ISSN 1678-765X doi 10.20396/rdbci.v20i00.8667925

Article

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

Information Storage and Retrieval System: an analysis of the impact of variables and measures aimed at the organization and retrieval of information centered on the user

Gercina Angela de Lima [1], Maria Luiza Almeida Campos [2]

ABSTRACT

Introduction: The effective performance of an Information Retrieval System depends on the quality with which the organization of information is performed, which will imply a retrieval of the most relevant and pertinent information, since these procedures are conditioned to each other, creating a bridge between input and output of information. Objective: To evaluate the impact of the exhaustiveness and specificity variables and the recall and precision measures, as well as the concepts of relevance and pertinence, in Information Retrieval Systems. Methodology: It is characterized as a descriptive and exploratory study, based on a narrative literature review aiming to present the different concepts, their converging and divergent points. Results: As a contribution, we present a proposal for a flow for an Information Storage and Retrieval System, centered on the user, bringing together several aspects related to measures of recall and precision, of relevance and pertinence. Conclusion: It is considered as the final contribution of this study to highlight the importance of a systemic view, in which all elements of an Information Storage and Retrieval System are in interrelation, having the user as the main element; and present the fundamental activities that are important for the training of professionals able to build consistent Systems.

KEYWORDS

Information storage and retrieval system. Recall. Precision. Relevance. Pertinence. Assessing the impact of measures on SRIs.

Sistema de Armazenamento e Recuperação da Informação: uma análise do impacto das variáveis e medidas visando à organização e recuperação de informação centrado no usuário

RESUMO

Introdução: O efetivo desempenho de Sistema de Recuperação da Informação depende da qualidade com a qual a organização da informação é realizada, o que implicará em uma recuperação da informação mais relevante e pertinente, visto que esses procedimentos são condicionados um ao outro fazendo uma ponte entre a entrada e a saída da informação. Objetivo: Avaliar o impacto das variáveis exaustividade e especificidade e das medidas de revocação e precisão,

Author's correspondence

[1] Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil / e-mail: limagercina@gmail.com

[2] Universidade Federal Fluminense, Niterói, RJ, Brazil / e-mail: marialuizalmeida@gmail.com

assim como os conceitos de relevância e pertinência, em Sistemas de Recuperação de Informação. Metodologia: Caracteriza-se como um estudo descritivo e exploratório, baseado em revisão de literatura narrativa visando apresentar as diversas conceituações, os seus pontos convergentes e divergentes. Resultados: Como contribuição apresenta-se uma proposta de fluxo para um Sistema de Armazenamento e Recuperação de Informação, centrado no usuário, reunindo diversos aspectos relacionados às medidas de revocação e precisão, de relevância e pertinência. Conclusão: Considera-se como contribuição final deste estudo evidenciar a importância de uma visão sistêmica, na qual todos os elementos de um Sistema de Armazenamento e Recuperação de Informação estão em inter-relação, tendo o usuário como elemento principal; e apresentar as fundamentais atividades que são importantes para a formação de profissionais aptos à construção de Sistemas consistentes.

PALAVRAS-CHAVE

Sistema de armazenamento e recuperação da informação. Revocação. Precisão. Pertinência. Relevância. Avaliação do impacto de medidas em SRIs.

CRediT

- Recognitions: The first co-author thanks the National Council for Scientific and Technological Development (CNPq), for the Research Productivity Grant (PQ-D1).
- Funding: This study was funded by the Brazilian agency Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) for the scholarships.
- 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: Not applicable.
- Availability of data and material: Not applicable.
- Authors' contributions: Conceptualization, Data Curation Formal Analysis, Acquisition of Financing, Research, Methodology, Resources, Supervision, Validation, Visualization, Writing - original draft: LIMA, G.A.; CAMPOS, M.L.A. Writing - review & edition: LIMA, G.A.; CAMPOS, M.L.A.

JITA: ID. Knowledge representation.

Article submitted to the similarity system

Submitted: 15/12/2021 - **Accepted**: 05/05/2022 - **Published**: 27/05/2022

1 INTRODUCTION

From the technical-scientific revolution after the Second World War, the large volume of information generated in the growing number of knowledge areas started to demand a higher level of informational organization, because the information must be ordered, structured or materially fixed, becoming a document, otherwise it will remain amorphous and unusable. It can be said that the advances that have occurred since the 1950s until the present day have been relevant and have marked the development both in the form of storage and representation and in information retrieval.

Thus, the needs of users have become the crucial point of studies in the area since information retrieval is the main objective of the entire field of Information Organization. Thus, librarians began to face new challenges with the changes in the conceptualization and delivery of and access to library services, and thus had to assume several roles, in addition to those already foreseen.

The storage of information, once only on the hard drives of computers with large processing capacity, now takes place in a unique way, so that it is possible to access files, data, and applications anywhere and anytime, using either a computer or mobile devices, provided there is an Internet connection, through storage in the so-called "clouds".

In the scope of the processes that involve the organization and retrieval of information, in the context of Librarianship and Information Science (LIS), the materiality of these processes is considered in the sphere of Information Retrieval Systems (IRS). For Lancaster (1978), the main function of an IRS is to act as an interface between a particular population of users and the universe of information resources in printed or other form. It is in this environment that cataloging, indexing, and classification processes subsidize the organization and retrieval of information from the various information supports. The product of these activities is the elaboration of catalogs of a given physical collection, or of a database of a digital library, or even of an online catalog and of digital repositories.

The effective performance of NIS depends on the quality with which the organization of information is carried out, which will imply a more relevant and pertinent information retrieval, since these procedures are conditioned to each other, bridging the gap between information input and output. In this context, one should pay attention to the variables exhaustiveness and specificity in the performance of indexing, as well as to the level of revocation and precision that the system proposes to serve its users. This will influence the system's ability to retrieve information, resulting in relevant and pertinent documents for the user. Relevance consists of the degree of similarity between terms that make up the user's search expressions and their occurrence in documents in the collection or in the indexing terms. Relevance is the relationship between the information obtained in a search that answers the user's need or demand for information, that is, information that is useful for the user.

This paper discusses the role of information organization and retrieval within an Information Retrieval System (IRS) to evaluate the impact of the variables exhaustiveness and specificity and the measures of revocation and precision, as well as the concepts of relevance and pertinence in IRSs. As a contribution, a flow proposal for a user-centered Information Storage and Retrieval System is presented, bringing together the various aspects presented.

From such discussions, it was considered to bring, in a didactic way, especially important concepts for the professional work and that may help in procedures related to the elaboration of indexing policies, fundamental procedures for the formation of professionals who dedicate themselves to information treatment. In this sense, a flow is presented in which the several aspects of variables and measures are gathered within a proposal of an Information Storage and Retrieval System (ISRS).

In the next sections, we first describe the methodology used to carry out this study and present the organization and retrieval of information as considered within the scope of this study. Subsequently, it is described about the NIS and its flows, and the processes to help users to achieve relevance and pertinence in their searches; soon after these discussions are brought and consolidated through the proposal of an EISR and, finally, the final considerations are presented.

2 METHODOLOGY

This study is characterized as descriptive and exploratory, with the purpose of understanding, through a flowchart, how the processes of organization and retrieval of information (ORI) impact on the variables exhaustiveness and specificity and, consequently, on the measures revocation and precision, to achieve greater relevance and pertinence in the results of user queries within a NIS.

To study how these variables and measures are evidenced in the structures of a NIS process flowchart in the context of ORI, it was proposed to first develop a narrative literature review, since the research is driven by more open questions, to: (1) map the main NIS flowcharts present in the literature and (2) collect initial inputs to support the theoretical and methodological proposal of a NIS that encompasses all these elements. These criteria are justified because the flowcharts representing the processes of a NIS do not explicitly place these variables and measures in this specific context. In this case, five flowcharts were selected based on the following criteria: (1) the seminal authors in the field and (2) the most cited in the literature.

For this, an exploratory search was made, without temporal definition, in Google Scholar, and in five specialized databases: Library and Information Science Abstracts (LISA), Information Science & Technology Abstracts (ISTA), Library, Information Science & Technology Abstracts with Full Text (LISTA), Scopus and Web of Science. These bases were chosen considering the relevance and the relationship they have with the area and sub-area of knowledge delimited for this review, using the following search expressions, presented in Chart 1.

Table 1. Expressions used in the literature search

Sign	Search expressions
E1	"Sistema de Recuperação da Informação" OR "SRI"
E2	"Information Retrieval Systems" OR "IR system"
E3	"Politica de indexação" AND "Organização da Informação" AND "Recuperação da Informação"
E4	"Indexing Policy" AND "Organization of Information" AND "Information Retrieval"
E5	"Exaustividade" AND "Especificidade" AND "Revocação" AND "Precisão" AND "Relevância" AND "Pertinência"
E6	"Exhaustivity and Specifity" AND "Precision and Recall" AND " Relevance and Pertinence"

Source: Prepared by the authors

At first, 48 documents were selected from the retrieved ones, using three criteria:

(1) documents that dealt with the information retrieval system and indexing policy; (2) documents that presented studies on the variables exhaustiveness and specificity, on the measures revocation and precision, and on relevance and pertinence within the scope of NIS; (3) documents that brought these terms in the title or in the keywords. From these 48 selected documents, we used 33 documents that deal specifically with the theme, among which we found the five flowcharts used to base a new structure that contextualizes the position of each of these variables and measures within a proposal of an Information Storage and Retrieval System (IRIS), as presented in section 6 of this article.

3 INFORMATION ORGANIZATION AND RETRIEVAL

This section presents the concepts information organization and information retrieval, which are considered relevant to the context of this paper.

3.1 Information Organization

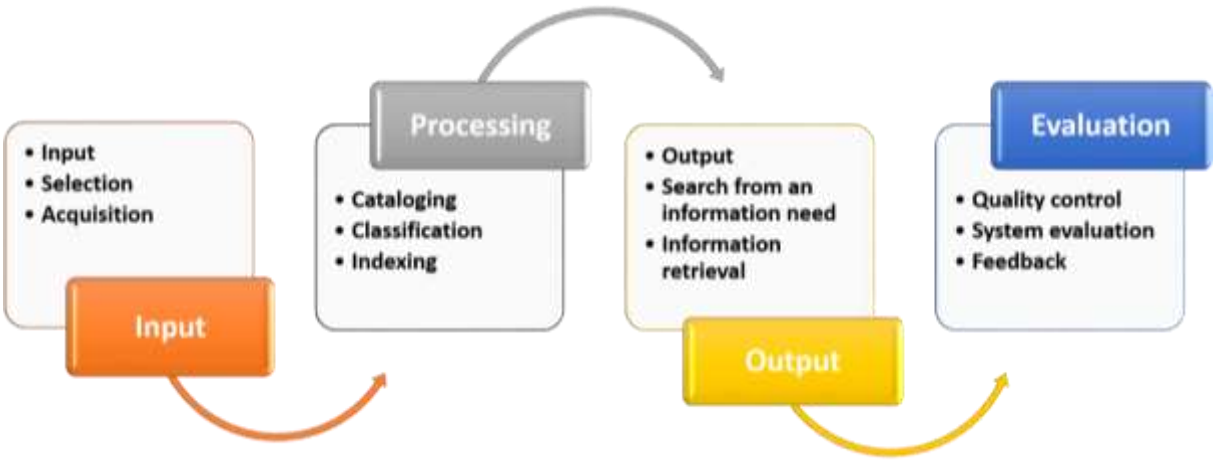
The field of Information Organization (IO) includes all studies related to the processes and instruments used in the organization of information resources of any nature, to meet the information needs of a given community of users.

Dahlberg (2006) defines information organization as the ordering of objects to create a link between the object of an area and its own activity. Novellino (1996) states that the process of information representation is characterized by the replacement of the descriptive and thematic content of a document by an abbreviated description, which will be stored for later retrieval.

Thus, OI aims at the representation, storage, and retrieval of information. According to Barreto (2002), the objective of the information organization process is to enable and facilitate the access to information, which, in turn, has the competence and intention to produce knowledge. To this extent, it is understood that the input of inconsistent data will imply the output of data that is also inconsistent. In this process, cognitive mechanisms are activated that influence both the input and output of the information retrieval system, because they are dependent on the way we use our mind to make abstractions.

In the context of Library and Information Science, the representation of information is accomplished through the processes of cataloging, indexing and classification in a NIS, as illustrated in Figure 1.

Figure 1. Information Retrieval System



Source: Adapted from Cesarino (1985, p. 161)

In the cataloging process, also known as descriptive representation, a bibliographic item is described to make it unique among the others in each collection, allowing it to be identified, located, and represented in the catalogs. Mey (1995, p. 5) considers that "cataloging is the study, preparation and organization of codified messages, based on existing or likely to be included in one or several collections, in order to allow intersection between the messages contained in the items and the internal messages of the users".

Indexing is another important representation process that occurs within a NIS, in which the indexer is expected to read the document and distinguish between relevant and peripheral information to better represent it for subsequent retrieval. According to ISO 5963-1985 (1985), indexing is seen as "[...] the representation of document content by means of special symbols,

either taken from the original text or chosen in an information or indexing language." This process is carried out in two stages: the first is the analysis of the document to identify its informational content; the second is the translation of the concepts into the terms of an indexing language, using

The second is the translation of the concepts into indexing language terms, using knowledge organization systems such as thesaurus and bibliographic classification systems.

While cataloging describes the physical characteristics of a bibliographic item and indexing is concerned with issues involving the intellectual content of the document, classification, as a process, involves the orderly and systematic assignment of each entity to only one class within a system of mutually exclusive and non-overlapping classes, based on similarities and differences. According to Tristan et al (2004, p. 163), "classification is a mental process by which we can distinguish things, beings, or thoughts by their similarities or differences." It is a fundamental activity of the human mind that processes ideas and distinguishes them based on common characteristics. According to Lima (2021), other conceptualizations can still be admitted for the word classification:

Depending on the point of view, classification is considered a discipline, but it can also be the product that results from the act of classifying and, simultaneously, it is the tool used to carry out the classification process (LIMA, 2021).

In the field of Library and Information Science, at least four conceptualizations can be attributed to the word classification: classification has been studied as a discipline, as a process of grouping and ordering knowledge, as the product of the grouping and ordering process, and as a tool for representing information.

3.2 Information Retrieval

Information Retrieval (IR) is an area originating from Computer Science (CC), and the expression was attributed to the American engineer Calvin Mooers, in 1951, who defined it, at the time, as a process that "[...] encompasses the intellectual aspects of describing information and its specifics for searching, in addition to any systems, techniques, or machines employed for the performance of the operation" (MOOERS, 1951, p. 51). Saracevic (1999), in the scope of Librarianship and Information Science, states that this conception of information retrieval brought by Mooers was centered on the construction of systems, but, as of the late 1970s, it expanded to a user-centered approach, taking into consideration the users' mental models.

For Ferneda (2003), the IR process consists in "identifying, within the set of documents (corpus) of a system, which ones meet the user's information needs". Thus, the IR is considered an important operation in an Information Retrieval System, which aims to relate the user's search with the items stored in the database, through a set of interconnected elements of processing routines of informational records, aiming to meet the information needs of a community of users. However, some authors define IR from different approaches.

For Saracevic (1999), information retrieval involves intellectual aspects of information description and the search specifications, as well as the systems, techniques, and equipment that are applied to carry out the whole process. While Baeza-Yates and Ribeiro-Neto (2011, p. 1) point out that:

> Information retrieval deals with the representation, storage, organization, and access to information items, such as documents, web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of information items should provide users with easy access to the information of their interest.

Salton (1968) considers IR as an area of research concerned with the structure, analysis, organization, storage, retrieval, and search of information. In turn, Lancaster (1993) considers

it as a process of searching a collection of documents to identify those texts that deal with a particular subject.

According to Rowley (1994, p. 113), IR is the "process of locating documents and items of information that have been the subject of storage". For the author, the process is composed of three elements: the query, the comparison, and the result. The query is the user's question transformed into a search strategy; the comparison is the action of checking whether the formulated question matches the stored items; and the result is the list or items that match(s) the user's search (ROWLEY, 2002). However, the process of information retrieval depends heavily on the steps taken in the input (cataloging, indexing, classification) and storage, which directly impact the searches performed in an Information Retrieval System.

4 INFORMATION RECOVERY SYSTEM (IRS): AN ANALYSIS FROM THE FLOWS

For Rowley (2002), NISs and computers have been used as synonyms, however, before the emergence of any computers and computer science itself, paper-based file and token systems already existed. A NIS is an integral part of a communication system and or information system. These systems allow users to search for information in a collection of documents (or other information sources) through queries usually formatted as a set of metadata and obtain it in a way that meets their needs with relevance and pertinence. But IRSs deal with at least two different problems regarding users' information needs or queries: (1) they must distinguish and identify the relevant information related to the query, and (2) they must get the answer quickly.

Salton and McGill (1983, p. xi) define a NIS as "a system that deals with the representation, storage, organization, and access to information items, and this may be in a physical or digital collection." The authors consider it as an interconnected set of routines for processing informational records, with its own purposes and criteria, aiming to meet the information needs of a community of users. For Silva, Santos, and Ferneda (2013, p. 29), information retrieval systems "have the function of representing the content of the corpus documents and presenting them to the user in a way that allows him to quickly select the items that totally or partially satisfy his information need [...]".

For Lancaster (1986, p. 1), every NIS is composed of two subsystems: information input and information output, represented in the form of a cycle, with three stages: representation, storage, and retrieval, characterized as a continuous and feedback process.

When documents are selected to compose a collection, whether physical or digital, they initially go through an organization considering the needs of the community to be served. This organization is performed by the processes of cataloging, indexing, classification, and abstracting. Thus, NIS encompasses from the organization processes to information retrieval by the user.

There are several proposals in the literature for representing the flow of these subsystems from information entry, representation, and retrieval in a NIS. In this paper, we will present some of these schemes, based on the literature, to verify the structure and processes of their subsystems. To this end, the following schemas have been selected.

Lancaster, already in 1978, in his book Information Retrieval Systems (LANCASTER, 1978), identifies the elements that compose the NIS. The subsystems he identified are selection and acquisition; indexing; vocabulary; search; user-system interaction (query negotiation); and match[1]. By calling all these elements subsystems, he brings in the concept of interrelationship,

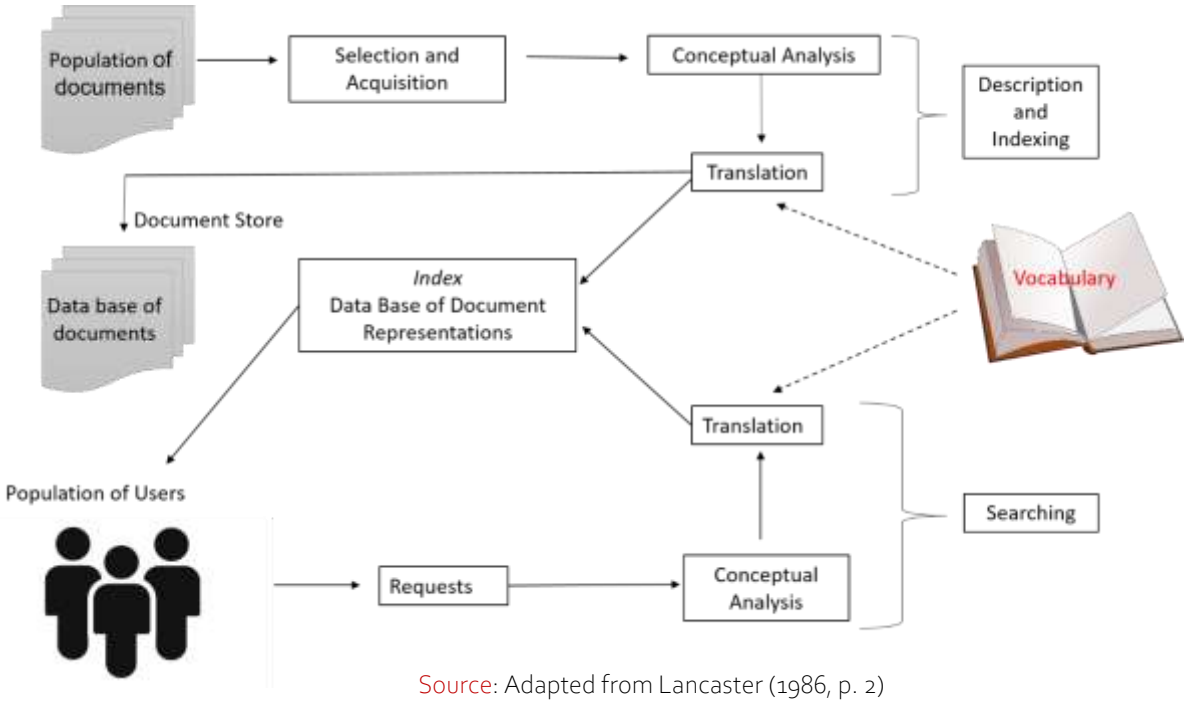
[1] We don't yet have a Portuguese term for this action. It is the matching of the profile of the question with the profile of the documents.

in which there is a relationship between users, collections, vocabularies, indexing and retrieval. This means that they affect each other. The group view of a NIS presented by Lancaster, as we will see later, is a view of the whole, in which it is of fundamental importance to understand the behavior of each element so that a NIS can achieve its function, that is, to enable consistent retrieval by its users.

To better understand this inter-relationship, and pointing out that all actions always start from the identification of the user, as we will see further on, we can present the following questions: what is your area of expertise (subject), what activity do you perform (teacher, student, researcher); what means of communication do you use; what language do you use (scientific, technical, you are not a specialist); and so on. The answers will provide elements to determine principles to form the collection, to treat the document, to determine its classification and even actions that have no relation, such as the number of documents that a user can borrow and the time allowed (GOMES; CAMPOS, 1998).

From this conception, Lancaster proposes two schemes in his publications (1986, 1993). In the first scheme, in his book Vocabulary Control for Information Retrieval (1986, p. 3), the author presents the components of a NIS. As can be seen, in his flow proposal, the realization of the representation of a document focused only on the indexing and cataloging processes. The author did not insert details of the cataloging elements, leaving the indexing steps at the same level of understanding. After this organization, the documents are stored in a printed or digital database, in which searches can be performed to meet user requests, as can be seen in Figure 2.

Figure 2. The components of an Information Retrieval System



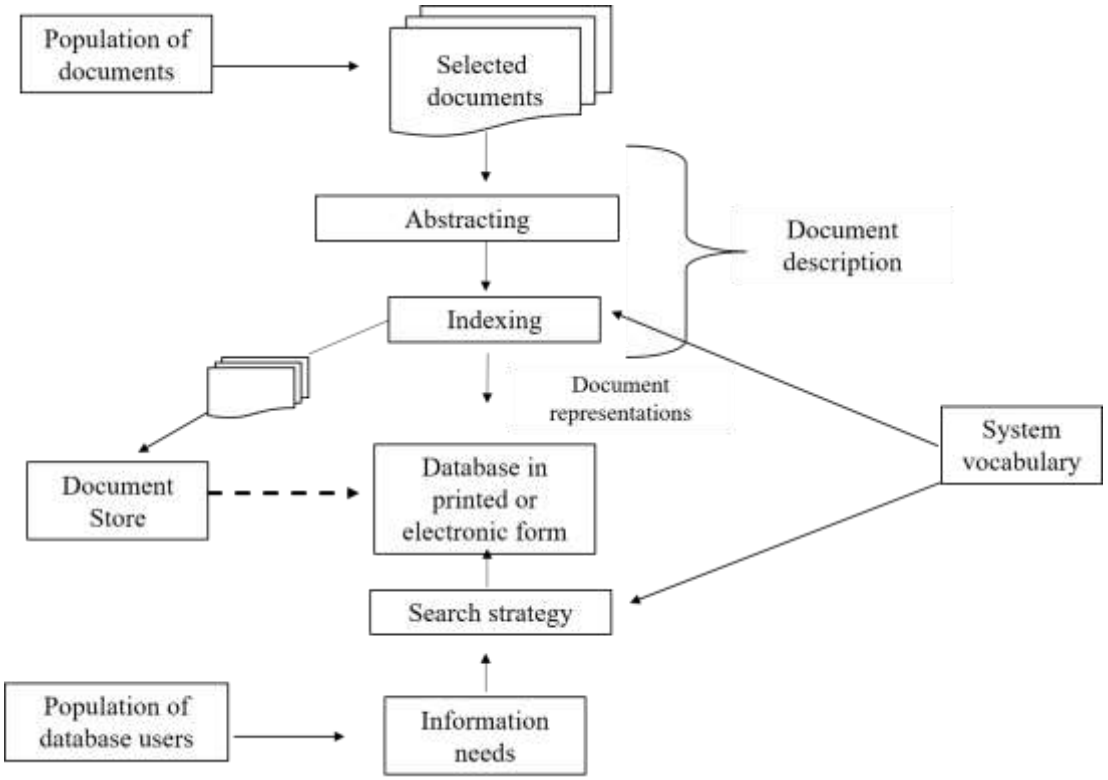
Source: Adapted from Lancaster (1986, p. 2)

The author considers that cataloging has two essential elements that must be considered: (1) the physical description of a document and (2) the choice of the access point to represent it. In relation to indexing, the author considers it as an intellectual process composed of two steps: conceptual analysis and translation, that is, the subject analysis of a document is done and, later, a controlled vocabulary is used to standardize the representation of this subject. However, in his flow proposal, the author considers the stages conceptual analysis and translation within the indexing and cataloging processes, although he has differentiated them at the time of his descriptions.

Lancaster, in his book Indexing and Abstracts: theory and practice (1993, p. 2), brings another proposal, more expanded, in which he also considers the preparation of index and

abstract, as shown in Figure 3. These changes are also considered due to technological advances at the time.

Figure 3. Role of indexing and summarization in the larger picture of information retrieval



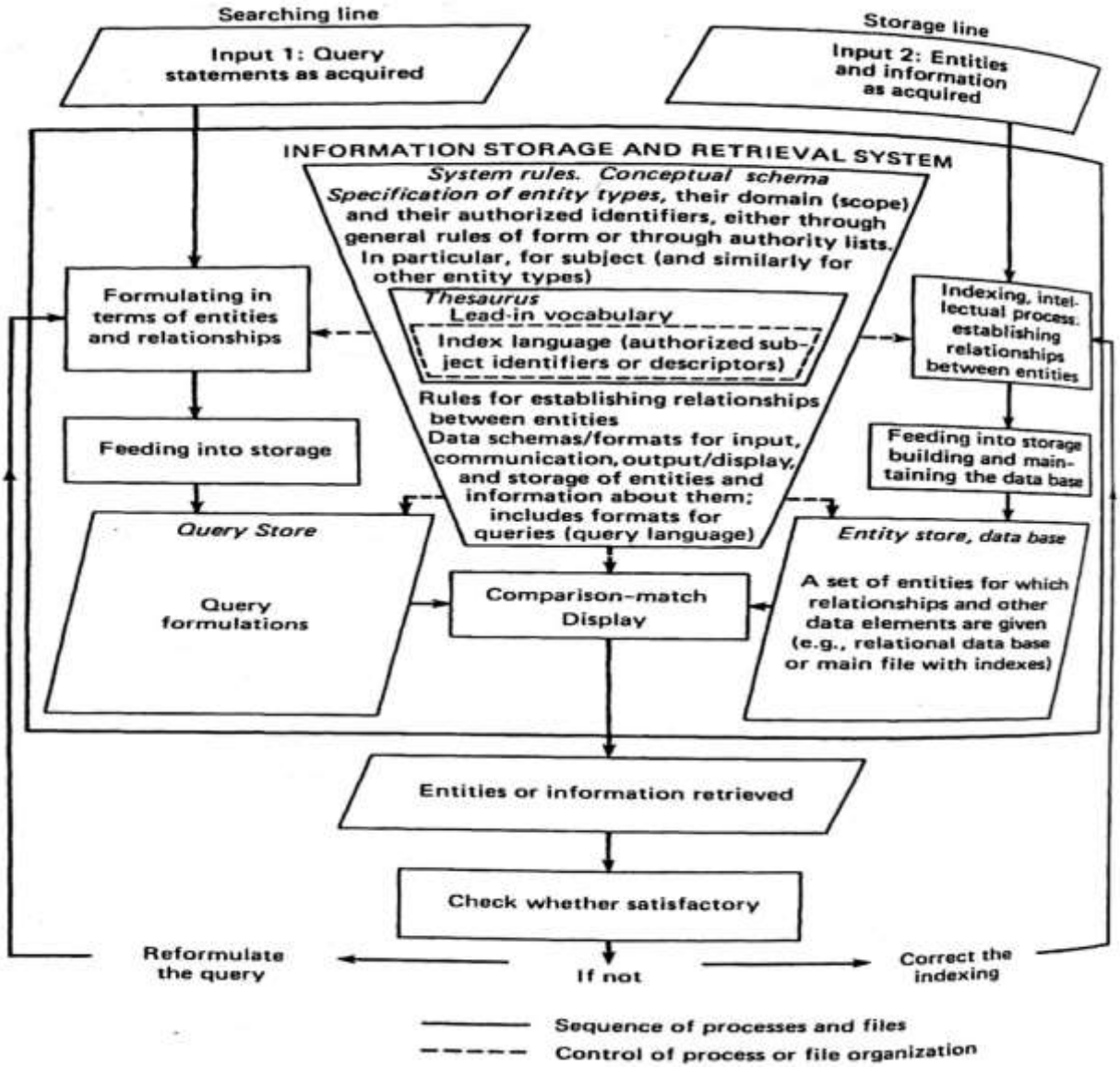
Source: Translated from Lancaster (1993, p. 2)

As can be seen, in this scheme, Lancaster precedes the writing of abstracts to indexing to help in condensing the representation of the document content and considers them as the process of describing the documents. On the other hand, the author points out the need to plan the search strategy according to the users' needs. In both subsystems, the author suggests the use of controlled vocabulary to standardize the terminology used by the author with that of the user in his search. These terms assigned by the indexer become access points used to retrieve bibliographic items. A common feature of these schemes are the processes of subject analysis and translation, which can occur both in the input, that is, in the representation, and in the output, which is the retrieval of information. Besides these aspects, one can also notice, in this 1993 flow, the inclusion of the specification and differentiation of databases in printed or electronic form.

We can see that Lancaster's flows alert us to the decision of which aspects of a document will be represented in a NIS, as well as which level of specificity or exhaustiveness will be attributed to the descriptors, and which are related to the set of decisions adopted by the NIS indexing policy. This decision goes through the type of vocabulary adopted in the system, because both, the indexing policy, and the vocabulary update, go together.

Soergel (1985) adds another aspect to be observed in NIS, the issue involving the storage of information, in which information items need to be processed, searched, retrieved, and disseminated to various user communities. In this sense, Soergel (1985) considers that an Information Storage and Retrieval System (ISAR System) is a subsystem of an information system. The author presents an Information Storage and Retrieval System structure, in which he emphasizes not only the information input and information output subsystems, but also the storage subsystem, as can be seen in Figure 4.

Figure 4. The structure of the Information Storage and Retrieval System



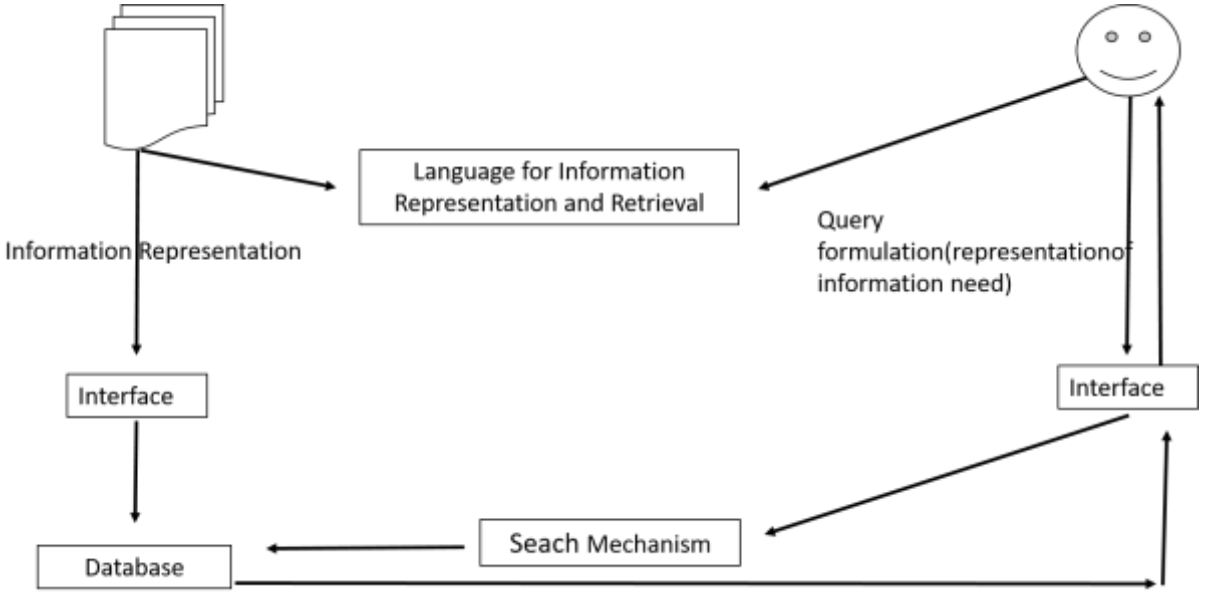
Source: Adapted from Soergel (1985, p. 19)

Unlike the schemes presented by Lancaster (1986, 1993), Soergel begins by first presenting the information output system, which he calls the Search Line, in which user profiles are studied and searches are formulated based on terms and relationships to arrive at the results represented at level 1. At the other end, we find the information entry system, with the documents and data, in which the information representation processes take place, starting with the indexing processes (descriptive and subject), so that the bibliographic items can be stored and made available to meet the needs of their community. The author adds to this flow a stage of comparison of results, when the relevance of the retrieved items in relation to the search strategy is evaluated, and the potential of relevance in the occurrence in documents of the collection.

Another crucial point in this scheme presented by Soergel (1985) is the suggestion to be made, based on the potential relevance or not of the document analysis, that the search be redone or that the indexing process performed for that document be corrected. This analysis, when done, contributes to maintaining the efficiency and accuracy of the system.

Chu (2005) presents the main components of an IRS, which he calls Information Representation and Retrieval (RRI) process. For the author, these components are the database, the search engine, the language, and the interface, as illustrated in Figure 5.

Figure 5. Information Representation and Retrieval Process



Source: Translated from Chu (2005, p. 18)

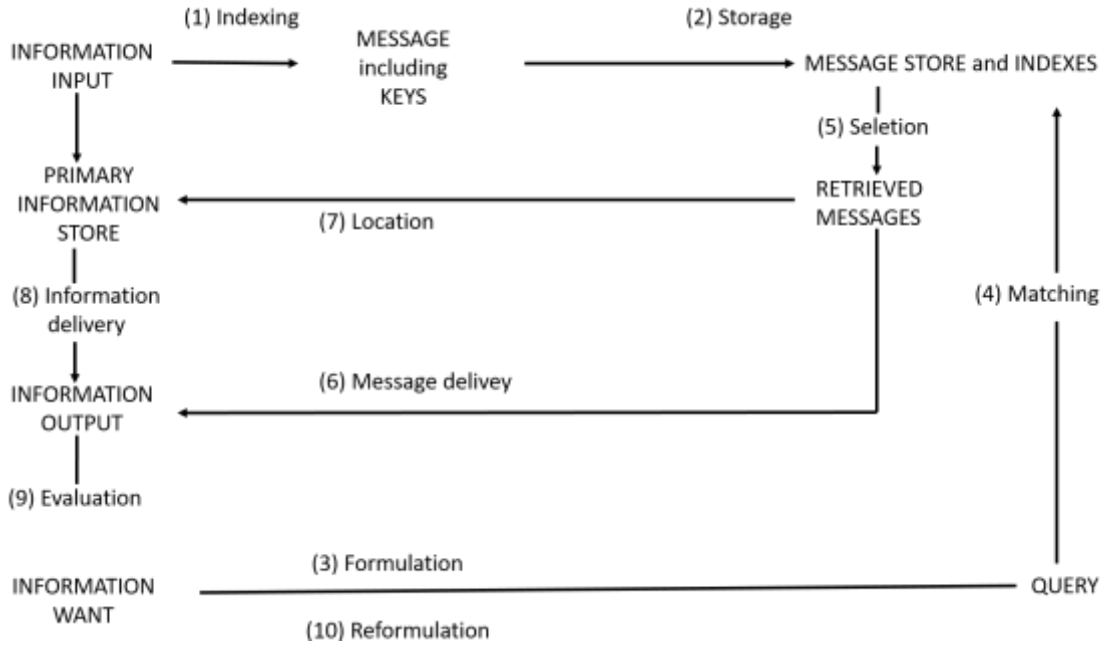
In this flow, we see that the process of representation of information appears with emphasis instead of naming the processes according to the previous schemes. The author also presents in a same level the languages and the information retrieval process and highlights the interface for accessing the database and the interface for formulating the user's search. Another critical point, highlighted by Chu (2005), is the role of the information professional in the representation of information with the use of controlled vocabulary, since discrepancies may occur during this process, which can cause problems in the return of searches, with low relevance and pertinence.

Chu (2005) points out that three problems may occur in the processes within a NIS: (1) the document typology, which may not be in a format that can be archived; (2) the difficulty of finding a term that exactly represents the content of the document with the thesaurus descriptor; and (3) the inconsistency of information representation, which may occur when more than one indexer performs the representation processes.

On the other hand, when formulating the search, the user must deal with the search in natural language which, most of the time, does not coincide with the descriptors used in the controlled vocabulary. In this case, the author suggests that the search may be more successful if there is a standardization between the decision making and the instruments used in the representation and retrieval of information (CHU, 2005).

The last scheme presented in this paper is that of Vickery and Vickery (2004). In this scheme, the authors present the flow from 10 processes (1) indexing, (2) storing, (3) formulating the search strategy, (4) matching the results, (5) selecting the relevant documents, (6) retrieving the documents, (7) locating the documents, (8) retrieving information, (9) evaluating the results. If the results have been satisfactory, it is finished at this stage. If the results are not satisfactory, (10) reformulation of the search is performed, as shown in Figure 6.

Figure 6. Information Storage and Retrieval



Source: Translated from Vickery and Vickery (2004, p. 118).

One can notice that, in the information entry, in this scheme, the authors emphasize only the indexing process, without mentioning the other processes, moving directly to storage and, with this, describing in more detail the information retrieval activities, which currently can be performed by means of a computer interface. Also, unlike the systems previously mentioned, this one does not present the issue of controlled vocabulary as an element of the system.

In general, what can be observed in the schemes presented is that there is no standardization in the flow of activities presented by the authors. All flows present the stages of input, storage, and retrieval; some specify the processes, and others do not. However, all have the objective of improving the effectiveness and efficiency in retrieval, because these measures directly impact the relevance and pertinence of the results of the searches performed by the users. Thus, it is inferred that the quality of information at the entrance of a NIS determines the quality of information at the exit.

The result of the analyses performed on these five flowcharts brought subsidies to base a new structure that contextualized the position of each of the variables and measures within a new proposal of an Information Storage and Retrieval System (IRIS), presented in section 5 of this article, but it was necessary to present, first, the importance of these procedures and how they impact user satisfaction in a NIS.

5 AN ANALYSIS OF THE IMPACT OF INDEXATION POLICY VARIABLES AND INFORMATION RECOVERY MEASURES ON AN IRS: IN SEARCH OF RELEVANCE AND PERTINENCE

The planning of indexing criteria is an important procedure in a NIS, which contributes to the performance of information retrieval. The consolidation of this planning should be materialized in a document - The Indexing Policy - that presents the principles and criteria that will serve to guide the decision-making to reach the objective of the NIS, that is, a consistent treatment for a quality retrieval of the requested information. The indexing policy can be considered as an administrative, management decision, considering the proposal of a NIS, in relation to the characteristics of the system, such as the type of information stored, the types of searches and the type of user.

Carneiro (1985, p. 221), in his article Diretrizes para uma política de indexação (Guidelines for an indexing policy), points out that an indexing policy

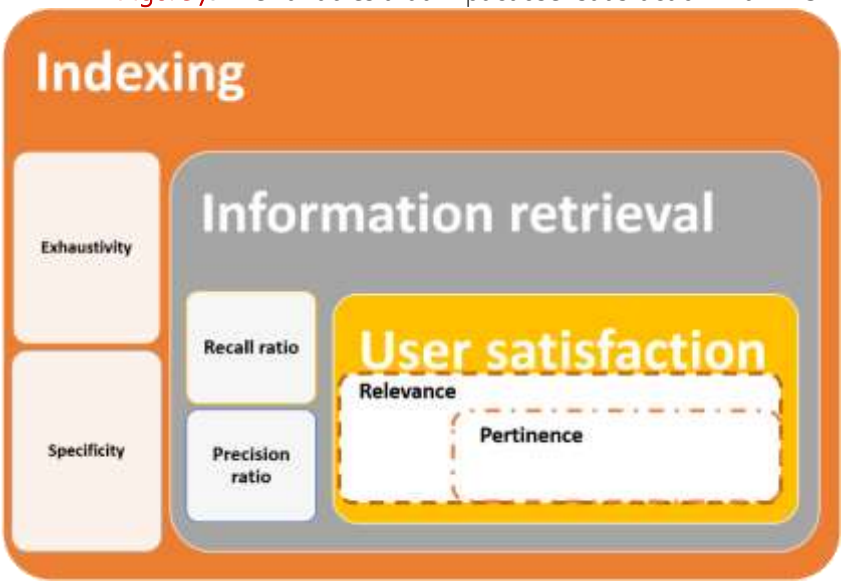
> [...] should serve as a guide for decision making; it should consider the following factors: characteristics and objectives of the organization, determining the type of service to be offered; identification of users, to meet their information needs and human, material, and financial resources, which delimit the operation of an information retrieval system.

Thus, the main objective of the indexing policy is to plan procedures to achieve efficiency in retrieving information from a NIS, that is, it is a set of principles for analyzing document content for its representation in a record.

All users expect that a NIS will be able to answer their searches with one or more documents pertinent to their needs. To know the efficiency of a NIS, Lancaster (1986, p. 131) points to three criteria, which are essential for this evaluation: (1) quality, (2) effort, and (3) system response time. Regarding quality, the author suggests observing "the coverage of the database, the ability of the system to retrieve the relevant documents in response to the user's question, the ability to retain the non-relevant documents at the same time." The evaluation of the performance of a SR, in relation to its efficiency, is measured from the user's satisfaction, that is, the ability of the system to answer the questions of its community.

For this, in the representation of information, one must observe the aspects of the level of exhaustiveness and specificity in the process of indexing, to achieve quality, and the measures of accuracy and revocation of the system in information retrieval, to achieve relevance and pertinence in the result of the search performed by users. The impact and relationship between these variables are depicted in Figure 7.

Figure 7. The variables that impact user satisfaction in an NIS



Source: Prepared by the authors

Dias and Naves (2013, p. 22) define exhaustiveness as "a previous decision taken by the system, to recognize, besides the main subject, all the secondary subjects contained in the document that is being indexed"; while specificity "refers to how precise one can be when specifying the subject of a document that is being indexed". In this sense, we consider that completeness in indexing is related to the decision, by the indexer, about the number of descriptors assigned to a document; and specificity is related to the degree of co-extensiveness between the descriptor assigned to the document and the concept it deals with.

Thus, we can define 'completeness' as the measure of the extent to which the different topics treated in a document are recognized and represented in the indexing. When this occurs, we say that the document has been exhaustively indexed. In this way, it can be retrieved by many index terms or many combinations of terms, which increases its retrievability. A document can be considered under multiple aspects, but only those that meet the purposes of the service should be represented. For example, in a library catalog specialized in hospitality, in a document about beverage services, there is no need to index the agricultural aspects of the plantations and/or regions where such beverages originate, even though the document provides such information. In a specialized beverage collection, by contrast, all aspects are useful. The opposite concept to exhaustivity is selectivity. When few index, terms are selected to represent the subject of a document, its retrieval potential is diminished. In the indexing policy, it must be clear which criterion is adopted (GOMES; CAMPOS, 1998).

Specificity, in turn, in summary, is a principle related to the level of representation of data in a record. Being specific means representing the content of the data at the same level as the document. If a document deals with the qualities of synthetic fabric, it is in this term that it should be indexed, and not in 'textile industry', for example, which would be a broad term that would not represent the subject treated in the document. The opposite concept is generality. This term should be understood as a generic term, that is, the most general term within the category to which the specific term belongs. Thus, the generic term for 'Synthetic Fabric' is 'Fabric' (GOMES; CAMPOS, 1998).

One should always consider that a high exhaustiveness, that is, many descriptors assigned to a document, can cause inadequate relationships and, consequently, false recoveries. Furthermore, factors of paramount importance that must be taken into consideration in these decisions are the user profile and the level of specificity of the subject treated in the document.

Two new concepts enter the scene, and they arise from exhaustiveness and specificity. They have implications for retrieval. They are Revocation and Accuracy.

When a service decides to adopt an exhaustiveness policy, there is the possibility of retrieving more items in response to a search request, but these are retrieved among others that do not satisfy the search. The extent to which this occurs is called revocation.

Lancaster (2004, p. 4) defines revocation as "the ability to retrieve useful documents", and precision "the ability to avoid useless documents". In other words, we consider that revocation expresses how well the system can retrieve all relevant items, that is, the ratio between the number of non-relevant records retrieved and the total number of records retrieved in a search. Precision, on the other hand, expresses how well the system can retrieve only relevant items, that is, the ratio of the number of relevant records retrieved to the total number of records retrieved in a search.

Lancaster (1986, p. 133) proposes a table of criteria to evaluate the efficiency of the system in returning the user's search, in which the author suggests (a) which documents are relevant, (b) which are irrelevant, (c) which are relevant and (d) which are not, according to Table 1, below.

Table 1. Proposed judgment of relevance in information seeking by the user.

		USER RELEVANCE JUDGMENT		
	Documento	**Relevant**	**Not relevant**	**TOTAL**
SYSTEM RELEVANCE PREDICTION	**Retrieved**	a (Hits)	b (Noise)	a+b
	Not Retrieved	c (Misses)	d (Correctly Rejected)	c+d
	TOTAL	a+c	b+d	a+b+c+d

Source: Adapted from Lancaster (1986, p. 133).

The revocability and accuracy of the system, as cited here, are related to completeness and specificity. These measures can be calculated by the expressions:

$$Revocação = \frac{nº\ de\ docs\ relevantes\ recuperados}{nº\ total\ de\ docs\ relevantes\ no\ sistema} \rightarrow R: \frac{a}{a+c} x100$$

$$Precisão = \frac{nº\ de\ docs\ relevantes\ recuperados}{nº\ total\ de\ docs\ recuperados} \rightarrow P: \frac{a}{a+b} x100$$

It is known that exhaustiveness results in high revocation and, consequently, low precision. This relationship is inversely proportional, a high specificity results in high precision and low revocation.

When the user searches an information retrieval service, if he selects few terms to represent his query, there is likelihood of high revocation; for example, if the user searches only for the term 'Ornamental Plants' in a specialized service on Ornamental Plant Cultivation, the revocation will be extremely high and so will be the number of irrelevant documents. However, if he includes one or more aspects of his interest, for example, 'Grafting', then the revocation will be lower, but the probability of retrieving relevant documents increases. Therefore, when the revocation is large, the probability of unwanted documents increases, i.e., it introduces imprecision. So, revocation and precision are in inverse relation: the higher the revocation, the lower the precision.

But exhaustiveness is not the only responsible for imprecision. The indexer can also introduce imprecision in indexing by:

(a) omission of a descriptor or descriptors important to the representation of a subject. For example, the indexer failed to index some relevant aspect; and,

b) use of an inappropriate term to represent an idea. For example, the indexer failed to correctly identify the meaning of the term assigned to a document. In this case, the vocabulary contributed to the indexer's failure (GOMES; CAMPOS, 1998).

The levels of completeness and specificity are related to the system; the measures of revocation and precision, to the retrieval process; besides, there are the measures of user satisfaction - relevance and pertinence.

According to Lancaster (2004), the terms relevance and pertinence are employed to refer to useful items, but defined in diverse ways, demonstrating that there is a controversy in the literature regarding the definition of these terms.

Cooper (1971, p. 19), despite considering the concept relevance inexplicable, points out that "relevance is one of the most fundamental, if not the fundamental, concept found in

information retrieval theory [...] whatever it is, it is at the heart of the problem of intellectual accesses". Also, Silva, Santos and Ferneda (2013) consider that the concept of relevance is subjective and inexact and cannot be defined by mathematical formulas and implemented in computational systems. For the authors, relevance consists of "[...] showing the possibly most relevant results in the form of a ranking, from the most relevant to the least relevant[2]".

One of the first definitions of relevance in the literature was presented by Cuadra and Katter (1967, p. 51), in which the authors point out that "relevance is the correspondence in context between an [information] requirement and an article, that is, the extent to which the article's coverage material is appropriate for the requirement statement."

Vickery and Vickery (2004, p. 265) present a more comprehensive definition of relevance, portraying it as "a measure of the effectiveness of contact between a source and a destination (recipient) in a communication process." Dias and Naves (2013) emphasize that relevance is the judgment made by the user in relation to the search result in a NIS. In short, relevance is the degree of similarity between terms that make up the users' search expressions and the occurrence in documents in the collection or in the indexing terms, being considered a comparative relationship between question and document.

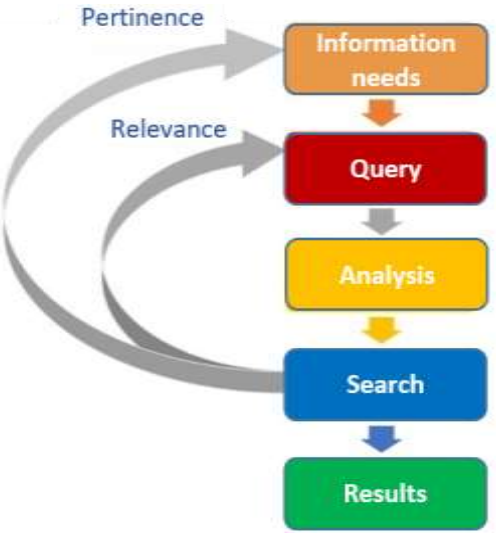
Ingwersen and Järvelin (2005) define relevance as an evaluation of the timeliness, relevance, or usefulness of information sources, carried out by a cognitive actor(s) or algorithmic devices, according to the informational need in a situation; a situation being perceived as a work task, which may be a problem or formulation of an information need, at a specific time. This process is dynamic, the results may change over time, even for the same actor. Relevance can be of a lower or higher order objective nature, that is, of a subjective multidimensional nature, and its measurement is binary or graduated.

While relevance is related to the system's answer to the user's question in a NIS, relevance is the relationship of those retrieved relevant documents to the user's question. Lancaster (2004) views relevance as the usefulness of information in retrieving an item from the library, which aims to fulfill the user's information need. Similarly, Kemp (1974, p. 37) points out that the relevance of a particular document to a particular need is something that can only be decided by the person with the need for that information. For Fosket (1972), pertinence means adding new information to that already stored in the user's mind, which is useful to the informational needs, which motivated that search.

Saracevic (1975) demonstrates, by means of a diagram, Figure 8, the difference between relevance and pertinence. As can be seen, relevance relates the results of the search to the question asked, while relevance relates the results to the information need of that specific user.

[2] In this definition, we could replace the term relevant with appropriate.

Figure 8. Relevance versus pertinence



Source: Adapted from Saracevic (1975, p. 331)

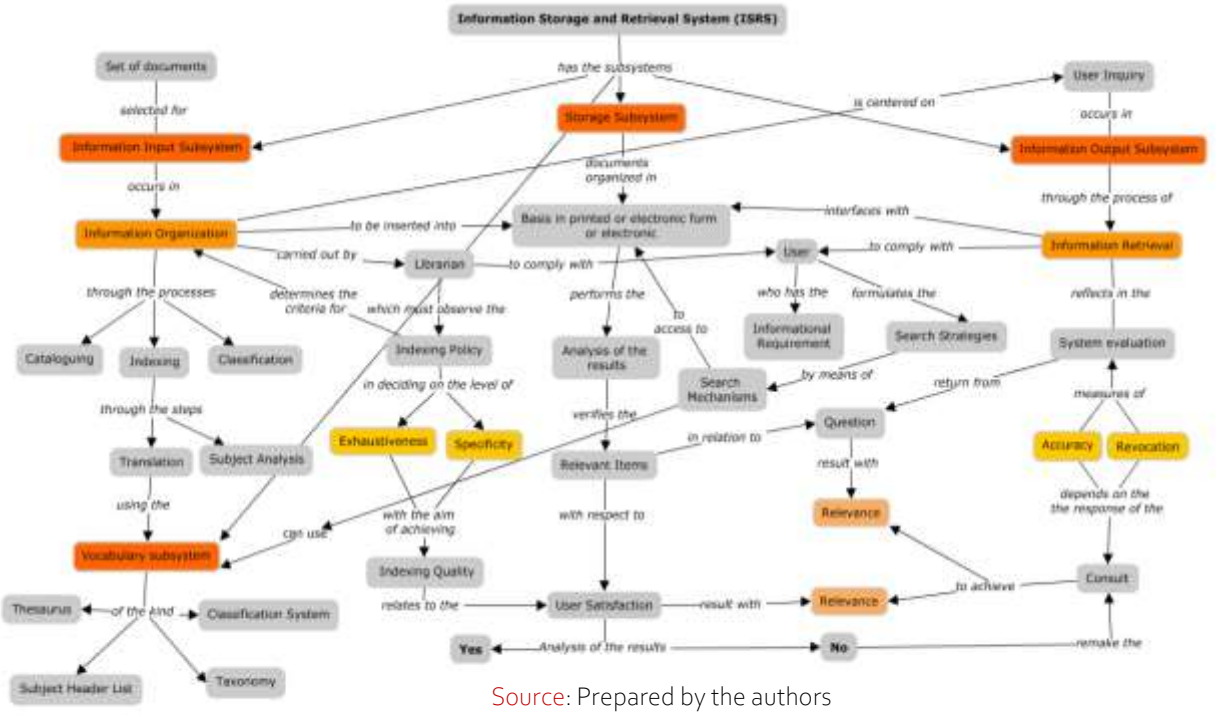
It should be noted that NIS cannot evaluate the relevance of the results, since only the users can do this evaluation. On the other hand, it is up to the NIS to answer the questions asked by users, calculating the relevance between the user's question and the feedback given by the system. In this way, it is inferred that the decision about the relevance of the results is made by the user; in this case, documents considered impertinent in a specific search cannot be considered as a fault of the NIS.

In the next section, we present a proposal for a NIS model, the Information Storage and Retrieval System Model, with the purpose of incorporating all the variables and measures presented above.

6 PROPOSALS OF A ISRS MODEL FROM THE CONCEPT MAP CONSIDERING THE VARIABLES AND MEASURES

From the analysis of the impact of the variables of the indexing policy and the information retrieval measures on a NIS presented in the previous section, it was considered to contextualize the position of each of these variables and measures within a proposal of an Information Storage and Retrieval System. The following is a graphical representation of the relationships of these concepts, by means of a concept map, as shown in Figure 9.

Figure 9. Proposed ISRS model from the concept map



Source: Prepared by the authors

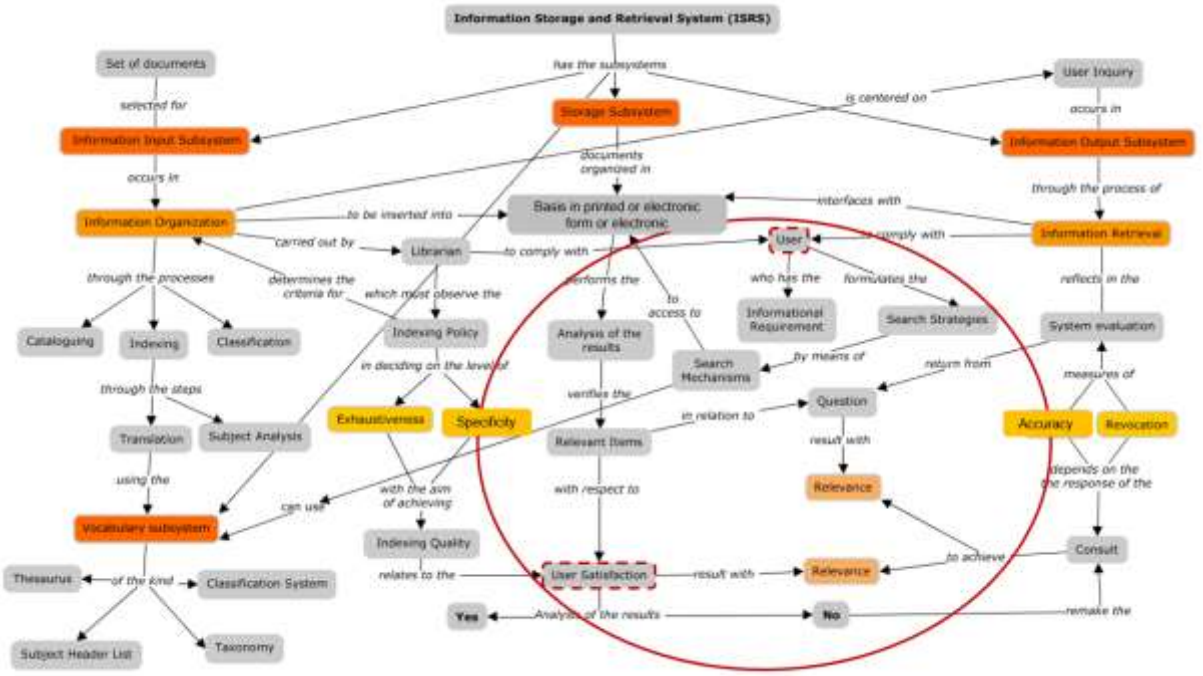
In this map, the propositions make explicit the relationships among the concepts that compose an EWRS, positioning the exhaustiveness and specificity in the input subsystem, the revocation and precision as evaluation measures; the relevance and pertinence among the information storage and retrieval subsystems, relating them to the analysis of the requests made by the users. Note that these variables and measures are intrinsically related.

There are several NIS flowcharts models that depict the activities and processes that occur within an information system. In this case, they are called the system-centric models. These systems explore the relationships between techniques and processes but exclude many variables that are related to information retrieval, which are essential for NIS to be efficient. Among these are knowledge of the community to be served, the user's real needs, and the context in which this NIS is inserted.

The cognitive structure is different for everyone since each person has his or her own world model and processes information in a variable way. Thus, the structure of the EIS, as proposed in this study, considers the world model of its users and how they process information, and how the system interacts with the user in terms of their emotions, intuitions and experiences that represent the user's world model. Therefore, the information professional who mediates between the informational content of the documents and the user who needs the information are seen as parts of the IR systems, not just the processes and technical components.

From this perspective, the following map, Figure 10, shows the beginning of ISRS centered on the System, but with an interrelation with the Information Retrieval part, which is found in both. However, the procedures are circulated, in the context of the users, in relation to the search in the system and their degree of satisfaction, regarding the result achieved. In this case, culminating in the analysis of the degree of relevance or pertinence in their results.

Figure 10. Proposed ISRS model from the user-centered concept map



Source: Prepared by the authors

Based on Lancaster's (1978) perspective, as previously presented (section 3), it is inferred, in this case, that the four subsystems are related; there is no possibility of one existing without the other, and each one has its functionality and role within an EWRS. It is noticeable that the user's role is concentrated between the user-centered subsystem and the storage subsystem, in which they act as the main actors, as well as in the vocabulary subsystem[3], because if the language used is not in line with the user's speech, there will be a gap in the system/user communication. The information professional has a key role in relation to the transfer of information from input to output, as he or she is committed to getting the authors' ideas as reliably as possible to their users.

Thus, it is assumed that the type of system, its characteristics, how and which components and structures make up an EWRS should be contextually related to each other. That is why it is necessary to decide initially which type of variable to give to the system: if the system will adopt a greater exhaustiveness in terms of representation or if the controlled vocabulary adopted will allow a representation of concepts with greater specificity, which is related to the degree of co-extensiveness between the descriptor attributed to the document and the concept treated therein. This will impact the measures of revocation and precision and, consequently, will affect the evaluation of the results of user searches. Relevance, if the system met your need (from the analysis of your question and the system's feedback), and Relevance, from the analysis of these results in relation to your question.

7 FINAL CONSIDERATIONS

Information processing and retrieval are fundamental steps for the understanding of the complex informational process. Such processes have, besides a series of steps, a range of

[3] With current technologies, the vocabulary subsystem has not only the "vocabulary document", but a whole computational environment in which these vocabularies are embedded, in which it is possible to establish definitions and conceptual relationships, knowledge discovery, among others, depending on the type of SOC being used.

interrelated aspects, such as: the perspective of the informational context, the role of the user within this context, the role of the information professional and his/her systemic vision, the issue of the type of knowledge production and its materialization in documents, the type of controlled vocabulary adopted, the technological issues added and, consequently, the way to treat and retrieve the information. If indexing is an activity that occurs in the data entry stage, and retrieval, in the output, in a systemic view, the input affects the output. Therefore, discussions involving the factors, measures, and variables that should be considered in an indexing policy are of fundamental importance.

In this article, we consider highlighting these issues by showing how in the information flows discussed these aspects can be aggregated into a consistent whole, embodying what is called Information Storage and Retrieval System (ISRS). This systemic vision in which all elements of a EWRS are interrelated and in which the user is the center of attention, in our conception, is fundamental for the training of professionals capable of building systems that aim at a treatment with quality in its representations and that will result in a more relevant and pertinent information retrieval.

Thus, we highlight, as a final contribution of this study, the importance of a systemic vision, in which all elements of an Information Storage and Retrieval System (ISAR) are related, having the user as the main element; as well as the activities that are considered important for the formation of professionals capable of building consistent Systems.

REFERENCES

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: Addison Wesley, 2011.

BARRETO, A. A. A condição da informação. **São Paulo em Perspectiva**, v. 16, n. 3, p. 67-74, jul./set. 2002.

CARNEIRO, M. V. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, v. 14, n. 2, p. 221-241, set. 1985. Available at: https://periodicos.ufmg.br/index.php/reb/article/view/36523/28575. Access on: 26 nov. 2021.

CESARINO, M. A. N. Sistemas de recuperação da informação**. Revista da Escola de Biblioteconomia da UFMG**, v. 14, n. 2, 1985. Available at: https://periodicos.ufmg.br/index.php/reb/article/view/36507/28553. Access on: 24 nov. 2021.

CHU, H. **Information representation and retrieval in the digital age**. Melford: Information Today, Inc., 2005.

COOPER, W. S. A definition of relevance for information retrieval. **Information Storage and Retrieval**, v. 1, n. 7, p. 19-37, 1971.

CUADRA, C. A.; KATTER, R. V. **Experimental studies of relevance judgements**. Santa Monica: Systems Development Corporation, 1967. (NSF Rep. TM-3520/001, 002, 003, 3 volumes).

DAHLBERG, I. Knowledge organization: a new science? **Knowledge Organization**, v. 33, n. 1, p. 11-19, 2006. Available at: https://www.ergon-verlag.de/isko_ko/downloads/ko3320061c.pdf. Access on: 26 nov. 2021.

DIAS, E. W.; NAVES, M. L. **Análise de assunto**: teoria e prática. Brasília: Briquet de Lemos, 2013. 115 p.

FERNEDA, E. **Recuperação da Informação:** análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**.** 2003. 147 f. Tese (Doutorado em Ciência da Comunicação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003.

FOSKET, D. J. A note on the concept of "relevance". **Information Storage and Retrieval**, v. 2, n. 8, p. 77-78, 1972.

GOMES, H. E.; CAMPOS, M. L. A. **Política de indexação.** SESC, 1998. (Material didático apresentado no Curso de Capacitação na área de indexação.)

INGWERSEN, P.; JÄRVELIN, K. **The turn:** Integration of information seeking and retrieval in context. Dordrecht: Springer, 2005. 448 p.

INTERNATIONAL STANDARD ORGANIZATION. *ISO 5963-1985* – Documentation: methods for examining documents, determining their subjects, and selecting indexing terms. Suíça: ISO, 1985.

KEMP, D. A. Relevance, pertinence and information systems development. **Information Storage and Retrieval**, v. 10, n. 2, p. 37-47, 1974.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 1993.

LANCASTER, F. W. **Information retrieval systems**. 2. ed. New York: Wiley, 1978.

LANCASTER, F. W. **Vocabulary control for information retrieval**. 2. ed. Arlington: Information Resources Press, 1986. 270 p.

LIMA, G. A. Gênesis da classificação: uma análise de conteúdo a partir da definição. **Perspectivas em Ciência da Informação**, v. 26, n. 1, p. 197-237, mar. 2021. DOI: https://doi.org/10.1590/1981-5344/32686. Available at: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/4402/2463. Access on: 26 nov. 2021.

MEY, E. S. A. **Introdução à catalogação**. Brasília: Briquet de Lemos, 1995.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, p. 20-32, 1951. DOI: https://doi.org/10.1002/asi.5090020107.

NOVELLINO, M. S. F. Instrumentos e metodologias de representação da informação. **Informação & Informação**, v. 1, n. 2, p. 37-45, jul./dez. 1996. Available at: https://www.brapci.inf.br/_repositorio/2010/05/pdf_0e3cc20139_0010458.pdf. Access on: 26 nov. 2021.

ROWLEY, J. **A biblioteca eletrônica.** 2. ed. Brasília: Briquet de Lemos, 2002. 399 p.

ROWLEY, J. **Informática para bibliotecas.** Brasília: Briquet de Lemos, 1994.

SALTON, G. **Automatic information organization and retrieval.** New York: McGraw-Hill, 1968.

SALTON, G.; McGILL, J. M. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.

SARACEVIC, T. Information science. **Journal of the American Society for Information Science**, v. 50, n. 12, p. 1051-1063, 1999. DOI: https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z.

SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science. **Journal of the American Society for Information Science**, v. 26, n. 6, p. 321-343, 1975. DOI: https://doi.org/10.1002/asi.4630260604.

SILVA, R. E.; SANTOS, P. L. V. A. C.; FERNEDA, E. Modelos de recuperação de informação e web semântica: a questão da relevância***. Informação & Informação***, v. 18, n. 3, p. 27-44, 2013. DOI: http://dx.doi.org/10.5433/1981-8920.2013v18n3p27. Available at: https://www.uel.br/revistas/uel/index.php/informacao/article/view/12822/pdf_3. Access on: 26 nov. 2021.

SOERGEL, D. **Organizing Information:** principles of data base and retrieval systems. California: Academic Press, 1985.

TRISTÃO, A. M. D. *et al.* Sistema de classificação facetada: instrumento para organização da informação sobre cerâmica para revestimento. **Informação e Sociedade:** Estudos, v. 14, n. 2, 2004. Available at: https://www.proquest.com/docview/1494045851. Access on: 26 nov. 2021.

VICKERY, B. C.; VICKERY, A. **Information science in theory and practice**. 3. ed. rev. aum. Munique: KG Saur, 2004. 400 p.