

Los problemas terminológicos de la organización y del acceso a la información

NAIR YUMIKO KOBASHI

Escola de Comunicações e Artes - Universidade de São Paulo, Brasil

Departamento de Biblioteconomia e Documentação

1 INTRODUCCIÓN

La información es un concepto analizable desde distintas perspectivas. En este texto nos acercamos a tal concepto en el contexto de los archivos; es decir, de “la memoria institucionalizada y funcionalizada como información” (Abril, 2004, p.9). Así, remitimos el término al dominio de la Ciencia de la Información, el campo que estudia la información y los dispositivos de almacenamiento y recuperación de información para garantizar flujos.

García Gutiérrez (2002, p. 9) nombra esos dispositivos como “exomemorias”; esto es, memorias exteriores en las que se encuentran inscritos “los pensamientos, experiencias, ilusiones, emociones”.

La interacción con las exomemorias, como también su producción, es afectada por la intervención mediática y tecnológica (Orozco Gómez, 2006). Las tecnologías de la información determinan, en gran medida, las formas de almacenar y hacer disponible la información.

En principio facilitadoras de los procesos informacionales, las tecnologías ubican nuevos problemas: se necesitan mediaciones cada vez más complejas para reconocer, encontrar y comprender la información. Las mediaciones permiten que las inscripciones del pasado sean

reinterpretadas para la construcción/reconstrucción “de la cultura, la conciencia y las identidades del presente” (García Gutiérrez, 2002, p. 13). La producción de informaciones, su almacenamiento en las exomemorias y el acceso no son, por tanto, procesos triviales.

2. PRODUCCIÓN DE EXOMEMORIAS

La producción y uso de las exomemorias pueden analizarse en el cuadro de las reflexiones acerca de la comunicación y del lenguaje. Para aclarar el acercamiento se definen preliminarmente los términos directamente involucrados en el campo de la producción y el uso de las exomemorias:

- a) *Información* es representación de contenidos inscritos en artefactos de información. En este sentido, información es una forma simbólica que representa los documentos recoletados y preservados en instituciones de memoria (bibliotecas, archivos y museos, físicos o electrónicos).
- b) Documento e información no son términos equivalentes. El documento es un soporte de mensaje; esto es, un objeto que tiene una dimensión material. La *información*, a su vez, es un objeto simbólico al que se accede por el lenguaje.
- c) *Información documental* es una información formateada para las exomemorias (Kobashi, 1994, 2006). Aunque la información documental es una forma simbólica, se trata de un tipo particular de representación que, en sus distintas formas de expresión (conjunto de descriptores, resúmenes, símbolos de un sistema de clasificación) busca responder a las funcionalidades de los sistemas de recuperación de información.

En la información, obtenida por un lado, por la neutralización del poder expresivo del texto y, por el otro, moldeada de acuerdo con reglas previamente determinadas, prevalece la idea de algo que, aunque formalmente distinta del original, y por lo tanto “representación”, es

equivalente al original desde el punto de vista del contenido. En este sentido la información documental es una representación construida a partir de un objeto efectivamente presente, el documento. Es, así, información segunda, de naturaleza referencial, producida por operaciones metonímicas (Kobashi, 2006).

Svenonius afirma que la organización de la información en sistemas presenta tres dimensiones: la social, la teórico-metodológica y la operacional. La dimensión social es del orden de la institucionalización de las actividades informacionales; la dimensión teórico-metodológica, del orden del conocimiento crítico acumulado que ofrece soluciones para los problemas de la producción y el flujo de la información, y la dimensión operacional es de naturaleza técnica y tecnológica, y se asocia a las técnicas e instrumentos utilizados en estos procesos (SVENONIUS, 2000).

Hay por tanto tres aspectos indisolubles en la producción de las exomemorias: la *esfera teórica*, en la cual se determina el objeto, las funciones, los instrumentos y los procedimientos metodológicos; la *esfera de la producción*, donde se engendran las reglas de generación de las diversas modalidades de información documental y la *esfera pragmática*, en la cual se verifican las condiciones de adherencia entre los usuarios y los sistemas de información (Kobashi, 1994, 2006).

2.1 Organización y recuperación de información en exomemorias

Organizar información en exomemorias es, en una primera aproximación, una operación de clasificación que reúne por semejanza y diferencia. Este criterio, en su aparente simplicidad es fuente de muchos problemas, uno de los cuales es la necesidad de contar con un sistema de referencia para representar las informaciones. Cada clase del sistema de referencia debe ser auto-excluyente, y además debe ser lo suficientemente discriminante como para abrigar un conjunto manipulable de objetos que pueda responder a las necesidades de información de distintos usuarios. En otras palabras, clases muy genéricas no son aptas para discriminar información; por otro lado, clases muy específicas pulverizan en lugar de reunir. Los extremos no son útiles para organizar la información.

La información puede ser recuperada de distintas maneras. Inspeccionar cada documento de una colección es una forma elemental de hacer búsquedas. Recurrir a los dispositivos que reúnen y sistematizan información en clases, como las bases de datos bibliográficos, es, en principio, mucho más eficaz (Wellisch, 1987).

Clasificar es también estandarizar. Por medio de la estandarización se obtienen conjuntos de documentos con cierta homogeneidad. Esta operación le otorga previsibilidad a las búsquedas, lo que concurre simultáneamente a la estabilidad y a la economía de los sistemas.

La estandarización se aplica a las cuatro funciones prioritarias de los sistemas de información: a) en la descripción de la forma del documento; b) en la descripción de contenido; c) en la transcripción de los datos para un formato estandarizado —los metadatos— y, finalmente, d) en la búsqueda (Wellisch, 1987).

Un sistema de información no opera, por tanto, sin la mediación de un lenguaje de codificación y de decodificación. Esto implica afirmar que el lenguaje de un sistema de información, para cumplir sus funciones, debe ser moldeado de acuerdo con la comunidad discursiva que va a utilizarlo. Un sistema simbólico compartido es, por tanto, capital en las exomemorias.

El sistema de referencia de los sistemas de información es lenguaje, y, como parte constitutiva de la realidad, sus conceptos presentan dinamicidad; o sea, están en conexión dinámica con la realidad y los sentidos.

2.2 Lenguajes documentales

La construcción de lenguajes de codificación y búsqueda de información es una de las tareas más complejas de los sistemas de información. Las categorías aristotélicas definen el principio de estructuración de gran parte de los lenguajes documentales. Las categorías promueven la estabilidad de un universo de conocimiento y permiten segmentar y jerarquizar los conceptos. Los lenguajes documentales (LD), son, además, un conjunto de términos relacionados semánticamente. En los tesauros documentales y ontologías, por ejemplo, la jerarquización es apenas uno de los modos posibles de relacionar

conceptos. Las relaciones son, por tanto, multidimensionales: los términos se asocian por sinonimia o por otros tipos de relaciones espacio-temporales (causa y efecto, proceso y producto, agente y paciente, por ejemplo).

Para que los LD sean funcionales, deben integrar los términos de los lenguajes de especialidad. Así, los LD son afectados, necesariamente, por las asimetrías que se desprenden de la dinámica de la producción de conceptos.

Los lenguajes documentales no están constituidos únicamente de unidades terminológicas; su eficacia se vincula a la asociación de diversos paradigmas de designación: el lenguaje de los dominios técnico-científicos, el lenguaje común o natural y el lenguaje de los documentos.

En ese conjunto, la terminología es el referente fundamental, el sistema a partir del cual se establecen las equivalencias entre los diferentes lenguajes en juego. Aunque el referente del descriptor sea el término, el descriptor no presenta el rigor del término (Cintra, 2003). Su valor es instituido igual por operaciones orientadas por la economía de los sistemas de información.

En efecto, es común en los lenguajes documentales, establecer equivalencias entre términos de mayor intensión y términos de mayor extensión. Éste es un recurso para limitar la cantidad de unidades del léxico documental y dotarlo de mayor eficacia.

Por otro lado, la producción de nuevos términos crea tensiones permanentes en los sistemas de información. Los LD deben por tanto, actualizar permanentemente su base léxica para que no se cristalice y pierda su potencialidad de codificar y transmitir información.

El lenguaje documental, en esa medida, solicita apertura permanente para incorporar nuevas unidades y nuevas relaciones, sin perder, con todo, el cierre de un sistema. No es difícil comprender semejante dinámica al establecerse una analogía entre lenguaje natural y lenguaje documental. El lenguaje natural incorpora nuevas nociones de forma permanente sin perder su condición de sistema. Este atributo es también constitutivo de los lenguajes documentales.

3. ESTUDIO EMPÍRICO DE LA CODIFICACIÓN DE INFORMACIÓN EN BASES DE DATOS

La elaboración de repositorios de alta calidad es condición *sine qua non* para preservar y hacer recuperable la memoria colectiva. El estudio de dos bases de datos de tesis brasileñas, presentado a continuación, demuestra algunos de los problemas de indización encontrados.

Los resultados más amplios y los aspectos teóricos, metodológicos y operacionales de estos estudios se detallan en otros trabajos de los investigadores del Grupo Scientia¹ (Eliel, 2006; Igami, Nozaki, Kobashi, 2007; Kobashi, Santos, 2005, 2006a, 2006b, 2006c; Santos, Kobashi, Igami, Bressiani, 2006; Igami, Nozaki, Kobashi, 2007).

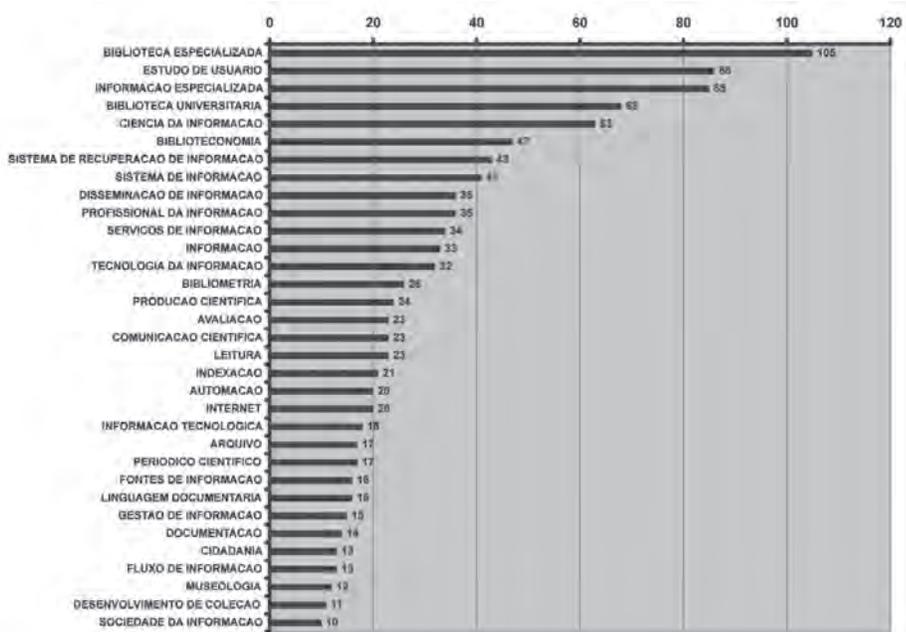
3.1 Análisis temático de las tesis de la Ciencia da la Información

Entre 1978 y 2001, han sido defendidas **833 tesis** en el área de Ciencia de la Información. En la indización de las tesis le han sido atribuidos **1929 descriptores** al conjunto. Para obtener un mapa temático del área se han seleccionado 33 descriptores cuya frecuencia de uso está por arriba de 10. Los 33 descriptores (2% del total) han sido atribuidos a 78% del total de documentos del *corpus*. La Fig. 1, abajo, presenta la distribución de los descriptores.

Se han detectado, en el análisis, los problemas siguientes: a) descriptores muy genéricos o, al contrario, muy específicos; b) ausencia de estandarización de conceptos equivalentes y c) uso excesivo o insuficiente de descriptores en cada documento.

1 El Grupo Scientia es constituido de investigadores de la Universidade de São Paulo, Universidade Federal de Santa Catarina, Instituto de Pesquisas Energéticas y Nucleares de São Paulo, estudiantes de grado y de posgrado.

Figura 1: Temas de las tesis de Ciencia de la Información (1971-2001)



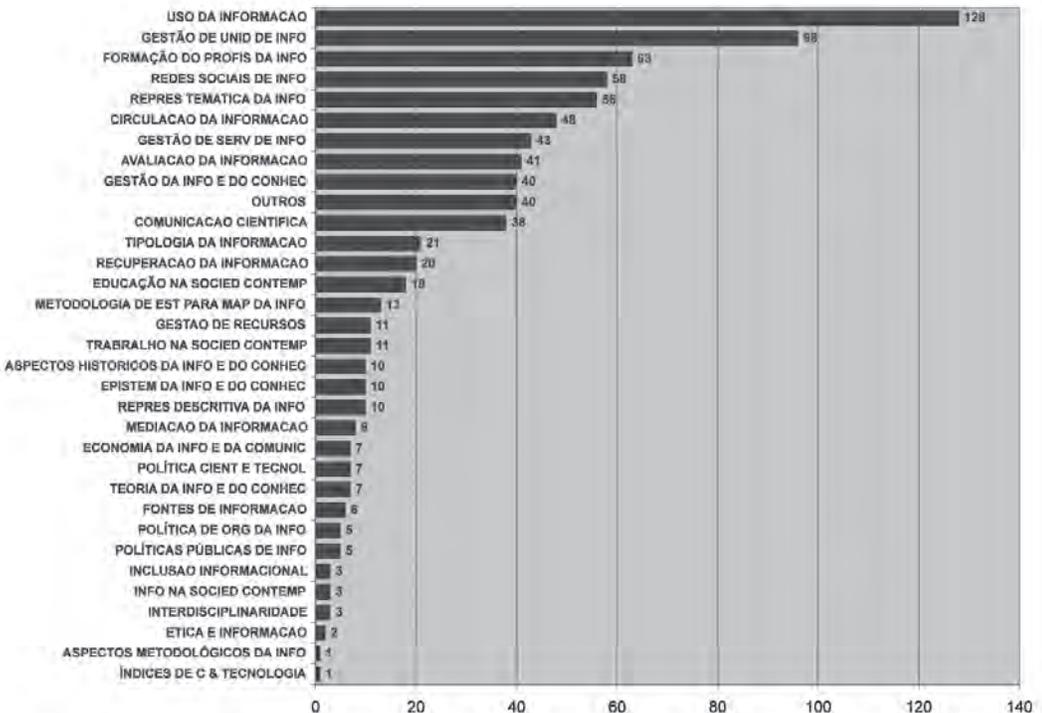
En la indización se han utilizado términos como “Producción científica” y “Comunicación científica” que, en ciertos contextos, representan un mismo proceso. “Información especializada” e “Información tecnológica” pueden ser también términos equivalentes. La ausencia de estandarización sintáctica de los descriptores (singular y plural, por ejemplo) aliada a la falta de uso de un vocabulario controlado de referencia para la indización, como se observa en la figura de arriba, indica que hay disparidades que comprometen la confiabilidad de la recuperación de información.

Para obtener una visión más exacta de los temas y simultáneamente promover diferentes efectos (*precision y recall*), se creó una base de datos *ad hoc* estructurada en: *a) campo de descriptores*: atribuidos por el autor de la tesis; *b) campo de macro-descriptores*: atribuidos *ad hoc*.

Primer Simposio Internacional sobre Organización...

Para obtener una visión más global de los datos, se adoptaron las denominaciones de los Grupos de Trabajo de la Associação Nacional de Pesquisa em Ciência da Informação e Biblioteconomia (Ancib) igual como criterio de agrupamiento temático de las tesis (Eliel, 2006). Los nombres de los Grupos de Trabajo, que agrupan las ponencias de los encuentros nacionales, se presentan como categorías temáticas reconocidas y legitimadas por investigadores de Ciência de la Información de nuestro país. Operan, por tanto, como marcos de referencia adecuados para clasificar las tesis en grupos homogéneos.

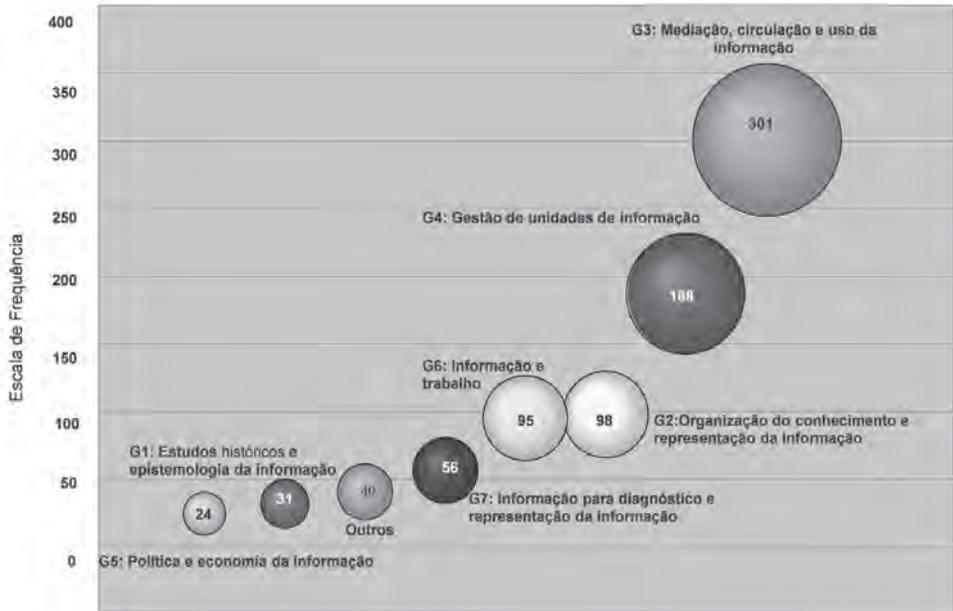
Figura 2 – Distribución de las tesis según términos descriptivos de los Grupos de Trabajo de ANCIB.



La Fig. 2 muestra la distribución de las tesis en un primer nivel de la descripción de los GT. El sistema de referencia utilizado permite ob-

tener una visión más exacta del área, y evitar la dispersión provocada por la atribución de descriptores por parte de los autores de las tesis.

Figura 3 - Tesis de Ciencia de la Información (1978-2001), según los Grupos de Trabajo de la ANCIB.



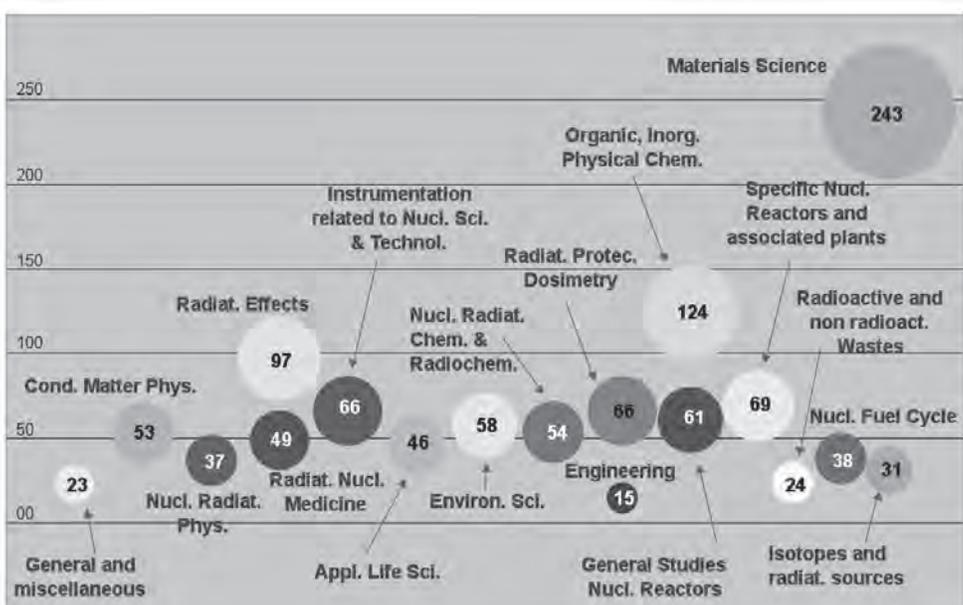
La Fig. 3 presenta un grado aún mayor de generalización de los temas del *corpus* considerado. En orden decreciente vemos: 1) Mediación, Circulación y Uso de la información; 2) Gestión de Unidades de Información, Organización del Conocimiento e Información y Trabajo. Las categorías con menor cantidad de ítems son Información para Diagnóstico, Estudios Históricos y Epistemológicos y Política y Ética de la Información. En la categoría Otros, están las tesis que no se identifican con ninguna de las 7 categorías.

3.2 Análisis de las tesis de Energía Nuclear

El *corpus* del estudio del área de Energía Nuclear está constituido por las tesis producidas en los últimos 30 años (1978-2005) en el Programa de Posgrado del Instituto de Pesquisas Energéticas e Nucleares (Santos, Kobashi, Igami, Bressiani, 2006). Los descriptores originales se han integrado a categorías más amplias para obtener conjuntos homogéneos y significativos. Se redondea, así, la dispersión provocada por descriptores muy específicos.

La Fig. 4 presenta la distribución temática de las tesis. La terminología adoptada representa el lenguaje del *Subject categories and scope descriptions* de INIS (2002, 2004).

Figura 4 – Distribución temática de tesis del Programa de Pos-rado de IPEN (1978-2005).



Es visible, en la Figura 4, la prevalencia de tres áreas: *a) Materials Science*, *b) Organic, Inorganic and Physical Chemistry* y *c) Radiation Effects*. Los resultados han sido considerados útiles para el planeamiento de la política de investigación del Programa Institucional de posgrado.

4. CONSIDERACIONES FINALES

Clasificar es esencialmente un proceso descriptivo, que consiste primeramente en elegir las dimensiones de los datos empíricos que serán luego divididos en elementos discretos y enumerables (Granger, 1974, 1989). Para los fines específicos de esta investigación, fue necesario neutralizar la dispersión y constituir nuevos conjuntos de informaciones.

Las clases son, en definitiva, aproximaciones que expresan “semejanzas de familia” (Wittgenstein, 1996), una operación lógico-semántica que permite reunir objetos que presentan trazos comunes. Son, con todo, agrupamientos con alguna inexactitud y fluidez, a semejanza de los descriptores de los lenguajes documentales. La abstracción conceptual es, en ese sentido, la operación que permite caracterizar los datos del *corpus*, y luego compararlos y resignificarlos.

El análisis ha demostrado que las bases de datos de tesis de Ciencia de la Información no presentan la misma calidad y consistencia que la base de tesis de IPEN. Algunos factores explican las diferencias encontradas. En el caso de la Ciencia de la Información, los metadatos han sido producidos descentralizadamente. Cada curso de posgrado envía sus datos a un repositorio central, pero no se hace un trabajo preliminar de darle consistencia a los datos antes de ser incluidos en la base de datos.

Para reorganizar los datos de la Ciencia de la Información, medida necesaria para obtener resultados significativos, fue necesario hacer un trabajo intenso de estandarización. Por esta razón los datos producidos sin recurrir a un sistema de referencia compartido provocaron dispersión de información.

En el caso de IPEN, la consistencia de la base de datos es resultado de la política de documentación de la biblioteca, unidad responsable de la catalogación e indización de la producción científica local.

Quedó evidenciada en este estudio la necesidad de establecer políticas de estandarización y de consistencia de los repositorios producidos cooperativamente. Es necesario destacar que la ausencia de directrices puede comprometer la calidad y la confiabilidad de los sistemas cooperativos. El proceso conocido por auto-archivamiento, si no es controlado por directrices, puede, a largo plazo, comprometer la interoperabilidad de los sistemas.

Los dispositivos de información son sistemas dinámicos continuamente modificados por la introducción de nuevas informaciones. En definitiva, la utilidad de los sistemas de información se establece en la tensión entre la manutención de la estructura y la modificación de sus constituyentes. Los constituyentes son representaciones documentales de discursos de la ciencia, cuya característica fundamental es la creación conceptual, expresada en términos. Por tanto, un sistema de información que no incorpora la dinámica de la creación terminológica, está condenada al fracaso. Se evidencia, de ese modo la importancia de incorporar los métodos de la Terminología en la construcción de los lenguajes de representación de la información y en las operaciones de exploración de las memorias colectivas.

En el caso específico de la Ciencia de la Información, es evidente la necesidad de directrices y de un instrumento compartido de control terminológico para estandarizar la base de datos nacional de tesis. Un Tesoro de Ciencia de la Información actualizado es imprescindible para establecer los parámetros de indización.

Un tesoro documental, en la conceptualización que adoptamos, tiene como referente fundamental los términos producidos en la actividad discursiva de los dominios científicos. Esto es señal de la necesidad del diálogo permanente que debe existir entre la Ciencia de la Información y la Terminología para producir instrumentos de control terminológico adherentes a la dinámica de producción de conceptos.

La Ciencia de la Información reflexiona sobre la producción de las exomemorias, y presenta estructuras lógicas y significantes adecuadas para integrar flujos. El reconocimiento de que la información participa de diferentes estructuras significantes ha sido factor de motivación para renovar los métodos de elaboración de lenguajes documentales, ontologías y taxonomías. Sin esos instrumentos, se estrangula el flujo.

La Ciencia de la Información debe, por tanto, reflexionar sobre las prácticas informacionales y proponer acciones relativas a la producción de la información, su distribución y formas de acceso, en contextos mediáticos y tecnológicos. Por eso es urgente abandonar las visiones simplistas acerca del lenguaje, la representación de información y la producción de sentido.

Agradecimientos: expreso mi gratitud a Raimundo N. M. Santos, Priscila Nozaki y Mery P. Z. Igami, compañeros del Grupo Scientia de Investigación; al CNPq y a la Pró-Reitoria de Pesquisa-USP por las becas y el apoyo a la investigación. Un agradecimiento especial a Mery P. Z. Igami por su revisión del texto.

BIBLIOGRAFÍA

- Abril, G. Prólogo, en GARCÍA GUTIÉRREZ, A., *Otra memoria es posible: estrategias descolonizadoras del archivo mundial*, Sevilla, Un. de Sevilla; Buenos Aires, La Crujia, 2004, pp. 9-13.
- Cintra, A.M.M. *et al.*, Do termo ao descritor: estudo exploratório, en *Revista Comunicações e Artes*, São Paulo, 1993.
- Cintra, A.M.M. *et al.*, *Para entender as linguagens documentárias*, São Paulo, Polis, 2005.
- Ducrot, O., Todorov, T., *Dictionnaire encyclopédique des sciences du langage*. París, Seuil, 1972.
- Eliel, R.A., *Institucionalização da Ciência da Informação no Brasil: estudo da convergência entre a produção científica e os marcos regulatórios da área*, Campinas, PUC-CAMPINAS, 2007, 115f, (dissertação).

Primer Simposio Internacional sobre Organización...

- García Gutiérrez, A., *La memoria subrogada: mediación, cultura y conciencia en la red digital*, Granada, Universidad de Granada, 2002.
- García Gutiérrez, A., *Otra memoria es posible: estrategias descolonizadoras del archivo mundial*, Sevilla, Un. de Sevilla; Buenos Aires, La Crujia, 2004.
- Granger, G., *Pensamento formal e ciências do homem*, Lisboa, Presença, 1974, (v. 1 e 2).
- Granger, G., *Por un conhecimento filosófico*, São Paulo, Contexto, 1989.
- Hutchins, W.J., *Languages of indexing and classification: a linguistic study of structures and functions*, Herts, Peter Peregrinus, 1975.
- Igami, Mery P. Z. ; Nozaki, Priscila, y Kobashi, Nair Y., A gestão de qualidade em bases de dados da produção científica, en Congresso brasileiro de biblioteconomia, documentação e ciencia da informação, 12, 8-11 julho, 2007, Brasília, DF. *Anais*, (1cd).
- International Nuclear Information System - Inis, *Joint thesaurus*, Vienna, International Atomic Energy Agency, 2004, (ETDE/INIS Joint reference series, nº 1, rev. 1).
- International Nuclear Information System - Inis, *Subject categories and scope descriptions*, Vienna, International Atomic Energy Agency, 2002, (ETDE/INIS Joint reference series nº 2).
- Kobashi, N. Y., *A elaboração de informações documentárias: em busca de uma metodologia*, São Paulo, ECA-USP, 1994 (Tesi de doutorado).
- Kobashi, N.Y., *Exomemórias*, São Paulo, ECA-USP, 2006, (Tese de livre-docência) .

Kobashi, N.Y.; Santos, R.N.M., Information science institutionalization in Brazil: a bibliometric study, en Guerreiro-Bote, Vicente P. (Ed.), Current research in information sciences and technologies. Multidisciplinary approaches to global information Systems, International conference on multidisciplinary information sciences and technologies, 1st, october 25-28th, 2006, Merida Spain, en *Proceedings...* Badajoz, España, Open Institute of Knowledge, 2006c. v.1 pp. 610-614.

Kobashi, N.Y.; Santos, R.N.M., Institucionalização da pesquisa científica no Brasil: cartografia temática e de redes sociais por meio de técnicas bibliométricas., en *Transformação*, v.18, p.27-26, 2006b.

Kobashi, Nair Y.; Santos, Raimundo N. M., dos. Arqueologia do trabalho imaterial: uma aplicação bibliométrica à análise de dissertações e teses, en Encontro Nacional de Pesquisa em Ciência da Informação - 7.,2006, Marília, *Anais...* Marília: FFC/UNESP, 2006b, 1 CD ROM, (VII ENANCIB).

Lara, M.L.G., *A representação documentária: em jogo a significação*, São Paulo: ECA-USP, 1993, (Dissertação de Mestrado).

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*, São Paulo: USP, 2004.,(Tese de Doutorado).

Orozco Gómez, G., Aprendiendo con video juegos. III Congreso Educarede, São Paulo, maio/2006, disponível em: www.educarede.gov.br

Santos, R.N.M.; Kobashi, N.Y.; IGAMI, M.P.Z.; Bressiani, J.C. Brazil's insertion into nuclear energy research: bibliometric analysis of dissertations developed at the Instituto de

Primer Simposio Internacional sobre Organización...

Pesquisas Nucleares- São Paulo, en Guerreiro-Bote, V.P. (Ed.), Current research in information sciences and technologies. Multidisciplinary approaches to global information Systems, International conference on multidisciplinary information sciences and technologies, 1st October 25-28th, 2006, Merida Spain, en *Proceedings...* Badajoz, España: Open Institute of Knowledge, 2006. v.1, pp.553-556.

Svenonius, E., *The intellectual foundation of information organization*, Cambridge, Mass, MIT, 2000.

Wellisch, H.H., *Uma teoria dos sistemas de recuperação da informação*, Brasília, IBICT, 1987.

Wittgenstein, L., *Investigações filosóficas*, Petrópolis, Vozes, 1996.