercadoEditorial.org. The lack of consistency can cause disparities in document retrieval and cause several errors in the database of the platform, which must be corrected in the future.

Librarians are professionals trained to perform subject indexing, however, those responsible for this task have the possibility of being trained and provided with standards for the insertion of subject metadata. Standardization in the number of terms used, guidelines establishing the degree of specificity or even the definition of an indexing language are elements of an indexing policy that can benefit the platform and the customers who use it, such as publishers, second-hand bookstores, distributors and libraries.

Conclusions

MercadoEditorial.org is a free Brazilian initiative for consultation, and institutions can register to use the other services offered by the platform. With this startup, it is clear that the importance of technology in favor of information is actively present, such as in the exchange of metadata.

Consistency in indexing promotes database standardization and, consequently, better information retrieval indexes. It brings benefits to all parties involved in the work performed.

In the library environment, there have always been efforts to organize and make information available in the best possible way to the user, creating strategies and methodologies to improve librarian activities. From the study herein, it is clear that there is a demand on the part of the publishing market that can benefit from procedures already consolidated in Library Science.

Acknowledgments: We thank Johnathan Dabney for the English language editing.

References

Alves, R. C. V. (2018). Metadados editoriais e livreiros: algumas considerações e relações com os padrões de metadados do domínio bibliográfico. *Informação & Tecnologia (ITEC),* 5(2), 238–252.

Baca, M. (2016). *Introdution to metadata*. Getty Research Institute.

Gil Leiva, I. (1999). *La automatización de la indización de documentos*. Trea.

Gil Leiva, I. (2001). La consistência en la asignación de materiais en Bibliotecas Públicas Españolas. *Boletín de La Asociación Andaluza de Bibliotecarios*, 63, 69–86.

Gil Leiva, I. (2002). Consistencia en la indización de documentos entre indizadores noveles. *Anales de Documentación*, 5, 99–111

Gil Leiva, I. (2008). *Manual de indización*: teoría y práctica. Trea.

Gil Leiva, I., Rubi, M. P., & Fujita, M. S. L. (2008). Consistência na indexação em bibliotecas universitárias brasileiras. *Transinformação,* 20(3), 233–253.

Hooper, R. S. (1965). Indexer consistency tests: origin, measurement, results, and utilization. IBM Corporation.

Publishnews. (2016). *Metadados, o ovo e a galinha*. Disponível em: https://www.publishnews.com.br/materias/2016/08/26/metadados-o-ovo-e-a-galinha. Acesso em 18 abr. 2022.

Riley, J. (2009–2010). *Seeing Standards*: a visualization of the metadata universe. Indiana University Libraries.

Tartarotti, R. (2019). *Avaliação do processo de indexação de assuntos em repositórios institucionais pela abordagem da recuperação da informação.* [Doctoral thesis, Universidade Estadual Paulista Júlio de Mesquita Filho]. Repositório Unesp.

Zeng, M. L., & Qin, J. (2016). *Metadata*. 2. ed. Neal-Schuman, American Library Association.