Biases in Knowledge Representation: an Analysis of the Feminine Domain in Brazilian Indexing Languages†

Abstract: The process of knowledge representation, as well as its tools and resulting products are not neutral but permeated by moral values. This scenario gives rise to problems of biases in representation, such as gender issues, dichotomy categorizations and lack of cultural warrant and hospitality. References on women’s issues are still scarce in the literature, which makes it necessary to analyze to what extent the terms related to these particular issues are inserted in the tools in a biased way. This study aimed to verify the presence of the terms female, femininity, feminism, feminist, maternal, motherly, woman/women within the following Brazilian indexing languages: Subject Terminology of the National Library (STNL), University of Sao Paulo Subject Headings (USPSH), Brazilian Senate Subject Headings (BSSH) and Law Decimal Classification (LDC). Each term identified in the first three alphabetical languages generated a registration card containing both its descriptors and non-descriptors, as well as scope notes, USE/UF, RT, and BT/NT relationships. As for the analysis of LDC, the registration card was filled out by following the categories proposed by Olson (1998). The results showed signs of biases, which enabled the proposition of guidelines that may contribute to minimize them and open the way for further discussions.

†Sponsor: Sao Paulo State Research Foundation - FAPESP

1. Introduction

Presently, Library and Information Science (LIS) is reaching beyond the traditional issues of retrieval, access and dissemination of information, and has turned its investigative focus to epistemological questions. There has been a growing reflection on the way knowledge organization processes (e.g., subject analysis) are supported by tools (e.g., indexing languages, classifications, ontologies, etc.), so as to generate reliable and trustful products, namely, products that are both authoritative and defendable before the user community (e.g., indexes, abstracts, classification numbers, etc).

In this context, the study of ethical aspects of knowledge organization and representation (KOR) responds to concerns that have been guiding the area for nearly three decades (Dahlberg 1992, López-Huertas 2008b), but further research is still needed because, as highlighted by Fernández-Molina and Guimarães (2002), this theme has been traditionally addressed in terms of the professional practice in information production and use, often merged with measures of information retrieval, or even perceived as inherent to a generic concept and common sense.

For the past 10 years, LIS throughout the world has particularly focused on research involving ethical issues in KOR. There has been an increasing interest in deontological ethics, mainly professional codes of ethics and conduct, as reported in a significant international survey of key institutions from several countries (Vaagan 2002).

Moreover, authors like Froehlich (1994), Gorman (2000), Koehler and Pemberton (2000) among others have approached ethical issues from an axiological perspective by emphasizing the actual professional practice as a whole and in a vertical direction. However, while such concern has been more evident in professional activities related to information
production and use as well as management of information units and systems, a fundamental intermediate issue, i.e., organization process, remains to be explored.


Amongst the various themes (political, religious, racial aspects, etc.) discussed from an axiological perspective, gender is treated in isolation and with limitations, which poses the question: to what extent are the terms concerning female issues embedded with biases in the tools of knowledge representation?

Knowledge representation is developed within a cultural context and aims to provide information to users, so it becomes essential to ensure that they can recognize themselves in the subject representations. The existence of biases in these products, which should act as surrogates of knowledge (Olson 2002), could lead the users away from the information system as a whole, because they do not feel familiar with nor included in it.

This paper is based on studies by Guimarães (2003, 2006a) and aims to analyze biases in relation to female issues by comparing four Brazilian indexing languages with the international situation described in the literature.

2. Theoretical approaches to women issues

The ethical values - and problems, eventually - that are present in the processes of knowledge representation were identified and grouped in three sets (Guimarães et al. 2008):

a) superior values that must guide all the informational activity, such as Respect to Privacy, Authorship (Copyright), Accessibility, Freedom, Safety, Equity, Diversity and Risk Minimization;

b) values previously recognized as professional requirements because they are essential to the action of an information professional, such as Competence-Skill, Efficiency, Flexibility, Reliability, Professional Recognition, Up-to-dating, Autonomy, Power Awareness and Cooperation;

c) values previously considered as mere information retrieval measures, but currently recognized as part of the axiological universe of KOR, such as Precision, Recall, Cultural Warrant, Exhaustivity, Consistency, Usability, Hospitality.

In this context, five fundamental values, which together correspond to 57% of the whole content analyzed, were observed: Privacy, Precision, Cultural Warrant, Authorship (Copyright) and Exhaustivity.

Following the same methodological procedure, ethical problems were grouped into two categories:

a) Problems that are not specific to the KOR activities context but are present all over the world today, such as Digital Segregation, Pornography, Discharge of electronic garbage, Professionals being replaced by technology, Violence;

b) Problems that directly relate to KOR professional activities, such as Surveillance, Censorship, Lack of cultural warrant, Negligence, Informational Directness, Professional Inefficiency, Misrepresentation, Racism, Ambiguity, Marginalization, Impartiality or Neutrality Belief, Idiosyncrasy, Inaccessibility to information, Biased Terminology, Inadequate Translations.

Furthermore, while a great dispersion of problems was found, Surveillance itself showed a significant incidence (12%).
By taking the notion of cultural warrant proposed by Beghtol (2002) as a synthesis of the ethical values identified and the problem of biased representation in ethical problems related to misrepresentation, racism, marginalization and impartiality or neutrality belief, the present investigation aimed to analyze how women were represented in the four Brazilian indexing languages, as well as to find out whether biases were embedded in such context.

The problem of biases towards women in knowledge representation has been addressed from various standpoints: Milani and Guimarães (2010) refer to this issue as being related to prejudice and discrimination (Berman 1993); Olson (2002) and López-Huertas and Torres (2005), to gender questions; Guimarães (2006b), to dichotomy categorizations in classification systems or thesauri, and (Beghtol 2002), to lack of cultural warrant.

Olson (2002, 15) believes that - when naming is biased, when it leaves out diverse features, it disenfranchises groups and topics outside of an accepted norm. Therefore, I expect the greatest problems of subject access to occur in the representation of groups and topics that are marginalized from that norm or that seek to change it (Olson 2002, 15).

Bias can be conceived as - a negatively loaded word, as something to be avoided or minimized (Hjørland 2008, 256), and is found in knowledge representation usually related to gender, sexuality, race, age, ability, ethnicity, language and religion matters, - described as limits to the representation of diversity and to effective library service for diverse populations (Olson 2002, 7).

Brey (1999) argues that virtual representations are morally problematic due to misrepresentation (when they misrepresent reality and are clearly inadequate to standards of accuracy), and biased representation (cases in which the values and interests may unfairly disadvantage certain individuals or groups or unjustifiably promote certain values or interests over others). According to the author, producers and programmers of virtual reality games, for example, are responsible for the consequent biases generated by the thematic approach and representation used, e.g., how stereotype characters of bandits and/or dominant groups are represented. Representation involves a constant decision-making process and it must be reliable (or defensible). In this sense, Brey (1999) emphasizes the need for a proper methodology for the identification and prevention of biases.

An observation deserves attention: biases may be more evident in virtual environments because they are simulations of reality, but are not easily identifiable in knowledge representation.

Gender and discrimination of minority groups have been the focus of several studies: in the Universal Decimal Classification (Santos et al. 1999); on indexing languages and social exclusion (Caro and San Segundo 1999); on woman’ representation in the Universal Decimal Classification (Morán Suárez and Rodríguez Bravo 2001); on KOR of gender issues on the internet (López-Huertas and Barité 2002); on the analysis of terminological representations of women’s issues in four thesauri (López-Huertas et al. 2004); a theoretical approach in which it was observed that in relation to the domains involving minority groups, specialists make up for lack of knowledge by adding an unequivocal interdisciplinary expression like gender, women, etc. (López-Huertas 2006); on a comparison between the terminology of gender in Spain and Uruguay (López-Huertas 2008a); and on the feminine issue in lists of subject headings and in thesauri (Rodríguez Bravo 2007).

Another problem presented by Milani and Guimarães (2010) is the non-uniformity of thematic representation and the emphasis on representation of gender aspects only related to sexuality (specially to homosexuality), family life (including maternity) and personal relations.

The identification of the problem in the international scenario and the evidence of lack of knowledge representation studies in the literature, particularly on women’s issues, were the
starting points of the present research aimed to verify how these issues are represented in four Brazilian indexing languages mostly used in the country.

3. Methodology

A conceptual domain comprising the terms female, femininity, feminism, feminist, maternal, motherly, woman(en) as being related to feminine issues was applied in four Brazilian indexing languages: Subject Terminology of the National Library (STNL), University of Sao Paulo Subject Headings (USPSH), Brazilian Senate Subject Headings (BSSH) and Law Decimal Classification (LDC) (Carvalho 2002) on account of their wide use and subject range in Brazilian libraries.

Regarding the first three languages, each term identified generated a registration card containing its descriptor, scope note, USE/Used For (UF), Related Term (RT), and Broader Term (BT)/Narrower Term (NT) relationships.

As for the analysis of Law Decimal Classification, the registration card was filled out by answering the categories proposed by Olson (1998): notation, classification concept, hierarchical context and remarks.

Both BSSH and USPSH are the two most widely used indexing languages. The first one has approximately 9,500 descriptors; the second has about 30,796 subject descriptors plus 11,432 geographical and historical; 43 gender and form; 1,070 professions and occupations, and 575 qualifiers. BSSH is used by a cooperative library network located in Brasilia including the Executive, Legislative and Judiciary powers of the Federal Administration as well as the Federal District Government. USPSH serves the largest academic library system in the country and covers knowledge areas related to teaching, research and community services of the University of Sao Paulo.

STNL has about 37,949 descriptors plus 128,077 subject topics, 20,360 geographical and 2,819 subject subdivisions. It is an adaptation of the Library of Congress Subject Headings to fulfill the informational needs in Brazil. The National Library Foundation, in Rio de Janeiro, is in charge of its maintenance and management.

LDC, which was introduced by Carvalho (2002), is an expansion of class 340 (Law) of Dewey Decimal Classification. This indexing language is widely used in this particular field, since it predicts the application of the Dewey Decimal Classification law system (based on the Common Law) to the Civil Law Code prevailing in Latin America countries and Continental Europe.

After the collection and systematization of all the terms found in the four languages, each context was analyzed following the dimensions proposed by Olson (2002) and López-Huertas and Torres (2005). Data were analyzed taking into consideration issues on women, the literature available and knowledge organization, not from a feminist theory perspective.

4. Analysis of the results

The results revealed a set of 360 descriptors (68,33% in STNL, 16,39% in USPSH, 11,67% in BSSH, and 3,61% in LDC) from which it was possible to extract the following subject dimensions: Crime, Culture, Law, Education, Feminism, History, Individual, Maternity, Religion, Health and Sport, Sexuality, Sociology and Work.

In this universe, it was observed that 108 descriptors (30% of total) presented more than one incidence in indexing languages.

The analysis of these results showed specific aspects related to biases that are commented below.
4.1 Subject Terminology of the National Library (STNL)

In STNL, there is a wider range of biases, such as:

a) the specific descriptor Journals for women without the corresponding Journals for men, which suggests the idea of an exceptional treatment given to females;

b) the descriptor Sanitary education appears as a narrower term of the descriptor Female education when aspects referring to sanitary education should not include gender qualifiers;

c) the descriptor Lesbianism is used in the context of sexual behavior of women, which reveals two biases: the use of suffix -ism (denoting vice) rather than lesbian women, and the relation of this issue only to sexual aspect, disregarding affective and legal dimensions;

d) a social stereotype is reinforced from the prediction of remitting to Femme Fatale, a term used to refer to seductive women;

e) specificities connected to religious issues lead to the idea of women as exceptions to masculine norms, as revealed in the scope note of the descriptor Woman (Christian theology): used for resources that deal with females’ Christian theology. Resources on Theology of the human race and males from the point of view of two or more religions enter Man (Christian theology). Resources on Christian theology of the human race enter Man (Theology);

f) in relation to sports and physical education, STNL attempts to include women in some specificities, such as Soccer, Gymnastics, Physical modeling, Weight training and Gliding, leaving other sports blank, thus revealing a tendency to treat women as exceptions to male norms and to consider generalities as masculinity;

g) in the professional area, some professions are directly assigned to Females (Architects, Lawyers, Composers, Philosophers, etc.), while in other activities the term Women is used (Women in aeronautics instead of Women aeronauts; Women in public service instead of Public servants; Women journalists instead of Journalists, etc.), which shows that women’s professional performance traditionally takes place in some privileged areas.

4.2 University of Sao Paulo Subject Headings (USPSH)

As for USPSH, the following biases were particularly observed:

a) Women who have been hit and Abused women appear related to Violence in the family, which indicates that violence against women occurs only within the family and usually referring to married women, thus excluding other female circumstances. Moreover, a generic descriptor family is found representing attacks on a specific individual (woman);

b) Women delinquent described next to descriptors such as Juvenile delinquent, Habitual delinquent, Passionate delinquent, Political delinquent, Sexual delinquent, which leads to the idea of a particular form of crime committed by a woman;

c) Feminism hierarchically subordinated to Women, where this currently refers to a movement that supports both women and men in their struggles and reflections.

4.3 Brazilian Senate Subject Headings (BSSH)

The following biases were found in BSSH:
a) the adjective masculine appears only related to the descriptor Homosexuality throughout the language, thus showing the issue is considered a characteristic of males (or the norm) and females are the exception;

b) the descriptor Women's Health reveals a tendency to emphasize only aspects related to sexual education, hygiene and maternity.

4.4 Law Decimal Classification (LDC)

In LDC, these biases were clearly observed:

a) Women delinquent (alcoholics, alienated, prostitutes, etc.) in Criminal Law appears under delinquents or criminals and the like, revealing that women are considered in a specific or exceptional situation, whereas the male is protected by a generic descriptor. It is also emphasized the fact that alcoholics should be assigned to classes related to health, and only prostitutes could be inserted under crime;

b) Women Rights in general appears only in the context of family law, subordinate to Married Women, Marital authority, Marital authorization, which creates a false sense of dependency among the themes;

c) the notation Sex-Woman Condition is presented subordinate to the notation Circumstances that influence the status and legal capacity, which lends women an inferiority status;

d) within the penitentiary sphere, Penal institutions for women and girls is subordinate to Other penitentiary institutions, which explicitly shows an exceptional and marginal situation, for in a context of equity one may think of Penitentiary institutions for men and boys next to Penitentiary institutions for women and girls. Another bias previously pointed out by Santos et al. (1999) is the lack of criteria to differentiate women from girls, as well as to determine the age of admission into these institutions.

5. Discussion

Since STNL is a list of subject headings, all the terms related to women are under the subject headings woman and women. Although the number of terms under the subject headings man and men is significantly smaller than that assigned to woman/women, this indexing language does not give an equal treatment to both genders. STNL offers various terms on women’s issues, but on the other hand they are inherently biased towards the natural language itself.

Considering that STNL aims to embrace all knowledge areas, it showed the highest rates of biases referring to women’s journals, sanitary education, lesbianism, femme fatale, woman (Christian theology), sports and physical education and professional performance.

As for USPSH, its primary purpose is geared to teaching, research and community activities in the Sao Paulo State University, but is also used by other Brazilian universities. Biases were identified in this indexing language referring to women who have been hit, women delinquent and feminism.

In BSSH the adjective masculine and the relations in women’s health are omitted in their descriptors, a shortcoming that should be revised. In spite of these biases, there seems to be a concern on the part of BSSH managers to contemplate feminine issues and to make use of the mechanisms available in the language, thus revealing a more equitable conception of relations between men and women.
The purpose of LDC, as an expansion of class 340 of Dewey Decimal Classification is to insert Brazilian Law specificities in this context. However, the signs of biases in the language were found to have a most negative impact as to women’s issues, such as women delinquent, women’s rights in general, sex-woman condition and penitentiary institutions for women and for girls.

On comparing the indexing languages analysed, it was found that only one instance of bias was repeated in USPSH and LDC referring to *Women delinquent*. In the first one, it appears as a criminal biotypology and in the second one, it is linked to alcoholics, alienated and prostitutes as types of delinquents or criminals.

The analysis of the corpus confirmed that the major Brazilian indexing languages

- treat women as exceptions to masculine norms or ignore these issues entirely (Olson 2002); and
- generality is considered as masculinity; masculine terms are used in the plural form with the semantic content of men and women; there is a tendency to use female descriptors and to omit their corresponding male terms (López-Huertas and Torres 2005).

These possibilities of biases are a reflection of the sexism inherent to human language.

On the other hand, the four Brazilian indexing languages

- do not isolate women’s issues to separate them from knowledge as a whole (Olson 2002);
- do not consider that the subject of contraception refers to women only (López-Huertas and Torres 2005).

In view of these observations, it is believed that reflection and application of mechanisms, such as syntagmatic relationships, paradigmatic relationships and scope notes, remarks or definitions would create and improve contexts, making them more inclusive and equitable, as shown in the following examples:

**Subject Terminology of the National Library (STNB)**
Descriptor: Sterilization of women

- UF Female sterilization
- UF Women, Sterilization
- UF Sterilization, Female
- RT Female sterility
- NT Tubarian sterility
- NT Hysterectomy

**Brazilian Senate Subject Headings (BSSH)**
Descriptor: Sterilization (birth control)

- UF Women, Sterilization
- UF Sterilization, Female
- RT Birth Control
- RT Family planning
- BT Anti-contraceptive
These two examples show that some mechanisms available in the language were used. Nevertheless, the first one characterizes an approach in which the subject is exclusively referred to as the medical procedure of sterilization, relating it to sterility. In the second example, sterilization is viewed as an anti-contraceptive possibility related to family planning.

In both cases, syntagmatic relationships should have been used in the indexing languages by inserting terms like Male sterilization or Vasectomy as related terms (RT) of Sterilization of women and Sterilization (birth control). It was found that in BSSH Vasectomy appears as a narrower term (NT) of Anti-contraceptive and as related term (RT) of Birth Control. As for STNL, paradigmatic relationships are used, for example: Vasectomy and Anti-contraceptive as narrower terms (NT) of Sterilization.

6. Conclusion

The analysis of the four Brazilian indexing languages and the examples discussed in this investigation confirm the need for mechanisms to prevent biases, such as: insertion of notes that allow the use of the feminine and/or the insertion of syntagmatic descriptors in order to distinguish the masculine from the generic when no other option is available; also, the insertion of gender qualifiers in descriptors (M) and (W) whenever the subject matter does not represent humanity in general (Rodríguez Bravo 2007).

Indexing management mechanisms such as these confer dynamism to indexing languages, and are currently viable through information and communication technologies (ICT) as well as digital processing, storage and retrieval environments. Therefore, it is crucial that information professionals keep in mind that an effective knowledge representation is key to the access and appropriation of an ever-increasing number and variety of resources in all fields.

Furthermore, it is the information professionals’ responsibility not only to make information available to the end-user, but also to enable him/her to understand the underlying mechanisms of the indexing language, the standpoint and mission of the unit it represents and the community discourse embedded in the products. In this context, it is not possible to avoid biases, but there are mechanisms to elucidate or exteriorize them, otherwise the user’s needs will not be fulfilled.

References


Olson, Hope A. 2002. The power to name: locating the limits of subject representation in libraries. Dordrecht: Kluwer Academic Publisher.


