GARDEN JOURNALS
Journal of Qualitative Research in Business Law, Eco-Fin, Accounting, and Statistics
ISSN : 1533-1520

Issue 22 . No.2 2022

Strategic Approach To Retrieving Scientific Information On Neglected Tropical Diseases: A Comparative Analysis Of Scopus, Pubmed, And Web Of Science

Owen B. Cooper & Natalie A. Morris
Department of Chemistry, University of Buenos Aires, Argentina

ABSTRACT

The objective is to develop a methodological path to retrieve scientific information about Neglected Tropical Diseases (NTDs) in international databases (Scopus, Web of Science and PubMed). The methodological procedures of this research were: a) identification of the main tropical diseases according to the World Health Organization; b) standardization of disease names based on the Health Sciences Descriptors of the Virtual Health Library (DeCS/VHL); c) construction of two lists: one with controlled terms associated with Boolean operators, and another with uncontrolled terms, also structured with Boolean operators; d) search for information in the main databases mentioned above in a controlled and uncontrolled manner; e) comparison of the results obtained between the controlled and uncontrolled search; f) construction of a methodological flow for replication purposes. It was noticed that the use of controlled search increased the recall of the information retrieval process, specifically, the Scopus Base showed a more significant increase in the retrieval of results (11.28%). PubMed had an increase in recall of 8.92% and Web of Science of 2.64%. Furthermore, this work proposes a flow for retrieving information in the aforementioned databases, which serves researchers on this topic, as it is understood that the high degree of variability of words and the dispersion in the representation of the theme in the databases constitutes important impediment to information retrieval.

Keywords: Tropical Medicine; neglected diseases; data base; information storage and retrieval; controlled vocabulary.

INTRODUCTION

Tropical diseases are infectious diseases that proliferate in hot and humid climatic conditions, common in countries located close to the Equator, between the tropics of Cancer and Capricorn, [1] exactly where a large part of the Brazilian territory is located, providing favorable conditions to the rapid spread of these diseases, requiring continuous attention from the scientific community.

Knowing that the most advanced resources (human, financial and technological) in Medicine are concentrated in developed countries, it is understood that it is essential to build partnerships between developing and developed countries, with the aim of advancing research in Tropical Medicine, stimulating scientific collaboration for the production of medicines and effective solutions to these problems. On the one hand there are consolidated resources to deal with advanced problems in the field of health, on the other there are challenges and researchers eager for fruitful partnerships that contribute to the viability of creative solutions designed in developed countries for developing countries. Such a scenario presupposes possibilities for exchanging resources and knowledge, which are positive for both.

In general, it is admitted that the involvement of developed countries in solving challenges linked to Tropical Medicine is commonly associated with the following factors: 1) the possibility of profiting from developing countries with the sale of vaccines and medicines; 2) outbreaks that affect their territories sporadically, such as the cases of Zika in 2016 in the United States, which resulted in congenital problems in 5% of babies and fetuses of women affected by the disease. [two]

From the pharmaceutical industry, strongly influenced by market forces, there is an inert stance regarding the production of medicines for tropical diseases. Of the 850 medicines and vaccines approved for all diseases between 2000 and 2011, only 4% were for Neglected Tropical Diseases (NTDs), including malaria, tuberculosis, diarrheal diseases and other diseases of poverty. [3]

Having realized the social and academic relevance that surrounds the topic of NTDs, and in particular, the importance of the responsibility of participation of researchers from Latin America in carrying out research on this subject, it is noted that the solution to the questions in this field require interdisciplinary approaches , considering that many areas of knowledge, in addition to the medical domain, have contributions to offer. In this context, as an interdisciplinary area, Information Science (IS) establishes relationships with the most diverse areas of knowledge, aiming to contribute to the solution of different problems, however, centered on informational issues.

Thus, evidently, the intention of a work in the IC area is not to present solutions for NTDs, [4] however, the aim is to understand issues linked to the processes of searching, monitoring and analyzing information located in the field of Tropical Medicine, taking into account that the difficulty of retrieving scientific information about NTDs constitutes an important problem, especially for researchers who do not belong to the medical field, and who, consequently, do not master its terminology.

Thus, it is assumed that IC has as its object the production, selection, organization, interpretation, storage, recovery, dissemination, transformation and use of information, [5] delimiting in this scope its contribution to the most varied themes and areas of knowledge , including for herself. While problematic, from a diachronic perspective, it appears that the discussion of the relationships between the accumulation of information and its subsequent selection and use constitute a phenomenon of interest to IC throughout its short history.

In the context of NTDs, this problem directly interferes with the way and speed in which scientific information is generated, systematized, retrieved and used by researchers, especially considering that in many cases, there is the involvement of scientists outside the medical field who wish to study the topic, however, they have little knowledge about the alternatives and criteria for retrieving scientific information about the illnesses that make up the scope of NTDs. It is understood here that the aforementioned reality provides an obstacle to these researchers, which prevents the advancement of interdisciplinarity in the field in question and the development of creative solutions that can only be developed based on the relationship of diverse disciplinary skills and abilities.

This study justifies the need to use controlled vocabularies complemented by other term discovery methods. [6] The lack of terminological dexterity encourages researchers to limit themselves to general terms, compromising their retrieval rates and harming the composition of the bibliographic corpus. In a previous study, it was possible to verify that the choice of the correct set of controlled terms favored the retrieval of information in the biomedical area, both in precision and recall.[7] For this reason, the need to think about the entire recovery cycle is reinforced, from delimiting the concept of what is intended to be recovered, to its use.

Therefore, the purpose of this study is to develop a strategy for retrieving scientific information about NTDs in the main international scientific information bases that index content in this area. In short, this research aims to develop a methodological path for retrieving scientific information about NTDs in the following databases: Scopus, Web of Science and PubMed.

The aforementioned bases were chosen due to the eminent prestige they have in the academic-scientific environment, and for their vast collection of scientific information in the field of NTD with high quality and impact, thus constituting an indispensable set of information for carrying out research involving this topic. This strategy should advance to other bases and repositories of scientific knowledge as the doctoral project to which it is linked evolves and carries out experiments in other contexts. It is understood in this research that the retrieval of scientific information on a given subject is the starting point for the production of knowledge and the advancement of discoveries and academic dialogues in any area.

The strategies presented here will be used in the future for the Retrieval of Information on NTDs, aiming at bibliometric and informetric applications, which provide comparative panoramas of national and international scientific production on the topic in question. The initial proposal of the doctoral thesis is the exploration of Information Retrieval methods in databases and the classification of scientific information in DTN, and this article is presented as the result of these studies.

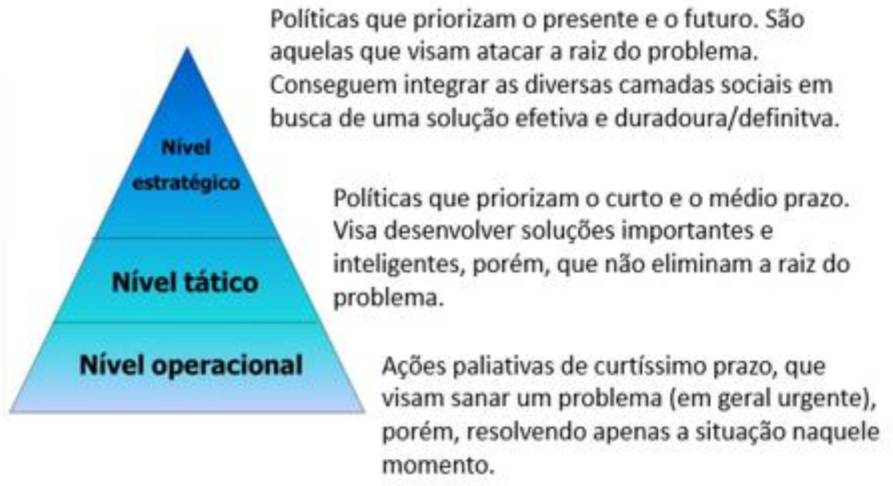
NEGLECTED TROPICAL DISEASES AND PUBLIC POLICIES

Tropical Medicine is an area of scientific knowledge that invests a large part of its efforts in the study of neglected diseases, typical of developing countries, located close to the Equator, specifically located in places with severe poverty. This line of studies, in general, is not part of the agenda of developed countries, which have greater economic and scientific power, due to the fact that they are not frequently affected by such illnesses. Therefore, developing countries bear the greatest responsibility for carrying out this research.

The regions of Latin America, the Caribbean, Africa and Asia are the most affected by tropical diseases. [8] This occurs mainly due to the following causes: enormous ecological diversity, ongoing environmental changes, massive migrations, socioeconomic issues, proximity to the equator (hot and humid climate), sanitary conditions, and lack of effective policies aimed at combating these diseases .

Here, attention is drawn to the fact that in addition to NTDs, there is a neglected tropical science, which reinforces the difficulties of Latin American, African and Asian countries. However, it is admitted that these nations alone will find it difficult to achieve the solutions they need, given their current stage of development in Science, Technology & Innovation (ST&I). To mitigate the problem, the recommendation is to form a collective effort that involves health workers from local communities and researchers interested in tropical diseases from around the world. [9] Without an organized, systematized and monitored global effort, it is unlikely that the necessary solutions to NTD problems will be achieved.

Considering this reality, it is stated that the fight against public and tropical health problems must be carried out at a strategic level, aiming to attack the causes and mitigate the variables that cause diseases. Thus, figure 1 and table point out the levels of action of public health policies regarding their effectiveness. [10]



Políticas que priorizam o presente e o futuro. São aquelas que visam atacar a raiz do problema. Conseguem integrar as diversas camadas sociais em busca de uma solução efetiva e duradoura/definitva.

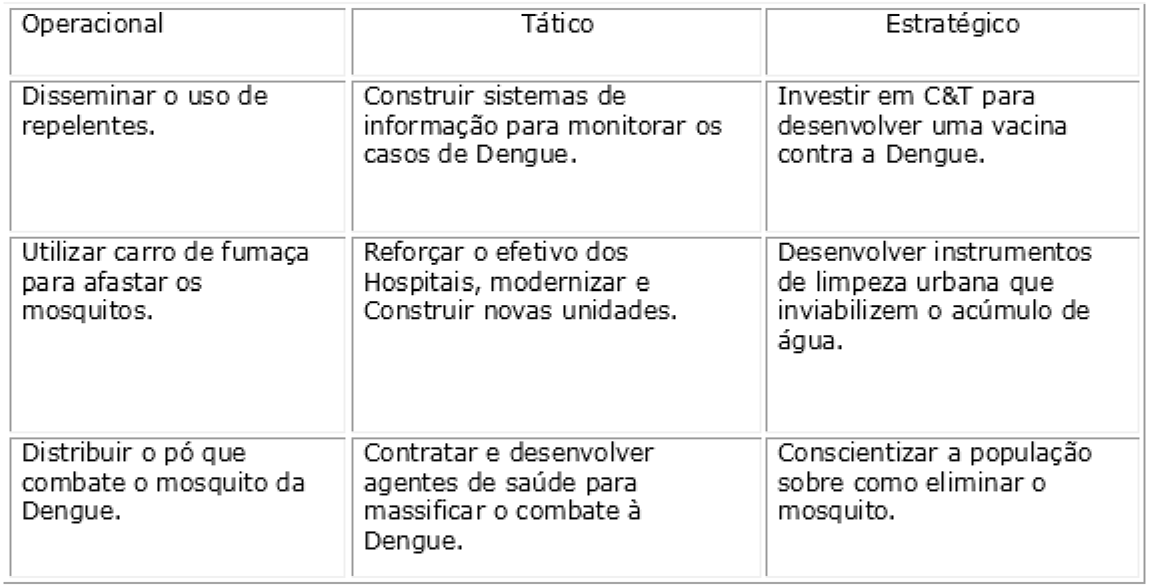
Políticas que priorizam o curto e o médio prazo. Visa desenvolver soluções importantes e inteligentes, porém, que não eliminam a raiz do problema.

Ações paliativas de curtíssimo prazo, que visam sanar um problema (em geral urgente), porém, resolvendo apenas a situação naquele momento.

Fonte: Sobral; 2015.

Fig. 1. Níveis de atuação das políticas públicas em saúde.

Quadro. Exemplo de ações para cada nível de atuação das políticas públicas em saúde – combate à Dengue.



Operacional	Tático	Estratégico
Disseminar o uso de repelentes.	Construir sistemas de informação para monitorar os casos de Dengue.	Investir em C&T para desenvolver uma vacina contra a Dengue.
Utilizar carro de fumaça para afastar os mosquitos.	Reforçar o efetivo dos Hospitais, modernizar e Construir novas unidades.	Desenvolver instrumentos de limpeza urbana que inviabilizem o acúmulo de água.
Distribuir o pó que combate o mosquito da Dengue.	Contratar e desenvolver agentes de saúde para massificar o combate à Dengue.	Conscientizar a população sobre como eliminar o mosquito.

Fonte: Sobral - 2015.

Actions of an operational nature occur in a very short term, in general, they act within the scope of the consequence of the problem in a corrective way, and should serve as a palliative while preventive, tactical and strategic actions are being developed. Tactical actions employ a higher degree of intelligence than operational intelligence. Although it does not serve the purpose of a solution, its nature is closely related to processes of knowledge generation, use of technologies, modernization and planning and management techniques, which allow progress towards full control of a given problem, and also, mitigate the consequences caused by diseases. As for the strategic level, its action is visibly linked to the causes of the problem, and not to the effects. Its essence is oriented towards the construction of effective and lasting solutions, which precisely reach the conditions in which the problem is produced.

This time, it is noted that the first column on the left is concerned with presenting urgent actions to prevent and treat the effects of the proliferation of the mosquito that causes the disease, carrying out acts of a simple and immediate nature. The second column focuses on management and control actions, which aim to expand prevention and treatment conditions, strongly supported by administrative techniques and information and communication technology tools. Meanwhile, the third column emphasizes actions that combat the cause of the problem, bearing in mind that the creation of vaccines, the elimination of mosquito breeding conditions and public awareness, if fully achieved, would eliminate or drastically reduce cases of Dengue in Brazil.

In general, the schemes presented in figure 1 and the table can be used to propose public policies to prevent and combat NTDs. It is admitted that action at the three levels presented is important to combat current health problems, therefore, even operational actions are considered valid, as they help to save time while more elaborate, effective and lasting actions are developed.

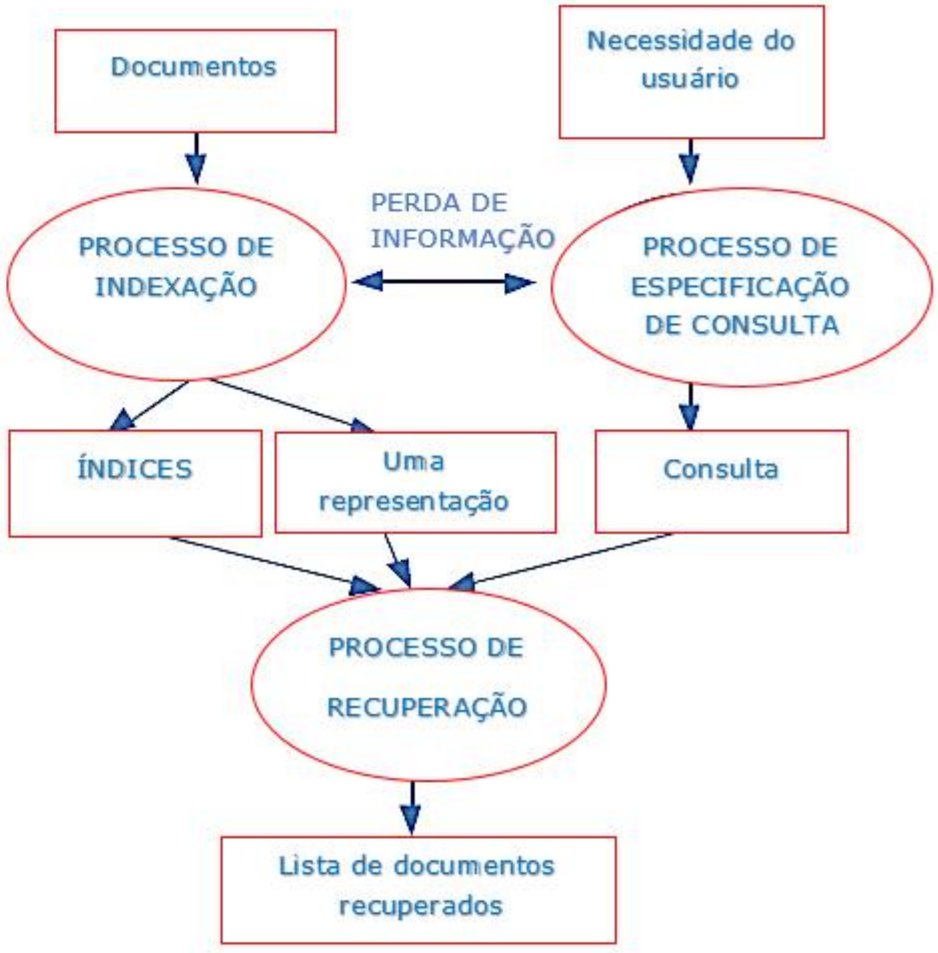
In addition to the assumptions described above, it is recognized that, in order to advance solutions focused on NTD problems, regardless of the level of political action, it is beneficial to apply interdisciplinary knowledge, which goes beyond the limits of the medical field, involving other areas of knowledge and their possible contributions to perceived problems, dealing in greater depth with the complexity that the topic requires.

Contextualizing the aforementioned statement with the propositions presented in the table , the question arises: is it coherent to formulate a policy to raise public awareness about the elimination of the Dengue mosquito without consulting experts in the Human and Social Sciences to understand issues of motivation and communication? Is it plausible to imagine the construction and improvement of urban cleaning instruments without the collaboration of professionals and researchers in the field of Engineering and the Environment? Are the theoretical and technical contributions of Information and Computer Sciences useful for the design of techniques and methods aimed at systematizing and retrieving scientific information in order to favor processes of generating innovation in NTD? Such questions open space for constructive reflections on the role of other domains in solving problems in the NTD area, which will not be explored in depth in this article as it is not its primary focus, however, they are discussions that deserve attention and greater space in reflection circles. academics.

RECOVERY OF SCIENTIFIC INFORMATION AND ITS RELATIONSHIP WITH INFORMATION METRIC STUDIES

Search strategy is the process of translating a search question into a format that the search engine can understand. [11] This strategy is part of the Information Retrieval process. Information Retrieval is the process of finding information, in general, texts in large collections, satisfying users' information needs. [12] *Mooers* , founding author of Information Retrieval, conceptualized it as the process or method by which a potential user of information is able to convert his need for information into an actual list of information about documents of interest to him. [13]

From a practical aspect, the items that make up an Information Retrieval System (Fig. 2) include documents, user needs that trigger the formulation of queries, and finally, Information Retrieval, which depends on the alignment between the indexing of documents and the search carried out. [14,15] As a result of this, a list of documents considered relevant is presented to the requesting user. If there is a significant divergence between the indexing terms and the users' search, the possibilities of information loss are increased, and the Information Retrieval process tends to be less effective. [14]



Fonte: Gey; 1992.

Fig. 2. Componentes de um sistema de recuperação de informação.

Historically, studies linked to Information Organization and Retrieval are present at the center/core of CI. In Web of Science, the first works linked to the topic were published by *Mooers* , who studied the mechanisms of information retrieval and the relationships between communication theory and retrieval theory. [16] In 1956, the same *Mooers* already expressed concern about the rapid development of information retrieval devices, stating that, before a librarian or research administrator understood a process, several new methods would have already been announced, a similar panorama to that found today . [17]

At the end of the 1970s, *Hawkins* developed pioneering work relating Information Retrieval to Information Metric Studies (more specifically bibliometrics), however, such work had the purpose of studying the literature on Online Information Retrieval, and not necessarily understanding the interrelations between subjects. [18] The conclusion of the aforementioned author was that scientific production on the aforementioned topic was dispersed across several journals, many of them not dedicated to Librarianship and IC, which explicitly highlighted the importance of the topic for other areas of knowledge.

Within the scope of Librarianship and IC, some instruments have received greater attention in recent years, among them, the thesaurus, which is one of the most important forms of documentary language, and, like others, appears as a response to the inefficiency of library organization resources. information unable to meet the demands imposed by the specialized document production environment [19] . [19] Dodebei reports that the thesaurus evolves from the need to work with more specific vocabulary than that present in the subject headings (references and cross-references, such as: see and see also). [20] Its main objective is terminological control, which can be achieved with modifiers that contextualize the intended meaning, and with definitions and scope notes that avoid two occurrences: polysemy (depending on the context a word can have more than one meaning), and homonymy (different objects designated by the same word). [21]

The thesaurus can assist the user in informational searches, as well as help the indexer during the classification process. *Moreira* , *Alvarenga* and *Oliveira* [22] consider that the thesaurus is a very important component in a retrieval system as it fulfills the role of: determining which terms can be used in indexing; establish which terms can be used in the search so that it has a satisfactory result; and allow the introduction of new terms and relationships, in order to bring the languages of the user and the system closer together. In the area of Biomedical Sciences, the following stand out as controlled languages: MeSH (*Medical Subject Heading*); and DeCS (Health Sciences Descriptors), the latter, strongly present in this article, was created by BIREME (Latin American and Caribbean Center for Medical Sciences Information) for use in indexing journal articles, books, conference annals, as well as to assist in the retrieval of scientific information in bases such as LILACS and MEDLINE.

Another important function of controlled languages is to avoid the dispersion of information in bibliometric, informetric and scientometric studies, thus allowing the grouping of what is similar, and the separation of what is different, contributing to reliable and precise metric studies. More recently, prominent studies have linked information metric studies with Information Retrieval. *Leydesdorff* and *Bornmann* discussed the categorization of subjects that the Web of Science offers to the scientific journals indexed in its database. Among the numerous concerns, the authors' attention was drawn to the importance of this classification, given its use by the main scientific rankings in the world. [23] Therefore, any problem with subject classification, in addition to causing harm to Information Retrieval, can also compromise the quality of scientific rankings, which are, in short, of bibliometric, scientometric and informetric essence.

Information Retrieval is not just restricted to academic communication contexts and bibliographic databases. Its area of application encompasses artificial intelligence, commercial information, library catalogues, museum and library collections, and the world wide web (web) as a whole (search engines). [24] However, it is reinforced that within the scope of Information Metric Studies, within IC, in general, Information Retrieval is studied, mainly in the universe of magazine articles, and only recently, space has been opened for communication academic on the web (webometrics and altmetrics). [25]

Glänzel , when observing the relationships between Information Metric Studies and Information Retrieval, discusses the importance of metrics to adjust search strategies, and remembers that bibliometric retrieval *is* a powerful tool for developing and adjusting the strategy of search at any level of aggregation, but there will always be noise in the search process. [19] In particular, these noises are caused by indexing problems, synonyms, duplication, incomplete metadata, typing errors, and obviously, it can also occur due to users' lack of dexterity.

Therefore, it is noted that the Information Retrieval process is not trivial, different from a simple consultation process, and therefore, it depends on some specific knowledge to obtain a satisfactory result. Among the various skills required, the following stand out:

- Knowledge of the database structure and its functional requirements.
- Understanding of the topic and the variety of terms that represent it.
- Understanding of the indexing policy and the method used to assign terms.
- Perception of the search interface.
- Skill with operators and use of advanced search.

- Use of search filters.

- Understanding of the possibilities of graphical representation of the information retrieved.

- Measurement of informational statistics for the purpose of validating the search and defining new sections.

Furthermore, as an automated alternative, there is the possibility of constantly searching and monitoring information via API (*Application Programming Interface*). This technology, which uses query commands, requests information from databases to extract data, records and information. Such monitoring can serve different purposes, ranging from monitoring updates on a specific production to obtaining real-time indicators.

On the Web of Science, for example, it is possible to access formatted and updated information to improve the institution's repository; automatically consult multiple records in real time, eliminating the need for manual searches. Using this technique, it is also possible to develop advanced search strategies and request information requests from databases, through computerized routines, in which external applications will have electronic access to the requesting system (http://wokinfo.com/products_tools /products/related/webservices/) .

This study, of a methodological nature, aimed to develop a set of procedures for retrieving information in the main international scientific information bases (Web of Science, Scopus and PubMed). In Brazil, access to these databases is via the Capes Periódicos Portal (PPC). This portal is linked to the Coordination for the Improvement of Higher Education Personnel (Capes), being a virtual library that brings together and makes available to teaching and research institutions in Brazil the best of international scientific production. It has a collection of more than 38 thousand titles with full text, 134 reference bases, 11 bases dedicated exclusively to patents, as well as books, encyclopedias and reference works, technical standards, statistics and audiovisual content (http://www. periodicos.capes.gov.br/index.php?option=com_pcontent&view=pcontent&alias=missao-objetivos&Itemid=102) .

METHODOLOGICAL PROCEDURES

To understand the process of searching and retrieving information, a bibliographical research was carried out, consulting national and international literature that discussed the subject. *Aiming to structure the result in a visually understandable scheme, the flowchart technique, widely adopted in Business Process Management* , was applied . For this purpose, the Bizagi® tool was used (https://www.bizagi.com/pt).

After structuring the flow and carrying out the retrieval process, it was possible to compare the results and verify the percentage increase in information retrieved in each database analyzed, based on the parameter "controlled search using a thesaurus" *versus* "uncontrolled search". The experiments were carried out in 2017, and had technological support from the Otlet CI Laboratory at the Federal University of Pernambuco, Northeast, Brazil.

RESULTS AND DISCUSSION

The aim here was to develop a strategy for retrieving scientific information about NTDs in the main international databases that index content in this area. Therefore, the results obtained by the study are summarized in two topics: 1) systematization of the methodological flow of information retrieval (Fig. 3); and 2) comparison of search and retrieval with and without controlled terms (Fig. 4).

This methodological proposal, presented as part of the research result, served the purpose of testing, aiming to evaluate the effect of information retrieval from obtaining specific words, based on the names of diseases, obtained from the WHO website. Then, the words were converted into controlled terms using DeCS/VHL. In this way, the methodology is intended to be replicable, aiming to prove the benefits of using the instruments: WHO website, DeCS/BVS and Scientific Databases, for composing precise terms in NTDs, according to the steps described below:

a) identification of the main tropical diseases according to the World Health Organization (WHO) website (it is worth highlighting that this list is constantly updated: http://www.who.int/neglected_diseases/diseases/en/): Buruli ulcer, Chagas disease, Dengue and Chikungunya, Dracunculiasis (guinea-worm disease), Echinococcosis, Yaws (Endemic treponematoses), Foodborne trematodiases, Human African trypanosomiasis (sleeping sickness), Leishmaniasis, Leprosy (Hansen's disease), Lymphatic filariasis, Onchocerciasis (river blindness) , Rabies, Schistosomiasis, Soil-transmitted helminthiases, Taeniasis/Cysticercosis and Trachoma.

b) conversion of disease names into controlled terms based on the Health Sciences Descriptors of the Virtual Health Library (DeCS/VHL). To do this, the VHL website (http://decs.bvs.br/) was accessed , typing the name of each disease in the query field, and obtaining the equivalent names.

c) construction of two lists: one with controlled terms associated with Boolean operators, and another with uncontrolled terms, also structured with Boolean operators. The lists were structured as follows:

- No control (24 words): ("Buruli ulcer" OR "Chagas disease" OR "Dengue" OR "Chikungunya" OR "Dracunculiasis" OR "guinea-worm disease" OR "Echinococcosis" OR "Endemic treponematoses" OR "Yaws" OR "Foodborne trematodiases" OR "Human African trypanosomiasis" OR "sleeping sickness" OR "Leishmaniasis" OR "Leprosy" OR "Hansen disease" OR "Lymphatic filariasis" OR "Onchocerciasis" OR "river blindness" OR "Rabies" OR "Schistosomiasis " OR "Soil-transmitted helminthiases" OR "Taeniasis" OR "Cysticercosis" OR "Trachoma").
- With control (47 words): ("Buruli Ulcer" OR "Mycobacterium ulcerans Infection" OR "Chagas Disease" OR "South American Trypanosomiasis" OR "Dengue" OR "Dengue Fever" OR "Chikungunya virus" OR "Chikungunya" OR " Dracunculiasis" OR "Dracunculosis" OR "Guinea Worm Infection" OR "Echinococcosis" OR "Hydatid Cyst" OR "Hydatidosis" OR "Cysts, Hydatid" OR "Yaws" OR "Frambesia" OR "Trematoda" OR "Flukes" OR "African Trypanosomiasis" OR "African Sleeping Sickness" OR "Nagana" OR "Leishmaniasis" OR "Leprosy" OR "Hansen's Disease" OR "Filarial Elephantiasis" OR "Lymphatic Filariasis" OR "Bancroftian Elephantiasis" OR "Bancroftian Elephantiasis" OR "Lymphatic Filariasis" OR "Onchocerciasis" OR "Rabies" OR "Hydrophobia" OR "Schistosomiasis" OR "Bilharziasis" OR "Helminthiasis" OR "Nematomorpha Infections" OR "Helminth Infestation" OR "Parasitic Worm Infections" OR "Parasitic Worm Infestations" OR "Vermination" OR "Taeniasis"

OR "Taenia Infections" OR "Cysticercosis" OR "Taenia" OR "Trachoma" OR "Egyptian Ophthalmia").

d) search for information in the main databases (Web of Science, Scopus and PubMed) in a controlled and uncontrolled manner;

e) comparison of the results obtained between the controlled and uncontrolled search;

f) construction of a methodological flow for the purpose of replicating the methodological strategy.

Based on these steps, the scheme presented in figure 3 was created , highlighting a technical process aimed at researchers from any area of knowledge who wish to seek scientific information on the topic of NTD, whether for the purpose of metric information studies, or for the composition of bibliographic corpus. Below, the step "application of filters specific to the databases to refine the results obtained" was added, which is related to the refinement of the records recovered, using categories pre-established by the databases selected for the research.

The Tropical Medicine literature lacks a clear delimitation as to which diseases belong to the list of NTDs. As *Camargo* states, a priori, all human diseases are tropical, since the human species originated in the tropics, and with it, its diseases, except the diseases that humanity has acquired, throughout its history as companions of journey like dogs, cats, rodents, birds and even close relatives, primates. [1]

By way of example, *Silva* and *Domingues* point out that it is necessary to observe the different dimensions (political, economic and social) to determine whether a disease is in fact tropical. [26] In general terms, it is understood that the factors causing diseases are diverse, which makes it difficult to determine the disease as a "tropical disease". Thus, it is noted that there are several classifications for the subject, and it is a process that is constantly updated, given the frequent emergence of new diseases. In any case, in agreement with *Camargo* , it is understood that the use of the term "neglected diseases" appears as the most appropriate, as it does not single out the tropics as a causative factor, and strongly refers to the issue of poverty, which is a key factor for the lack of prevention and treatment of these diseases. [1]

Therefore, it is thought that the systematization proposed by the WHO is a good starting point from the classification aspect of diseases, above all, due to the institution's egregious position as a promoter and encourager of knowledge on this topic. Thus, in the process expressed in figure 3 , preference was given to mapping the list of diseases belonging to the list of NTDs from the WHO's perspective, this being a legitimate path to be taken by researchers who wish to obtain information on the subject.

To solidify the search strategy, we chose to use the DeCS/VHL thesaurus. Frequently used in scientific works, the study by León et al. stands out here, who used this instrument to standardize terms and carry out structured searches using algorithms based on Boolean language, [27] similarly to the present work.

With this, the information retrieval process becomes more qualified, as in addition to being based on the WHO NTD classification, it also adopts a thesaurus to list equivalent terms, aiming to increase query recall. In general, recovery processes are based only on general terms, such as "tropical medicine", "tropical health" and "tropical diseases", as can be seen in previous work. [28] This strategy, as the authors themselves criticize, may prove to be inefficient as it restricts the amount of information retrieved. The ideal is to use, in addition to general terms, the names of the diseases (specific terms), their equivalent terms (synonyms), and, when applicable, the disease-causing agents. This strategy ensures that the maximum number of records will be recovered, avoiding the loss of information due to the use of a restrictive search strategy.

Aware that one of the problems of information retrieval in DTN is the loss of information due to the use of uncontrolled search expressions, we sought to compare the results of searches with and without the use of a thesaurus in percentage terms (Fig. 4). To avoid retrieving results for the topic consulted in other domains, in addition to Medicine, the search in each database was limited to the most specific domain linked to the topic. Example: In Web of Science, it was possible to limit it to the area of Tropical Medicine; in Scopus, simply Medicine; and in PubMed, as it is essentially a health database, it was decided not to make delimitations.

As shown in figure 4 , Scopus was the database that showed the most significant increase in the retrieval of results using controlled terms (11.28%). Then, PubMed (8.92%) and Web of Science (2.64%). In general, with the proposed strategy it was possible to obtain an increase in information retrieval. Thus, it is assumed that the search without the use of a thesaurus generates loss of information, and with the use of a thesaurus it generates gains, considering that the thesaurus expands the possibilities of recall by indicating a set of synonymous terms that represent a certain theme, and also certifies that the keywords used are legitimate and present in the literature.

For researchers interested in NTDs, this result is encouraging, as it allows: 1) the conceptual delimitation of NTDs and the perception of the set of diseases that make up this theme; 2) obtaining equivalent keywords, solving problems of an informational and linguistic nature that hinder the query; 3) strategy for obtaining metainformation about NTDs, which can be exported for bibliometric purposes, and also; 4) the acquisition of documents for literature review purposes, which is fundamentally useful for researchers in the Human and Social Sciences who wish to gather scientific information about the NTD umbrella.

FINAL CONSIDERATIONS

The proposed strategy can be effective, especially for researchers who do not have command of the terminology in the area of Tropical Medicine, but wish to develop studies in the area. The difficulty in obtaining terms that represent the scientific production on NTDs motivated this research, given the project's future need to collect metainformation from documents to carry out metric studies involving the NTD theme.

In short, Scopus and PubMed presented very significant results regarding the advantages of using a thesaurus to search for information. The results achieved in the Web of Science should not be overlooked, given that depending on the context of the results, a document can represent a significant success for a researcher, therefore, the maximum amount of relevant information on a given topic always favors an effective strategy .

In future studies, we intend to carry out a more qualitative analysis of the results found, aiming to understand in which items the loss of information occurs based on the individual analysis of each keyword used. Another future strategy is the use of other thesauri in addition to the DeCS/BVS. It is suggested here to further explore MeSH (*Medical Subject Headings*), the controlled vocabulary of articles indexed in PubMed, which, by the way, underpins the DeCS/VHL.

Finally, it should be noted that the results were generated at the beginning of 2017 and are in the process of being improved based on monitoring and testing in other databases, with the aim of building indicators linked to Brazilian and international scientific production on NTDs, including , using the ScriptLattes tool (http://scriptlattes.sourceforge.net/) and automatic techniques for generating and processing keywords that will make up chapters of the research author's doctoral thesis. Thus, the findings presented here will guide the progress of the project, establishing a starting point for obtaining the corpus to be studied.

Thanks

Our gratitude to the Otlet CI Laboratory at the Federal University of Pernambuco and the Coordination for the Improvement of Higher Education Personnel (Capes) for their history of supporting this project. Special thanks to the evaluators of this edition who helped raise the level of the article.

Authors' contributions

All authors participated in the preparation of the article according to the order of authorship indicated: supervisee, supervisor and co-supervisor.

Conflict of interests

We declare that there are no conflicts of interest in this article.

REFERENCES

1. Camargo EP. Tropical diseases. Advanced studies internet]. 2008 cited 10 November 2017];22(64):95-110. Available at: http://www.producao.usp.br/handle/BDPI/11791

2. Centers for Disease Control and Prevention. CDC analysis of data from US territories finds serious birth defects in about 1 in 12 fetuses or infants of pregnant women with Zika infection in the first trimester. United States of America: Department of Health & Human Services ; 2017 cited 10 November 2017] . Available at: https://www.cdc.gov/media/releases/2017/p0608-zika-data-first-trimester.html

3. Pedrique B, Strub-Wourgaft N, Some C. The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment. The Lancet Global Health. 2013;1(6):e371-e9.

4. Sobral NV. Alignment of the scientific production of the postgraduate program in tropical medicine at UFPE with the social needs of tropical health in Pernambuco: scientometric analysis dissertation]. Recife: Federal University of Pernambuco; 2015 cited 10 November 2017]. Available at: http://repositorio.ufpe.br/handle/123456789/13842

5. Griffith BC. Key papers in information science . New York: Knowledge Industry Publ; 1980.

6. Reyes AAM, Peña CN. Methods and trends of biomedical and genomic information retrieval based on semantic relations of thesauri and

MeSH. Bibliotecological Research: Archivonomy, Bibliotecology and Information internet]. 2016 cited 2 December 2017];30(68):109-23. Available at: http://www.sciencedirect.com/science/article/pii/S0187358X16300326

7. Su Y, Andrews J, Huang H, Wang Y, Kong L, Cannon P, Xu P. Reengineering of MeSH thesauri for term selection to optimize literature retrieval and knowledge reconstruction in support of stem cell research. BMC Medical Informatics and Decision Making. 2016 cited 2 December 2017];16(54):1-10. Available at: https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-016-0298-z?site=bmcmedinformdecismak.biomedcentral.com

8. Dujardin JC, Herrera S, Rosario V, et al. Research Priorities for Neglected Infectious Diseases in Latin America and the Caribbean Region. PLoS Negl Trop Dis internet]. 2010 cited 10 November 2017];4(10). Available at: http://dx.doi.org/10.1371/journal.pntd.0000780

9. World Health Organization (WHO). The Millennium Development Goals Report internet]: United Nations; 2015 cited 10 November 2017]. Available at: http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

10. Sobral NV. Social Responsibility and Public Policies in the field of Health. Brazil: Imip; 2015. p. 1-32.

11. Gillespie LD, Gillespie WJ. Finding current evidence: search strategies and common databases. Clin Orthop Relat Res 2003;413:133-45.

12. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press; 2008.

13. Mooers CN. Zatocoding applied to mechanical organization of knowledge. J Assoc Inform Scien Technol. 1951;2(1):20-32.

14. Gey F. Models in Information Retrieval. 19th ACM Conference on Research and Development in Information Retrieval: Proceedings of the 19th ACM ; 1992.

15. Cardoso ONP. Information Retrieval. INFOCOMP Journal of Computer Science internet]. 2004 cited 10 November 2017];2(1):33-8. Available at: http://infocomp.dcc.ufla.br/ojsfiles/journals/1/articles/46/submission/proof/46-1-64-1-10-20140917.pdf

16. Mooers CN. Choice and coding in information retrieval systems. Transact IRE Profess Group Inform Theory. 1954;4(4):112-8.

17. Mooers CN. Zatocoding and developments in information retrieval. Aslib proceedings: MCB UP Ltd; 1956. p. 3-22.

18. Hawkins DT. Bibliometrics of the online information retrieval literature. Online Review. 1978;2(4):345-52.

19. Lopes IL. Use of controlled and natural languages in databases: literature review. Information Science. 2002;31(1):41-52.

20. Dodebei VLD. Thesaurus: Documentary memory representation language. Niterói: Intertext; 2002.

21. Cintra AM, Thalamus MFGM, Lara MLG, Kobashi NY. To understand documentary languages. São Paulo: Polis; 2002.

22. Moreira A, Alvarenga L, Oliveira A. The level of knowledge and instruments of representation: thesauri and ontology. DataGramaZero. Rev Ciênc Inform internet]. 2004 cited March 27, 2017];5(6). Available

at: http://www.brapci.inf.br/index.php/article/view/0000007546/48a6a7587e86e3ae4285329027026973

23. Leydesdorff L, Bornmann L. The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: the cases of "library and information science" and "science & technology studies". J Assoc Inform Scien Technol. 2016;67(3):707-14.

24. Mayr P, Scharnhorst A. Scientometrics and information retrieval: weak-revitalized. Scientometrics. 2015;102(3):2193-9.

25. Glänzel W. Bibliometrics-aided retrieval: where information retrieval meets scientometrics. Scientometrics. 2015;102(3):2215-22.

26. Silva PCV, Domingues ALC. Epidemiological aspects of hepatosplenic schistosomiasis in the State of Pernambuco, Brazil. Epidemiol Serv Saúde. 2011 cited 10 November 2017];20(3):327-36. Available at: http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742011000300007&lng=pt&nrm=iso

27. León LHB, Barrera AAL, Rodríguez AMA, Vega LCE. Analysis of scientific production published between 2008 and 2013 on suicide in children, girls and adolescents through a bibliometric study. Rev Hosp Psiquiátr Habana internet]. 2015 cited 10 November 2017];12(2):1-16. Available at: http://www.medigraphic.com/pdfs/revhospsihab/hph-2015/hph152i.pdf

28. Sobral NV, Silva FM, Miranda ZD. Profile of scientific production in Tropical Medicine in Latin America: analysis of the term "Tropical Medicine" in the Web of Science. Question. 2017 cited 10 November 2017];23:31-49.

GARDEN JOURNALS

Journal of Qualitative Research in Business Law, Eco-Fin, Accounting, and Statistics

ISSN : 1533-1520

Enhancing Data Normalization Requirements in Information Metric Analysis Software

Isaac Z. Peterson & Elizabeth L. Morgan

1Department of Psychology, University of Toronto, Canada

2 College of Medicine and Health Sciences, Sultan Qaboos University, Oman

ABSTRACT

Due to the diversity of forms in the entry of the author-affiliation fields, the normalization of bibliographic data is one of the problems that limit the analysis of metric information at run time, reliability of the indicators and size of the data *corpus* . This work aims to propose the requirements for improving data normalization in metric analysis software. To achieve the objective, a diagnosis of the main methods and techniques that are used worldwide in this type of study was carried out. As a main result, the requirements for an application of automated data preprocessing for metric purposes are related. The database, tasks, steps and algorithms that this application will contain are proposed. A combination of algorithms must be used to disambiguate the affiliation and author fields.

Keywords: data processing; data mining; bibliometrics; literature-based discovery; analysis of data.

INTRODUCTION

There are three types of scientific concepts: classificatory, comparative and metric. The first two are qualitative, while the metrics are quantitative. Obtaining quantitative concepts is based on the measurement of the corresponding magnitudes. [1] In that sense, forming quantitative concepts in social sciences, and specifically in Information Science, lies in taking qualitative concepts to quantitative ones; That is, in searching for measurement units for the different problems associated with information activities, such as the analyzes associated with the study of scientific and technological activity at various levels of complexity. That is why it is about identifying the scientific development of a country, representing a domain of knowledge, characterizing the flows in collaboration, etc., based on metric indicators such as article count, patent count or the amount of collaboration. that an institution records, among other magnitudes. [2]

The obtaining of metric concepts currently has several meanings, such as bibliometrics, informatics, scientometrics, patentometrics, etc. depending on the object and topic of study of each of them. However, in general it could be stated that metrics are an instrumental discipline, which applies metric indicators to the information recorded in different media, using techniques from any analysis and visualization algorithm; That is, it involves the application of an algorithm to any meaningful data set. This set of data is contained on a medium and comes from a specific source such as digital or paper databases (DB). Likewise, these supports have different structures and purposes. In this work, the problems associated with bibliographic digital databases will be addressed, regardless of their content (biological, patent, press, etc.).

These databases have a high probability of recording "dirty" or "noisy" data, due to the way the information is collected, the different citation formats, violations of integrity restrictions and standards, the names of very frequent or ambiguous authors, abbreviations of the names of publication sources and large volumes of citation data, etc. Ambiguity in author names in bibliographic databases has long been recognized as a major problem.

The result of metric analyzes based on these "noisy" values can lead to erroneous and unrealistic interpretations. Likewise, it can lead to masking useful behavioral patterns that are hidden in the data, as well as poor performance and quality of output. All of these elements are important causes to concentrate efforts on data preparation. This problem coincides with what was reported in the literature consulted. [3-6] According to *Spinak* [7] and *Lardy,* [8] these problems constitute a major drawback for the metric exploitation of DBs.

Terms such as preprocessing, normalization and data preparation are often used interchangeably to name the data preprocessing stage. In Abstract, the preprocessing of the data contained in the databases, up to the current moment, has the following limitations and particularities:

1. There are numerous causes that cause "dirt" in the system records, which results in a large amount of stored data that lacks adequate quality to be used reliably and it is necessary to treat it in different ways.

2. Data cleaning is divided into several steps: separating elements, standardizing, verifying, comparing, grouping and documenting.

3. Data correction includes removing duplicate records or records with invalid values. In many cases, the information and knowledge available is insufficient to determine the transformations necessary to eliminate anomalies; We are only left with the deletion of these records as the only practical solution, even though it may lead to the loss of information.

4. It is the stage in which the analyst invests the greatest amount of time and effort.

5. It is highly expensive in computing time.

Currently, problems in DBs can be recognized and resolved in two ways:

a) Manually with the participation of a specialist. Many of the tasks in this stage are performed manually or with a very low level of automation by experts in the topic in question, since prior knowledge of the topic and certain expertise and skills are required. [9]

b) Automatically, to a lesser or greater degree, with the use of tools for the detection of particular values that are in contradiction with some functional dependencies implicit in the DB. [10] However, existing software tools aimed at metric analysis do not sufficiently support preprocessing.

Given the context explained above, the objective of this work has been to propose the requirements to improve data normalization in metric analysis software.

RATIONALE OF THE PROPOSED METHOD

DIAGNOSIS

A diagnosis of the methods and trends used in data preprocessing was carried out. For this purpose, a search was carried out in the multidisciplinary database Scopus and 486 records were obtained in the period 1985-2015. The strategy followed was to search for possible terms associated with the object of study, since there is no single descriptor defined for this area of knowledge. This analysis served as the basis for identifying the research field and the algorithms that can be used for preprocessing.

The 486 documents were analyzed using the classic documentary analysis technique. Based on the methods used internationally for preprocessing and the algorithms described, the fundamental requirements that a computer application for data preprocessing must have were obtained as results. The design of the database, the tasks that the application must perform and how they are articulated are explained.

The map obtained was processed following the ViBlioSOM methodology, which allowed, based on the Self-Organizing Maps (SOM) algorithm , to organize the input information automatically and visualize important relationships between the data. In this case, the descriptors associated with the analyzed articles are represented.

The map represented in Figure 1 shows that data preprocessing techniques range from decision trees (*Decision Trees* , *cluster* C8) to genetic algorithms (*Genetic Algorithms* , *cluster* C2). Other techniques can be identified such as *Artificial Neural Networks, Association Rule Mining, Bayes Theorem, Clustering, Discretizations, Heuristic Methods, Hierarchical Systems, Feature Selection, Learning Algorithms and Machine-Learning* . These techniques fall within the field of artificial intelligence and are mainly applied for the disambiguation of author and affiliation data, as well as for natural language processing in text processing.

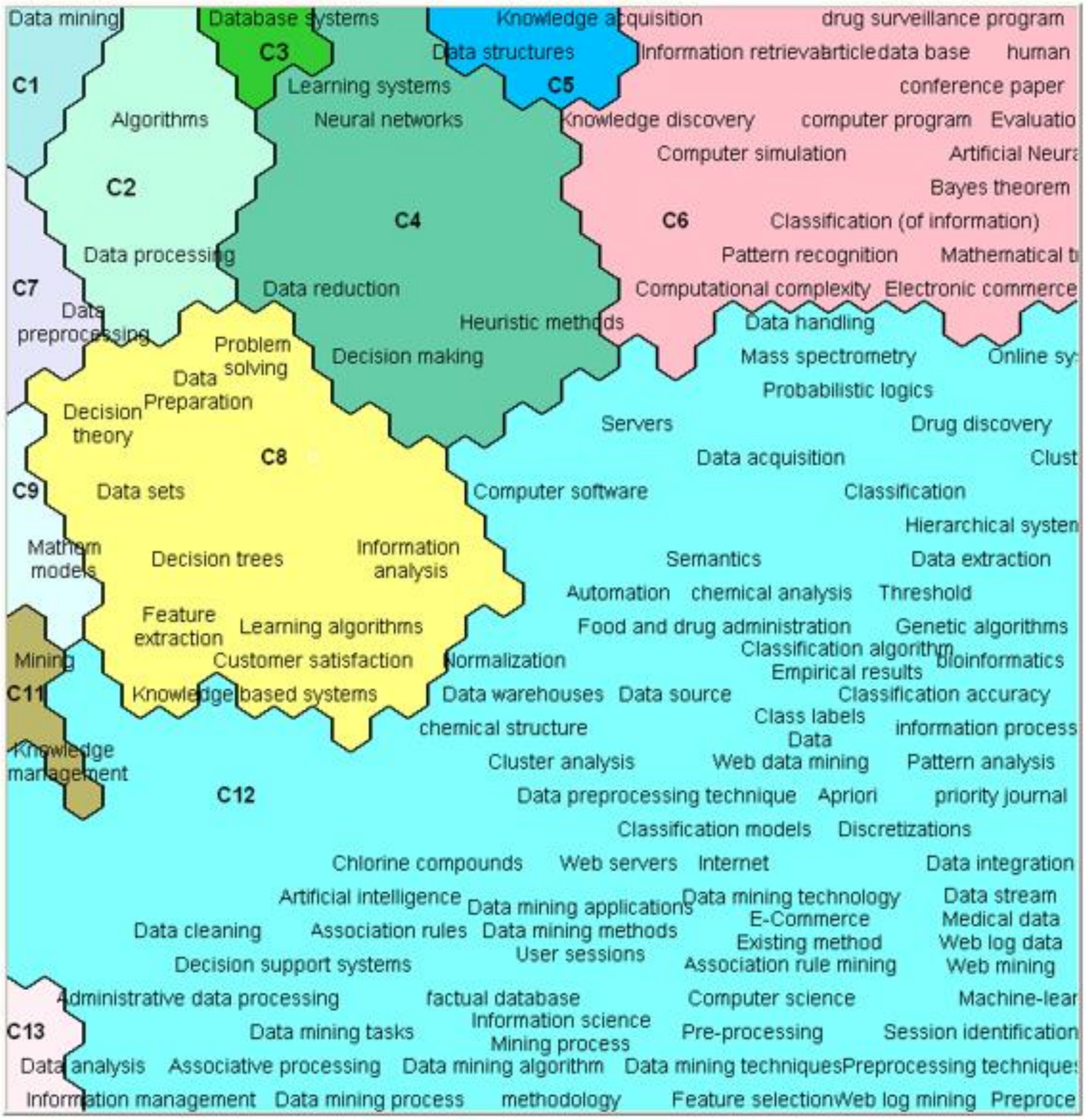


Fig. 1. Mapa de las líneas de investigación relacionadas con el preprocesamiento de datos.

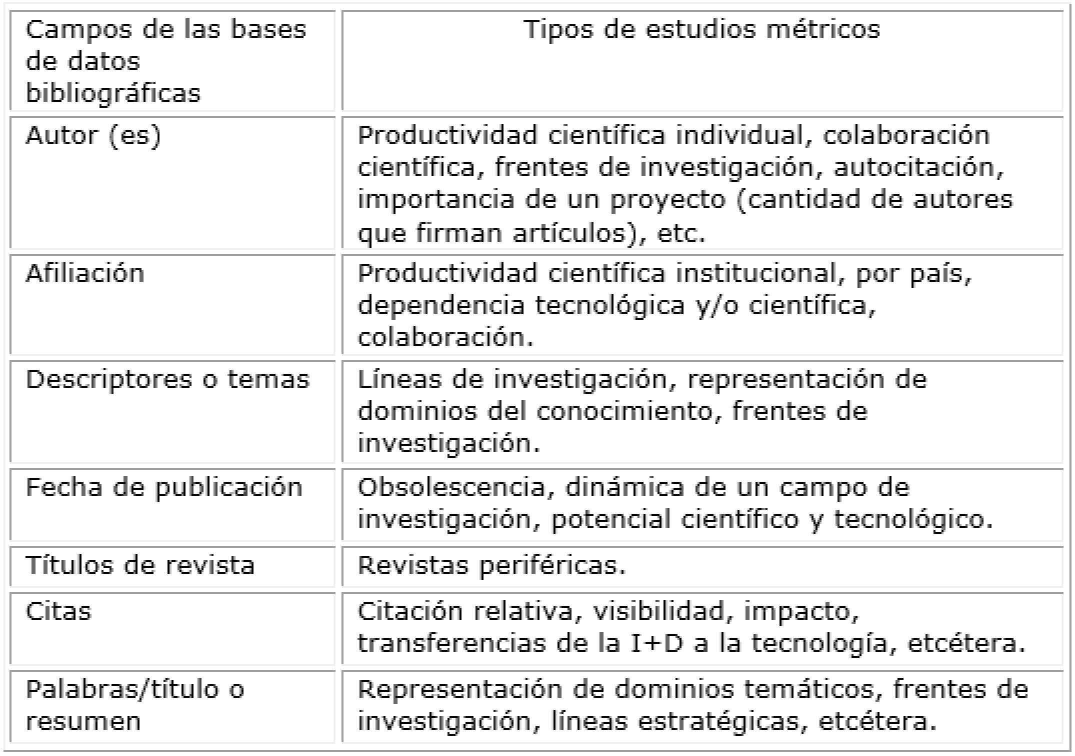
The map does not explicitly show those techniques that are part of statistics and that are fundamentally applied to the treatment of errors and resolution of conflicts; that is, for the treatment of missing values, outliers, duplicates and noisy data, although all these techniques are closely interrelated with each other.

In the database analyzed, no studies were found in Cuba that addressed the problems related to data preprocessing focused on metrics. This contrasts with the increase reported in the literature in the use of metric techniques and their current level of applicability (intelligence studies, scientific-technological surveillance, project evaluation, etc.). [2]

BIBLIOGRAPHICAL DATABASE FIELDS TO NORMALIZE

Table 1 details the fields to be normalized and the types of related metric studies. The fields from bibliographic databases that have high relevance and the greatest need for preprocessing for metric studies are the authors, the affiliation of origin of the author or signatory of a patent, as well as thematic data (descriptors, MeSH terms or Medical Subject Headings, topics, etc.). [eleven]

Cuadro 1. Campos a normalizar y estudios relacionados



Campos de las bases de datos bibliográficas	Tipos de estudios métricos
Autor (es)	Productividad científica individual, colaboración científica, frentes de investigación, autocitación, importancia de un proyecto (cantidad de autores que firman artículos), etc.
Afiliación	Productividad científica institucional, por país, dependencia tecnológica y/o científica, colaboración.
Descriptores o temas	Líneas de investigación, representación de dominios del conocimiento, frentes de investigación.
Fecha de publicación	Obsolescencia, dinámica de un campo de investigación, potencial científico y tecnológico.
Títulos de revista	Revistas periféricas.
Citas	Citación relativa, visibilidad, impacto, transferencias de la I+D a la tecnología, etcétera.
Palabras/título o resumen	Representación de dominios temáticos, frentes de investigación, líneas estratégicas, etcétera.

GENERAL METHODS FOR DATA PREPROCESSING

The best results of preprocessing methods and algorithms depend on the nature of each data set. The role played by the data analyst's experience is also relevant. The general methods identified that were appropriate and feasible to employ in this study are explained below.

Methods for disambiguation of "author" and "author affiliation" data

The problem of associating names with actual entities is known as name disambiguation. The disambiguation of authors' names is a process that aims to simultaneously separate cases of ambiguous names referring to different individuals and merge cases of variant names referring to the same individual. The problem of disambiguation of author names includes:

1. *Synonymy:* the same individual can publish under multiple names. This includes:

a) Spelling variants and letter changes.

b) Typing errors.

c) Name changes over time as occurs through marriages, divorces, religious conversion or sex change.

d) Use of pseudonyms, aliases and variants of names and surnames.

2. *Homonymy:* many different individuals have the same name.

3. The necessary metadata is often incomplete or missing.

4. Many publications have multiple authors, who also represent multiple institutions.

The field "institutional affiliation" presents an analogous ambiguity. Many affiliations can appear with different variants in the DB. Ambiguities also occur as a result of hierarchy, since some institutions belong to others and can be interpreted as if they were different. The extensive use of acronyms and acronyms to identify institutions is also the source of ambiguities.

Disambiguation of author names and affiliations is a fundamental step for the identification of knowledge domains and for other metric analyses. [12] In this study it was proven that there is no disambiguation method that should be taken as a paradigm. Each research task, each database, each data set has its own particularities. The flexibility of the method and the appropriate balance between accuracy, scalability and computing time must be sought. Furthermore, it was appreciated how each of the different authors consulted experiments with various variants and approaches, combining different functions and disambiguation algorithms in different DBs and then compare the efficiency and results obtained by other authors with their own. Table 2 includes a Abstract of the methods that appear in the name disambiguation literature.

Cuadro 2. Métodos generales de preprocesamiento de datos

Método	Comentarios
Tratamiento a valores atípicos (*outliers*) o con ruido	
Eliminación	Puede provocar pérdida de información.
Agrupamiento	Se agrupan los valores de los atributos en grupos, se detectan y luego se eliminan o imputan (sustituyen) los *outliers* por un valor.
Muestreo	Se considera solamente una parte de los valores de los atributos y luego se estima un valor representativo.
Ordenamiento	Consiste en ordenar los valores de atributos, dividirlos en intervalos y luego escoger para cada intervalo un valor representativo.
Data *scrubbing*	Consiste en limpieza de errores tipográficos.
Tratamiento a datos faltantes	
Ignorar y descartar datos faltantes	Se eliminan del conjunto de datos los registros que contengan valores faltantes de un atributo. Es apropiada para los conjuntos de datos con aleatoriedad de la clase MCAR (*Missing Completely At Random*), ya que no se introduce sesgo en los datos.
Imputar los valores faltantes	Consiste en sustituir los valores faltantes mediante alguno de los métodos de imputación.
Imputación de valores atípicos y datos faltantes	
Regresión	Se remplazan todos los valores faltantes o atípicos con un estadígrafo y se asume una relación lineal entre los atributos.
Hot-deck	Se remplazan todos los valores faltantes o atípicos con una distribución estimada de los datos reales (o sea, no imputados previamente).
Cold-deck	Es similar al *hot-deck*, pero el valor que se imputa tiene que ser tomado de una fuente de datos diferente.
Estimación de parámetros	Mediante procedimientos que usan variantes del algoritmo esperanza-maximización se estiman valores para sustituir los faltantes o atípicos.
k-NN	Consiste en remplazar todos los valores faltantes o atípicos con el k-vecino más próximo (kNN) mediante un algoritmo que determina la similitud de dos instancias con el empleo de una función de distancia.
Imputación múltiple	Se remplaza cada valor faltante por diferentes valores y posteriormente, con el uso de todo el conjunto de datos, se calcula el promedio de los valores y así se obtiene el conjunto de datos completo.
Imputación	Se remplazan los valores faltantes o atípicos por un valor

One of the most used proposals is that of *Torvik,*[12] who states that most disambiguation methods summarize the scores of all characteristics in a single number, which indicates the degree of similarity of a pair of articles. *Torvik*[26] developed a model to automatically generate data sets for training and subsequent estimation of the probability that a pair of Medline articles that have the same last name and the first letter of the first name are from the same author, based on other metadata (title, publication name, MeSH, co-authors, affiliation). [11.27]

Han, Zha and Giles[15] consider that the "k-way" application of the spectral clustering method with QR decomposition provides better results than traditional clustering methods, for example, k-means. *Giles* and others [16] first apply a pruning method by author and then a clustering using SVM as a distance function. On the other hand, *Bhattacharya*[12] proposes an adaptation of Latent Dirichlet Allocation (LDA). Authors may belong to one or several groups of individuals who tend to write together. This method simultaneously discovers individual-author *clusters and article clusters* , which has a high computational cost. They employ an unsupervised training method and the “expectation-maximization” algorithm (Table 3). A first step of cleaning, standardization and pruning by the last name field is proposed by *Pino-Mejías* . [20] Subsequently, comparisons must be made between six fields of each pair of articles and a similarity index between 1 (the strings are the same) and 0 (the strings are completely different) is calculated using similarity functions (exact string, Levenshtein, Jaro , Winkler). The data set obtained is subjected to clustering.

When analyzing the methods summarized in the previous paragraph, it is corroborated that these authors apply common principles, which must be taken into consideration in the proposals of this study, such as, among others: applying a prior "cleaning", a pruning mechanism to reduce complexity, distance functions and clustering algorithms.

Ordoñez [28] has experienced that the best way to perform preprocessing is to take advantage of the advantages of database management systems (DBMS), such as the SQL language (*Structured Query Language*). This is an important aspect to consider for the proposed application, since it is considered more convenient to have an integrated environment supported by a DBMS than to use external applications known as ETL (*Extract-Transform-Load*) designed for a broader spectrum of specific problems and environments. .

PROPOSAL OF THE PROCEDURES AND METHODS TO BE USED IN PREPROCESSING

From the analysis of each of the methods, the following premises were established, which could be guidelines to follow in the design of the application:

1. The disambiguation algorithm must have at least three main components, which are executed sequentially:

- A selection or pruning mechanism (*blocking*) using a *hash function,* to divide the data set and thus reduce the computational cost. An example of this is the one proposed by *Bilenko,* [12] who separates authors who have the same last name and first initial initial into groups.

- A similarity comparison function to analyze pairs of records or strings. This function must determine if two records or strings refer to the same entity based on some attributes or characteristics, so its output must be a binary decision (yes or no) or a similarity index, which is generally between 0 and 1 This function can be based on *tokens* or edit distances.

- A classification algorithm that can be supervised, unsupervised or hybrid. Some supervised classifiers such as SVM and decision trees can also be used as a comparison function. [17]

2. The disambiguation algorithm must summarize the scores of all characteristics to indicate the degree of similarity of a pair of records, but taking into account that these are independent of each other. [12]

3. Errors involving letter changes in authors' names and surnames can be ignored. *Torvik* and *Smalhaiser* [22] showed that they appear in approximately 1.8% in BD Medline.

4. Feature selection is the most important aspect in designing a disambiguation model, because it determines the upper limit of accuracy. A good approach is to use as many useful features as possible, because using only one or a few features will likely limit the results of the method. [12]

5. The disambiguation algorithm must be flexible enough so that the user can adjust it to the needs of the analysis being performed, the characteristics of the data set, etc.

The preprocessing of the "affiliation" and "author" fields is more complex than the processing of others such as "year" and "MeSH", and may depend on them. Therefore, it is not enough to select the methods to use, but a logical sequence is required in their use to obtain better results. A proposal appears in table 4 .

An order is recommended in the application of the methods by fields of a record (in this case some fields of the Medline DB). It should be noted that some methods require input parameters. For example, imputing missing values can be by the mode, mean, etc., a value given by the user.

To carry out the disambiguation it is recommended:

- Use different methods for the author and author affiliation fields because they have different characteristics. These differences consist of the fact that the author is made up of three fields (name, surname and initials), while the affiliation is a single field, but it can contain country and email data. These data must be separated before. Furthermore, to disambiguate the author field, metadata is included, but not for affiliation, where only the country field is relevant, and similarity between strings must be used. It is evident that the disambiguation of affiliation must be undertaken first, since it provides characteristics that can contribute to the disambiguation of the author.
- Use cleaning procedures for the data that are going to be disambiguated - in this case author and affiliation - as well as the rest of the fields that are going to be used as relevant characteristics for the disambiguation method or are relevant for the bibliometric study that is being carried out. performed by the user.

- Have an application that is sufficiently flexible, since there is no ideal procedure, and that is why it is required that the user (who is the expert in the data domain) can configure what he needs.

The selection of the methods used in the application was based on the following criteria:

- That they are appropriate, effective and efficient to solve the problem posed.
- That they are sufficiently documented to be understood and used fully.
- That they exist in open source libraries and their programming complexity is not very great.
- That they are flexible and adjustable to the problem posed.

To disambiguate the author's affiliation, it is proposed to use the *Magnani* and *Montesi* method . [24] These authors highlight the comparison of company names (*Company Name Matching*) as another form of the disambiguation problem, applied to patent databases such as Amadeus and Patstat. As a result, the legal name of the company is obtained. The data is cleaned by eliminating punctuation marks, stop words, multiple spaces, etc. *Duplicate rows are then removed and edit distance and term- or token-* based functions are applied to company names . Through another function they determine the weight that each *token* has proportional to its significance or importance. The pruning technique is used using the "country" field to reduce complexity, that is, by first obtaining the country it can then be divided into subsets and reduce the calculation time.

Based on the methods studied and the user's problems, it is considered that in the case of disambiguation of author names, the most flexible and convenient method is the *Bolikovski method,* [22] which consists of three steps:

1. Using pruning, documents are separated into groups using a hash function.
2. For each pair of documents in the same group, their total affinity (similarity) is calculated, which is the sum of the atomic affinities for each of the characteristics (attributes) to be considered. These, in turn, are obtained as a result of a function that returns a value between -1 and 1, which represents the contribution of that characteristic in the comparison and which is subsequently multiplied by the weight corresponding to it. The value 1 indicates that according to this function it is certain that the attributes correspond to the same individual. A value of zero indicates that this function cannot determine whether or not those characteristics correspond to the same person. The value -1 indicates that these two attributes correspond to two different people. Some characteristics provide a high weight when they coincide, for example the email, as they demonstrate that it is the same person; others are of weak importance, for example, the name of the publication. Sometimes it happens that the same characteristic is strong for coincidence, but weak for difference (or vice versa), for example, email. An important element of this author's method is that the weights can be adjusted flexibly; They can also be determined using a computer application. The result of this step is a matrix of the total affinities.

3. The last step consists of clustering based on the matrices obtained in the previous step and using the " *Single-Linkage Hierarchical Agglomerative Clustering* , SLHAC" algorithm, [23] which compares with an established threshold.

A significant advantage in this method is its flexibility, since the *hash* function , the similarity function and the clustering algorithm can be replaced by others at convenience. This would allow us to adapt the method to the needs, experiment with different variants, new algorithms and create our own theoretical-practical base on the subject of preprocessing.

DESIGN OF THE TOPIC BASE

These recommendations and requirements are specific to the case of our study. They are the most important for the design of a computer application for preprocessing, which does not exclude that a more detailed document must be prepared that meets the requirements of Software Engineering to design, program and exploit said application. It is foreseeable that in this process and with the experiences of working with the application, aspects that must be perfected, more efficient programming methods, existing routines and functions that can be used and others will be detected.

The scope of this work must be considered in stages. In the first stage of the application:

- It will be limited to the preprocessing of data obtained from the Medline database in XML format through ViBlioSOM Software, which has its own database design.

- It will be limited to preprocessing (normalization according to the ViBlioSOM methodology) of the data, without interfering in the creation of the DB.

- It will function as an application designed for a single user and a database on a local area network. It is not relevant whether it is programmed as a web or desktop application.

- It will not require security measures against unauthorized access to the information, but it will require security measures to prevent the possible destruction of the information, such as performing backups, implementing consistency checks in transactions and recovery (rollback *)* in case that they cannot be completed successfully, limited rights for users and others. In this way, greater reliability will be guaranteed in the results of the tasks carried out in the application.

- It will have an interface that must be easy to understand, intuitive and friendly, in line with current standards, considering that the user of the application is an information worker, not necessarily an IT expert.

- It must provide clear messages in the event of application or user errors and solutions that allow the latter to continue working and not lose information. All errors must be recorded.

- It must interact with the ViBlioSOM Software in a modular way, without interfering in any way with its current operation, that is, it will import the data from its database, pre-process it and subsequently restore it. Depending on the results obtained with this application, suggestions will be made so that this preprocessing module is fully integrated in the design of new versions of ViBlioSOM.

- It will be programmed in open source. There are repositories of algorithms and code that can be evaluated for use in programming the application. Among these are:

SecondString: Open source Java package with fuzzy string comparison techniques.

Simmetrics: Open source Java library with similarity techniques.

Febrl: Open source Python library with disambiguation techniques.

ATLaS: Extension to the SQL language to enhance aggregate functions and data mining.

- The DBMS to be used will be PostgreSQL v.8.4, as it is where ViBlioSOM is supported and it guarantees that most processing is carried out on the server side. It has the PL/pgSQL language, which is also open source and has extensive possibilities, such as its extensibility with Java, Python and other languages. Furthermore, among its advantages it has facilities for manipulating arrays, which in our application are necessary to create similarity matrices and token processing.

- Although it will not have critical demands in terms of performance, it will try to reduce execution times as much as possible, use the server's calculation potential, rationalize disk writing and take other programming measures (Fig. 2).

- The main objective of the application will be to improve data preprocessing in a methodology taken as a reference framework such as

ViBlioSOM. Another objective will be to obtain a development and experimentation platform that allows evaluating the potential and results of the various algorithms and methods for preprocessing. This work will allow us to develop our own conception on this topic.

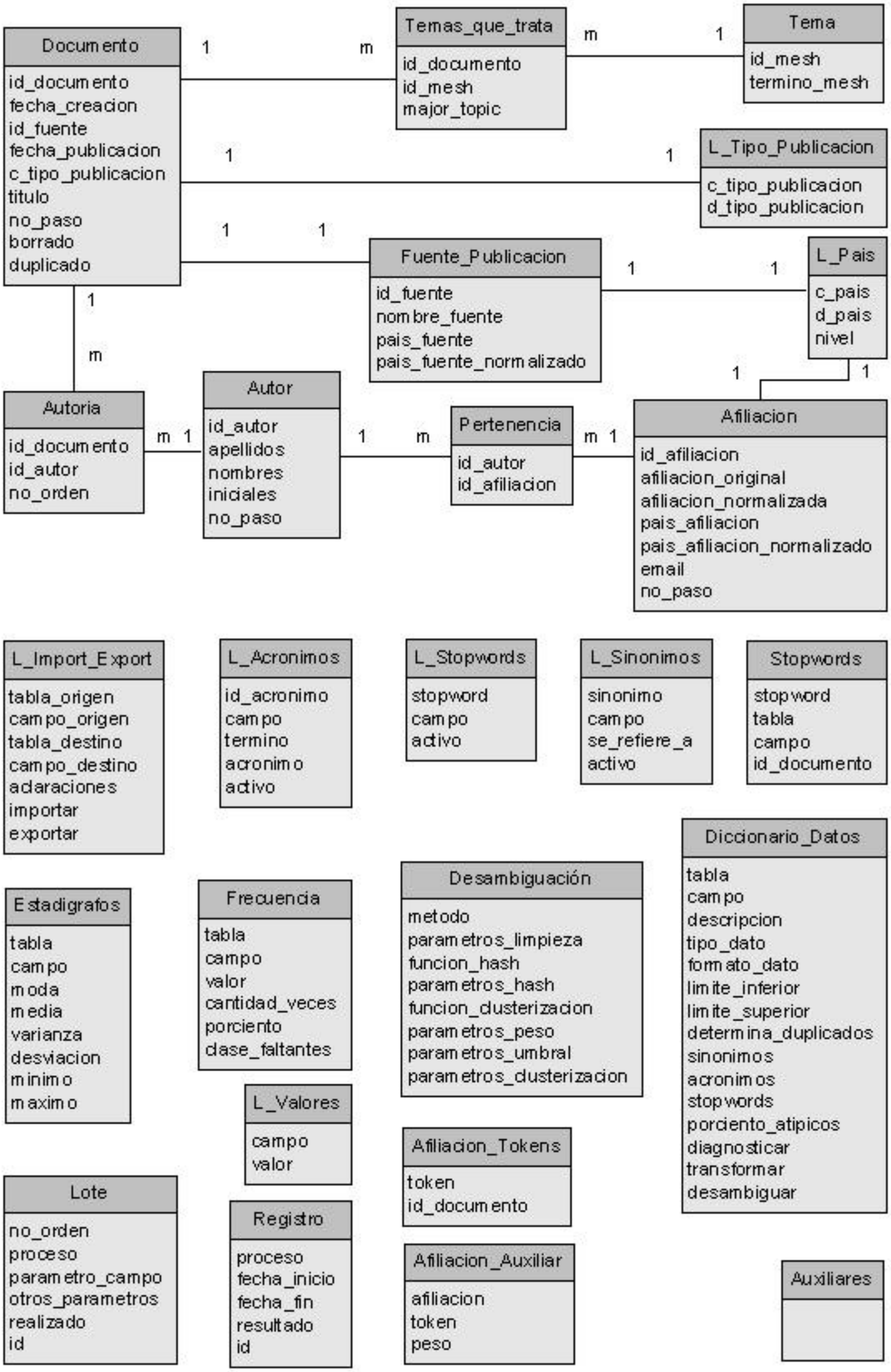


Fig. 2. Estructura de la base de datos.

CONCLUSIONS

The proposal of an algorithm for the preprocessing of the "author-affiliation" fields with a metric approach and that could also be implemented in a metric information analysis system, is not a trivial problem. There is little accumulated knowledge on this topic at the international level and in Cuba no published documents on the topic were found. This made it possible to confirm that, despite the increase in the use of metrics, the development of applications that solve data preprocessing problems has not been addressed, so the techniques for data preprocessing in its broadest sense, do not They are being used sufficiently in Cuba for metric purposes.

The study of the literature on preprocessing has allowed us to establish that it can be defined as a process aimed at optimal transformations of data, aimed at obtaining significant knowledge and can be composed of one or several methods. Although preprocessing is associated with the computer application of mathematical algorithms, it is also linked to a cognitive process in terms of its purpose of “discovering new knowledge.”

Metric analyzes can improve their results by using statisticians before making data transformation or elimination decisions, as well as applying methods for disambiguation of author names and affiliations, as a solution to one of the problems that most deteriorates the quality of data. metric studies.

A viable approach to optimize data preprocessing for metric studies can be like the one proposed here: modular, based on the potential of the DBMS, taking advantage of the existing open source, configurable and flexible and where the information specialist does not lose control of what is happening.

It was possible to establish a group of requirements for the design of a computer application for data preprocessing that can be made up of a set of tasks, steps and algorithms. It is suggested to use not a single algorithm, but a combination of these to disambiguate the affiliation and author field. These can be selected based on the data and needs of the user.

For other stages of application development it is recommended :

- Consider the need to process other bibliographic databases.
- Consider parallel processing (e.g. Google's Map Reduce).
- Consider the possibility of group work.
- Extend preprocessing to other fields (for example, titles, summaries, etc.).

REFERENCES

1. Omelianovsky ME. The methods of contemporary mathematics and the mathematization of knowledge. In: Omelianovsky ME, editor. Dialectics and general scientific research methods (Volume I). Havana: Social Sciences; 1981. p. 179-243. 2. Guzmán MV. Vibliosom: Methodology for the visualization of metric information with self-organized maps Doctoral Thesis]. Havana: University of Havana; 2009. 3. Kimball R. Dealing with dirty data. DBMS. 1996;9(10):55-60.

4. Müller H, Freytag JC. Problems, methods and challenges in comprehensive data cleaning. Berlin: Professoren des Inst. Für Informatik; 2005:23.

5. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE DEBU. 2001;23(4):3-13.

6. Ontalba-Ruipérez J. Normalization of fields in bibliometrics: Fecyt actions. Prof Inf. 2007;16(4):381-3.

7. Spinak E. Spelling errors when entering databases. Rev Esp Doc Cient. 1995;18(3):307-19.

8. Lardy J, Herzhaft L. Bibliometric treatments according to bibliographic errors and data heterogeneity: the end-user point of view. In: 16th international online information meeting. London. Oxford: Learned Information; 1992. p. 547-56.

9. Anguita A, Pérez D, Crespo J, Maojo VM. Automatic generation of integration and preprocessing ontologies for biomedical sources in a distributed scenario. In: Proceedings of 21st International Symposium on Computer-Based Medical Systems (CBMS-2008). Washington DC: IEEE Computer Society; 2008. p. 336-41.

10. Zimei S. KDDML: Extend the Preprocessing Phase Graduate Thesis]. Pisa: University of Pisa; 2004.

11. Bordons M, Costas R. Algorithms to solve the lack of standardization of author names in bibliometric studies. Library Research. 2007;21(42):13-32.

12. Smalheiser NR, Torvik VI. Author name disambiguation. Annu Rev Inform Sci. 2009;43(1):1-43.

13. Han H, Giles CL, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In: Proceedings of Joint

Conference on Digital Libraries (JCDL 2004). Tucson, USA: ACM; 2004. p. 296-305.

14. Han H, Zha H, Giles CL. A model-based k-means algorithm for name disambiguation. In: Proceedings of 2nd International Semantic Web Conference (ISWC-03) Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. Sanibel Island FL, Germany: Springer; 2003.

15. Giles CL, Han H, Zha H. Name disambiguation in author citations using a K-way spectral clustering method. In: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries (JCDL '05). Denver, New York: ACM; 2005. p. 334-43.

16. Huang J, Ertekin S, Giles CL. Efficient name disambiguation for large-scale databases. In: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin: Humboldt-Universität zu Berlin; 2006. p. 536–44.

17. Treeratpituk P, Giles CL. Disambiguating authors in academic publications using random forests. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL-09). New York: ACM; 2009. p. 39-48.

18. Ferreira AA, Veloso A, Gonçalves MA, Laender AHF. Effective self-training author name disambiguation in scholarly digital libraries. In: Proceedings of the 10th annual joint conference on Digital libraries (JCDL'10). Gold Coast, New York: ACM; 2010. p. 39-48.

19. Torvik VI, Smalheiser NR. Author name disambiguation in Medline. ACM Trans Knowl Discov Data. 2009;3(3):1-29.

20. Pino R, Cubiles MD, Caballero E. A comparison of probabilistic record linkage techniques in the Institute of Statistics of Andalusia (ISI' 2011). In: 58th World Statistics Congress of the International Statistical Institute. Dublin: ISI; 2011.

21. Jijkoun V, Khalid MA, Marx M, Rijke M. Named entity normalization in user generated content. In: Proceedings of the second workshop on Analytics for noisy unstructured text data (AND'08). Singapore, New York: ACM; 2008. p. 23-30.

22. Bolikowski L, Dendek PJ. Towards a flexible author name disambiguation framework. In: Sojka P, Bouche T, editors. Towards a digital mathematics library. Brno: Masaryk University Press; 2011. p. 27-37.

23. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press; 2008:482.

24. Magnani M, Montesi D. A study on company name matching for database integration. Bologna: University of Bologna. 2007. Technical Report: UBLCS-07-15.

25. Ferreira AA, Laender AHF, Gonçalves MA, Cota RG, Santos RLT, Silva AJC. Keeping a digital library clean: new solutions to old problems. In: Eighth ACM symposium on document engineering (DocEng '08); 2008 Sep 16-19, Sao Paolo. New York: ACM; 2008. p. 257-62.

26. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: A model for author name disambiguation. J Am Soc Inf Sci Technol. 2004;56(2):140-58.

GARDEN JOURNALS

Journal of Qualitative Research in Business Law, Eco-Fin, Accounting, and Statistics

ISSN : 1533-1520

Health Hazards Linked To Applications On 'Smartphones': A Case Study Of Pokémon Go

Gabriela F. Santos,

School of Business, São Paulo State University, Brazil

Dear director:

In the second half of 2016, the Pokémon Go application for smart phones was launched and some pointed out that it could have some health benefits and not just represent a trend, but we disagree with this perception, since the use of smart phones could have effects negative effects on the health of the population.

This game generates different changes in human behavior. [1] Unlike other video games that keep players immobile in front of monitors—or while holding portable players—or locked in their rooms, Pokémon Go forces its players to go outside. [2] Thus, the sedentary population socializes, but also changes their routine and has more physical activity outdoors, [3-5] with the subsequent cardiovascular benefits and reduction of the individual's obesity and depression rates. [4,6] However, it can also bring negative effects such as increased risk of exposure to solar radiation, as well as vector-borne diseases, [7] traffic accidents, [8,9] injuries, kidnappings, intrusions and violence, among others. [2] Furthermore, being a "game", its use decreases as interest in playing it dissipates, often influenced by climatic conditions; [4] Consequently, physical activity is also reduced. [10] That is why the change in behavior would not be permanent.

Many of the applications for smart cell phones have positive impacts on the activities of daily life, but we must not ignore the negative impacts that they could generate on their users. The use of this type of cell phone has been associated with dry eye in children,[11] as well as temporary blindness,[12] and could lead to addiction, especially in adolescent women;[13-15] Even using an application that allows us to establish a route to go to a certain place can expose us to a greater risk of interpersonal violence if—when following it—we circulate through areas with high rates of violence.[16,17]

In conclusion, positive behavioral changes would be mediated by "fashion" and would not be sustainable over time, while negative changes could generate immediate exposure to risk situations with a greater impact on the individual's health. Currently, there are few studies on this topic, so research must also be aimed at demonstrating and characterizing the risk generated by the use of "smart cell phones" and the applications available on them, in different situations.

Cordially,

Alfredo Enrique Oyola-García, Melisa Pamela Quispe Ilanzo
Abraham Valdelomar Housing Complex C-201. Ica, Peru.
Email: aoyolag@gmail.com

REFERENCES

1. Quinn J. Identity of Pokémon Go players: how social gaming affects behavior. Advanced Writing: Pop Culture Intersections. Paper 19. 2016 (citado 22 de diciembre de 2016). Disponible en: http://scholarcommons.scu.edu/engl_176/19

PubMed 2. Serino M, Cordrey K, McLaughlin L, Milanaik RL. Pokémon Go and Augmented Virtual Reality Games: A Cautionary Commentary for Parents and

Pediatricians. Current Opinion in Pediatrics. 2016 (cited December 22, 2016);28(5):673-7. Available at: http://journals.lww.com/co-pediatrics/Abstract/2016/10000/Pok_mon_Go_and_augmented_virtual_reality_games_a.17.aspx

3. Althof T, White RW, Horvitz E. Influence of Pokémon Go on physical activity: study and implications . . . . J Med Internet Res. 2016 (cited December 22, 2016);18(12):e315. Available at: https://arxiv.org/pdf/1610.02085v2.pdf

4. Chunara R, Bouton L, Ayers JW, Brownstein JS. Assessing the online social environment for obesity prevalence surveillance. PLoS One. 2013 (cited December 22, 2016);8:e61373. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061373

5. Gladwell VF, Brown DK, Wood C, Sandercock GR, Barton JL. The great outdoors: how a green exercise environment can benefit all. Extrem Physiol Med. 2013 (citado 22 de diciembre de 2016);2:3. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3710158/

6. Beyer KMM, Szabo A, Nattinger AB. Time spent outdoors, depressive symptoms, and variation by race and ethnicity. Am J Prev Med. 2016;51(3):281-90. Disponible en: https://www.ncbi.nlm.nih.gov/pubmed/27320702

7. Oidtman RJ, Christofferson RC, ten Bosch QA, Espana G, Kraemer MUG, Tatem A, Barker CM, Perkins TA. Pokémon Go and exposure to mosquito-borne

diseases: how not to catch'em all. PLOS Currents Outbreaks. 2016 (citado 22 de diciembre de 2016). Disponible en: http://currents.plos.org/outbreaks/article/Pokémon-go-and-exposure-to-mosquito-borne-diseases-how-not-to-catch-em-all/

8. Joseph B, Armstrong DG. Potential perils of peri-Pokémon perambulation: the dark reality of augmented reality? Oxford Medical Case Reports. 2016 (citado 22 de diciembre de 2016);10:265-6. Disponible en: http://omcr.oxfordjournals.org/content/2016/10/omw080.full

9. Ayers JW, Leas EC, Dredze M, Allem JP, Grabowski JG, Hill L. Pokémon GO-a new distraction for drivers and pedestrians. JAMA Intern Med. 2016 (citado 22 de diciembre de 2016);176(12):1865-6. Disponible en: http://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2553331

10. Howe KB, Suharlim C, Ueda P, Howe G, Kawachi I, Rimm EB. Gotta catch'em all! Pokémon GO and physical activity among young adults: difference in differences study. BMJ. 2016 (citado 22 de diciembre de 2016);355:i6270. Disponible en: http://www.bmj.com/content/355/bmj.i6270.full

11. Moon JH, Kim KW, Moon NJ. Smartphone use is a risk factor for pediatric dry eye disease according to region and age: a case control study. BMC Ophthalmol. 2016 (citado 31 de marzo de 2017);16(1):188. Disponible en: https://www.ncbi.nlm.nih.gov/pubmed/27788672

PubMed 12. Alim-Marvasti A, Bi W, Mahroo OA, Barbur JL, Plant GT. Transient Smartphone "Blindness" N Engl J Med. 2016 (cited March 31, 2017);374:2502-4. Available at: http://www.nejm.org/doi/full/10.1056/NEJMc1514294

13. Körmendi A, Brutóczki Z, Végh BP, Székely R. Can smartphone use be addictive? The case report. J Behav Addict. 2016 (citado 31 de marzo de 2017);5(3):548-52. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5264424/

14. Randler C, Wolfgang L, Matt K, Demirhan E, Horzum MB, Besoluk S. Smartphone addiction proneness in relation to sleep and morningness-eveningness in German adolescents. J Behav Addict. 2016 (citado 31 de marzo de 2017);5(3):465-73. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5264414/

15. Sam-Wook C, Dai-Jin K, Jung-Seok C, Heejune A, Eun-Jeung C, Won-Young S, et al. Comparison of risk and protective factors associated with smartphone addiction and Internet addiction. J Behav Addict. 2015 (citado 31 de marzo de 2017);4(4):308–14. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4712765/

16. The Argentine tourist shot after entering a favela in Rio de Janeiro died. RPP News. March 25, 2017 (cited March 31, 2017). Available at: http://rpp.pe/mundo/actualidad/murio-la-turista-argentina-baleada-tras-ingresar-a-favela-en-rio-de-janeiro-noticia-1039530

GARDEN JOURNALS
Journal of Qualitative Research in Business Law,
Eco-Fin, Accounting, and Statistics
ISSN : 1533-1520

Knowledge Management Contexts In Cuban Organizations: Exploring The 'Ba' Concept In Cuba

Samuel K. Weaver
Institute of Mathematics, Bulgarian Academy of Sciences, Bulgaria

ABSTRACT

The results of a research carried out in 15 Cuban institutions are analyzed, which represent four groups of organizations that work intensively with information. Its objective was to analyze the behavior of different variables in the particular contexts of each one and assess whether or not the type of institution favored the behavior of the different contexts for the creation of knowledge. An interview in the form of a questionnaire was applied to managers or specialists of the intentionally selected institutions. The behavior of four types of Knowledge Management contexts in the institutions under study is presented. Considerations are offered about the behavior of some variables in the institutions studied and it is verified that the identified behavior corresponds to what is identified in the literature on the subject.

Keywords: Knowledge Management; "Ba"; contexts; Cuba.

INTRODUCTION

Currently, the value of knowledge as an organizational resource has increased. More and more actions are being applied focused on increasing its profitability as a factor that makes the difference in competitiveness. Knowledge is constantly created in organizations; It is important that contexts that facilitate activities and interactions between its members are available. These are given by time, space and relationships; That is, they vary according to when and where the creation of new knowledge takes place, and who and how participate in the process. In an organization, knowledge is the fundamental resource that allows the rational use of other resources; It is present in each of its workers. This social knowledge when managed by groups of an organization is called organizational knowledge and is created when they solve problems using their knowledge and making it available to the organization. The future of the organization itself will depend on its correct management.

Alavi and *Leidner* [1] explain that interest in organizational knowledge has prompted efforts to manage knowledge for the benefit of the organization. If managed correctly, the organization will be able to identify and elevate collective knowledge and thus increase its competitiveness and innovation. The creation of organizational knowledge should be considered as a process that "organizationally" amplifies the knowledge created by people and crystallizes it as part of an organization's knowledge network. [2] Knowledge occurs in a specific context, is dynamic and is created from social interactions, which is why it has a humanistic character. [3] The contexts for the creation of organizational knowledge may vary depending on the characteristics of the organization itself; These can be reflection meetings, discussions, professional training, training programs, use of intranets and forums, among many other actions that can be carried out (Fig. 1).

Nonaka and *Konno* [4] proposed the use of the concept Ba, which can be translated as *place* , with the purpose of applying it in their model related to the creation of knowledge. The concept comes from philosophy; It was proposed by *Nishida* and developed by *Shimizu* . These authors consider that "it can be conceived as a shared space where relationships emerge. This space can be physical (offices, various work spaces), virtual (email, teleconferences), mental (shared experiences, ideas, ideals) or any combination of them". What differentiates Ba from regular human interaction is the concept of knowledge creation. It proposes a platform for individual and collective knowledge to develop and from this platform all the information needed is integrated. The Ba allows solving problems in the organization or in work teams and provides new ways to analyze

them. [5] *Nonaka* , *Toyama* and *Konno;* [6] *Acosta* , *Zárate* and *Fischer;* [5] *Choo* and *Ne to* ; [7] *Hansson;* [8] *Neto* and *Choo;* [9,10] as well as *Nonaka* and *Konno,* [4] consider that there are four types of *Ba* that are related to the SECI model. The SECI Model proposed by *Nonaka* , proposes the 4 forms of knowledge conversion based on processes of socialization, externalization, combination, and internalization, whose initials give the model its name.

The different types of Ba are listed below. Each of them accelerates the corresponding knowledge conversion processes. The types of Ba correspond to two types of dimensions: the type of interaction, which can be individual or collective, and the dimension according to the media, which can be face to face or virtually: [6]

Original Ba: corresponds to the socialization process of the SECI model. It is the primary way in which the knowledge creation process begins. [4] In this case, the interaction between individuals is face to face, where experiences, feelings and expressions that can be identified when people share a physical space are mainly shared. This type of *Ba* is an effect of the reformulation and maintenance of physical environments. [eleven]

Ba of interaction (or dialoguing with the Ba): is related to the process of externalization of knowledge. The type of interaction is collective. The *Ba* of interaction is consciously constructed from the original *Ba* [6] . Its context may be work groups with selected individuals who share skills and knowledge. The objective is to achieve moments for dialogue, as well as meetings and effective communication channels between workers. Through dialogue, the individual's mental models and abilities become common terms and concepts [4] . The knowledge of the group members is externalized and at the same time each one takes data and information from the others that are understood and internalized and are integrated into individual knowledge.

Cyber Ba: is the place of interaction in the virtual world before the physical world. [12] It also suggests that it can involve hundreds of individuals in the organization through technologies. It corresponds to the combination stage of the SECI model and is effective in collaborative contexts. The combination of explicit knowledge is more efficient if it is supported by the information and technology provided by the network. [13] To develop it, it is necessary that the organization has a good technological infrastructure that favors and promotes virtual interaction environments, such as digital libraries, intranets, virtual forums, the use of databases, virtual learning spaces (EVA), etc.

Exercising Ba: it develops in the process of internalization of knowledge. It focuses on managers and their co-workers. It consists of continuous exercises where certain patterns are developed and enhanced. These exercises provide training in individuals and reflection is achieved. *Choo* and *Neto* [7] state that the objective is teaching based on analysis and continuous learning based on active participation.

The types of Ba can be present in organizations even if they are not identified with the name presented above. Recognizing these spaces allows knowledge management processes to be promoted and makes it possible to create the conditions to generate these contexts more effectively. Figure 2 summarizes these four contexts associated with knowledge conversion processes. [14]

There is no limit of physical space, since it represents a subjective context, which is delimited by a specific time and space that is given by the context in which it develops. Context plays an essential role in Ba and provides meaning to the information that is transformed into knowledge; Its main components are individuals, the organizational environment and the relationships between them.

Nonaka and *Konno* [4] argue that in *Ba* , knowledge can be acquired and shared in different ways in which individuals interact within the organization, such as work teams, meetings, communities of practice or other less formal ways, such as sharing. experiences, reflections and personal criteria, among others. The *Ba* is present in any space of the organization, since knowledge is not only shared in places arranged for this, but can be found in hallways, telephone conversations, discussions, chats, on intranets, among others, where opinions are shared. and information. It occurs during the work day and at times of meetings, such as lunches, informal meetings, etc.

This article analyzes certain perspectives on the contexts of Knowledge Management in Cuban organizations of different types, which work intensively with information, as well as delving into its characteristics, discussing different perspectives, clarifying the importance of managing and raising awareness among them about the need to take into account their contexts to improve them, and outline ways that allow advantages to be introduced to these organizations.

METHODS

A descriptive research from a qualitative methodological perspective was carried out:

- *Descriptive.* Analyzes and characterizes the phenomenon of contexts (Ba) for Knowledge Management in a set of Cuban organizations of different types.

- *Qualitative.* Obtains and analyzes data from the organizations under study to determine the behavior of the situational variable in the identified unit of analysis, using induction to derive possible explanations based on the observed phenomena.

Regarding research methods, classic documentary analysis was used for theoretical-conceptual characterization. For the sample, 15 Cuban organizations of different types, representative of the Cuban environment, were intentionally selected. These national organizations are all located in the province of Havana. For this, four inclusion groups were created:

1. Libraries (4 libraries: one national, one university and two specialized).

2. Information centers (3 information centers: one national and two corresponding to prioritized sectors).

3. Research institutions of the Scientific Pole (4 research centers of the Western scientific pole).

4. Other information institutions (4 institutions: 1 archive, a specialized publishing house, a museum and an institution specialized in documentary heritage).

The study did not aim to analyze particular cases, but rather to identify possible distinctions or differences between the contexts of different types of institutions. The interview was used, and specifically the questionnaire as one of its instruments. The questions in this questionnaire were mostly closed, which were applied to managers of the selected entities. In the case of groups 3 and 4, the interview was applied to the information managers of these institutions, trying to obtain the most homogeneous analysis possible from the responses obtained. The analysis of the results obtained allowed us to delve deeper into the characteristics of the contexts for Knowledge Management (Ba) in the organizations under study.

The variables were taken into account based on expert criteria and what is described in the international literature on the components of the different types of Ba, referenced in this work.[4-10] The operationalization of the variables was specified in the design of the questionnaire used to collect data and information about the presence of certain conditions in the particular contexts of the groups of institutions analyzed.

RESULTS

The aspects related to the contexts for Knowledge Management (Ba) in the observed institutions are analyzed below. For the analysis of the results, the variables used to obtain information in the interviews carried out were taken. For each variable, the criteria collected in the interviews carried out in the selected institutions have been incorporated. The number assigned to each group of institutions used in the research has been respected: 1) Libraries; 2) Information centers; 3) Research institutions of the scientific pole and 4) Other information institutions. In this way, each table represents the qualitative criteria related to a variable, based on the considerations of the four groups of institutions represented in the study.

Table 1 reflects the absence of specialized bodies to attend to CG functions in institutions. Since there is no functional body, it is difficult to specify strategies and policies on the subject, which generally rely on manuals and regulatory documents. Group 3 is the one that has the most control over the topic, not formally but in the approaches they reflect in the interviews.

Table 2 reflects how different groups of institutions value human resources. It can be seen that this variable highlights the role of training as an approach that tends to develop knowledge.

Group 3 reflects other more advanced views related to knowledge retention and the need to align knowledge with institutional strategies. In relation to the presence and use of technology, it reflects strengths in all groups, although group 3 stands out with a view not only to communication, but also to its integration into the main processes of the institutions (table 3).

Table 4 reflects how group 3 shows maturity in terms of access to institutional information linked to the system's communication strategy. However, group 4 shows greater weaknesses in this aspect, due to an undervaluation of the importance of this variable. It is interesting to see how attempts are made to improve the communication of internal information through the creation of functional work groups, especially in group 2.

Table 5 analyzes the behavior of this variable in the groups of institutions analyzed. This variable, perhaps one of the most complex, requires certain conditions and attitudes that favor its development. However, there is still no notable development, and it is group 3 that analyzes it under better conditions.

Table 6 reflects the behavior of different types of Ba based on the *organizational communication* variable . There is an appreciation of this variable by group 1 and group 2, in contrast to what is reflected in other variables. This does not mean that other groups do not consider this variable, but perhaps currently it is already in a stage of maturity, and communication flows spontaneously from the cultural behavior of the group.

Table 7 shows the important valuation that the 4 groups have of the with their internal context and with their environment. Each group interacts with different contexts, such as professional associations (group 1), technology and strategic alliances (group 2), product marketing and professional exchanges (group 3), and informal physical exchange in their own contexts (group 4). Group 4 is the one that reflects the least activity in this variable.

Table 8 shows in general the domain and quality of this variable. Given the professional culture of these institutions, it was to be expected that they would know and execute actions related to scientific and professional communication. This study did not aim to analyze quantitative elements, but rather the qualitative behavior of these variables, which is demonstrated in the recognition of their conditions in the four groups of institutions.

Table 9 reflects the mastery and use of the training variable as an enriching aspect of the work of the different groups analyzed. However, the absence of knowledge management policies reflected in variable 1 can influence the quality of the training actions carried out.

FINAL CONSIDERATIONS

The statement by Nonaka , *Toyama* and *Konno* is ratified in that the contexts for the creation of organizational knowledge can vary depending on the characteristics of the organization itself. Distinctions or differences were identified between the contexts of different types of institutions, although it was possible to specify the characteristics of these contexts in the four groups of institutions chosen. The effect of leadership on these approaches was also clear.

In general, training, exchange at events and meetings, and the use of technologies are the aspects that have the greatest uniformity in the four groups analyzed. Some institutions stand out more within their group, mainly due to the policy of their leaders and their vision in relation to this approach, as highlighted in the analysis of each variable. As weak aspects, it can be seen that there is still no culture about the importance of Knowledge Management and, except in a few cases, there is no functional entity that comprehensively conceives and develops actions to raise and better take advantage of the available knowledge.

In the information centers (group 2) strengths are seen in terms of the development of knowledge management. The development of Knowledge Management projects or the establishment of this function at the level of organizations would contribute considerably to increasing the employment and protection of this institutional asset. The research institutions of the Scientific Pole (Group 3) stand out, which have an infrastructure and policies that promote the use of Ba in their spaces to develop knowledge. The potential and use of knowledge in institutions and information centers is also appreciated, but not in libraries, which show less mastery and application of these approaches. It was possible to identify that among the four groups of institutions created there are differences in the characteristics of the contexts for Knowledge Management, and within each group there are variations regarding the characteristics of the *Ba* they use.

The importance attributed to vital components is appreciated, such as innovation, communication, transfer of information, training and dissemination of results through scientific publication. The level of development of cyber Ba in most of the institutions analyzed is very significant.

Conflict of interests

The authors declare that there is no conflict of interest in this article.

REFERENCES

1. Alavi M, Leidner D. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues . My Quarterly. 2001;25:107-36.

2. Ponjuán G. Introduction to Knowledge Management. Havana: Ed. Félix Varela; 2006.

3. Nonaka I, Konno N, Toyama R. Emergence of Ba: a conceptual framework for the continuous and self-transcending process of knowledge creation. In: Nonaka I, Nishiguchi T (Eds.) Knowledge emergence: social, technical and evolutionary dimensions of knowledge creation. New York: Oxford University Press; 2001.

4. Nonaka I, Konno N. The Concept of Ba: Building a foundation for knowledge creation. Calif Manag Review. 1998;40:40-54.

5. Acosta JC, Zárate RA, Fischer AL. *Ba* : knowledge spaces. Context for the development of innovation capacity. An analysis from knowledge management. Rev Esc Administr Neg. 2011;76:44-63.

6. Nonaka I, Toyama R, Konno N. SECI, *Ba* and leadership: a unified model of dynamic knowledge creation. Long Range Planning. 2000;33:5-34.

7. Choo CW, Neto RC. Beyond the *Ba* : managing enabling contexts in knowledge organizations. J Knowl Manag. 2010;14:592-610.

8. Hansson F. Science parks as knowledge organizations – the *Ba* in action? Eur J Innov Manag. 2007;10:348-66.

9. Net RC, Choo CW. The Post Nonaka Concept of Ba: eclectic roots, evolutionary paths and future advancements. 2010 cited March 26, 2017]. Available at: http://choo.ischool.utoronto.ca/fis/respub/asist2010.pdf 2010

10. Net RC, Choo CW. Expanding the concept of Ba: managing enabling contexts in knowledge organizations. Perspectives in Information Science. 2011;16:2-25.

11. Senoo D, Magnier-Watanabe R, Salmador MP. Workplace reformation, active *Ba* and knowledge creation from a conceptual to a practical framework. Eur J Innov Manag. 2007:10:296-315.

12. Nonaka I, von Krogh G, Voelpel S. Organizational Knowledge Creation Theory: evolutionary paths and future advances. Organize Stud. 2006;27:1179-1208.