From Document Analysis to terminology: theoretical and methodological trajectory

Vânia Mara Alves Lima
University of São Paulo
vamal@usp.br

The aim of this paper is to present the theoretical and methodological trajectory on the use of knowledge domain terminology as a reference to the construction and management of documentary languages, specifically controlled vocabularies and thesauri.

This research problem begins to emerge in the 1990s during our professional practice in the technical processing of one of the largest library collections of São Paulo University, (USP) the Library and Documentation Department of Philosophy, Languages and Human Sciences, which provided us with daily issues about information organization, representation and retrieval, produced on the same subject in different areas of knowledge, and which consequently addressed different views.

When we were transferred to the Library and Information Service of Faculty of Architecture and Urbanism of USP, we faced another problem related to information representation, the knowledge produced in the same domain was represented by different designations, often depending on the support on which it was registered, i.e., books, maps, architectural designs, slides, photos, journal articles.

The search for theoretical foundation to solve practical problem led us to the works of Smit, Tálamo, Cunha, Lara, Kobashi, Guimarães and Fujita, founders of Temma Group, which have the work of Jean Claude Gardin as theoretical foundations, but which considerably expanded the horizons of the so-called document analysis, its processes and its products, especially as regards the construction of documentary languages.

The studies by Temma Group showed that Document Analysis defined by Gardin (1981) as "L'expression désigne, on le sait, un ensemble de procédures pour exprimer le contenu des documents scientifiques sous des formes destinées à en faciliter le dépistage ou la consultation" is a methodological discipline that suggests procedures for text analysis in order to select information contents which may be represented, retrieved and disseminated (TÁLAMO, LARA, Kobashi, LIMA, 1992, LARA, 2011) regardless of the support on which they are recorded. To the concept of documentary language as "Un ensemble de termes utilisés pour représenter certains contenus de documents scientifiques avec des fins de classification ou de l'information de recherche rétrospective" (CROS, GARDIN, LEVY 1968)", Temma Group adds other characteristics that emphasize its character of "language" as a structure, whose terms must necessarily be related so that they may mean, in a certain way, working as a communication vehicle that represents the conceptual domains, respecting the community culture which it serves (VOGEL, 2004).

In this framework, we address our linguistic and terminological approach both on the development of documentary information, understood as the result of content representation of a document and the use of domain terminology as a reference for the construction of documentary languages.

Regarding documentary information, we try to outline it from the approaches of linguistic references, identifying it as a documentary sign, similar to linguistic sign, as defined by Peirce (1977,) as something that, in one sense or way, represents something to someone and is in place of something else in some respect or capacity, i.e., documentary information is proposed in place of recorded knowledge and, therefore, in the same way the linguistic sign is capable of semiotic process denominated by Lara (2006) as documentary semiosis.

Understanding semiosis as the construction of meaning by the interpreter, within a given context, it can be inferred that documentary semiosis is the construction of meanings based on terminological references which refer to the conceptual structures of domains. Therefore, contextual references of documentary information production are essential so that they effectively represent a set of true statements about the recorded knowledge, which is compiled in the definitions of each term present in the terminology of a domain.

We must clarify that when we talk about terminologies we are referring to terminology as a product, that is, the set of terms of a specialty, which is developed through the use of terminological standards proposed by Terminology as a discipline that addresses specialized terms (CABRÉ, 1995).

While domain terminology is the formal reflection of the conceptual organization of a specialty, and inevitable means of expression and professional communication which ensures the transmission of knowledge, documentary language has the function to normalize the search and ensure the retrieval of this recorded knowledge by preparing documentary information. Thus, it is understood that these two instruments are complementary.

According to Gardin (1981), documentary information is the product of document analysis, the result of a semantic operation formulated within a documentary language that transforms an original text in one or more keywords, and that even presenting - in the documentary language - the same form in natural language, does not necessarily have all the meanings present in a general language dictionary.

In this context, we infer documentary information as the content representation of a document, from the domain concepts to which it belongs, which designated by the terms of this domain serve as a reference for the descriptors of documentary language because they contextualize its meaning from a practice in this domain (LIMA, 1998).

In the following figure, we depict documentary information represented by the triad concept/term/descriptor, result of semiosis documentary process coupled with the practice in a knowledge domain.

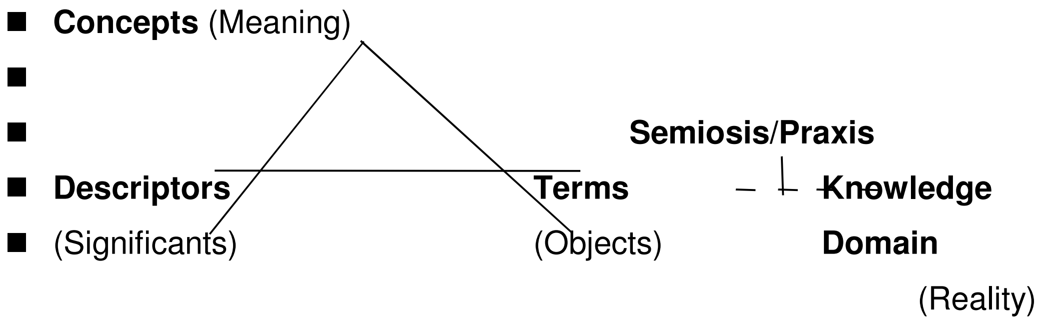

Figure 1 – Documentary information

This scheme of documentary information meets Smit's (2000, p.28) arguments that, for documentary information to work effectively and correctly as a mediator between the user and information stock, the user must be able to contextualize it and decode it, because when using the domain terminology as reference for documentary language, the user may recognize the terms that are part of his/her practice. In other words, coding documentary information from references that will be recognized, decoded and interpreted by the user in semiosis documentary process.

Guarantee of quality in information representation and retrieval depends on the effective coding and decoding of documentary information, but it is not limited to understanding its structure, it should also consider some dialectical tensions between existing opposing forces when management and maintenance of documentary language, such as **conservation x mutation** and **consensus x specificity** (LIMA, 2004).

Conservation ensures the understanding among subjects, and **mutation** meets the changing needs of society, i.e., in the management and maintenance process of documentary language, one should maintain access points from previous systems and at the same time that should enable the addition of new access points (LIMA, 2004).

Documentary language must also meet both **consensus** and **specificity**, that is, on the one hand, it should meet its characteristic of being institutional, as it is always built to be used under the objectives of a particular institution ensuring mutual understanding of the subjects, and on the other hand, it should provide elements of a specific experience to each user (LIMA, 2004).

The need for documentary language to ensure consensus at the same time that meeting specificity demands precision and consistency from descriptors that can only be achieved from the compilation of true statements that will disclose their meaning and disclose semantic relationships that articulate their conceptual network. In other words, it is necessary to identify the characteristics that make up the concept, whose designation (term) serves as reference for the descriptor who constitute documentary information when assigned to a document. Finally, the sense of documentary information is expressed by the definition of the concept it denominates.

In turn, the definition of the concept gathers attributes or characteristics that allow to determine the categories of a documentary language. These categories are defined by the common trait of a whole class of concepts/terms/descriptors which, for this reason, are associated. The limit of each category is established by specific traits that allow to individualize each concept/term/descriptor establishing the disjuncture among the elaborated documentary information.

Here, we include in the trajectory the notion of semantic class that was also addressed by Gardin (1966) when describing documentary lexicon referring to the need to organize its terms to disclose the existing hierarchical relationships among them, either by affinity or by semantic difference.

Determining the semantic domain in linguistics is, according to the epistemological assumptions, seeking to discover the structure of a given domain, or propose a structure to it (Dubois et al, 2011). Trier, cited by Lopes (1987), observes that the lexical units of a language are organized into structured groups in such a way that each unit is defined therein by the position it occupies in relation to the others. Thus, we infer that the meaning of a concept/term/descriptor is specified by its similarity and its difference in relation to other relevant elements of the semantic domain, as one word only acquires its meaning as in opposition to other units in the same field (Germain, 1981).

Genouvrier and Peytard (1974), with regard to the semantic domain, defined it as the set of employments of a word (or syntagm or lexia) where and through which the word acquires a specific semantic load and the delimitation of these employments would occur by recognizing all immediate contexts that the word receives in a given text.

According to Hernando Cuadrado (1995) the minimum condition for the words to belong to the same domain is that they have a significant common trait (sema) (the higher the number of semas, the more coherent the semantic field will be, and in general, the fewer words integrates it); a word can take part of all semantic domains that are built over any significant traits that are discovered in it; when a word has several meanings, each of them belongs to a different semantic domain.

For example, while we can identify as belonging to the semantic domain of the word *table*, due to the common characteristic "object that allows gathering around", the following words: dining table, round table, assembly table, operating table, each may form part of a different category in a documentary language, due to delimiting characteristics such as "for foods"; "for discussion"; for surgery", etc. In another case, the word *iodine* may to be included in different categories of documentary language from delimiting characteristics, namely, *to be a raw material; to be a product, to be a reactant.*

At the moment one has discussed interoperability among documentary languages in semantic web, we believe it is necessary to deepen discussions on mapping of semantic fields that make up a domain, as only from concept characteristics listed in the definitions of terms, that will serve as referent and will contextualize the descriptors of documentary languages, it will be possible to make information representation and retrieval more effective.

References

The references was made following the ABNT rules.

CABRÉ, M. T. La terminologia hoy: concepciones, tendências y aplicaciones. *Ciência da Informação*, Brasília, v.24, n.3, p.289-298, set./dez. 1995.

CROS, R.C.; GARDIN, J-C; LEVY, F. *L' automatisation des recherché documentaries: un modele general: le Syntol*. Paris: Gauthiers-Villars, 1968.

DUBOIS, J. ET AL. *Dicionário de linguística*. São Paulo: Cultrix, 2011.

GARDIN, J. C. Éléments d'un modele pour La description des lexiques documentaires. *Bulletindes Bibliothèques de France (en ligne). n.*5, 1966. Disponível em http://bbf.enssib.fr/consulter/bbf-1966-05-071-001, Acessado em 20 jun. 2015.

GARDIN, J. C. *La logique du plausible*. Paris: Editions de la Maison des Sciences de l´homme, 1981.

GENOUVRIER, E., & PEYTARD, J. *Linguística e ensino do português*. Coimbra: Almedina, 1974.

GERMAIN, C. (1981). *La sémantique fonctionnelle*. Paris: PUF, 1981.

Hernando Cuadrado, L. A. *Introduccíon a la teoria y estrutura del lenguage*. Madrid : Editorial Verbum, 1995.

LARA, M. L. G. Conceitos de organização e representação do conhecimento na ótica das reflexões do Grupo Tema. *Inf. Inf.,* Londrina, v. 16. n. 3. p. 92 – 121, jan./ jun. 2011.

LARA, M. L. G. É possível falar em signo e semiose documentária? *Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.,* Florianópolis, 2° número esp., 2º sem 2006

LARA, M. L. G. *Representação e linguagens documentárias: bases teórico-metodológicas.* (doutorado), Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 1999.

LIMA, V.M.A. *Da classificação do conhecimento científico aos sistemas de recuperação de informação: enunciação de codificação e enunciação de decodificação da informação documentária.* (doutorado), Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo

LIMA, V.M.A. *Terminologia, comunicação e representação documentária*. 1998. Dissertação (Mestrado) Escola de Comunicações e Artes, Universidade de São Paulo, 1998

LOPES, E. *Fundamentos da linguística contemporânea*. São Paulo: Cultrix, 1987.
SMIT, J. W. Informação. In: LIMA, Y.D.; SMIT, J.W. (Org.) Organização de arquivos. São Paulo: IEB/ECA, 2000. p.19-31

PEIRCE, C.S. *Semiótica.* São Paulo: Perspectiva, 1977.

TÁLAMO, M.F.G.M.; LARA, M.L.G.; KOBASHI, N.Y. & LIMA, V.M.A. Instrumentos de controle terminológico: limites e funções. *Actas do II Simpósio Latino-americano de Terminologia,* 1990. Disponível em: <http://www.riterm.net/actes/2simposio/indice90.htm>. Acesso em 20 de junho de 2015.

VOGEL, M.J.M. A influência de Jean-Claude Gardin e a linha francesa na evolução do conceito de linguagem documentária. *Perspectivas em Ciência da Informação*, v.14 , número especial, p.80-92, 2009.