SCIRE

Representación y organización del conocimiento

SCIRE

Representación y organización del conocimiento

Vol. 28, n.º 2, julio-diciembre 2022

ISSN 1135-3716

ISSN (e) 2340-7042

Scire:
knowledge representation and organization
Vol. 28, n. 2, July-December 2022

Ibersid:
Red de Investigación
en Sistemas de Información
y Documentación

ISSN: 1135-3716 = Scire (Zaragoza)
ISSN (e): 2340-7042
Depósito legal: Z. 1.790 — 1995

Edita: Ibersid® con la colaboración de Prensas de la Universdad de Zaragoza

Imprime:
Servicio de Publicaciones. Universidad de Zaragoza.
Edificio de Ciencias Geológicas, C/ Pedro Cerbuna, 12.
50009 Zaragoza, España. Tel.: 976 761 330. Fax: 976 761 063.

Scire

representación y organización del conocimiento

Alcance y objetivos

Scire: representación y Organización del Conocimiento es una publicación semestral de carácter interdisciplinar sobre la representación, normalización, tratamiento, recuperación y comunicación de la información y el conocimiento.

Difusión

Scire tiene difusión internacional. Agradecemos la inclusión en los siguientes servicios de referencia: Scopus, ESCI, Information Science Abstracts, Information Services in Physics, Electronics and Computing, Library and Information Science Abstracts, Sociological Abstracts, ERIH Plus, Knowledge Organization Literature, Base de Datos ISOC y Catálogo Latindex.

Instrucciones para los autores y procedimiento de evaluación

La última versión de las instrucciones para presentación de trabajos y del procedimiento de evalaución editorial están disponibles en: https://www.ibersid.eu/ojs/index.php/scire/about/submissions

Agradecimientos

Agradecemos el apoyo del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón con su subvención a grupos de investigación S6520D, al Vicerrectorado de Investigación y a la Facultad de Filosofía y Letras de la Universidad de Zaragoza.

Redacción, distribución y canje

Revista Scire
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Facultad de Filosofía y Letras
Universidad de Zaragoza
C/ Pedro Cerbuna 12,
E-50.009 Zaragoza (Spain)

Tfno: int + 34 976 762239. Fax: 34 976761506.
E-mail: mailto:scire@unizar.es

Suscripciones y números sueltos

Suscripción anual: 30 €. Número suelto: 20 €. (IVA inc.)

Scire

knowledge organization and representation

Scope and aims

Scire: Representación y Organización del Conocimiento is an interdisciplinary journal published twice a year that is devoted to the representation, standarization, treatment, retrieval and communication of information and knowledge.

Dissemination

Scire has international distribution. We acknowledge its inclusion in the following reference services: Scopus, ESCI, Information Science Abstracts, Information Services in Physics, Electronics and Computing, Library and Information Science Abstracts, Sociological Abstracts, ERIH Plus, Knowledge Organization Literature, Base de Datos ISOC and Catálogo Latindex.

Instructions for authors and evaluation process

The last version of the instructions for authors and assessment process is available at: https://www.ibersid.eu/ojs/index.php/scire/about/submissions

Acknowledgments

We acknowledge the help of the Department of Science, University and Knowledge Society of the Government of Aragon (grant S6520D to research groups), and of the Research Vice Rectorate and the Faculty of Philosophy and Arts of the University of Zaragoza.

Contact address

Revista Scire
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Facultad de Filosofía y Letras
Universidad de Zaragoza
C/ Pedro Cerbuna 12,
E-50.009 Zaragoza (Spain)

Tel.: int + 34 976 762239. Fax: 34 976761506.
E-mail: scire@unizar.es

Subscriptions

Annual subscription: 30 €. Issue: 20 €. (VAT included)

Tabla de contenidos en español

Table of contents in Spanish

Tabla de contenidos en inglés

Table of contents in English

Google Discover: entre la recuperación de información y la curación algorítmica

Google Discover: between information retrieval and algorithmic curation

Carlos Lopezosa, Javier Guallar, Gema Santos-Hermosa

Facultad de Información y Medios Audiovisuales, Centro de Investigación en Información, Comunicación y Cultura CRICC, Universitat de Barcelona, Melcior de Palau, 140. 08014 Barcelona, lopezosa@ub.edu, jguallar@ub.edu, gsantos@ub.edu

Resumen

La motivación de este estudio es el análisis de Google Discover tanto desde el punto de vista del usuario que lo utiliza como del webmaster que trata de posicionar su contenido y con ello conseguir más visitas. Asimismo, el objetivo principal es caracterizar y sintetizar la visión experta sobre Google Discover, a partir de una revisión de la literatura. La metodología empleada es la revisión sistematizada, más específicamente, la *scoping review* (revisión sistematizada exploratoria) de la literatura gris recuperada utilizando el buscador generalista de Google en su dominio google.es, que se completa con un análisis funcional de la herramienta desde las dos visiones (usuario y webmaster). Gracias a este análisis se ha podido comprobar qué es Discover, cómo usarlo y cómo optimizar el contenido para aparecer en su feed.

Palabras clave: Google Discover. Visibilidad web. Curación de contenidos. Curación algorítmica. SEO. Scoping review. Recuperación de información.

Abstract

The motivation main of this research is the analysis of Google Discover from the point of view of the user who uses it and the webmaster who tries to rank his content and thereby get more visits. Likewise, the main objective is to characterize and synthesize the expert vision on Google Discover, based on a review of the literature. The methodology used is the systematized review, more specifically, the scoping review of the gray literature retrieved using Google's general search engine in its domain google.es, which is completed with a functional analysis of the tool from both perspectives (user and webmaster). Thanks to this analysis it has been possible to verify what Discover is, how to use it and how to optimize the content to appear in the Google Discover feed.

Keywords: Google Discover. Web visibility. Content curation. Algorithmic curation. SEO. Scoping review, Information retrieval.

1. Introducción

La visibilidad web, por un lado, y la curación de contenidos, por otro, se han consolidado en los últimos años como actividades relevantes en el acceso a la información.

Ello ha comportado que los buscadores como Google hayan decidido desarrollar herramientas como Google Discover (Google, 2020a). Conocida también como Google Feed hasta 2017, este servicio de Google permite a los usuarios recibir noticias de actualidad relacionadas con sus intereses sin tener que realizar una búsqueda en Google. Se trata de un servicio que ofrece resultados de búsqueda noticiosos, en teléfonos inteligentes, atendiendo a los intereses de los usuarios, y que toma en consideración las estrategias de posicionamiento en buscadores (SEO por sus siglas en inglés) y de curación de contenidos algorítmica (Diakopoulos y Koliska, 2017).

Bajo esta circunstancia, nace este trabajo, que propone estudiar las características generales de Google Discover desde dos puntos de vista: (1) el del usuario que utiliza esta herramienta para informarse y, (2) el del webmaster que crea contenido en su sitio web con la intención de aparecer en los resultados de Google Discover.

En consecuencia, el objeto de estudio de este trabajo es Google Discover como herramienta a la vez para consumir información (usuario) y para mejorar la visibilidad de un sitio web intensivo en contenidos (webmaster) y conseguir, con ello, más tráfico orgánico. Por lo tanto, el objetivo principal de esta investigación es analizar los principales elementos técnicos y conceptuales del uso y de la optimización de contenido en Google Discover proponiendo así el primer estudio académico existente hasta la fecha sobre Google Discover y su funcionamiento.

De acuerdo con este objetivo principal, surgen las siguientes preguntas de investigación:

PI1) ¿Cuáles son los aspectos que caracterizan Google Discover desde el punto de vista del webmaster y del usuario final?

PI2) ¿Es posible determinar e identificar estrategias procedentes del SEO que ayuden a posicionar contenido web en Google Discover?

Para presentar respuestas a las tres preguntas anteriores, se realiza, en primer lugar, una revisión sistematizada de la literatura, más concretamente se aplica una *scoping review* (Arksey y O'Malley 2005); y en segundo lugar se realiza un análisis funcional y de interfaz de usuario de este servicio de Google para entender su funcionamiento.

En la siguiente sección, se presenta el marco teórico y se describen las metodologías empleadas. Seguidamente, se exponen los resultados de la *scoping review* y del análisis funcional de Google Discover y se presenta la discusión en torno a los datos obtenidos. Por último, se desarrollan las conclusiones, las limitaciones y las líneas de investigación futuras.

2. Marco teórico

La visibilidad web y la curación de contenidos están siendo objeto de estudio por parte de los académicos de manera creciente en los últimos años.

Sobre la visibilidad web se pueden encontrar tanto estudios cuantitativos como cualitativos, principalmente relacionados con estudios de caso, que analizan sectores empresariales, tipos y técnicas de posicionamiento y herramientas de auditoría SEO.

En lo que se refiere a sectores, destacan los estudios relacionados con la visibilidad web de: medios de comunicación (Giomelakis y Veglis, 2016; Lopezosa et al., 2020; Pedrosa y Morais, 2021); universidades (Gonzalez-Llinares et al., 2020; Vállez et al., 2022); bibliotecas (Onaifo y Rasmussen, 2013; Vállez y Ventura, 2021); o sitios web de turismo (Fernández-Cavia et al., 2013; Pedraza-Jiménez, 2018), entre otros.

Respecto a los tipos de SEO podemos encontrar, por poner algunos ejemplos, estudios destacados sobre SEO off page (Lopezosa et al., 2019), SEO semántico (Lopezosa et al., 2018), o Search Experience Optimization (Lopezosa et al., 2021).

Por último, se observa un predominio de estudios cuantitativos que utilizan herramientas de posicionamiento en buscadores. Principalmente las investigaciones operan con los servicios de SEMrush (Vállez, 2011), Sistrix (Vállez y Ventura, 2020), Ahrefs (García-Carretero, 2022) y Majestic (Orduña, 2021).

Por otro lado, la investigación sobre curación de contenidos se focaliza, entre otras temáticas, en la curación en medios digitales y en redes sociales y la curación algorítmica.

Sobre curación y contenidos digitales destacan entre otros los estudios de Dale (2014), Cui y Liu (2017) y Guallar et al., (2021) que proponen, entre otras cosas, estudios de caso y aplicación de protocolos de análisis.

En cuanto al uso de la curación de contenido en las redes sociales se encuentran estudios generales como los de López-Meri et al. (2017), Bruns (2018), Kümpel (2019) o Chagas (2018).

Como se ha mencionado anteriormente, otro de los grandes campos de investigación sobre esta disciplina es el de curación de contenidos algorítmica. En esta línea destacan los trabajos de Diakopoulos y Koliska (2017), los de Chakraborty et al. (2018) y los de Zubiaga (2019).

Como demuestra esta aproximación sobre visibilidad web y curación de contenidos, los campos estudiados son amplios; sin embargo, hemos detectado que existe un hueco de investigación sobre Google Discover, razón por la cual decidimos desarrollar esta investigación.

3. Metodología

A continuación, se describe la metodología mostrada para el desarrollo de esta investigación. En primer lugar, se explica de forma detallada cómo se aplican las revisiones sistematizadas exploratorias (en adelante *scoping review* por su nombre estandarizado en inglés), cómo se obtiene la batería de documentos y también el proceso seguido. Es importante recalcar, como se verá seguidamente, que esta *scoping review* no recoge artículos académicos sobre Google Discover ya que, como se indica en el marco teórico, no se ha encontrado ninguna investigación específica sobre este servicio de Google en las principales bases de datos académicas (Web Of Science y Scopus), razón por la cual, la *scoping review* aquí presentada utiliza principalmente literatura gris (Pons y Monistrol, 2017), concretamente, documentos divulgativos e informes sectoriales publicados en español.

Procede señalar que una *scoping review* equivale a una investigación en sí misma, en la cual la base de la evidencia son los documentos seleccionados. Adicionalmente, este método está respaldado metodológicamente por el trabajo teórico sobre revisiones sistematizadas de Codina (2020a).

Para el desarrollo de *scoping review* se adapta el Framework SALSA (Hart. 2008, Booth et al., 2012). SALSA es el acrónimo de:

Search (fase de búsqueda): Esta se resuelve mediante la propuesta de las ecuaciones de búsqueda acordes a la investigación, su aplicación en base de datos principalmente académicas,

pero también extensibles a buscadores comerciales (como por ejemplo Google) y la selección de las referencias considerando, para ello criterios de exclusión e inclusión (Codina, 2020b).

Appraisa**L** (fase de evaluación): Esta se desarrolla realizando una re-revisión del banco de documentos final atendiendo a criterios de inclusión, exclusión y verificación de la calidad del documento identificado (Codina, 2020b).

Synthesis (fase de síntesis): En esta se realiza y estructuran resúmenes tomando en consideración parámetros como resultados, limitaciones, tipo de estudio etc. (Codina, 2020c).

Analysis (fase de análisis): En esta se realiza la extracción de datos e informaciones de manera sistemática sobre los aspectos a recoger como resultados de la *scoping review* (Codina, 2020c).

En este sentido, para la elección del corpus de análisis de esta investigación se han seleccionado una serie de palabras clave y ecuaciones de búsqueda, que se aplican a Google.es (ya que se pretende conocer qué se ha publicado en España sobre Google Discover) para obtener así la conceptualmente llamada literatura gris (Pons y Monistrol, 2017) como opuesta a las publicaciones académicas.

Las ecuaciones de búsqueda aplicadas toman en consideración dos elementos esenciales de Google Discover, (1) el usuario final y (2) el webmaster. Dado que Google es especialmente eficaz para interpretar lenguaje natural, hemos usado frases en lugar de ecuaciones booleanas en su búsqueda simple. En este sentido, las frases de búsquedas son las siguientes:

Motor de búsqueda	*Consulta de búsqueda*
Google.es	"¿Qué es Google Discover?"
Google.es	"¿Cómo usar Google Discover?"
Google.es	"¿Cómo posicionar tu sitio web en Google Discover?"
Google.es	"¿Cómo optimizar tu contenido para Google Discover?"

Tabla I. Ecuaciones de búsqueda aplicadas a la scoping review

Por lo tanto, el idioma y los términos de la búsqueda han sido el español. A continuación, se muestran los criterios de exclusión tomados en cuenta:

- Documentos resultantes de falsas coordinaciones de frases y que el contenido no resuelva la intención de la ecuación de búsqueda.
- Documentos realizados por fuentes poco fiables, es decir, solo se recogerán documentos de empresas, profesionales y/o medios de comunicación vinculados con el SEO y otros sistemas de información que tengan una trayectoria reconocida dentro del sector.
- Documentos más allá de los 10 primeros ítems en el listado de resultados de Google.es.

Tras la aplicación de estos criterios de inclusión, se obtuvo una batería final de 20 documentos.

Para el análisis de estos documentos se plantea una tabla estructurada y sistemática que permite realizar una síntesis heterogénea.

Parámetro	*Descripción*
Tipo de documento	Noticia, informe, manual de uso, etc.
Enfoque	Usuario, Webmaster o ambos
Temática	Se identifica la temática atendiendo a (1) qué es Google Discover (2) cómo usar Google Discover (3) cómo optimizar contenido para Google Discover
Relevancia de la fuente	Empresa, persona o medio de comunicación que publica el documento y su posición en el sector
Principales ideas	Se incorpora un resumen de máximo 300 palabras

Tabla II. Tabla de síntesis aplicada a los 20 documentos para obtener un análisis estructurado

A continuación, se presenta, en primer lugar, una panorámica sobre Google Discover resultado de la *scoping review* y, en segundo lugar, una guía de las principales funciones de Google Discover basada en capturas de pantalla rotuladas y eventualmente con anotaciones.

Para poner en práctica Google Discover se realiza, por un lado, un análisis funcional como usuario de esta herramienta de Google, utilizando para ello un teléfono inteligente Android Xiaomi y, por otro lado, un testeo del Google Search Console (herramienta de Google para auditar la

visibilidad web de un sitio web) con datos cuantitativos específicos sobre Discover.

4. Resultados

Google Discover es una herramienta que permite a los usuarios recibir novedades sobre sus intereses sin tener que realizar una búsqueda en Google. El usuario puede configurar la herramienta a partir de cada noticia señalando si es (o no) de su interés y filtrando las fuentes que quiere consultar. (Google 2020b)

Google Discover muestra un contenido que se basa en aquello que los algoritmos de Google consideran que puede interesar más a cada usuario (González, 2022). Este algoritmo puede ser afinado en parte por el usuario, si este utiliza las opciones disponibles en cada noticia para confirmar su interés (o su rechazo).

De este modo, es una herramienta que permite al usuario ver noticias, desde el móvil, vinculadas con sus intereses. Como hemos señalado, dichas noticias se determinan, para cada usuario a partir de su perfil. Concretamente, de sus búsquedas y lectura de noticias anteriores (Infobae, 2020). La inteligencia artificial de Google utiliza, por un lado, las búsquedas registradas en el perfil del usuario y por otro lado, las acciones manuales que realiza el usuario cuando interactúa con las noticias (Linares, 2020),

Según declara Google, sus sistemas automatizados muestran en Discover contenido de sitios web que tienen buenos niveles de autoridad, conocimiento y fiabilidad. Por lo tanto, de acuerdo con Google, Discover muestra, filtra y omite el contenido que "no es adecuado o que podría confundir a los lectores" (Google, 2020a).

Entre los resultados que ofrece Discover se incluyen noticias, vídeos, novedades relacionadas con entretenimiento (como estrenos de cine), resultados deportivos, cotización de acciones e información sobre eventos culturales (como los nominados a un premio importante o carteles de festivales de música) así como información meteorológica, entre muchas otras cosas (Ramos, 2019; Andrés, 2019; Fernández, 2020).

El definitiva, este *feed* de Google destaca por ser una sección interactiva que se nutre de las interacciones que realizan los usuarios sobre las portadas del mismo (Linares, 2020).

Esto hace que Google Discover tenga un enorme potencial ya que el usuario se libera de hacer búsquedas para informarse, puesto que recibe lo que (se supone que) quiere sin tener que buscar (Coppola, 2020).

De hecho, como ya se ha indicado anteriormente, los contenidos que aparecen en Discover son seleccionados a partir del perfil del usuario, por lo que la posibilidad de que interactúe en ellos será más alta que con una búsqueda en el buscador general. De este modo, en realidad se trata de contenido curado semi algorítmicamente que no necesita responder a una necesidad inmediata. A partir de aquí, el uso de Discover se convierte en un hábito, ya que muchos usuarios lo utilizan rutinariamente para descubrir nuevos contenidos (Vicent, 2021).

Es así como Discover ha sido utilizado por 800 millones de personas en el mundo en 2018 (Marco, 2020), o como la revista Vogue ha recibido más tráfico de Google Discover, a finales de octubre de 2019, que del buscador general en el mismo periodo de tiempo (Coppola, 2021).

4.1. Google Discover y los webmasters

Discover representa una de las principales fuentes de tráfico actualmente para los medios digitales, Aunque todavía no es posible medirlo a la perfección (Soteras, 2021), se reconoce que a muchos sitios web (principalmente medios de comunicación), Discover les envía picos del 40%, 50% o incluso cerca del 70% de tráfico en momentos puntuales (González, 2020).

Ante esta circunstancia, resulta esencial que los webmasters traten de posicionar su contenido en este servicio de Google.

Para ello, algunos expertos consideran que es muy importante tener un número mínimo de visitas en la web durante los últimos 16 meses (Santiago, 2021), e incluso tener más de un 10 % de CTR —tasa de clics— (Soteras, 2021) para que un sitio sea candidato a aportar contenidos a Discover.

Además, existe una serie de requisitos que se pueden aplicar al contenido de un sitio web para tener mayores posibilidades de aparecer en el feed de Google y que tienen relación directa o en gran medida con el SEO. Por lo tanto, el posicionamiento en buscadores y el posicionamiento en Discover van muy de la mano (Natale, 2020; Vicent, 2021; Santiago, 2021)

A continuación, se presenta un decálogo que incluye los consejos de posicionamiento en Discover que gozan de mayor consenso, a partir de la batería de documentos analizados en la *scoping review*:

1. Publicar contenidos noticiosos basados en tendencias y actualidad, pero al mismo tiempo también contenidos atemporales conocidos a veces como *evergreen*, ya que en

Discover ambos son casi igual de importantes (Ramos, 2019; Natale, 2020; Soteras, 2021). Además, dichos contenidos deben ser de calidad (Santiago, 2021), útiles, con información relevante y que se alinee con los intereses del público objetivo (Vicent, 2021).

2. Introducir contenido continuo, atractivo y en diversos tipos de formatos (Santiago, 2021). Además, se recomienda crear tendencia (Misa, 2021), es decir hacer que un contenido se viralice y se convierta en noticioso.
3. Crear noticias frecuentes en torno a temas generales (Coppola, 2021).
4. Usar imágenes de alta calidad, grandes 1200 píxeles de largo (Google, 2020a; Natale, 2020; Seguí, 2020; Santiago, 2021; Soteras, 2021) y 200 píxeles de ancho (Marco, 2020), atractivas (Ordoñez, 2019) y únicas (Pecánek, 2020).
5. Optimizar el contenido para AMP (accelerated mobile pages) de Google, por lo tanto, tener un sitio web mobile friendly o responsive (Google, 2020b; Marco, 2020; Seguí, 2020, Antevenio, 2020). En definitiva, tener un sitio web que tenga muy buena velocidad de carga (Misa, 2021).
6. Cumplir con las políticas sobre contenido de Google News y Google Discover (Marco, 2020; Vicent, 2021; Misa, 2021; Satiago,2021; Antevenio, 2021). Y en definitiva, intentar siempre desarrollar las señales que favorecen el E-A-T (siglas de Experiencia, Autoridad y Confianza) del medio (Soteras, 2021).
7. Publicar páginas con títulos que describan la esencia del contenido (Ordoñez, 2019). Dichos titulares deben ser llamativos, pero evitando el clic cebo (también conocido como *clickbait*) (Soteras, 2021). Se recomienda añadir el mes y año en el título SEO (Ramos, 2019). Adicionalmente, se aconseja afinar la entradilla, como así se optimiza en SEO (Gonzalez, 2020; 2022) y publicar más videos (Coppola, 2020). Todas estas noticias deben situarse en la homepage —página de inicio— (Soteras, 2021).
8. Es necesario monitorizar el rendimiento de Discover para tener una idea de los resultados obtenidos y a partir de ahí plantear nuevas estrategias (Ordoñez, 2019).
9. Realizar una campaña de captación de tráfico como, por ejemplo: compartir en redes sociales, notificaciones de email, grupos de Telegram, WhatsApp y otras apps (Seguí, 2020).
10. Por último, es importante rellenar un formulario en el que se autoriza a Google a que utilice las imágenes del sitio web (Santiago, 2021).

4.2. Google Discover y los usuarios

Una vez vistos los puntos anteriores, a continuación, se presenta, una breve aproximación de la guía de utilización de Google Discover basada en capturas de pantalla rotuladas y, eventualmente, con anotaciones desde el punto de vista del usuario final y del webmaster de un sitio web que ha conseguido posicionar su contenido en la herramienta de Google.

En la imagen que se muestra a continuación se observan los resultados de búsqueda de Google.es tras aplicar la consulta “scire journal” en un teléfono inteligente. En la parte inferior izquierda se puede identificar un asterisco que es el símbolo de Google Discover.

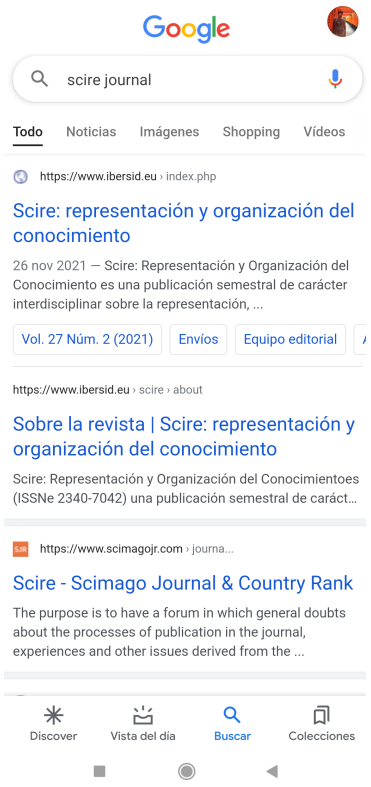

Figura 1. *Resultados de búsqueda en móvil de la consulta “scire journal” realizada en Google.es*

Al ingresar en Google Discover, se abre su feed en donde se muestra su contenido según las preferencias y perfil del usuario. En la imagen que se muestra a continuación (figura 2) se observa, el

resultado del tiempo en Barcelona, seguido de una información sobre posicionamiento en buscadores. En dicho resultado se observa, en primera instancia, la imagen de la noticia, seguida del título y una pequeña descripción de la misma. Por último, de izquierda a derecha se muestra (1) la fuente de la entrada —en este caso Search Engine Journal—, (2) la fecha de publicación —hace 6 horas—, (3) un icono de un corazón —interacción de me gusta de la noticia que ayuda a Google a entender los gustos del usuario y mostrarle contenido relacionado—, (4) el icono para compartir —permite enviar el contenido a través de redes sociales, mensajería, correo electrónico, etc.— y (5) los tres puntos que, al seleccionarlos, permiten realizar una serie de acciones sobre la noticia. Estas acciones se muestran en la figura 3 y posibilitan que el usuario informe a Google de que: no le interesa la noticia asociada; no le interesa el tema de la noticia que a su vez sirve para identificar la entidad que ha trabajado esta noticia (en este caso el posicionamiento en buscadores); no le interesa recibir en el feed noticias de la fuente/sitio web (en este caso, Search Engine Journal) o no le interesa recibir noticias en diferentes idiomas y/o el contenido es inadecuado. También ofrece la opción de interactuar y dejar comentarios sobre la noticia e incluso gestionar intereses (figura 4). Con todo ello Google puede enviar al feed del usuario contenido más afinado con sus gustos e intereses.

Figura 2. Interfaz principal de Google Discover

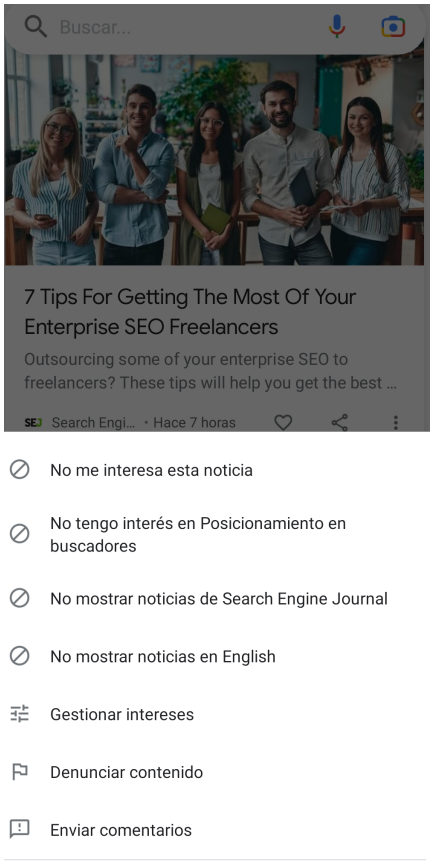

Figura 3. Interfaz de segundo nivel (icono de los tres puntos) de Google Discover

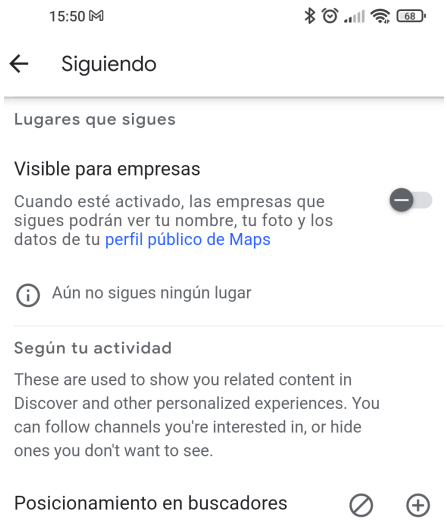

Figura 4. Interfaz de la categoría "gestionar intereses"

4.3. Google Discover, métricas y Search Console

Una vez analizadas las interfaces principales que componen Google Discover desde el punto de vista del usuario, en lo que sigue realizamos una pequeña descripción de la monitorización del

rendimiento de Discover a través de la herramienta de Google Search Console para tener la visión del webmaster.

Google Search Console es un servicio gratuito de Google que ayuda a monitorizar, mantener y solucionar problemas relacionados con la presencia de un sitio en los resultados de búsqueda de Google (Google, 2022) y Google Discover; siempre y cuando se consiga posicionar el contenido del sitio en el feed de Discover (Figura 5).

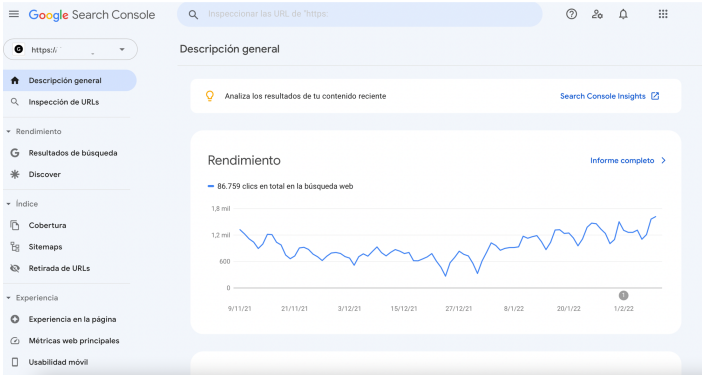

Figura 5. *Página principal de Google Search Console de un sitio web posicionado para Google Discover*

Al ingresar en Google Search console, se observa por defecto el rendimiento de un sitio web posicionado en el buscador de Google. Sin embargo, cuando el sitio web consigue posicionarse en Discover, se muestra en la parte de la izquierda la funcionalidad de Discover (esta es la que tiene el icono del asterisco).

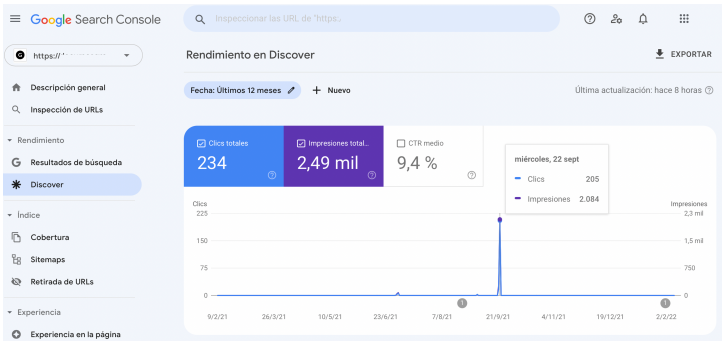

Figura 6. Interfaz del servicio de análisis de Google Discover que se encuentra en Google Search Console

Como muestra la figura 6, al acceder a Discover desde la Search Console se pueden ver tres indicadores (1) clics totales, (2) impresiones totales y (3) CTR medio.

Los Clics totales son el número de veces que se ha accedido a un contenido desde el Google Discover de los usuarios. Las impresiones totales, son las veces en las que ha aparecido este contenido en el feed de Google. Y el CTR (por sus siglas en ingles *Click Through Rate*) es el número de clics obtenidos por un enlace respecto a su número de impresiones.

Debajo de estos tres indicadores se visualiza la línea temporal en la que se muestra cuándo se consiguió posicionar un contenido específico en Google Discover y qué contenido fue el que se consiguió posicionar (figura 7).

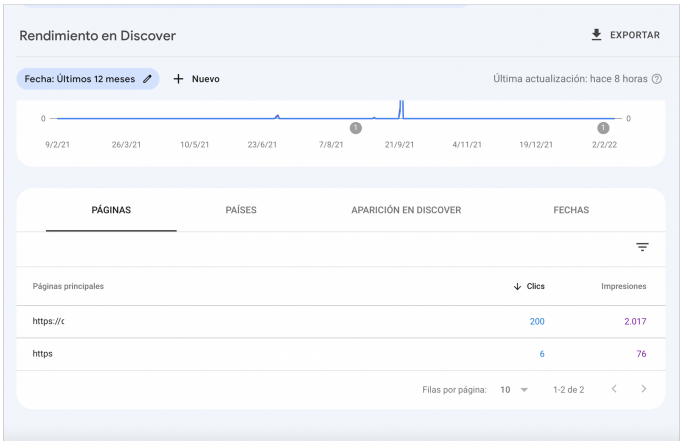

Figura 7. *Interfaz del servicio de análisis de Google Discover en Google Search Console*

Como se muestra en la figura 7, los resultados tras posicionar el contenido en Google Discover demuestran su potencial para mejorar la visibilidad web de un sitio web.

5. Discusión y conclusiones

A continuación, se examinan las preguntas de investigación para comprobar su grado de cumplimiento, así como se presentan también las discusiones y sugerencias para nuevas investigaciones.

PI1. Desde el punto de vista del usuario, Discover es una herramienta que ofrece contenido personalizado sin que tengamos que pasar por el buscador y hacer una consulta de búsqueda para informarnos. Desde el punto de vista del webmaster, aparecer en el feed de Discover se convierte en una necesidad estratégica para tener mayor visibilidad web y por tanto conseguir más lectores.

PI2. Esta investigación confirma que sí es posible determinar estrategias de SEO que ayuden a posicionar contenido web en Google Discover. La *scoping review* ha identificado ciertas prácticas y

consejos de expertos en el campo con un alto grado de consenso.

Si se toma en consideración el grado de consenso de las publicaciones obtenidas de la *scoping review*, destaca la coincidencia en el uso de fotografías de alta calidad, el contenido de interés que cumpla con las políticas de calidad de Google, y tener un sitio web responsive.

Este trabajo exploratorio es un punto de partida para abrir nuevas líneas de investigación dentro de la visibilidad web y la curación de contenidos algorítmica, tales como estudios cuantitativos (y estudios de caso analizando los resultados de Discover) hasta estudios cualitativos, (como los estudios de caso, entrevistas a expertos, observación participante en medios de comunicación centrados en posicionarse en Google Discover, etc.).

Asimismo, se considera especialmente relevante hacer un seguimiento de la evolución de las estrategias de posicionamiento en Discover desde la Academia; ya que se trata de un producto en constante actualización, que puede provocar una rápida obsolescencia de las pautas establecidas.

Por último, otras investigaciones futuras de interés pueden ser estudios comparativos sobre las ventajas y desventajas de Google Discover frente a alternativas similares, el análisis de las aportaciones de Google Discover respecto a otros sistemas, como los de alertas como Google Alerts, o los agregadores de feeds como Feedly o Netvibes, o el estudio de la integración de esta herramienta con otras de analítica web como Google Analytics.

Agradecimientos

Este trabajo forma parte del proyecto "Parámetros y estrategias para incrementar la relevancia de los medios y la comunicación digital en la sociedad: curación, visualización y visibilidad (CUVICOM)". PID2021-123579OB-I00 (MICINN), Ministerio de Ciencia e Innovación (España).

Actividad financiada por la Unión Europea-NextGenerationEU, Ministerio de Universidades y Plan de Recuperación, Transformación y Resiliencia, mediante convocatoria de la Universidad Pompeu Fabra (Barcelona).

Referencias

Andrés, Rubén (2019). Qué es Google Discover y cómo sacarle el máximo partido. // ComputerHoy. https://computerhoy.com/reportajes/tecnologia/google-discover-como-sacarle-maximo-partido-416731 (2022-02-09).

Antenvenio (2020). Google Discover: qué es y cómo te ayuda a incrementar el tráfico web. // Antevenio. https://www.antevenio.com/blog/2021/03/google-discover/ (2022-02-10).

Arksey, Hilary; O'Malley, Lisa (2005). Scoping Studies: Towards a Methodological Framework. // Int. J. Social Research Methodology 8:1. 19-32, https://doi.org/10.1080/1364557032000119616

Booth, Andrew; Papaionnou; Sutton, Anthea (2012). Systematic Approaches to a Successful Literature Review. London: Sage.

Bruns, Axel (2018). Gatewatching and news curation: Journalism, social media, and the public sphere. Peter Lang, 2018. https://doi.org/10.3726/b13293

Chagas, Luan (2018). Gatewatching and Collective Curation: Selecting Popular Radio Journalism Sources at Bandnews Rio FM. // Brazilian journalism research. 14:3.

Chakraborty, Abhijnan; Luqman, Mohammad; Satapathy, Sidhartha; Ganguly, Niloy (2018). Editorial Algorithms: Optimizing Recency, Relevance and Diversity for Automated News Curation. // Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 77–78. https://doi.org/10.1145/3184558.3186937

Codina, Lluís (2020a). Revisiones bibliográficas sistematizadas en Ciencias Humanas y Sociales. 1: Fundamentos. En: Lopezosa C, Díaz-Noci J, Codina L, editores Methodos Anuario de Métodos de Investigación en Comunicación Social, 1. Barcelona: Universitat Pompeu Fabra; 2020. p. 50-60. https://doi.org/10.31009/methodos.2020.i01.05

Codina, Lluís (2020b). Revisiones sistematizadas en Ciencias Humanas y Sociales. 2: Búsqueda y Evaluación. En: Lopezosa C, Díaz-Noci J, Codina L, editores Methodos Anuario de Métodos de Investigación en Comunicación Social, 1. Barcelona: Universitat Pompeu Fabra; 2020. p. 61-72. https://doi.org/10.31009/methodos.2020.i01.06

Codina, Lluís (2020c). Revisiones sistematizadas en Ciencias Humanas y Sociales. 3: Análisis y Síntesis de la información cualitativa. En: Lopezosa C, Díaz-Noci J, Codina L, editores Methodos Anuario de Métodos de Investigación en Comunicación Social, 1. Barcelona: Universitat Pompeu Fabra; 2020. p. 73-87. https://doi.org/10.31009/methodos.2020.i01.07

Coppola, María (2021). Qué es Google Discover y por qué te importará en 2021. // hubspot. https://blog.hubspot.es/marketing/que-es-google-discover (2022-02-07).

Cui, Xi; Liu, Yu (2017). How does online news curate linked sources? A content analysis of three online news media.// Journalism. 18:7, 852–870. https://doi.org/10.1177/1464884916663621

Dale, Stephen (2014). Content curation: The future of relevance. // Business Information Review. 31:4, 199–205. https://doi.org/10.1177/0266382114564267.

Diakopoulos, Nicholas; Koliska, Michael (2017). Algorithmic Transparency in the News Media. // Digital Journalism. 5:7, 809-828, DOI: 10.1080/21670811.2016.1208053

Fernández, Yúbal (2020). Google Discover: 18 trucos y consejos para dominar las recomendaciones en Android de la app de Google. // Xataka https://www.xataka.com/basics/google-discover-trucos-consejos-para-dominar-recomendaciones-app-google (2022-02-03).

Fernández-Cavia, Jose; Díaz-Luque, Pablo; Huertas, Assumpció; Rovira,Cristòfol; Pedraza-Jimenez, Rafael; Sicilia, María; Gómez, Lorena; Míguez, María (2013). Marcas de destino y evaluación de sitios web: una metodología de investigación. // Revista Latina de Comunicación Social. 68, 622-638.

García-Carretero, Lucía; Codina, Lluís; Díaz-Noci, Javier; Iglesias-García, Mar (2016). Herramientas e indicadores SEO: características y aplicación para análisis de cibermedios. // Profesional de la Información. 25:3, 497–504. https://doi.org/10.3145/epi.2016.may.19

Giomelakis, Dimitrios; Veglis, Andreas (2016). Investigating Search Engine Optimization Factors in Media Websites. // Digital Journalism. 4:3, 379-400, https://doi.org/10.1080/21670811.2015.1046992

Gonzalez-Llinares, Juan; Font-Julián, Cristina; Orduña-Malea, Enrique (2020). Universidades en Google: hacia un modelo de análisis multinivel del posicionamiento web académico. // Revista Española De Documentación Científica, 43:2, e260. https://doi.org/10.3989/redc.2020.2.1691

Google (2020a). Discover y tu sitio web // Google Developers https://developers.google.com/search/docs/advanced/mobile/google-discover (2022-02-01).

Google (2020b). Personalizar el contenido de Discover. // Suport.Google https://support.google.com/websearch/answer/2819496?hl=es&co=GENIE.Platform%3DAndroid (2022-02-02).

Google (2022). About Search Console. // Suport Google https://support.google.com/webmasters/answer/9128668?hl=en (2022-02-16).

González, David (2020). Cómo detectar noticias para aparecer en Google Discover. // Red de periodistas https://www.reddeperiodistas.com/como-titular-para-google-discover/ (2022-02-09).

González, David (2022). Cómo detectar noticias para aparecer en Google Discover. // Red de Periodistas https://www.reddeperiodistas.com/como-crear-noticias-para-aparecer-en-google-discover/ (2022-02-09).

Guallar, Javier; Anton, Laura; Pedraza-Jiménez, Rafael; Pérez-Montoro, Mario (2021). Curación de noticias en el correo electrónico: análisis de newsletters periodísticas españolas. // Revista Latina de comunicación social. 79, 47-64. https://doi.org/10.4185/RLCS-2020-1488

Hart, Chris (2008). Doing a Literature Review: Releasing the Social Science Research Imagination. London: Sage, 2008.

Infobae (2019). Cómo usar Google Discover para ver las noticias que te interesan en tu celular. // Infobae. https://www.infobae.com/america/tecno/2019/09/02/como-usar-google-discover-para-ver-las-noticias-que-te-interesan-en-tu-celular/ (2022-02-05)

Kümpel, Anna (2019). Getting Tagged, Getting Involved with News? A Mixed-Methods Investigation of the Effects and Motives of News-Related Tagging Activities on Social Network Sites.// Journal of Communication, Volume 69, Issue 4, 373–395, https://doi.org/10.1093/joc/jqz019

Linares, Iván (2020). Cómo activar las noticias de Google en Android: así puedes tener Discover aunque no esté accesible. XatakAndroid. https://www.xatakandroid.com/sistema-operativo/como-activar-noticias-google-android-asi-puedes-tener-discover-no-este-accesible (2022-02-10)

López-Meri, Amparo; Casero-Ripollés, Andreu (2017). Las estrategias de los periodistas para la construcción de marca personal en Twitter: posicionamiento, curación de contenidos, personalización y especialización.// Revista Mediterránea de Comunicación. 8:1, 59-73. https://www.doi.org/10.14198/MEDCOM2017.8.1.5

Lopezosa, Carlos; Codina, Lluís; Caldera-Serrano, Jorge (2018). SEO semántico: framework ISS para la optimización de sitios intensivos en contenidos. // Cuadernos de documentación multimedia. 29:97-123. https://doi.org/10.5209/CDMU.60607

Lopezosa, Carlos; Codina, Lluís; Gonzalo-Penela, Carlos (2019). SEO off page y construcción de enlaces: estrategias generales y transmisión de autoridad en cibermedios. // Profesional de la Información. 28:1. https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/66177

Lopezosa, Carlos; Codina, Lluís; Díaz-Noci, Javier; Ontalba, Jose (2020a). SEO y cibermedios: de la empresa a las aulas // Comunicar. 63, 65-75. https://doi.org/10.3916/C63- 2020-06

Lopezosa, Carlos; Iglesias-García, Mar; González-Díaz, Cristina; Codina, Lluís (2020b). Experiencia de búsqueda en cibermedios: análisis comparativo de diarios nativos digitales. // Revista Española de Documentación Científica. 43:1, e254. https://doi.org/10.3989/redc.2020.1.1677

Marco, Joan (2020). Google Discover: qué es y cómo sacarle partido. // Semrush https://es.semrush.com/blog/google-discover/ (2022-02-02).

Misa, Víctor (2021). Cómo aparecer en Discover con estos 8 simples pasos. // Víctor Misa. https://victor-misa.com/seo/salir-en-discover/ (2022-02-02).

Natale, Cecilia (2020). Google Discover: qué es y por qué es la clave para incrementar tu tráfico. // Inboundcycle https://www.inboundcycle.com/blog-de-inbound-marketing/google-discover-que-es (2022-02-03).

Onaifo, Daniel; Rasmussen, Diane (2013). Increasing libraries' content findability on the web with search engine optimization. // Library Hi Tech. 31:1, 87-108. https://doi.org/10.1108/07378831311303958

Ordoñez, Jordi (2019). ¿Cómo usar Google Discover para eCommerce? // Jordiob. https://jordiob.com/google-discover-ecommerce/ (2022-02-13).

Orduña-Malea, Enrique (2021). Dot-science top level domain: Academic websites or dumpsites? // Scientometrics. 126, 3565–3591 (2021). https://doi.org/10.1007/s11192-020-03832-8

Pedraza-Jiménez, Rafael (2018). Analysis of destination search in Google, IPBA, 90

Pedrosa, Leyberson; de Morais, Osvando (2021). Visibilidad web en buscadores. // Estudios sobre el Mensaje Periodístico, 27(2), 579-591. https://doi.org/10.5209/esmp.71291

Pecánek, Michal (2020). Google Discover: Cómo Posicionarte y Atraer Tráfico. // Ahrefs. https://ahrefs.com/blog/es/google-discover/ (2022-02-16).

Pons, Mariona; Monistrol, Olga (2017). Técnicas de generación de información en investigación cualitativa II. // Calderón C, Conde F, Fernández de Sanmamed MJ, Monistrol O, Pons M, Pujol E, Sáenz de Ormijana A. Curso de Introducción a la Investigación Cualitativa. Máster de Investigación en Atención Primaria. Barcelona: semFYC. Universitat Autònoma de Barcelona. Fundació Doctor Robert.

Ramos, Bruno (2019). Google Discover, Qué Es y Cómo Dispara tus Visitas con las Recomendaciones del Buscador. // Agencia SEO EU. https://agenciaseo.eu/google-discover/ (2022-02-02).

Santiago, Ignacio (2021). Qué es Google DISCOVER y cómo mejora tu SEO. // Ignacio Santiago https://ignaciosantiago.com/google-discover/ (2022-02-19).

Seguí, Eric (2020). Google Discover: qué es, cómo aparecer y mejorar tu visibilidad. // Publisuites. https://www.publisuites.com/blog/google-discover/ (2022-02-17).

Soteras, Clara (2021). Google Discover y… ¡gas a fondo!: Guía práctica con todo lo que debes saber. // Blogger3cero. https://blogger3cero.com/google-discover/ (2022-02-03).

Vállez, Mari (2011). Keyword Research: métodos y herramientas para identificar palabras clave.// BiD: textos universitaris de biblioteconomia i documentación. 27.

Vállez, Mari; Ventura, Anna (2020). Analysis of the SEO visibility of university libraries and how they impact the web visibility of their universities. // The Journal of Academic Librarianship. 46:4, 102171.

Vállez, Mari; Lopezosa, Carlos; Pedraza-Jiménez, Rafael (2022). A study of the Web visibility of the SDGs and the 2030 Agenda on university websites.// International Journal of Sustainability in Higher Education. 23:8, 41-59. https://doi.org/10.1108/IJSHE-09-2021-0361

Vicent, Jaume (2021). Qué es Google Discover y cómo funciona. // Trecebits. https://www.trecebits.com/2021/08/27/que-es-google-discover-y-como-afecta-al-seo/ (2022-02-06).

Zubiaga, Arkaitz (2019). Mining social media for newsgathering: A review. // Online Social Networks and Media 13:100049 DOI: 10.1016/j.osnem.2019.1.

Enviado: 2022-03-02. Segunda versión: 2022-07-17.
Aceptado: 2022-10-27.

Modelo y algoritmo para identificación y clasificación de afectaciones en un corpus de testimonios sobre desaparición forzada

Model and algorithm for identifying and classifying affectations in a corpus of testimonies on enforced disappearance

Ana-María TANGARIFE-PATIÑO (1), Sandra-Patricia ARENAS-GRISALES (1), José-David RUIZ ÁLVAREZ (2), Fabián BAENA-HENAO (1), Natalia MUÑOZ-OSORIO (1), Tatiana TIRADO-TAMAYO (1), Brayan-Alexánder MUÑOZ-BARRERA (2)

(1) Escuela Interamericana de Bibliotecología, Universidad de Antoquia. ana.tangarife@udea.edu.co, sandra.arenas@udea.edu.co, natalia.munozo@udea.edu.co, tatiana.tirado10367@udea.edu.co (2) Instituto de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Antioquia. josed.ruiz@udea.edu.co, balexander.munoz@udea.edu.co

Resumen

Se presenta, en este artículo, la metodología para la construcción de un modelo y un algoritmo para la identificación y clasificación de afectaciones (económicas, físicas, políticas, psicológicas y socioculturales) en un corpus de testimonios de familiares de víctimas de desaparición forzada. Se describe, inicialmente, el contexto y las características del fenómeno de desaparición forzada en Colombia y se enuncian las preguntas que guían la investigación en torno al procesamiento y la modelación de la información contenida en relación con el volumen de datos, con los modelos conceptuales para la clasificación y con las métricas para medir la eficiencia del algoritmo implementado. Luego, se detalla la metodología y la base conceptual sobre la que se inscribe el modelo, para discutir, posteriormente, las métricas y los criterios de valoración de estas, para comparar la eficacia entre la clasificación automática y la manual. Por último, se ofrecen reflexiones y líneas de trabajo futuro sobre el uso de técnicas de procesamiento de lenguaje natural en textos de memoria histórica.

Palabras clave: Algoritmos. Corpus. Desaparición forzada. Lingüística computacional. Procesamiento de lenguaje natural. Testimonios. Colombia.

Abstract

This article presents the methodology for the construction of a model and an algorithm for the identification and classification of affectations (economic, physical, political, psychological and sociocultural) in a corpus of testimonies of relatives of victims of enforced disappearance. Initially, the context and characteristics of the phenomenon of enforced disappearance in Colombia are described and the research questions guiding the processing and modeling of the information contained in relation to the volume of data, the conceptual models for the classification and the metrics to measure the efficiency of the implemented algorithm are stated. The methodology and the conceptual basis on which the model is based are then detailed, followed by a discussion of the metrics and their evaluation criteria to compare the efficiency between automatic and manual classification. Finally, reflections and lines of future work on the use of natural language processing techniques in historical memory texts are offered.

Keywords: Algorithms. Enforced disappearance. Computational linguistics. Natural language processing. Testimonies. Colombia.

1. Introducción

La clasificación y extracción de información en corpus de dominios especializados es una tarea de gran relevancia en el procesamiento de lenguaje natural. Distintas aproximaciones han propuesto métodos y técnicas para reconocimiento automático en análisis de contenido (Eriksson, 2007), (Aguado y otros, 2002), (Buendía, 2010). Se reconoce un amplio avance en muchas lenguas en donde se recogen aspectos semánticos del lenguaje en corpus como Wordnet.

Este artículo plantea una reflexión en un dominio concreto para atender a necesidades de clasificación y representación de información en el campo específico de las afectaciones de familiares de víctimas de un fenómeno violento.

La desaparición forzada es una de las formas de violencia más recurrentes en Colombia. El Observatorio de Memoria y Conflicto, del Centro Nacional de Memoria Histórica (CNMH), en agosto de 2021 tenía información de 80 674 personas desaparecidas (Observatorio de Memoria y Conflicto, CNMH, 2021). Por su parte, la Unidad de Búsqueda de Personas dadas por Desaparecidas[(2)] tiene datos sobre más de 120 000 personas desaparecidas (Ortiz Fonnegra, 2019).[(3)] Para Gonzalo Sánchez, exdirector del CNMH, la desaparición forzada es un proceso inverso al re-

velado de una fotografía: lentamente, va borrando la imagen de la víctima hasta hacerla invisible (citado en CNMH, 2016, p. 14). El autor del crimen intenta eliminar todo rastro, todo registro, se esfuerza por ocultar su existencia, establece un manto de duda e incertidumbre sobre la víctima y su familia.

La lucha de familiares de víctimas y organizaciones de derechos humanos ha logrado visibilizar el fenómeno y ha presionado para la creación de leyes que lo tipifiquen como delito.[4] Desde los años setenta, han formado comunidades afectivas y organizaciones sociales de víctimas para denunciar, visibilizar y exigir justicia. La agregación en comunidades y organizaciones es necesaria, por las características del crimen: impunidad de los responsables, riesgo de vida y estigmatización de miembros de la familia, e invisibilidad de las víctimas.

La desaparición forzada es definida por la "Convención internacional para la protección de todas las personas contra las desapariciones forzadas" (Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, 2006), en su artículo 2, como:

> El arresto, la detención, el secuestro o cualquier otra forma de privación de libertad que sean obra de agentes del Estado o por personas o grupos de personas que actúan con la autorización, el apoyo o la aquiescencia del Estado, seguida de la negativa a reconocer dicha privación de libertad o del ocultamiento de la suerte o el paradero de la persona desaparecida, sustrayendo a la protección de la ley.

Además, es considerada por el "Estatuto de Roma" como crimen de lesa humanidad, cuando cumple con características como ser generalizada y sistemática, es decir, cuando deja multiplicidad de víctimas, y cuando se ha llevado a cabo de manera frecuente y reiterada (Corte Penal Internacional, 1998).

La desaparición forzada es una agresión que implica múltiples vulneraciones a los derechos humanos de víctimas directas, sus familiares y comunidades de las cuales hacen parte. Entre estas vulneraciones se evidencian principalmente la violación al derecho a la libertad, a la vida, a la seguridad, a la integridad, entre otras, y en algunos casos se suman a este hecho otros delitos, como tortura y tratos crueles, inhumanos y degradantes. Por esto, y por ser un hecho que se extiende en el tiempo, es uno de los crímenes más atroces, pues genera un sufrimiento que solo puede cesar en el momento en que se obtiene información sobre las circunstancias en las que sucedió y el paradero de la persona desaparecida.

Según el CNMH (2016), al interior de las familias, el daño sufrido es diverso, dependiendo del lugar de la víctima en la familia y la comunidad. Si bien todos estos comparten la vulnerabilidad ante la violencia, algunas personas son aún más vulnerables, debido a factores como la pobreza, el desempleo o el analfabetismo, lo cual agrava las condiciones de inseguridad y la falta de acceso a instituciones de apoyo y protección.

Ante un delito que pretende borrar la existencia de un ser humano, un recurso para traer de vuelta al desaparecido es la palabra, el testimonio, la denuncia. Estos relatos permiten una aproximación a la experiencia y a las afectaciones que sufren los familiares y las comunidades que lo padecen.

Uno de los problemas que enfrentan hoy los investigadores que analizan este fenómeno en el marco del conflicto armado es cómo procesar ese volumen de información derivada de testimonios, entrevistas, escritos, videos, entre otros. En ese sentido, el objeto de estudio de esta investigación es el procesamiento, la modelación y la comunicación-difusión de ese cúmulo de testimonios sobre afectaciones en familiares de víctimas de desaparición forzada.

La investigación surge de preguntas relacionadas con el fenómeno y con la disposición de información sobre memoria; así como sobre el procesamiento y la modelación de esta información, de manera que puedan ofrecerse formas de comprensión y difusión desde los análisis textuales. Así pues, se busca responder a las preguntas: ¿cómo pueden ser analizados grandes volúmenes de información en relación con los testimonios de la desaparición forzada en Colombia? ¿Es posible crear herramientas automáticas para facilitar la labor de análisis y clasificación de esos testimonios? ¿Cuáles serían las métricas para evaluar la eficacia de un algoritmo de clasificación en comparación con lo que hace una máquina y lo que haría un analista?

En esta investigación se conforma un corpus de testimonios sobre afectaciones de los familiares de las víctimas de desaparición forzada en Colombia, con el objetivo de reconocer patrones para construir una base conceptual que permita el diseño de un algoritmo orientado a la clasificación automática, a partir de la identificación de afectaciones de la desaparición forzada mencionadas en testimonios. Si bien existen otros análisis de esta información concreta en términos de recolección de hechos, el aporte de este trabajo radica en el diseño de una herramienta para realizar análisis automático sobre un aspecto específico de estos testimonios abriendo la posibilidad de facilitar el análisis de amplios volúmenes de información sobre afectaciones a víctimas de violencia.

Este artículo presenta el proceso de construcción de dicho algoritmo y las métricas para evaluar su eficacia, haciendo hincapié en el ejercicio conceptual para la construcción de nodos de clasificación. Se presentan también los resultados que genera el algoritmo comparando sus resultados en relación con el análisis manual.

Para ello, inicialmente, se identifica el campo del procesamiento de lenguaje natural como punto de partida para analizar el corpus conformado. A continuación, se describe la metodología para analizar los testimonios y la construcción de un diccionario de afectaciones, que sirvió como base conceptual para el diseño de un modelo y de un algoritmo para la identificación y clasificación automática de las afectaciones. Después, se presenta el desarrollo del algoritmo y las métricas implementadas para medir la eficacia del mismo. Finalmente, se enuncian algunas reflexiones sobre las limitaciones, posibilidades y futuros desarrollos en la implementación del análisis de datos lingüísticos en corpus de memoria.

2. Procesamiento de lenguaje natural y analítica de datos

El corpus documental responde a un tipo de información que da cuenta de relatos y testimonios de familiares de víctimas en los cuales se aplicaron técnicas de procesamiento de lenguaje natural para la extracción de información y la relación existente entre distintos elementos.

Para ofrecer una plataforma de consulta sobre un corpus de memoria es necesario plantearse, por un lado la estructuración y anotación de los textos para permitir la consulta estructurada y, por otro lado, el desarrollo de herramientas de análisis conceptual que específicamente permitan analizar aspectos de los textos, y por ello nos centramos concretamente en las afectaciones.

El *procesamiento de lenguaje natural* vincula saberes de la lingüística, la analítica de datos y la informática, con el objeto de generar modelos computacionales que reproduzcan uno o más aspectos del lenguaje natural. Se plantea, igualmente, comprobar los modelos lingüísticos y las teorías, diseñando algoritmos y sistemas que puedan ser evaluados y comprendidos en máquinas computacionales.

La *lingüística computacional* está focalizada en el conjunto de procedimientos o métodos para el estudio del lenguaje, con miras a facilitar su exploración, de modo que pueda contribuir en análisis de textos leídos por máquina, a fin de leer, buscar y manipular información en donde la cantidad es relevante (McEnery & Hardie, 2012). Su campo de desarrollo se encuentra igualmente entre la oralidad y la textualidad, para lo cual requiere disponer de recursos lingüísticos, entre los que se cuentan corpus, diccionarios y gramáticas.

Este proyecto de investigación emplea tecnologías del lenguaje para la construcción de un corpus cuyo análisis sirvió como muestra para la construcción conceptual y la posterior modelación.

Los *corpus* son definidos como un conjunto estructurado de textos que constituyen una muestra lo más realista posible del uso lingüístico. Debe tener un diseño coherente, introducción de marcas en los textos que definan su estructura según estándares comúnmente aceptados, y documentación completa que permita conocer la procedencia y las características de los materiales.

El concepto *corpus de memoria* hace referencia al conjunto de recursos de información relacionados con violaciones a derechos humanos y a procesos de defensa y lucha por la verdad y la justicia en el marco de conflictos armados. Son documentos y objetos de información (tanto digitales como analógicos) que son recopilados o producidos por instituciones u organizaciones de investigación, atención o defensa de los derechos humanos. Para este caso, son recursos dispuestos en la web y enriquecidos con técnicas de procesamiento de lenguaje natural para recuperar contenido significativo de modo que puedan relacionarse elementos como actores, entidades, impactos, modalidades de victimización, entre otros. Estos recursos tienen diversas características formales y de contenido pues van desde documentos relacionados con la denuncia y el registro de hechos hasta testimonios textuales, audiovisuales, y producciones escritas a partir de talleres o ejercicios que son realizados con personas que son víctimas o familiares de víctimas del conflicto.

Como evidencia lingüística, el corpus recoge muestras del uso del lenguaje natural y permite realizar diversos tipos de estudios de su comportamiento. Las colecciones de datos deben conformarse a partir de textos producidos en situaciones reales de comunicación, sean orales o escritas, lo cual dará un carácter de fiabilidad, en tanto esas situaciones sean una muestra representativa de la lengua. “La lingüística de corpus es importante porque permite la construcción de elementos de estudio que luego pueden servir para la comprobación de teorías lingüísticas” (Duque, 2009, p. 28).

Los *corpus textuales* están conformados por colecciones que pueden ser de tipo literario, periodístico, científico. Recopilan información lingüística para el procesamiento de grandes cantidades textuales, que son utilizadas en distintos recursos y aplicaciones. A partir de los corpus, se pueden

obtener conclusiones relacionadas con un escritor, una época, una variedad lingüística, cambios lingüísticos, diferencias en la adquisición y el uso de la lengua según un grupo social, un género, un tema, etc. (McEnery & Hardie, 2012). También pueden conformarse, como en el caso de esta investigación, a partir de asuntos más específicos relacionados con las afectaciones que se expresan en testimonios a través del discurso.

Por su parte, la *analítica de datos* hace referencia al proceso de tomar un conjunto de datos agrupados bajo unos criterios para el estudio de un proceso o fenómeno natural, y el posterior procesamiento para identificar patrones, recurrencias o características que terminen dando una explicación descriptiva y predictiva del proceso o fenómeno estudiado. Típicamente, este término se utiliza cuando el conjunto de datos es muy grande y se necesita desarrollar herramientas para hacer un análisis asistido por sistemas informáticos. En este último caso también se utiliza de forma más precisa el término en inglés *big data* o *minería de datos*.

3. Categorías para analizar el dominio

El análisis de testimonios se plantea a partir de algunas categorías para encontrar otras dimensiones del fenómeno, más allá de datos relacionados con los hechos. Por esa razón se decide abordar los testimonios desde las afectaciones.

Para comprender la dimensión de la afectación producida por la desaparición forzada es necesario visibilizar la vulnerabilidad a la que sus familiares se ven expuestos y que se manifiesta en los relatos sobre la manera como el evento transformó sus vidas.

Para efectos de esta investigación, se entiende por *afectaciones* las consecuencias o efectos de la desaparición forzada que alteran, perjudican, cambian de manera negativa y abrupta el curso de vida de individuos, familias y comunidades. Se agruparon las afectaciones en: económicas, físicas, políticas, psicológicas y socioculturales. Esta definición de afectaciones se basó en trabajos sobre el tema (Castrillón Rivera *et al.*, 2007; Charry Lozano, 2016; CNMH, 2014, 2016; Durán Pérez *et al.*, 2004; Rebolledo y Rondón, 2010). A continuación, se describen cada una de ellas.

Afectaciones económicas: derivadas del deterioro de condiciones económicas y materiales en las cuales se sustenta la calidad de vida y el desarrollo de individuos, familias y comunidades. Estas condiciones pueden estar relacionadas con el hecho de que la víctima directa de la desaparición es quien proveía el sustento familiar, o con la pérdida de propiedades como la vivienda, tierras, fuentes de ingresos como negocios propios o el empleo. Sobre estas afectaciones, es pertinente mencionar que pueden ser producto de hechos asociados a la desaparición forzada en un contexto específico de violencia, que dé cuenta de intereses de actores armados, y que se caracterizan por procesos de desplazamiento forzado, donde familias y comunidades se ven obligadas a abandonar sus tierras y bienes por amenazas, o por temor a ser objeto de nuevas agresiones.

Afectaciones físicas: se relacionan con el deterioro de las condiciones de salud física y se pueden expresar en la aparición o agravamiento de enfermedades. Si bien las víctimas manifiestan un deterioro de la salud física en general y un decaimiento asociado con la ocurrencia del hecho victimizante, se puede hablar de algunas enfermedades comunes para los casos específicos de desaparición forzada, como alteraciones del sueño, cáncer, enfermedades cardiovasculares, desnutrición, dolores, entre otras (CNMH, 2016).

Afectaciones políticas: son aquellas que debilitan, desestructuran o eliminan identidades políticas, expresiones de movimientos o procesos políticos que se identifican como diferentes o contrarios a los intereses del perpetrador. Pueden estar relacionadas con prácticas dirigidas a desincentivar la participación político-electoral, a acallar la oposición política, a imponer el silenciamiento de voces disidentes y obstaculizar la organización comunitaria.

Afectaciones psicológicas: efectos de la desaparición forzada relacionados con el deterioro de la salud psíquica y emocional, que impiden el desenvolvimiento adecuado de los individuos en los distintos ámbitos de la vida. Estas afectaciones pueden expresarse en el desencadenamiento de trastornos mentales o en la aparición de síntomas, emociones, comportamientos y sentimientos que generan sufrimiento. En el campo de las emociones, son frecuentes las expresiones como angustia, culpa, impotencia, incertidumbre, miedo, ira, y la persistencia de un duelo alterado que no se concreta por la inexistencia de un cuerpo o la esperanza de encontrarlo con vida.

Afectaciones socioculturales: hacen referencia a los efectos de las desapariciones sobre el tejido social, el sistema de creencias y tradiciones de las comunidades, y los elementos que se desprenden de estos, como las formas de organización, la cosmovisión, los rituales y, en general, todo tipo de acuerdos tradicionales construidos sobre la base de la identidad cultural. Entre las afectaciones socioculturales más frecuentes en los testimonios se encuentra la estigmatización, entendida como la valoración que se hace de una

persona o colectivo con la que se busca identificar y resaltar su condición de diferente no aceptado, “a partir de un rasgo desacreditador que se contrapone al estereotipo acerca de cómo deben funcionar los hechos de la realidad” (Castrillón Rivera *et al.*, 2012, p. 45). Para el caso de víctimas de desaparición forzada, esta afectación se origina en discursos que circulan socialmente, que las asocian con conductas y comportamientos reprochables, o que las acusan de hacer parte de grupos armados, o tener algún tipo de vínculos, con lo que se busca justificar su victimización. También es frecuente la referencia al *descrédito*, es decir, el no reconocimiento del daño producido, asociado a la indiferencia, la estigmatización y falta de atención del Estado.

La memoria es construida a partir de los relatos y por tanto está en el mundo del lenguaje. Estos testimonios han sido generados en espacios de conversación dentro de investigaciones, atención psicosocial, instancias judiciales, entre otros escenarios. Son relatos que cuentan, además de las circunstancias en las que se dan los hechos para aportar en el proceso de búsqueda, la manera como éstos han marcado las vidas de familiares y comunidades, porque cuando se da la desaparición hablamos de varios niveles en los que se están afectando las vidas de las personas, que van desde lo emocional con relación a sentimientos de vacío, angustia, frustración y rabia, pasando por el cuerpo y por lo material, pues muchas familias, luego de la desaparición de un ser querido, se ven enfrentadas a cambiar su vida, a desplazarse, a vivir con miedo y temor del señalamiento.

La decisión de abordar el asunto de la clasificación de textos desde las afectaciones está relacionada con el interés de construir un léxico sobre emociones presentes en testimonios de víctimas de conflicto armado que sirva como herramienta para tareas de clasificación y análisis por parte de investigadores y como aporte a formas de difusión de la memoria. Y el trabajo en un corpus pequeño con una sola modalidad de victimización brinda las pistas para la construcción de corpus más amplios sobre violación de los derechos humanos en el marco del conflicto armado.

Aquí entran en consideración elementos para pensar el campo del trabajo con afectaciones, específicamente emocionales. El análisis de sentimientos, según Liu (2017), no es solo un problema de clasificación y por tanto estudiar la subjetividad presente en textos puede ayudar a reconocer variaciones y construir herramientas más sofisticadas que diferencian aspectos como el afecto, la emoción, el humor. El tratamiento de los sentimientos/emociones en el lenguaje incluye la preocupación por la subjetividad, los sentimientos y las creencias (Lui, 2015). Este autor también diferencia la orientación, intensidad y calificación del sentimiento y aporta una relación de las emociones básicas planteadas desde distintas teorías.

El léxico o diccionario emocional puede luego ser usado en sistemas automáticos de clasificación con el propósito de ampliar las herramientas léxicas para el español en un dominio específico: la comprensión de las emociones representadas en los discursos o testimonios de víctimas de conflicto armado.

Joshi et. al. (2017) definen el *Sentiment resource* (lexicon) como un repositorio de unidades textuales marcadas con etiquetas que representan un sentimiento y que está compuesto por una unidad textual y etiquetas correspondientes.

4. Método

El método implementado en esta investigación es mixto. Toma elementos de la lingüística de corpus y del procesamiento de lenguaje natural, para analizar un conjunto de testimonios y construir una base conceptual, con el objetivo de proponer un modelo de clasificación automática. Se usan también métodos cuantitativos, para establecer las métricas de evaluación del algoritmo.

Para el desarrollo de esta investigación, se hizo una descripción de las afectaciones de los familiares, enunciadas en testimonios como consecuencia de la desaparición forzada. Los 391 testimonios del corpus fueron recopilados de informes de memoria o investigaciones académicas realizadas por organizaciones oficiales, investigadores y asociaciones de víctimas. Se encuentran recursos de prensa, informes de instituciones jurídicas de defensa de derechos humanos, colectivos y organizaciones sociales y de víctimas, así como información producida en procesos académicos e investigativos.

El principal criterio de selección de testimonios para el corpus consistió en que el texto diera cuenta de alguna de las afectaciones definidas. Generalmente, los testimonios refieren información sobre los hechos y las circunstancias de desaparición de personas o de aspectos relacionados con el proceso de búsqueda de las víctimas. Por tal motivo, se tuvieron en cuenta sólo aquellos fragmentos que hacían referencia a algún tipo de afectación, considerando, además, elementos que aportaran contexto dentro del mismo fragmento. En este punto interesaba especialmente reconocer las distintas formas en las que en el texto se manifiestan afectaciones, en

tal sentido se reconocía tanto la palabra que explícitamente aparecía como algunas figuras retóricas (metáfora, ironía, hipérbaton, eufemismo) para referir afectaciones o daños.

Por otra parte, como criterio de extensión mínima de los fragmentos, se tomó como base que el mismo se constituyera en una expresión cuyo conjunto de elementos diera cuenta del sentido en el marco de las afectaciones. De esta manera, la forma más simple de un fragmento podía corresponder a una única oración de tipo declarativo, desiderativo, dubitativo o interrogativo siempre y cuando representaran situaciones y emociones relacionadas con las afectaciones definidas. Se estableció este criterio debido a las dificultades para acceder a testimonios referidos al delito de desaparición forzada y sus afectaciones, ya fuera por causas de no accesibilidad por protección de la información o porque los testimonios disponibles, como ya se mencionó, referían principalmente a hechos y circunstancias de la desaparición de las personas y aspectos relacionados con su búsqueda. Es así como la variación de la cantidad de palabras de los testimonios del corpus se encuentra en un rango de 10 a 4981 y en total el corpus se compone de 171 364 palabras.

Una vez recopilado el corpus, se realiza la etapa de modelación, donde se hace anotación manual, que consiste en la lectura de testimonios y la selección de los apartados o fragmentos y la anotación de las afectaciones identificadas en dichos testimonios, con ayuda del *Software Nvivo*, herramienta para análisis cualitativo que permite codificar texto y explorar los datos para visualización y descubrimiento de temas y asociación entre datos. Este proceso fue realizado por cuatro integrantes del equipo, a partir del análisis de contenido, se procedió a un etiquetado exploratorio, sobre el 30% del corpus equivalente a 116 testimonios. Previa creación de los nodos y subnodos en Nvivo (codificación deductiva), cada etiquetadora debía realizar el proceso de anotación a 29 testimonios. Adicionalmente, y atendiendo a la técnica de codificación inductiva, también se podían crear nodos y subnodos según se complementaran nuevos niveles de la categorización. Con estas condiciones establecidas se procedió a reclasificar conforme a la nueva distribución de nodos y al ajuste de las anotaciones de párrafos, los cuáles se fragmentaron a partir de la extracción de las oraciones de interés. Estas oraciones o expresiones sirvieron para conformar el diccionario de afectaciones al igual que los nodos y subnodos bajo los cuales fueron clasificadas.

En cuanto a la distribución del corpus, este 30% fue destinado para la etapa de entrenamiento del algoritmo y un 5% (20 testimonios) para testeo. La distribución se realizó de manera aleatoria bajo el enfoque *Train-Test Split* (Browniee, 2020). De los testimonios de entrenamiento, 20 fueron analizados manualmente de forma paralela al algoritmo para contrastar resultados y medir la eficacia de la clasificación. Finalmente, una vez probado el algoritmo y sus métricas este fue aplicado a todo el corpus.

Por su parte, la elaboración del diccionario se desarrolló en función del trabajo conceptual y de los resultados de anotación. Para el trabajo conceptual se tomaron como principales referentes al CNMH (2016 y 2014) y a Pérez Sales, P., et.al. (1998). A partir de la elaboración de fichas analíticas se descubrieron y caracterizaron a través de definiciones, términos asociados y, en ocasiones, algunas expresiones, las categorías de *daños*, *impactos* y *afectaciones*.

El diccionario fue usado posteriormente en las fases de entrenamiento y testeo en la construcción del algoritmo, el cual fue desarrollado en Python 3.

Afectaciones	*Sinónimos-palabras*	*Expresiones*
Económicas	Desempleo, Desplazamiento, Escasez, Necesidad, Pérdida, Pobreza...	“La pobreza aumentó”, “No pude volver a trabajar”,, “Nos tuvimos que desplazar”...
Físicas	Achaque, Afección, Decaimiento, Dolencia, Enfermedad, Insomnio, Malestar, Maluquera...	“Comencé a enfermarme”, “He sufrido muchas enfermedades”, “Murió de pena moral”...
Políticas	Clandestinidad, Desconfianza, Impunidad, Inseguridad, Persecución, Represión, Señalamientos...	“Desaparecidos de la Unión Patriótica”, “La organización se debilitó”, “Se generaron desconfianzas entre las personas”...
Psicológicas	Ansiedad, Sufrimiento, Zozobra…	“Casi no me puedo acostumbrar a ver las fotografías”, “El dolor nunca pasa”, “Ella se decayó mucho”...
Socioculturales	Indiferencia, No reconocimiento, Revictimización	“Allá nadie ve nada”, “No recibí ninguna ayuda”, “Por algo les pasó lo que les pasó”...

Tabla I. Niveles de la estructura del diccionario

La *estructura del diccionario* se presenta en tres niveles. En el primero se encuentran las afectaciones según el aspecto de la vida en el que inciden de manera directa. Luego, en el segundo nivel, se ubican sinónimos que representan o están asociadas a cada una de las afectaciones. Y, finalmente, en el tercer nivel, están las expresiones (Tabla I).

5. Modelo, algoritmo y métricas

Para evaluar la clasificación de datos se emplean distintas métricas las cuales pueden ser: métricas para evaluar la capacidad de generalización del clasificador entrenado, métricas de evaluación que determinan cuál es el mejor clasificador de distintos modelos entrenados; y las métricas de evaluación que permiten elegir una solución óptima para clasificar (Hossin & Sulaiman, 2015).

En este trabajo se define como métrica de evaluación la eficacia para reconocer términos y expresiones relacionados con las afectaciones en comparación con la selección y clasificación que se haría manualmente.

A partir del diccionario construido con las palabras y expresiones clasificadas según la afectación, se construyó un modelo para la identificación de las afectaciones en un testimonio. El *modelo* hace referencia a la construcción conceptual por medio de la cual se da cuenta de las afectaciones como fenómeno ligado a las desapariciones forzadas y que se evidencia en las palabras utilizadas en los testimonios.

La identificación de afectaciones en un testimonio de desaparición forzada normalmente implica la lectura manual por parte de un experto en el área, capaz de identificarlas. Este trabajo, si bien es completamente realizable para un número limitado de testimonios, cuando el conjunto de testimonios aumenta, se convierte en una labor desafiante. Por tanto, se vuelve importante desarrollar herramientas de lectura e identificación automática de afectaciones. El *algoritmo* se constituye, entonces, en la serie de pasos automáticos mediante los cuales es posible identificar afectaciones que aparecen en uno o más testimonios.

Si bien los tipos de afectaciones han sido ya especificadas en la sección anterior a nivel textual, se escoge una unidad de medida que indique la presencia u ocurrencia de una afectación en un testimonio. Para este proceso, se optó por dos unidades de análisis, para combinarlas en el modelo.

- *Unidad 1: la palabra*. La aparición de una palabra en un testimonio marca la ocurrencia de una o más afectaciones. Es decir, se puede atribuir, a una palabra, una o varias afectaciones, y el hecho de encontrar esta palabra en un testimonio sería indicador suficiente de que en el testimonio acaece(n) la(s) afectación(es) relacionada(s). Es, por tanto, indispensable construir un conjunto de palabras asociadas a cada una de las afectaciones, o en otros términos, bajo esta definición de unidad, una afectación vendría determinada por un conjunto de palabras que marcan su existencia en un texto.
- *Unidad 2: la expresión*. Se reconoce que no solo en palabras está el contenido de las afectaciones, sino también en expresiones completas, frases que determinan una o varias afectaciones. Estas expresiones se tratan de forma rígida, es decir, se establece que dichas expresiones ocurren en el lenguaje bajo una única forma y, por tanto, para identificar la afectación o afectaciones asociadas, la expresión debe aparecer íntegramente en el texto.

Utilizamos ambas unidades para contar la cantidad de veces que ocurre cada afectación en un testimonio. Si aparece una palabra o una expresión de una afectación determinada, se cuenta por una ocurrencia. Con este método, se suma la cantidad de veces que aparece cada afectación y se establecen así varias métricas para categorizar los testimonios.

Todo el texto es procesado en Python 3 con la ayuda del paquete NLTK (Natural Language Toolkit), para hacer un preprocesamiento que incluye quitar las palabras irrelevantes, símbolos y signos de puntuación, así como la lematización.

Por otro lado, las *métricas* son la medida que se utiliza para poder categorizar cada testimonio de acuerdo con las afectaciones identificadas. Con el conteo de ocurrencias de las afectaciones por testimonio, se definieron dos métricas: la primera, para asignar al testimonio una única afectación como la de mayor ocurrencia. Si dos afectaciones ocurren la misma cantidad de veces, entonces se decide aleatoriamente entre ambas. Y la segunda, en la que se dividen las ocurrencias de cada afectación por el total de ocurrencias de todas las afectaciones, para obtener porcentajes de afectación por cada testimonio. Estas dos métricas permiten estudios diferenciales y un nivel de detalle diferente para entender el alcance y la precisión del algoritmo a la hora de encontrar afectaciones en testimonios.

Para medir la eficacia del algoritmo se hizo el testeo, separando un conjunto de testimonios que no habían sido etiquetados anteriormente y a los cuales se les identificaron las afectaciones manualmente, y se comparó con la identificación realizada por el algoritmo.

Utilizando la métrica 1, se pudo establecer que el algoritmo identifica la misma afectación predominante que la lectura manual en 56,25 ± 28,57 %; y si se agrupan en categorías de afectaciones, la eficiencia sube al 78,57 ± 21,43 % de las veces. Es importante precisar dos aspectos intrínsecos

al problema, en primer lugar la limitación estadística en *términos* del número de testimonios, la cantidad de palabras de los mismos y la población de afectaciones en función de los testimonios. Por lo tanto, se pueden esperar desviaciones estándar altas debido a la estadística del corpus y adicionalmente se entiende también que al agrupar los testimonios en categorías de afectaciones mejore la identificación puesto que se pueblan mejor cada categoría con la estadística limitada, a la vez que se pueblan también más ricamente los diccionarios por categoría de afectación. Finalmente, entiéndase las incertidumbres en este trabajo como intervalos de confianza máximos y no como mínimos dado que preferimos establecer un intervalo mayor pero con una mayor probabilidad de contener los valores verdaderos.

Ahora bien, con la métrica 2 se define una distancia promedio entre las clasificaciones para un mismo testimonio, de la siguiente forma:

$$D_T = \sum_{i=1}^{n} |A_i^M - A_i^A|$$

donde D_T es la distancia entre las clasificaciones A_i^M, el porcentaje de ocurrencia de la afectación *i* atribuido por la clasificación A_i^A, por la clasificación hecha por el algoritmo. Finalmente, *n* corresponde al total de afectaciones.

Mientras más pequeña sea esta distancia, quiere decir que ambas clasificaciones son más similares o se diferencian menos. La distancia media encontrada es 0,076. Dado que la definición de esta métrica se hace en función de las características del problema, resulta tremendamente difícil definir un punto de referencia para esta distancia sin incurrir en sesgos del problema mismo, por tanto los valores deben relativizarse al contexto de esta investigación.

Adicionalmente, se define una distancia entre todos los testimonios para cada afectación, con el fin de medir cómo el algoritmo puede reconocer cada una de las afectaciones. Esta *distancia por afectación* se define de la siguiente forma:

$$D_A = \frac{1}{m} \sum_{i=1}^{m} |A_i^M - A_i^A|$$

Aquí, la A_i^A es el promedio de porcentajes de todos los testimonios manuales para la afectación *i*, y A_i^M, el promedio de todos los testimonios del algoritmo para la afectación *i*. Además, *m* es el número total de testimonios.

La diferencia principal entre estas dos distancias es que la primera mide la similitud entre la clasificación manual y la del algoritmo en un solo testimonio, dando así una medida de la cercanía de ambas clasificaciones teniendo en cuenta todas las afectaciones. La segunda distancia establece una medida de similitud entre las clasificaciones teniendo en cuenta todos los testimonios, para tratar de disminuir el sesgo que podríamos tener por la variabilidad de los elementos del corpus.

Afectación		D_A ± *0,066*
Económicas		0,092
Físicas		0,039
Políticas		0,004
Psicológicas		0,128
	Angustia	0,118
	Culpa	0,009
	Impotencia	0,131
	Incertidumbre	0,094
	Miedo	0,023
	Ira	0,003
Socioculturales		0,041
	Estigmatización	0,028
	Descrédito	0,009

Tabla II. Distancia por afectación

En la Tabla II se muestran las distancias por afectación de acuerdo con lo definido. Mientras más pequeña la distancia, quiere decir que el algoritmo identifica la afectación en cuestión de forma más similar al procedimiento manual.

Este proceso de validación demuestra que el algoritmo y el modelo desarrollado son aceptables para clasificar testimonios en categorías de afectaciones, y moderadamente aceptable para identificar afectaciones particulares.

Utilizando la métrica 1, se puede concluir que, basados en la clasificación de una única afectación por testimonio, el algoritmo es eficaz y utilizable para grandes bases de datos, que pueden dar una confianza moderada en su clasificación. Por otro lado, utilizando la métrica 2, se puede ver que la diferencia en las clasificaciones es baja y se espera que el espectro de afectaciones identificadas para un testimonio no se aparte de forma significativa de la clasificación manual.

Por otra parte, este mismo conjunto de testimonios utilizados para la validación sirvieron para medir una incertidumbre sobre las asignaciones de afectaciones encontradas por el algoritmo. Para esto, se definió un valor *de identificación correcto por afectación*, denotado por ϵ_i^{ID}, como el número de veces en que fue identificada correc-

tamente la afectación, dividido por el total de veces que ocurre dicha afectación en relación con la clasificación manual. Y también se definió un valor *de identificación erróneo por afectación*, denotado por ϵ_i^{ER}, como la cantidad de veces que se asignó la afectación, pero que la afectación identificada manualmente era distinta, y dividido por el total de veces que ocurren las demás afectaciones.

Afectación	ϵ_i^{ID}	ϵ_i^{ER}	σ_i
Económicas	0,308	0,022	0,015
Físicas	0,286	0,005	0,004
Políticas	0,000	0,000	-
Psicológicas	0,659	0,368	0,126
Angustia	0,927	0,167	0,012
Culpa	0,000	0,000	-
Impotencia	0,333	0,067	0,045
Incertidumbre	0,158	0,028	0,023
Miedo	0,000	0,000	-
Ira	0,000	0,000	-
Socioculturales	0,200	0,005	0,004
Estigmatización	0,000	0,005	-
Descrédito	0,000	0,000	-
Promedio	0,410	0,075	0,03

Tabla III. Medición de eficiencia e incertidumbre por afectación

Se definió la *incertidumbre del algoritmo por afectación* como

$$\llbracket \sigma\_i = \epsilon \rrbracket \_i\hat{}ER \quad \llbracket (1-\epsilon) \rrbracket \_i\hat{}ID)$$

Dado que se requerirían demasiados testimonios para poder determinar correctamente estas cantidades y debido a que la estadística es limitada, se define una incertidumbre promedio entre las que se pudo medir y se le asigna por igual a todas las afectaciones. Adicionalmente, en los casos en que no se logra medir una eficacia debido a las limitaciones estadísticas, no se considera posible medir una incertidumbre asociada. Si bien es difícil establecer qué tan justa es esta estimación al tomar el promedio, se encuentra en un punto medio, que se puede esperar que no subestime o sobreestime por mucho los valores reales de la incertidumbre. Los resultados para las eficiencias e incertidumbre discutidas están en la Tabla III. En este caso, y de acuerdo con las decisiones explicadas en la metodología descrita, no aventuramos valores aceptables para la incertidumbre, dadas las limitaciones estadísticas del problema.

Una vez definidas las métricas y la eficacia, se aplica extensivamente el algoritmo sobre el conjunto completo de los 391 testimonios, de los cuales se muestran los valores sintetizados en las Figuras 1 y 2.

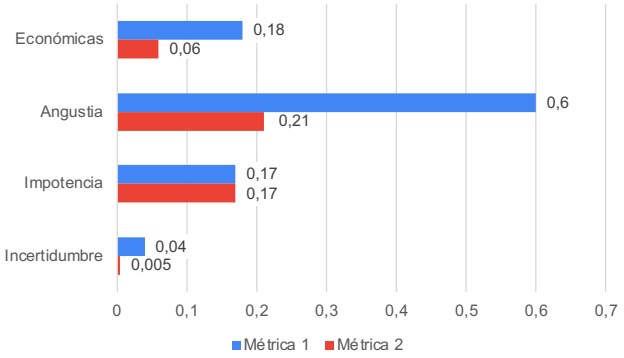

Figura 1. Comparativo de afectaciones utilizando las métricas 1 y 2

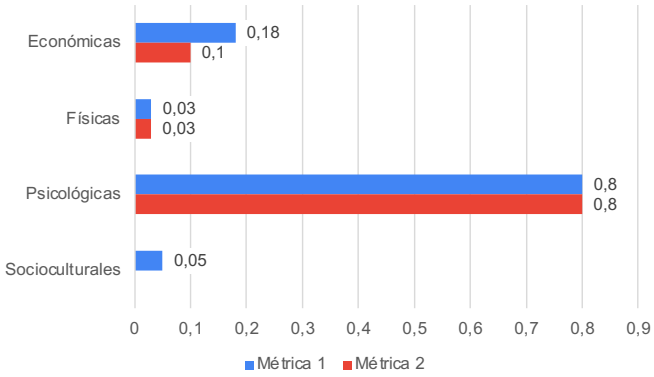

Figura 2. Categorías y su respectivo error en el total del corpus

De acuerdo con los resultados de nuestro modelo y algoritmo, y según se observa en las Figuras 1 y 2, las afectaciones psicológicas son fuertemente dominantes, seguidas por las afectaciones de orden económico.

En las Figuras 1 y 2 podemos ver que las afectaciones psicológicas y económicas tienden a ocurrir de manera simultánea en los testimonios; en otras palabras, típicamente los testimonios que reflejan afectaciones psicológicas también reportan afectaciones económicas. De las Figuras 1 y 2 podemos inferir que, en una muestra de testimonios, lo más probable sería encontrar afectaciones psicológicas y económicas.

Adicionalmente, de la Figura 1 identificamos que la mayoría de testimonios expresan angustia e impotencia ante las desapariciones forzadas. Estas dos afectaciones aparecen correlacionadas,

dándose de forma simultánea con facilidad en los testimonios analizados. Además notamos cómo, dentro de las afectaciones psicológicas, la angustia y la impotencia son dominantes sobre el resto de afectaciones encontradas.

6. Plataforma de divulgación

Con el fin de presentar los resultados y con la intención de visibilizar las afectaciones presentes en los testimonios se diseñó un sitio web[(5)] que recoge los procedimientos ejecutados en el proyecto, con el fin de disponer el algoritmo en su fase demo para potenciales usuarios, como investigadores en áreas de memoria y desaparición forzada, y con intereses en procesamiento de lenguaje en corpus de testimonios de víctimas, organizaciones defensoras de derechos humanos, víctimas relacionadas con desaparición forzada, y visitantes en general.

Se consideró además la posibilidad de validar testimonios, tomando como base los diccionarios y el algoritmo referenciados anteriormente. Así mismo, es una alternativa que permite aportar nuevos testimonios al proyecto, que puedan ser integrados al corpus para mejorar la precisión de los resultados del algoritmo.

7. Discusión y conclusión

Aquí se plantean algunas cuestiones en relación con el método y con la evaluación de los resultados obtenidos con el uso del algoritmo, en contraste de lo que sería una lectura manual, y con la perspectiva de ofrecer herramientas que les permitan a investigadores sobre memoria del conflicto reconocer afectaciones y determinar una clasificación de textos a partir de ellas, tanto para el fenómeno de la desaparición forzada como para otros relacionados. Estas reflexiones están planteadas en la mención de algunas limitaciones intrínsecas al trabajo con analítica de datos, la valoración de resultados del modelo para investigación sobre memoria y algunas líneas de trabajo futuro en la lingüística computacional aplicada a corpus de memoria.

En cuanto a las limitaciones, se encuentra la necesidad de incluir más volumen de testimonios, que si bien es un asunto intrínseco al análisis de datos, contar con más testimonios permitiría evidenciar nuevas categorías en los ámbitos individual, familiar o colectivo, o en relación con actores o formas de victimización.

En cualquier caso, dado que esta investigación implicaba una conceptualización no solo del fenómeno, sino también de las afectaciones mismas, y de que los datos obtenidos fueron cuidadosamente seleccionados, los resultados comparativos entre la clasificación manual y la clasificación automática son bastante aceptables.

También se hace necesario aplicar el algoritmo a nuevos textos para corroborar su eficacia con corpus de otras formas de victimización y hacer una validación. Si bien implementamos técnicas para evitar los sesgos, sería importante usar otros datos que previamente no hayan sido tratados.

La modelación realizada con el algoritmo permite identificar porcentualmente las diferentes afectaciones. En las publicaciones analizadas para esta investigación estaban presentes todas las afectaciones, eran descritas y analizadas. Sin embargo, el modelo de análisis nos permite identificar la prevalencia y recurrencia de afectaciones psicológicas muy por encima de otras de carácter físico, económico, político, social, y dentro de aquellas, la angustia y la impotencia como dos emociones preponderantes.

Herramientas como estas pueden aportar a la comprensión de afectaciones generadas por desaparición forzada sobre individuos, familias y comunidades, en la perspectiva de avanzar en la política pública relacionada con los derechos de las víctimas, especialmente con el carácter reparador y el principio de integralidad de las acciones de resarcimiento que debe garantizar el Estado.

Sobre las perspectivas de estudios futuros acerca de la desaparición forzada u otros hechos victimizantes, puede ser de interés el análisis de los efectos diferenciales en la población, acudiendo a la posibilidad de ampliar el corpus con testimonios de sectores específicos, como mujeres, hombres, jóvenes, niños y niñas, personas que pertenecen a organizaciones comunitarias o movimientos sociales.

En cuanto a la aplicación de modelos de analítica de datos y lingüística computacional en corpus de memoria, se destacan la necesidad de construir más diccionarios y desarrollar conceptualmente un campo que permita hacer mejores clasificaciones, por ejemplo, ampliando o bien la relación de las afectaciones con otros elementos presentes en los textos o bien la variedad de corpus, para explorar cómo se comportan estas mismas afectaciones expresadas en relación con otras victimizaciones o con testimonios recogidos bajo otro enfoque.

Otro elemento para explorar es el análisis automático de textos utilizando herramientas de procesamiento de lenguaje natural. Esto permitiría entrenar corpus más grandes y a partir de la construcción de diccionarios más automáticos, que posibiliten cruzar afectaciones con información geográfica, nombres de entidades y actores,

de manera que puedan hacerse otros análisis mediante categorización simultánea en múltiples niveles. Esto, nuevamente, requiere datos en cantidad, porque la recopilación y la validación de la veracidad y la calidad de los datos debe ser garantizada por medio de esquemas o guías estandarizadas.

Por último, un campo que resulta de gran interés es el procesamiento de audios, puesto que gran cantidad de testimonios son recogidos en estos formatos, y la lectura automática y clasificación a partir de estos esquemas puede aportar sin duda en tareas de análisis por parte de investigadores sobre memoria histórica.

Notas

(1) Este artículo es producto de la investigación "Bases para un modelo de análisis textual y etiquetado semántico de testimonios de víctimas y sobrevivientes del conflicto armado colombiano", financiada con recursos del Fondo Primer Proyecto. Vicerrectoría de Investigación. Comité para el Desarrollo de la Investigación de la Universidad de Antioquia. Código proyecto 2018-23597.

(2) En el acuerdo firmado entre el Estado colombiano y las Fuerzas Armadas Revolucionarias de Colombia en 2016, se creó la Unidad de Búsqueda de Personas dadas por Desaparecidas, que se inserta en el Sistema Integral de Verdad, Justicia, Reparación y No Repetición, junto con la Comisión de la Verdad y la Justicia Especial de Paz. La Unidad de Búsqueda es una instancia que tiene fines humanitarios y extralegales, orientada exclusivamente a la búsqueda de personas desaparecidas, secuestradas y reclutadas por grupos armados ilegales en el contexto del conflicto armado colombiano (Comisión de la Verdad, Justicia Especial para la Paz y Unidad de Búsqueda de Personas Dadas por Desaparecidas, s. f.).

(3) Si bien la Unidad de Búsqueda de Personas dadas por Desaparecidas recoge las cifras suministradas por el Observatorio de Memoria y Conflicto, las cifras aumentan, porque en sus registros incluyen a personas reclutadas, secuestradas y excombatientes dados por desaparecidos en el marco del conflicto armado.

(4) En Colombia, la Ley 589 de 2000 (Colombia, Congreso de la República, 2020) atribuye las responsabilidades a particulares e integrantes de grupos armados y a servidores públicos o particulares que cometan el delito bajo la determinación o con la aquiescencia de estos.

(5) Disponible para consulta en: https://entrelineas.com.co/

Referencias

Aguado de Cea, Guadalupe; Álvarez de Mon; Rego, Inmaculada; Pareja Lora, Antonio (2002). Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. 17, 37–49

Browniee, Jason (2020). Train-Test Split for Evaluating Machine Learning Algorithms. // Python Machine Learning

Buendía, Miriam (2010). Anotación semántica en el dominio especializado de la Meteorología. // Caballero, Rosario; Pinar Sanz, María Jesús (Ed.). Modos y formas de la comunicación humana (923-934). Cuenca: Ediciones de la Universidad de Castilla-La Mancha / AESLA.

Castrillón Rivera, D. G.; Liscano Pinzón, L. M.; Suárez Huertas, A. M. (2012). Contraste de las formas de estigmatización frente a la desaparición forzada: entre el discurso de las víctimas y la opinión de la sociedad civil. [Tesis de grado, Pontificia Universidad Javeriana]. Repositorio de la Pontificia Universidad Javeriana. https://repository.javeriana.edu.co/handle/10554/7936

Centro Nacional de Memoria Histórica (CNMH). (2014). Desaparición forzada. Tomo III: Entre la incertidumbre y el dolor: impactos psicosociales de la desaparición forzada. Imprenta Nacional.

Centro Nacional de Memoria Histórica (CNMH). (2016). Hasta encontrarlos: el drama de la desaparición forzada en Colombia. CNMH.

Charry Lozano, Liliana (2016). Impactos psicológicos y psicosociales en víctimas sobrevivientes de masacre selectiva en el marco del conflicto suroccidente colombiano en el año 2011. Colombia Forense, 3:2, 53-62. https://doi.org/10.16925/cf.v3i2.1756

Colombia, Congreso de la República (2000). Ley 589, por medio de la cual se tipifica el genocidio, la desaparición forzada, el desplazamiento forzado y la tortura; y se dictan otras disposiciones (2000, 6 de julio).

Comisión de la Verdad; Justicia Especial para la Paz; Unidad de Búsqueda de Personas Dadas por Desaparecidas (s. f.). Sistema Integral de Verdad, Justicia, Reparación y No Repetición. https://www.jep.gov.co/Infografas/SIVJRNR_ES.pdf

Corte Penal Internacional (1998). Estatuto de Roma de la Corte Penal Internacional. https://www.un.org/spanish/law/icc/statute/spanish/rome_statute(s).pdf

Duque, Erika Teresa (2009). Metodología para la extracción de metadatos semánticos de textos en español utilizando procesamiento de lenguaje natural: subaplicación para la identificación de contextos espaciales y temporales en textos que describan interacciones entre actores. Trabajo de grado, Universidad EAFIT, Medellín. https://repository.eafit.edu.co/bitstream/handle/10784/1261/erika_duque_2009.pdf

Durán Pérez, Teresa; Baci Herzfeld, Roberta; Pérez Sales, Pau (2004). Muerte y desaparición forzada en la Araucanía: una aproximación étnica. Universidad Católica de Temuco. http://www.pauperez.cat/wp-content/uploads/2017/Ω11/perez-sales-muerte-y-desaparicion-forzada.pdf

Eriksson, Henrik (2007). An Annotation Tool for Semantic Documents. // Franconi E., Kifer M., May W. (eds). The Semantic Web: Research and Applications. ESWC 2007. Lecture Notes in Computer Science, vol 4519. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72667-8_54

Hossin, M; Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5: 2, 01-11.

Joshi, Aditya; Bhattacharyya, Pushpak; Ahire, Sagar. (2017). Sentiment Resources: Lexicons and Datasets. // A practical guide to sentiment analysis. Springer.

Liu, Bing. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Estados Unidos: Camdrige University Press

Liu, Bing (2017). Many Facets of Sentiment Analysis. // A practical guide to sentiment analysis. Springer.

McEnery, Tony; Hardie, Andrew. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Observatorio de Memoria y Conflicto, Centro Nacional de Memoria Histórica (CNMH) (2021). El conflicto en cifras. http://micrositios.centrodememoriahistorica.gov.co/observatorio/portal-de-datos/el-conflicto-en-cifras/desaparicion-forzada/

Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (2006). Convención Internacional para la protección de todas las personas contra las desapariciones forzadas. https://www.ohchr.org/sp/professionalinterest/pages/conventionced.aspx

Ortiz Fonnegra, María Isabel (2019, mayo 29). “Hay que encontrar a más de 120.000 desaparecidos por el conflicto”. El Tiempo. https://bit.ly/32rRAYC

Perez Sales, Pau; Bacic, Roberta; Durán, Teresa (1998). Muerte y Desaparición Forzada en la Araucania. Una perspectiva étnica. Santiago de Chile: Editorial LOM.

Rebolledo, Olga; Rondón, Lina. (2010). Reflexiones y aproximaciones al trabajo psicosocial con víctimas individuales y colectivas en el marco del proceso de reparación. Revista de Estudios Sociales. 36, 40-50. https://journals.openedition.org/revestudsoc/13259.

Enviado: 2022-03-03. Segunda versión: 2022-10-27.
Aceptado: 2022-10-27.

Diálogos entre as questões socioculturais e os sistemas de organização do conhecimento

Diálogos entre las cuestiones socioculturales y los sistemas de organización del conocimiento

Dialogues between sociocultural issues and knowledge organization systems

Walter MOREIRA (1), Deise SABBAG (2)

(1) Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, Av. Higyno Muzzi Filho, 737, Marília – SP, Brasil – CEP: 17.525-900, walter.moreira@unesp.br. (2) Universidade de São Paulo (USP), Faculdade de Filosofia, Ciências e Letras, Av. Bandeirantes, 3900, Ribei-rão Preto – SP – Brasil, CEP 14040-901, deisesabbag@usp.br

Resumen

La base ontológica necesaria para los sistemas de organización del conocimiento (SOC) se estructura en una base epistemológica y, como se requiere cada vez más, en el contexto de los estudios críticos sobre la organización del conocimiento, en una base cultural. Los SOC deben ser discutidos en términos de sus impactos sociales, directos o indirectos, visibles o no. Así, se plantean como problemas generales de investigación las siguientes preguntas: ¿cómo reconocer e incorporar la diversidad cultural en los SOC? ¿Cómo reconocer su dimensión aplicada en la construcción de representaciones documentales? SOC que no son inclusivos fracasan en la socialización del conocimiento. El descuido de las variables culturales involucradas en la producción y organización del conocimiento hace que el sistema sea opresivo o irrelevante, en ambos casos prescindible. Así, el objetivo es comprender los requerimientos formulados a los SOC de acuerdo con los intereses de la perspectiva cultural de la organización del conocimiento. Para ello, se adopta como fundamento metodológico la construcción de un texto crítico-reflexivo con base en los elementos señalados por Hjørland y Pedersen (2005) y resumidos en Hjørland (2008) como fundamentos para una teoría de la clasificación. Así, los diez principios enumerados por estos autores son sistematizados en cinco dimensiones de análisis relacionadas con la concepción de la estructura clasificatoria como componente de los SOC: objetividad/subjetividad; base ontológica; base sociocultural; el dominio como elemento rector; efectos sociales de la clasificación. Se concluye que la incorporación de la diversidad cultural en los SOC requiere atención a tres elementos que, aunque fácilmente identificables, resultan extremadamente complejos en su aspecto pragmático: a) el mapeo y reconocimiento de las diferentes perspectivas socioculturales a través de las cuales un determinado concepto puede ser observado; b) la incorporación de esta diversidad a los SOC flexibilizando la estructura de clasificación que la soporta; y c) la explicación de los puntos de vista adoptados en la construcción del SOC.

Palabras clave: Sistemas de organización del conocimiento. Organización del conocimiento. Estudios culturales. Aspectos sociales.

Abstract

The necessary ontological basis for knowledge organization systems (KOS) is structured on an epistemological basis and as is increasingly required in the context of critical knowledge organization studies, on a cultural basis. Thus, KOS must be discussed in terms of their social impacts, either directly or indirectly, visible, or not. The following questions are raised as general research problems: how to recognize and incorporate cultural diversity in KOS? How do recognize its applied dimension in the construction of documentary representations? KOS that are not inclusive fail in their fundamental purpose, which is the socialization of knowledge. The neglect of the cultural variables involved in the production and organization of knowledge makes the system oppressive or irrelevant, in both cases expendable. Thus, the objective is to understand the new requirements formulated for knowledge organization systems in accordance with the interests of the cultural perspective of knowledge organization. To do this, the construction of a critical-reflexive text based on the elements indicated by Hjørland and Pedersen (2005) and summarized in Hjørland (2008) as foundations for a classification theory is adopted as a methodological parameter. Thus, the ten principles listed by these authors are systematized into five dimensions of analysis related to the conception of the classificatory structure as a key component of KOS, they are: objectivity/subjectivity; ontological basis; sociocultural base; the domain as a guiding element; social effects of classification. It is concluded that the incorporation of cultural diversity into KOS requires attention to, at least, three elements that, although easily identifiable, prove to be extremely complex in their pragmatic aspect: a) the mapping and recognition of the different sociocultural perspectives through which a given concept can be observed; b) the incorporation of this diversity into the SOCs by making the classification structure that supports it more flexible; c) the explanation of the points of view adopted in the construction of the SOC.

Keywords: Knowledge organization systems. Knowledge organization. Cultural studies. Social issues.

1. Introdução

Paralelamente à revolução digital, ou em consonância dialógica com ela, ocorre também uma reforma relativa aos processos de inclusão social, com maior respeito pela diversidade em todos os seus aspectos.

No que diz respeito à biblioteconomia, à arquivologia e à ciência da informação, caminha-se, há algumas poucas décadas, de uma postura com foco no acervo ou na custódia dos documentos para perspectivas mais abertas e inclusivas, focadas no acesso pleno ao conhecimento registrado e na promoção das condições para sua livre circulação e produção. Evidentemente que as tecnologias digitais facilitam a organização, circulação e compartilhamento de documentos, notadamente aqueles marcados pela originalidade, sem que a necessidade de preservação seja tomada como o elemento principal. Óbvio que ainda é preciso preservar, mas a ubiquidade do documento digital torna o processo menos oneroso.

Como é próprio das mudanças sociais, trata-se, evidentemente, de um processo lento, circunscrito compassadamente por avanços e por retrocessos, cujo maior desenvolvimento tem início a partir da segunda metade do século XX.

O conhecimento, tomado como insumo básico para o desenvolvimento das sociedades, possui tanto aspectos subjetivos (produzindo subjetividades resultantes de um processo de configuração sócio-histórico que pode ser elaborado pela noção de subjetivação foucaultiana e deleuziana (Foucault, 2014; Deleuze, 1992), ou seja, entendida como processo que nunca está acabado, um devir, como objetivos (coletivos ou sociais).

É preciso reconhecer, portanto, as duas faces, ambas em diálogo, pelas quais o conhecimento opera: numa face um conjunto de características que o tornam praticamente insondável e que o configuram como posse do sujeito cognoscente, na outra face o aspecto de socialização que lhe confere o seu registro. As duas faces não são antagônicas e nem poderiam ser, são complementares, uma vez que as relações entre sujeito e sociedade também o são.

Paralelamente a essa concepção, aponta-se, com apoio em Hjørland (2006, 2007), uma distinção entre a abordagem intelectual e a abordagem social da organização do conhecimento. Tais modelos dialogam com as relações entre "organização intelectual das ciências" e "organização social das ciências" conforme a compreensão desses conceitos por Whitley (1984).

A organização intelectual do conhecimento tem como base as descrições e representações relativas à estruturação da realidade em seus diversos elementos. Podem ser citados como exemplos o mapa geográfico, a tabela periódica da química e as taxonomias da biologia. A organização social do conhecimento, por sua vez, refere-se à organização das disciplinas e das profissões, sendo, desse modo, assentada em sistemas sociais de organização do conhecimento, dentre os quais podem ser citados como exemplo as universidades e seus modelos de organização disciplinar do conhecimento.

Considerando-se o aspecto social da organização do conhecimento, busca-se construir uma reflexão crítica sobre a função reservada aos sistemas de organização do conhecimento (SOC) nesse cenário, incluindo-se o problema da representação. Assim, os SOC são contemplados de modo dialógico: tanto como dispositivos que assentam e estabilizam, de certo modo, as representações, como dispositivos formadores dos modelos gerais a partir dos quais o conhecimento é socialmente organizado, faceta que acentua sua proatividade na construção social do conhecimento.

Coloca-se como problema geral de pesquisa os seguintes questionamentos: como incorporar a diversidade cultural aos SOC? Como reconhecer sua dimensão aplicada na construção de representações documentárias?

O argumento fundamental que justifica a compreensão do problema assenta-se no fato de que SOC que não sejam inclusivos, nos quais os usuários não se percebam representados, falham de modo contundente em seu propósito fundamental que é a promoção do acesso, da circulação e da produção do conhecimento.

Qualquer SOC que negligencie as variáveis culturais envolvidas na produção e organização do conhecimento está fadado a sofrer as consequências de sua disfunção: tornar-se-á opressor, tendencioso ou irrelevante; em quaisquer os casos mostrar-se-á inadequado e, portanto, dispensável.

SOC são sistemas de classificação, são artefatos, e, nessa condição, possuem efeitos sociais.

Propõe-se como objetivo para esta pesquisa discutir as novas e antigas exigências formuladas aos SOC em função dos interesses da perspectiva cultural interdisciplinar da organização do conhecimento.

Para tanto, adota-se como procedimento metodológico a construção de um texto crítico-reflexivo assentado em dimensões de análise construídas a partir dos elementos apontados por Hjørland e Pedersen (2005) e sumarizados em Hjørland (2008). Esses textos apresentam um conjunto de

fundamentos aplicáveis a uma teoria da classificação que visa a recuperação da informação e enumeram dez princípios orientados pela da semântica pragmatista, em que as expressões são ferramentas de interação cujos significados são as funções que assumem na interação, em oposição a uma semântica positivista em que "as expressões 'representa' entidades e seus significados são as entidades representadas por elas" (Hjørland, 2008, p. 372-373, tradução livre).

Assim, os dez princípios enumerados por esses autores são adaptados e sistematizados em cinco dimensões de análise clivadas pela abordagem de uma estrutura classificatória como componente dos SOC: objetividade/subjetividade; base ontológica; base sociocultural; domínio como elemento norteador; efeitos sociais da classificação. A aplicação dessas dimensões neste estudo assenta-se na suposição de que apenas SOC cujas estruturas classificatórias estejam teleologicamente orientadas para a inclusão de diferentes perspectivas culturais poderão efetivamente fazê-la.

2. A organização do conhecimento e sua representação

A organização do conhecimento, observada em diálogo com a ciência da informação, pode ser compreendida por meio de três eixos ou dimensões (Guimarães, 2015): a) eixo epistemológico: refere-se à construção da própria organização do conhecimento em termos de seus paradigmas e suas bases teóricas e metodológicas; b) eixo tecnológico: refere-se à dimensão aplicada do campo, notadamente em relação aos impactos de seus instrumentos, os quais são referidos de modo genérico neste artigo pelo termo "sistemas de organização do conhecimento"; c) eixo cultural: em que se consideram os aspectos socioculturais que interferem na socialização do conhecimento registrado, com favorecimento da mediação de universos culturais.

Após a ampla revisão dos conceitos de cultura e dos costumes, de modo geral, realizados no bojo dos movimentos sociais pós 1968, essa temática tem ganhado cada vez mais destaque em diversas manifestações das ciências sociais, incluindo-se, naturalmente, as ciências sociais aplicadas. Dentre as dimensões apontadas por Guimarães (2015), destacam-se nesta pesquisa, sem prejuízo das demais em favor da perspectiva integradora de abordagem na pesquisa, as discussões sobre os referentes socioculturais na organização do conhecimento, abordando-se de modo mais específico as manifestações desses referentes nos SOC.

Assim, discutem-se os SOC em perspectiva dialógica com seus próprios efeitos, isto é, ao mesmo tempo em que são instrumentos que refletem um determinado modo de organizar o conhecimento, também o influenciam, por acordo ou desacordo. Conhecendo-se a influência do poder das visões de mundo privilegiadas sobre as classificações, como apontadas em Bowker e Star (XXX), é preciso encarar a função social dos SOC e questionar, minimamente, o que ou quem determina quais pontos de vista deverão prevalecer ou, em outros termos, o que será silenciado e o que será revelado pela seleção dos conceitos e sua relações. São questões amplas, cujas respostas não cabem neste artigo cuja pretensão é apontar alguns questionamentos subsidiários.

Em referência específica aos interesses da comunidade brasileira de pesquisadores sobre organização do conhecimento nas questões culturais, citam-se como exemplos significativos os temas de alguns eventos nacionais (capítulos) e um evento internacional, promovidos pela International Society of Knowledge Organization (ISKO), todos realizados no Brasil (Quadro I).

Tema	*Natureza*	*Data*
Complexidade e organização do conhecimento, desafios de nosso século	capítulo Brasil	2013
Organização do conhecimento e diversidade cultural	capítulo Brasil	2015
Organização do conhecimento para um mundo sustentável: desafios e perspectivas para o compartilhamento cultural, científico e tecnológico em uma sociedade conectada	internacional	2016
Memória, tecnologia e cultura na organização do conhecimento	capítulo Brasil	2017
Organização do conhecimento responsável: promovendo sociedades democráticas e inclusivas	capítulo Brasil	2019

Quadro I. Temáticas dos congressos ISKO realizados no Brasil (2013-2019)

Os referentes socioculturais nos SOC estão presentes na sua construção e modulação conceituais, razão pela qual devem ser analisados em sua função precípua de recorte de uma realidade

que está assentada num tempo e espaço. Assim, tais sistemas, afastados da pretensiosa vaidade da universalidade, contemplam apenas uma ou, no melhor dos casos, algumas perspectivas de categorização da realidade como lentes objetivas constituídas de diversas camadas de vidro, e outros componentes complexos, que capturam elementos da realidade de um determinado saber-poder que é controlado, selecionado, organizado e distribuído por procedimentos de exclusão: interdição, separação e vontade de verdade (Foucault, 2013).

No momento da constituição das suas bases classificatórias já se evidenciam traços históricos, culturais e ideológicos que permitem aos classificacionistas, como sujeitos de um outro contexto, avaliá-los criticamente, podendo incorporar ou não seus traços, sendo indispensável que haja interesse permanente para percebê-los (Shera, 1959).

O que se quer evitar é que a advertência de Shera (1959, p. 120, tradução livre), relativa à concepção de classificação restrita ao seu aspecto utilitário de ferramenta de localização, ainda seja necessária no mesmo grau em que foi enunciada, sessenta e três anos depois:

Dizer que a classificação bibliográfica é utilitária não é, em si mesmo, depreciativo; ela deveria ser útil, mas hoje a classificação bibliográfica é utilitária no nível mais baixo de suas capacidades. Não estrutura o conhecimento registrado em padrões harmoniosos com os padrões de pensamento do usuário da biblioteca, serve principalmente como um dispositivo pelo qual se pode encontrar um determinado livro.

Considerando-se as expectativas decorrentes da evolução do conceito de "bibliotecas tomadas isoladamente" para o conceito de "bibliotecas como dispositivos que funcionam em redes colaborativas", as perspectivas de universalidade dos primeiros sistemas de classificação bibliográficos trazidas na esteira do positivismo que os inspirou já não tem mais sentido. A atual ausência de grandes projetos universais de organização do conhecimento como a que se verificou no final do século XIX e início do XX é um forte argumento a esse respeito (Moreira, 2018).

Observando-se os aspectos conceitual, terminológico e representacional dos SOC, o desafio que se impõe à organização do conhecimento é a construção de instrumentos atentos às questões de interoperabilidade não apenas em nível das diversas línguas, mas, principalmente, que possam representar e promover o diálogo entre as diferentes culturas, incluindo-se, claro as diferentes línguas.

Isto é, requerem-se SOC que avancem em relação à perspectiva, ainda absolutamente necessária, dos instrumentos multilíngues de organização e recuperação da informação em favor de instrumentos multi e transdisciplinares que ultrapassem as barreiras culturais. Em movimento de resistência e enfrentamento ao desejo de transposição, alcance e atravessamento que advém da revolução cultural e os processos de desenvolvimento do meio ambiente global que caminham para a compressão espaço-tempo, bem como para a homogeneização cultural (Hall, 1997) que trabalha para o único, o lugar único, um mundo único que desconsidera as diferenças, a multiplicidade.

Neste sentido surge a necessidade de que os SOC não sejam meras formas de regulação cultural que existem no interior dos sistemas classificatórios,

que delimitam cada cultura, que definem os limites entre a semelhança e a diferença, entre o sagrado e o profano, o que é 'aceitável' e o que é 'inaceitável' em relação a nosso comportamento, nossas roupas, o que falamos, nossos hábitos, que costumes e práticas são 'normais' e anormais', quem é 'limpo' e quem é 'sujo' (Hall, 1997, p. 42).

A norma mais recente da ISO aplicada à construção de tesauros (International..., 2011), que já manifesta de modo explícito preocupações com a interoperabilidade e que formaliza, inclusive, algumas orientações, também destaca o multilinguismo e a multiculturalidade como horizontes. O cenário que demanda e assegura essas necessidades é o da crescente internacionalização das ciências, das técnicas e das humanidades, além, evidentemente do próprio caráter multicultural da internet, como destaca García Marco (2016).

Considerando-se a perspectiva e as demandas do multiculturalismo em busca de uma sociedade igualitária, manifesta-se cada vez mais preocupação com os aspectos e as consequências sociais decorrentes das ações de construção e de uso dos SOC. Procura-se, portanto, compreender a organização do conhecimento pela perspectiva aplicada, principalmente ontológica, e cultural dos SOC, mantendo-se, naturalmente, o diálogo com suas bases epistemológicas.

A estrutura classificatória presente nos SOC faz de todos eles também agentes de produção, manutenção ou renovação dos mesmos efeitos de sentidos silenciados que, no bojo da intolerância e das desigualdades, trazem a compreensão do dito e do não-dito.

3. Posturas epistemológicas

O enfrentamento teórico dos mecanismos que colaboram para a manutenção e reprodução da intolerância e das desigualdades, tais como aqueles que se manifestam nas escolhas de itens lexicais adotados no âmbito das terminologias (termos) e das linguagens documentárias (descritores) são aqui problematizados como nucleares para que a organização do conhecimento possibilite olhares pela perspectiva cultural inclusiva.

Para tanto, a partir sumarização dos elementos fundadores de uma teoria da classificação, de sua estrutura classificatória e componentes, são cotejadas por meio de reflexão crítica as cinco categorias estruturais classificatórias inspiradas em Hjørland (2008) e consideradas essenciais para a promoção e movimento dialógico entre as questões socioculturais e a organização do conhecimento, cognominados de: a objetividade/subjetividade dos critérios classificatórios; a base ontológica dos SOC; o domínio como unidade de análise; a base sociocutural dos SOC; a classificação e seus efeitos.

3.1. A objetividade/subjetividade dos critérios classificatórios

Não existem e nem são desejáveis critérios puramente objetivos para a organização do conhecimento, pois as representações são sempre orientadas por pontos de vista. Isto é, olhares que são produzidos culturalmente por algo comum que tem forma, propósitos e significados, mas que estão situados em um determinado tempo e espaço e, por essa razão, não podem ser aplicáveis sem levar em consideração a sociedade em desenvolvimento que se constrói e reconstrói, que está em debate permanente de criação em cada pensar individual.

A aplicação de critérios puramente objetivos descarta os significados e novas observações sociais marginalizando o processo ordinário da natureza de uma cultura (Williams, 2015). A decisão que se requer, portanto, é de ordem ética e não está entre subjetividade e objetividade, mas entre viés e "inclinação" (*slant*) (Guimarães, 2017), isto é, a assunção de um determinado ponto de vista devidamente justificado, conscientemente assumido e eticamente comprometido. Como exemplos mais danosos de vieses, Guimarães (2017) destaca o preconceito e o proselitismo.

Embora as classificações também devam ser abordadas, como de fato são, por seu aspecto técnico, principalmente quando são discutidas no âmbito aplicado das profissões conectadas à organização da informação, é preciso enfatizar seu caráter epistemológico. O ato de classificar, como salienta García Gutierrez (2011, p. 6, tradução livre), "não é orientado apenas por um conjunto de regras organizacionais explícitas, mas também por padrões comportamentais cognitivos, inconscientes e automáticos ligados à ideologia, cultura, identidade e memória que confinam o pluralismo e a interpretação".

3.2. A base ontológica dos SOC

Embora SOC seja um termo genérico empregado para se referir a uma ampla gama de instrumentos diversamente estruturados, com funções específicas e diferentes maneiras de se relacionar com as tecnologias (Mazzocchi, 2018), há, pela própria condição lógica de ser esse um termo genérico, alguns componentes comuns entre seus elementos.

Assim, sistemas de classificação, tesauros e ontologias, entre outros, manifestam, todos, uma base ontológica (explicitamente ou não) e são projetados para dar suporte à organização do conhecimento e da informação. Torna-se imprescindível, portanto, compreender em que condições são assumidos os compromissos ontológicos que os sustentam.

A atividade de classificação, ratificam Durkheim e Mauss (2009)[(1)], não é inata, é construída socialmente, e não é, portanto, natural, mas cultural, de modo que a organização das ideias no pensamento é feita em estreita conexão com a organização social do conhecimento. A hierarquia lógica nesse caso "é apenas outro aspecto da hierarquia social, e a unidade do conhecimento nada mais é do que a própria unidade da coletividade social estendida ao universo" (Needham, 2009, p. x, tradução livre).

Para Durkheim e Mauss (2009) as pessoas primitivas não seriam tão diferentes daquelas "cultas", haveria apenas uma diferença de grau, mas não de qualidade, entre os sistemas totêmicos dos aborígenes australianos e as visões mais racionais e científicas dos europeus. Fundamentalmente, o impulso de organizar obedeceria aos mesmos esquemas e existiria em todos os lugares e culturas, independentemente de serem orientados pela ciência ou outro modelo epistemológico.

As classificações primitivas, portanto, "não são singulares ou excepcionais, sem nenhuma analogia com aquelas empregadas por povos mais civilizados; pelo contrário, parecem estar ligadas, sem quebra de continuidade, às primeiras classificações científicas" (Durkheim; Mauss, 2009, p. 48, tradução livre).

Durkheim e Mauss (2009) aproximam as classificações sociais dos fatos sociais, tomando ambos como externos aos indivíduos e que se lhe impõem de modo coercitivo. Adicionalmente, como destaca Herrera López (2006, p. 6, tradução livre), "a classificação das coisas reproduz a classificação da sociedade, a qual vincula o sistema social com o sistema lógico".

Ainda a respeito do pensamento de Durkheim e Mauss, concorda-se com a síntese de Siqueira (2010, p. 39) segundo a qual esses pensadores defendem que "a organização de uma classe está mais associada à observação direta do mundo real, ao invés de uma elaboração abstrata, o que resulta numa classificação moldada segundo as categorias sociais, reflexos das relações familiares, socioeconômicas, políticas e culturais". Nega-se, desse modo, a ideia de mimese e enfatiza-se o aspecto sociocultural e ideológico das classificações, os quais influem diretamente nos aspectos éticos da organização e representação do conhecimento, como já destacado em Hudon (1997), Beghtol (2002) e Guimarães (2017), entre outros.

3.3. O domínio como unidade de análise

A concepção expressa na subseção imediatamente anterior a esta dialoga com a noção de domínio entendido como "unidade de análise para a construção de um sistema de organização do conhecimento" (Smiraglia, 2012, p. 114, tradução livre). Um domínio "é um grupo que possui uma base ontológica que revela uma teleologia subjacente, um conjunto de hipóteses comuns e um consenso epistemológico sobre abordagens metodológicas [...]" (Smiraglia, 2012, p. 114, tradução livre).

As interações entre os diversos aspectos, elementos ou componentes dos domínios definem seus contornos e revelam seu papel crítico na evolução do conhecimento. Portanto a base ontológica revela uma materialidade linguística e histórica remetendo a um discurso produzido e aceito dentro do domínio pela comunidade discursiva. Na noção de domínio e seu discurso estão explicitadas as regularidades, as referências, as formações discursivas que entram em jogo pela formação ideológica (Orlandi, 2020).

Mai (2011) chama de "distância semiótica" o movimento necessário para que se proceda à classificação de modo a se abandonar a ideia de que se pode classificar as coisas como realmente são, no sentido do realismo, e, nessa mesma linha, de que seria possível classificar os documentos pela identificação de seu sentido real. Nada mais redutor à classificação bibliográfica, aliás, do que clivá-la por seu aspecto técnico de ordenação. Para compreender e ser capaz de avaliar uma classificação, é preciso compreendê-la pela perspectiva da sua relação com o domínio e com a cultura, isto é, contemplando o contexto social em que a classificação é utilizada, bem como o seu poder de representação.

3.4. A base sociocultural dos SOC

SOC são artefatos culturais, isto é, são construídos e utilizados em contextos sociais nos quais as pessoas compartilham diversidades de costumes, linguagens, religiões, posições políticas, orientações sexuais, ideologias etc. Assim, para a compreensão do tema desta pesquisa é preciso acrescentar à base ontológica e epistemológica dos SOC uma "base cultural", isto é, incluir discussões sobre as referências socioculturais que orientam as discussões sobre organização do conhecimento e, por extensão, os SOC.

A base cultural, e o próprio conceito de cultura na perspectiva antropológica, tem como característica a reconstrução e fragmentação devido a inúmeras reformulações, mas como sistemas adaptativos os SOC são sistemas culturais que produzem comportamentos socialmente transmitidos; necessitam acompanhar a mudança cultural que é um processo de adaptação atravessado pela tecnologia, economia, organização social e meios de produção; é um sistema afetado pelos componentes ideológicos dos sistemas culturais que contribui como dispositivo de poder (Laraia, 2001).

Nesse sentido, constituição e o emprego de uma classificação requerem, como condição *sine qua non*, a adoção de pontos de vista. Admitindo-se que o número de diferentes perspectivas por meio das quais se pode compreender a realidade é potencialmente infinita, espera-se que os SOC sejam capazes de revelar os aspectos convergentes, concorrentes ou conflitantes relativos aos diversos fenômenos, se não todos, pelo menos o maior número deles. Assim, como ensina Mai (2011, p. 717), "o desafio do catalogador e do indexador não seria extrair o conteúdo dos documentos, mas observar os movimentos que o documento faz em conversas, perspectivas e debates particulares". Isto é, compreender o documento e sua representação pela perspectiva científica da bibliografia.

3.5. A classificação e seus efeitos

A classificação tem sido discutida há um longo tempo e de modo bastante abrangente por diversas áreas de conhecimento. Tomada em seu sentido genérico, de atividade que resulta em parâmetros discriminatórios cujos resultados orien-

tam as mais diversas atividades humanas, a classificação ocupa posição de destaque, desde a abordagem realista de Aristóteles e sua busca pelas categorias universais até as discussões sobre a fenomenologia e a interdisciplinaridade como marcas dos mais recentes avanços das relações entre as diversas ciências.

Afastando-se da compreensão do conceito de classificação orientado pelo princípio do realismo, em que os SOC são instrumentos utilizados para descrever a realidade e que por ancorar na realidade o seu referente advogam a precisão na descrição, adota-se neste artigo o princípio do perspectivismo aplicado a tais sistemas; princípio anunciado por Nietzsche que afirmava não existir uma verdade absoluta, mas verdades subjetivas que são construídas pela multiplicidade de interpretações. Sustenta-se, portanto, que há diversos modos de descrever a realidade e que todos eles podem ser igualmente precisos. Nessa linha teórica, qualquer classificação é perspectiva, isto é, reflete, necessariamente um recorte, um ponto de vista, e os fenômenos de que se ocupam as diversas ciências jamais lhes serão objetos exclusivos.

Consequentemente, não existe a possibilidade de apenas uma única ciência resolver todas as questões que envolvem a solução para um determinado problema.

Como uma espécie de "antídoto" ao modelo tecnicista e realista, García Gutierrez (2007; 2011) vem desenvolvendo o conceito de "desclassificação". Desclassificar não se refere à simples negativa da classificação, como pode fazer supor o emprego do prefixo de negação que se antepõe ao termo. Tal postura, além de ingenuamente improdutiva, seria absolutamente impraticável, posto que a classificação, enquanto operação gnosiológica e epistemológica, "impregna a totalidade e de modo total, a nossa relação com o mundo" (García Gutierrez, 2011, p. 6, tradução livre).

Assim, desclassificar envolve uma lógica diferente, plural e não essencialista, implica uma tomada de consciência a respeito da incompletude, do viés, da arbitrariedade e da subjetividade presentes na classificação.

Algumas questões de ordem tecnológica também se colocam. Como resolver, por exemplo, no âmbito das bibliotecas, as diversas questões apontadas com o emprego de sistemas de classificação ou tesauros disciplinarmente estruturados que replicam um modelo positivista de descrição das relações entre as ciências?

Nesse sentido, há resultados de pesquisas teórico-aplicadas sobre construção de SOC que já dialogam de modo produtivo com a interdisciplinaridade, com o perspectivismo e com abordagens integradoras da cultura. No conjunto desses estudos, destacam-se, além dos autores citados no texto, alguns outros: Hudon (1997), Beghtol (2002), Olson (2002), Smiraglia (2014), El Hadi (2015), Gnoli (2016, 2017a, 2017b, 2018), Lara; Mendes (2017), Szostak; Gnoli; López-Huertas (2016).

No espectro das implicações correlacionadas aos referentes socioculturais dos SOC, destacam-se também as relações entre classificação e poder. Os exercícios de poder, dos quais a classificação é fonte e manifestação, ocorrem nos microcosmos das manifestações diárias,

> uma classificação poderosa e milenar protegida pela tradição, sabedoria, conhecimento, memória, identidade, estabilidade, religião, cultura, ciência e modo de vida, como se costuma dizer, todos cooperando na busca de uma classificação idêntica e imutável que divulga incessantemente suas estruturas. Uma classificação concebida como origem e destino do mundo, sempre submissa e reforçando a ordem estabelecida em espaços nos quais talvez nenhuma ordem seja necessária (García Gutierrez, 2011, p. 11).

As estruturas classificatórias que sustentam os diversos tipos de SOC são, afinal, orientadas e, de certo modo, reguladas por alguma estrutura classificatória mais ampla e ubíqua, composta por macro e microrrelações, ainda que os sistemas utilizados para classificar e os critérios que os nortearam não sejam necessária e suficientemente explicitados.

Além do mais, tem-se como agravante que nem todas as classificações são formalizadas e, quando tomadas como naturais, geram percepções e comportamentos que se reproduzem facilmente. Os sistemas bons e úteis à manutenção do *status quo*, aliás, são invisíveis por definição (Bowker; Star, 2000). Nesse sentido, é fundamental compreender como as categorias são construídas e até mesmo desafiar os silêncios que as cercam.

Considerando-se o contexto social, há sempre muitas dificuldades para encaixar a pluralidade dos fenômenos no modelo aristotélico de classificação e sua contradição interna com classes definidas em limites rígidos e mutuamente excludentes. A rigidez nas classificações provoca, invariavelmente, mais prejuízos que benefícios, alcançando, inclusive, momentos de paroxismo. Veja-se, por exemplo, o emprego da classificação a serviço dos interesses do *apartheid* na África do Sul em seus limites rígidos, evidentemente não assentados em bases científicas, para

a fixação do que se quis compreender como característica definitória do polêmico e contraditório conceito de “raça” (Bowker; Star, 2000).

4. Considerações finais

A relação dialógica entre as questões socioculturais e a organização do conhecimento aponta para a reflexão de que SOC podem ser artefatos atravessados pelo poder que contribuem para a manutenção, reprodução e sustentação de comportamentos, atitudes e ideias; manutenção e reprodução da intolerância, desigualdades e efeitos de silêncio.

Contra a homogeneização cultural e os usos dos SOC como reguladores culturais traz-se para o debate a classificação e seus efeitos de sentido adotando o princípio do perspectivismo como possibilidade para compreensão analítico-crítica das classificações.

Incorporar a diversidade cultural aos SOC requer, minimamente, a atenção a três elementos que, embora facilmente identificáveis, revelam-se extremamente complexos em seu aspecto pragmático: a) o mapeamento e o reconhecimento das diferentes perspectivas socioculturais pelas quais um determinado conceito pode ser observado; b) a incorporação dessa diversidade aos SOC pela flexibilização da estrutura classificatória que os sustenta; c) a explicitação dos pontos de vista adotados na construção dos SOC.

Não são exatamente novas todas as questões que se apresentam ao debate. Como se apontou neste artigo, desde o momento em que se reconheceu o aspecto teleológico da organização do conhecimento, *lato sensu*, seus sistemas e processos foram reorientados para a incorporação do contexto (cultura) e das necessidades de informação do usuário. Ainda são, contudo, questões oportunas e tornadas ainda mais complexas pelas possibilidades de alcance das tecnologias digitais e pelos efeitos da globalização.

Considerando a dimensão complexa do diálogo assumido neste trabalho, não se espera de forma alguma esgotá-lo na oportunidade, ao contrário, que provoque desdobramentos já que “as categorias do pensamento humano nunca são fixadas de forma definitiva; elas se fazem, desfazem e refazem incessantemente: mudam com o lugar e com o tempo” como disse Durkheim (1909, s. p.).

Notas

(1) No original “De quelques formes primitives de classification”, publicado em 1901.

Referências

Beghtol, C. (2002). A proposed ethical warrant for global knowledge representation and organization systems. // Journal of Documentation. 58:5, (2002) 507-532.

Bowker, G. C.; Star, S. L. (2000) Sorting things out: classification and its consequences. Cambridge: Mit Press, 2000.

Deleuze, G. (1992). Conversações. São Paulo: Ed. 34, 1992.

Durkheim, E. (1909). Sociologie religieuse théorie de la connissance. // Reveu de Métaphysique et de morale. 17:6 (Nov. 1909) 733-758. https://www.jstor.org/stable/40895159 (2021-04-09).

Durkheim, É.; Mauss, M. (2009) Primitive classification. Tradução de: Rodney Needham. London: Cohen & West, 2009.

El Hadi, W. M. Cultural interoperability and knowledge organization systems. // Guimarães, J. A. C.; Dodebei, V. L. D. L. M. (orgs.). Organização do conhecimento e diversidade cultural. Marília: ISKO-Brasil, 2015. 575-606.

Foucault, M. (2013). A ordem do discurso: aula inaugural no Collège de France, pronunciada em 2 de dezembro de 1970. São Paulo: Edições Loyola, 2013.

Foucault, M. (2014). História da sexualidade 2: o uso dos prazeres. São Paulo: Paz e Terra, 2014.

García Gutierrez, A. (2007) Desclasificados: pluralismo lógico y violência de la clasificación. Barcelona: Anthropos, 2007.

García Gutierrez, A. (2011). Desclassification in knowledge organization: a post-epistemological essay. // Transinformação. 23:1 (Jan./Abr. 2011) 5-14.

García Marco, FJ (2016). Normas y estándares para la elaboración de tesauros de patrimonio cultural. // ESPAÑA. Ministerio de Educación, Cultura y Deporte. El lenguaje sobre el patrimonio: estándares documentales para la descripción y gestión de colecciones. Madrid: Secretaría General Técnica, 2016. 29-46.

Gnoli, C. (2016). Classifying phenomena: part 1: dimensions. // Knowledge Organization. 43:6 (2016) 403-415.

Gnoli, C. (2017a). Classifying phenomena: part 2: types and levels. // Knowledge Organization. 44:1 (2017) 37-54.

Gnoli, C. (2017b). Classifying phenomena: part 3: facets. // Smiraglia, R; Lee, H-L. (eds.). Dimensions of knowledge: facets for knowledge organization. Würzburg: Ergon, 2017, p. 55-67.

Gnoli, C. (2018). Classifying phenomena: part 4: themes and rhemes. // Knowledge Organization. 45:1 (2018) 43-53.

Guimarães, J, A. C. (2015). Organização do conhecimento: passado, presente e futuro em um contexto de diversidade cultural. // Guimarães, J. A. C.; Dodebei, V. L. D. L. M. (orgs.). Organização do conhecimento e diversidade cultural. Marília: ISKO-Brasil, 2015. 13-19.

Guimarães, J, A. C. (2017). Slanted knowledge organization as a new ethical perspective. In: Andersen, Jack; Skouvig, L. (orgs.). The organization of knowledge caught between global structures and local meaning. Bingley: Emerald Publishing, 2017. 87-102.

Hall, S. (1997). A centralidade da cultura: notas sobre as revoluções culturais do nosso tempo. // Thompson, K. Media and Cultural Regulation. Inglaterra: Educação & Realidade, 1997. http://www.gpef.fe.usp.br/teses/agenda_2011_02.pdf (2021-08-04).

Herrera López, S. (2006). Sobre las formas de clasificación en Durkheim y Bordieu. // Voces y contextos. I:II (2006) 1-18.

Hjørland, B. (2006). Intellectual organization of knowledge. 2006. http://arkiv.iva.ku.dk/kolifeboat/CONCEPTS/intellectual_organization_of_knowledge.htm (2021-08-09).

Hjørland, B. (2007). Social organization of knowledge. 2007. http://arkiv.iva.ku.dk/kolifeboat/CONCEPTS/social_organization_of_knowledge.htm (2021-08-09).

Hjørland, B.; Pedersen, K. N. (2005) A substantive theory of classification for information retrieval // Journal of Documentation. 61:5 (2008) 582-597.

Hjørland, B. Semantics and knowledge organization. // Annual Review of Information Science and Technology. 41:1 (2008) 367-405.

Hudon, M. (1997). Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. // Knowledge Organization. 24:2 (1997) 84-91.

International Organization for Standardization. (2011) ISO 25964: information and documentation: thesauri and interoperability with other vocabularies - part 1: thesauri for information retrieval. Genebra.

Lara, M. L. G.; Mendes, L. C. (2017) Referências socioculturais nos sistemas de organização do conhecimento. // Iris: informação, memória e tecnología. 3:n.esp. (2017) 26-44.

Laraia, R. S. (2001). Cultura: um conceito antropológico. Rio de Janeiro: Jorge Zahar Ed., 2001.

Mai, J.-E. The modernity of classification. // Journal of documentation. 67:4 (2011) 710-730.

Mazzocchi, F. (2018). Knowledge organization systems (KOS): an introductory critical account. // Knowledge Organization. 45:1 (2018) 54-78.

Moreira, W. (2018). Sistemas de organização do conhecimento: aspectos teóricos, conceituais e metodológicos. Tese (Livre-docência em Sistemas de Organização do Conhecimento) – Universidade Estadual Paulista, Marília, 2018.

Needham, R. (2009). Introduction. // Durkheim, E.; Mauss, M. Primitive classification. Tradução de: Rodney Needham. London: Cohen & West, 2009. viii-xxxii.

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

Orlandi, E. P. (2020). Análise de discurso: principios e procedimentos. Campinas: Pontes Editora, 2020.

Shera, J, H. (1959). What lies ahead in classification. In: Eaton, T.; Strout, D. E. (Eds). The role of classification in the modern american library: papers presented at an institute conducted by the University of Illinois Graduate School of Library Science, November 1-4, 1959. Michigan: Edward Brothers, 1959. 116-128.

Siqueira, J, C. O. (2010). O conceito de classificação: uma abordagem histórica e epistemológica. // Revista Brasileira de Biblioteconomia e Documentação. 6:1 (Jan./Jun. 2010) 37-49.

Smiraglia, R. P. (2012). Epistemology of domain analysis. // Smiraglia, R. P.; Lee, H.-L. (eds.). Cultural frames of knowledge. Wurzburg: Verlag, 2012. 111-124.

Smiraglia, R. P. (2014). The elements of knowledge organization. Cham: Springer, 2014.

Szostak, R.; Gnoli, C.; López-Huertas, M. (2016). Interdisciplinary knowledge organization. Cham: Springer, 2016.

Whitley, R. R. (1984). The intellectual and social organization of the sciences. Oxford: Oxford University Press, 1984.

Williams, R. (2015). Recursos da esperança: cultura, democracia, socialismo. São Paulo: Unesp, 2015.

Enviado: 2022-04-12. Segunda versión: 2022-07-07.
Aceptado: 2022-10-27.

Classificação arquivística: a perspectiva da metodologia funcional vinculada ao tipo documental

Clasificación archivística: la perspectiva de la metodología funcional vinculada al tipo documental

Archival classification: a functional methodology perspective linked to document type

Fernanda BOUTH PINTO (1,2), Clarissa MOREIRA DOS SANTOS SCHMIDT (2)

(1) Instituto Nacional de Infectologia Evandro Chagas/Fundação Oswaldo Cruz, PPGCI/UFF. Avenida Brasil, nº 4365, INI, sala 103, Manguinhos, Rio de Janeiro, nandabouth@yahoo.com.br (2) Universidade Federal Fluminense , Rua Lara Vilela, 126, São Domingos, Niterói, Rio de Janeiro. E-mail: clarissaschmidt@id.uff.br

Resumen

La clasificación archivística tiene como propósito, en primer lugar, la representación del contexto en el que se producen los documentos, con el objetivo de demostrar las razones de su producción y no su contenido. Como operación esencial para la gestión documental, se materializa en instrumentos de gestión como los planes de clasificación, los cuales deben reflejar las funciones y actividades que desarrollan las instituciones, así como los documentos relacionados. En vista de ello, se discute la importancia de la clasificación mediante la metodología funcional ligada al tipo de documento. Se toman como referencia los conceptos de especie y tipo documental previstos por Heloisa Bellotto. Se constata la necesidad de establecer instrumentos de clasificación archivística que representen no solo las acciones institucionales sino también sus documentos, ante la necesidad de preservar el contexto de producción a lo largo del tiempo, fundamentalmente en entornos digitales.

Palabras clave: Classificación arquivística. Metodología funcional. Tipo documental. Gestión de documentos. Tipología documental.

Abstract

The archival classification aims, firstly, at the representation of the context in which the documents are produced, aiming to demonstrate the reasons for their production and not their content. As an essential operation for document management, it is materialized in management instruments such as classification plans, which must reflect the functions and activities developed by the institutions, as well as related documents. In view of this, this work discusses the importance of classification by the functional methodology linked to the document type. It takes as reference the concepts of species and documental type envisaged by Heloisa Bellotto. We reinforce the need to establish archival classification instruments that represent not only institutional actions but also their documents, in view of the need to maintain the context of production over time, fundamentally in digital environments.

Keywords: Archival classification. Functional methodology. Document type. Document management. Document typology.

1. Introdução

A classificação de documentos de arquivos pode ser considerada a base da gestão de documentos e está diretamente ligada à questão da organização de arquivos. Entende-se que deve ser realizada a partir do momento da produção do documento, levando em consideração a estrutura da instituição, suas funções, atividades e também seus documentos.

Segundo Maria Mata Caravaca (2017, p. 19),

> A records classification scheme, also known as a record plan, is a diagram or chart composed of abstract partitions, categories or classes, which aims to logically organize the records created and maintained by an institution. Classification schemes often categorize the creator's records by hierarchical classes (from general to specific), which are uniquely identified by a coding system. Generally, classification schemes are integrated with file plans, which identify the types of files (by business, activity, natural or legal person) to be created within the abstract scheme of classes, including information about file naming and arrangement.

A classificação, portanto, de acordo com a autora, visa organizar logicamente os registros criados e mantidos por uma instituição.

É possível afirmar que a literatura do campo dos arquivos registra diversos modelos de planos de classificação e de parâmetros conceituais para identificar os órgãos produtores, fato que explicita a ausência de padronização de procedimentos metodológicos para classificar os documentos. Insere-se nessa discussão a falta de concordância entre os teóricos sobre quais e quantos níveis são necessários para reconhecer os elementos que caracterizam a hierarquia da ação

propulsora do documento, os termos para qualificá-los, bem como a necessidade ou não em representar os documentos.

Na esteira dessa discussão, Heredia Herrera (2013, p. 145) questiona se existe um quadro geral de classificação arquivística, pois se cada instituição é única, com missão e competências únicas, apesar da possibilidade de existirem funções das atividades-meio em comum entre instituições diferentes, o plano de classificação deve ser específico para a realidade do órgão produtor (Heredia Herrera, 2013, p. 157, tradução nossa).

> (...) os documentos nascem organicamente cumprindo suas funções administrativas, estimando tal realidade como um processo natural, de tal maneira que o a rquivista integrará os documentos dentro das classes ou grupos que já estão determinados pela mesma atividade do organismo de onde procedem.

A autora esclarece, ainda, que na classificação de documentos, o princípio da proveniência determina a classificação, ou seja, a relação entre estes conceitos está colocada através da estrutura orgânica da instituição, onde são produzidos os documentos (Heredia Herrera, 2013, p. 150). Torna-se importante considerar a existência das proveniências estrutural e funcional, tendo em vista que a perspectiva de tipo documental preconizada por Bellotto (2018), como veremos adiante, conjuga ambas as perspectivas ao antecipar a espécie ao tipo documental.

Importante esclarecer que, independente de tratarmos sobre o documento em suporte papel ou já num ambiente digital, a definição de procedimentos e o estabelecimento de requisitos para a classificação funcional não devem ser alterados, respeitando-se apenas a montagem de sistemas informatizados de gestão de documentos, no caso dos documentos arquivísticos digitais.

Para além das questões a respeito do princípio da proveniência, ressalta-se a importância da construção de instrumentos de gestão que demonstrem a representação do documento, retratando, assim, os vínculos entre as funções e atividades e seus tipos documentais correlatos.

2. Metodologia e justificativa

Este artigo teve como procedimentos metodológicos a revisão de literatura em relação ao tema da classificação funcional, os conceitos de espécie e tipo documental e as discussões teóricas que contemplam as questões sobre o plano de classificação que vincula o tipo documental, abordando além dos documentos convencionais, os digitais.

A principal justificativa para esta investigação concentra-se na tentativa de refletir a respeito da classificação de documentos de arquivo pela metodologia funcional que perspectiva a vinculação com o tipo documental. Com este trabalho, espera-se contribuir para que as reflexões acerca da classificação de documentos sejam realizadas com o conhecimento adequado e que não haja dificuldades para o classificador nesta tarefa arquivística. Além disso, objetiva-se também atribuir ao tipo documental um lugar de destaque nos instrumentos de classificação, considerando sua capacidade em representar e materializar a atividade imediata da produção documental.

Assim, a relevância desta pesquisa reside na ampliação do diálogo e discussões na comunidade arquivística a respeito da abordagem da classificação funcional vinculada ao tipo documental, entendendo-se como fundamental a visualização do documento no plano de classificação, visto que seu entendimento enquanto uma representação facilita a organização de arquivos e a compreensão de sua capacidade probatória.

3. Classificação funcional

A classificação funcional, no âmbito deste trabalho, é considerada como parte integrante de um programa de gestão de documentos e permite a representação do contexto de produção documental e, logo, da organicidade, de acordo com o que preconiza o princípio da proveniência. Assim, é possível entender como o documento foi produzido, por quê e para quê.

Na teoria arquivística, a função classificação é destacada como indispensável à gestão dos documentos, já que visa, por meio do plano de classificação, ser um elo entre os tipos documentais e as necessidades burocráticas para a tomada de decisões da administração. Sendo uma etapa desta gestão, deverá estabelecer a imagem do contexto onde são produzidos os documentos, independente se convencionais ou digitais.

O conceito de gestão de documentos que utilizaremos neste trabalho é o apresentado pela Norma International Organization for Standardization 15489:2001, definida como “o campo da gestão responsável pelo controle eficiente e sistemático da criação, recepção, manutenção, uso e destinação de documentos, incluindo processos para captura e manutenção da evidência de informação sobre atividades empresariais e transações na forma de registros documentais”.

No cenário da gestão de documentos, o processo de classificação deve ocorrer na fase corrente, quando os documentos são produzidos (Mokhtar

et al. 2016). Independente do suporte em que as informações são registradas, o primeiro processo a ser estabelecido é o de sua produção. Veremos que o estudo do contexto, tanto de produção do documento (o que num ambiente digital destaca-se seus atributos e metadados para inserção num sistema) quanto da visualização das hierarquias de funções, facilita a classificação numa abordagem funcional.

Mokhtar et al. (2016, p. 626) defendem a classificação baseada em função (functional-based classification) e elencam uma série de outros autores e projetos com a mesma visão, por ser mais estável quando comparada à classificação por assuntos:

> This study adopts function-based classification (FBC) as suggested by Orr (2005), the National Archives of Australia (2003), and Mitchell (2003) because it is more stable compared with subject-based classification. FBC could also ease the process of classification and retrieval; provides context for records rather than content (Robinson, 1999; International Standard Organization, 2001; Library and Archives Canada 2006; National Archives of Australia, 2003; Shepherd & Yeo 2003) and could aid appraisal and disposal activities and support the proactive management of records (Bantin, 2002; National Archives of Australia, 2003).

O Records Classification Model – FRCM, proposto por Mokhtar et al.(2016) e que tem como ponto de partida a produção dos documentos, pretende assegurar que estes sejam concebidos no contexto funcional, mantendo seus metadados e vínculos. Nesse sentido, pensando num sistema informatizado de gestão, a classificação arquivística requer o estabelecimento de métodos de captura de documentos, elementos de metadados e estrutura tecnológica.

Já Fiorella Foscarini (2010), no artigo "A Classificação de documentos baseada em funções: comparação da teoria e da prática", parte do princípio de que um estudo das funções, atividades e transações de qualquer produtor de documentos é um pré-requisito para o correto desenho de sistemas de classificação. Para a autora, existe pouca elaboração teórica sobre o assunto da classificação funcional e aponta uma qualidade desigual nos métodos de classificação tanto europeus, quanto norte-americanos, percebida pela incompreensão dos princípios de classificação por quem elabora tais instrumentos.

Nesse sentido, compartilhamos dessa perspectiva, qual seja, que uma classificação por aproximação funcional está justificada pela própria natureza dos documentos. Embora o plano seja construído com base nas funções e atividades do organismo produtor, não há impedimento de que seja relacionado a ele a estrutura organizacional da instituição, tampouco a representação do documento manifestada no tipo documental.

Ainda acerca da classificação funcional, Schellenberg (2006, p. 83) (1) traz importantes contribuições para a área sobre os princípios de classificação e o tratamento dispensado na administração de documentos correntes.

Ao compreendermos que os documentos de arquivo são o produto de uma ação - ação esta inserida num conjunto de atividades e funções de um órgão- a correta classificação facilitará o processo de avaliação, destinação e/ou recuperação da informação.

Identificando a questão do acesso aos documentos como um problema básico, Schellenberg (2006, p. 83) deixa clara a necessidade dos órgãos manterem seus documentos bem classificados e bem arquivados:

> A classificação é básica à eficiente administração de documentos correntes. Todos os outros aspectos de um programa que vise ao controle de documentos dependem da classificação. Se os documentos são adequadamente classificados, atenderão bem às necessidades das operações correntes (...). Refletirão a função do órgão, no amplo sentido do termo, e, no sentido mais restrito, as operações específicas individuais que integram as atividades do mesmo órgão.

Desta maneira, o autor defende que é possível dispor os documentos em relação às funções, já que sua classificação reflete a organização e a missão do órgão produtor. Por outro lado, a classificação funcional proporciona as bases para a correta avaliação dos documentos, colaborando para a preservação ou destruição dos mesmos.

Renato Tarciso Barbosa de Sousa em sua tese de doutorado (2005), destaca que os elementos fundamentais da estrutura organizacional são as unidades organizacionais. Essas unidades se caracterizam pelo agrupamento de pessoas sob uma autoridade, de acordo com critérios específicos, realizando atividades e tarefas para cumprir determinada função direcionada a elas.

Ainda segundo o autor, existem vários tipos de estrutura organizacional – funcional, clientes, territorial, por processos, etc – o que irá influenciar nas atividades exercidas pela organização, pois as mesmas são agrupadas de acordo com as suas funções.

Por sua vez, as funções dentro da organização têm o objetivo de alcançar a competência maior para a qual foi criada e buscam atingir a missão da instituição. Cada função pode originar um

diferente setor, departamento ou seção, onde estarão presentes diversas atividades análogas e interdependentes para o cumprimento de determinada função. Ainda segundo Sousa (2005), a função tem um caráter duradouro, sem término previsto e viabiliza o alcance da missão da organização.

A decomposição de uma função em diversas atividades pode caracterizar o conceito de atividade: um conjunto de tarefas baseado no consumo de recursos com um objetivo, os procedimentos realizados para a execução de uma função. Para a realização de cada atividade é necessário executar ações ainda menores, não de menor importância, porém mais minuciosas, consideradas tarefas. As tarefas são uma sequência de passos predeterminados, fundamentais para a continuidade do trabalho como um todo (Sousa, 2022).

Cabe afirmar que a produção de tipos documentais se dá dentro da execução das tarefas, que por sua vez, são o cumprimento de atividades. Estas são realizadas de acordo com a função estabelecida, dentro de uma grande competência a ser atingida como objetivo final da existência da instituição (Sousa, 2005). À vista disso, justifica-se a necessidade dos instrumentos de classificação não apenas manifestarem essas ações geradoras dos documentos, quais sejam funções e atividades, como também representarem os produtos destas atividades, que são os documentos.

No contexto dos arquivos, a classificação é entendida a partir da lógica orgânica entre a natureza da ação que gera o documento e a forma a ele conferida. A classificação arquivística demarca a estrutura do produtor do arquivo em suas funções, na totalidade das responsabilidades e das finalidades dessa entidade, e em atividades, enquanto ações referidas nos documentos que as efetivam. (Pinto, 2017). Assim, entendemos que a classificação não deve ser realizada por assuntos, pois dá margens à subjetividade. Além disso, não expressa a atividade que gerou o documento, podendo descontextualizá-lo em relação a seu próprio histórico orgânico-funcional.

Interessante ressaltar aqui a visão de Henttonen e Kettunen (2011) quando, no âmbito de um projeto que pretendeu analisar o sistema de gerenciamento de documentos eletrônicos na Finlândia, buscaram verificar se as classes funcionais seriam mais fáceis de usar. Percorrendo o caminho de defesa da classificação funcional, os autores afirmam que esta abordagem é o núcleo de um AMS (abreviação de “arkistonmuodostussuunnitelma”, ou seja, uma combinação de esquema de classificação funcional, tabela de temporalidade e plano de arquivo).

Os autores finlandeses declaram, ainda, que diversas autoridades arquivísticas, tais como o Archives New Zealand, 2005; Arkistolaitos, 2008; International Council on Archives and Australasian Digital Records Initiative, 2008; DLM-Forum, 2008; International Organization for Standardization, 2001, defendem que a classificação funcional é a melhor metodologia para a gestão de documentos. E complementam (Henttonen & Kettunen, 2011, p. 88-89):

> Campbell (1941) had already advocated functional classification. Campbell believed that functional classification would be more understandable and easier for a researcher to use and for an arranger to create than a classification based on administrative units. Nevertheless, before the 1990s, records were commonly classified in creating organisations by subject and in archival institutions by organisational provenance. Today, classification schemes are usually functional and based on what an organisation does. Since the 1990s functional classification has been strongly promoted. Recent records management textbooks in the UK and Australia promote functional classification as the only or main means of classifying records (Orr, 2005).

Na classificação de documentos de arquivo classifica-se o contexto e não o conteúdo do documento, ou seja, não há representação temática do documento, mas sim contextual. De acordo com o princípio da proveniência e o princípio da ordem original, elementos basilares da classificação, o objetivo desta função arquivística é a manutenção da organicidade. Greg Bak (2010, p.63) reforça a ideia de que a classificação funcional está alinhada à teoria arquivística principalmente no que se refere ao conceito do princípio da proveniência, na medida em que também resulta numa avaliação documenal mais eficiente. Neste sentido, entende-se a metodologia funcional para a classificação como mais eficaz, destacando a atividade intelectual que representa o contexto de produção do documento de arquivo. O plano de classificação deve representar a instituição produtora dos documentos, além das funções e atividades que geram o documento, e o próprio tipo documental.

Stuart Orr (2009), em seu artigo “Functions-based classification of records: is it functional?”, examina tanto a teoria quanto a prática da classificação funcional em arquivos com o objetivo de identificar se esta é uma abordagem aplicável à classificação de documentos. O autor discorre por uma série de apontamentos e visões de teóricos como Schellenberg, Campbell e Bearman em relação à viabilidade desta metodologia e entende que, apesar de ajustes necessários ao

longo do processo, a classificação por funções traz muitos benefícios para uma gestão eficaz e para o controle dos documentos como provas, principalmente no que tange ao ambiente digital. Orr (2009) ressalta, entre outras, como vantagens da classificação funcional: a estabilidade das funções em comparação a uma abordagem por estrutura organizacional; a melhor compreensão da organização, o que ela é e que documentos ela deve produzir; a facilidade no momento da classificação e na recuperação posterior desses documentos; assim como a garantia de contexto aos documentos e o apoio à avaliação e à possível eliminação.

Ao refletir sobre as múltiplas dimensões dos sistemas de classificação, Alejandro Delgado Gómez (2010) apresenta uma inovação na classificação de documentos, uma vez que propõe classificar tanto as atividades quanto os documentos e seus produtores a partir de diversos pontos de vista simultaneamente. Pode-se pensar que esta possibilidade é facilitada, fundamentalmente, através do ambiente digital.

Os conceitos de relação – fundamental num sistema informatizado, já que tudo sugere relações entre os identificadores – e de função – pois sem as funções nenhuma outra entidade passa a existir – são utilizados por Delgado Gómez (2010, p. 128) para fundamentar os procedimentos de vinculação para que a classificação seja elaborada de acordo com a atividade, com os agrupamentos documentais (ou séries) e os agentes produtores de documentos. E nessa perspectiva, entendemos que os tipos documentais manifestam-se enquanto a representação destes agrupamentos, destas séries.

Baseado na premissa das relações, Delgado Gómez (2010, p. 128) amplia a noção de classificação:

> Nuestra noción de la clasificación, (...), no pasa por poner unas cosas dentro de otras, sino más bien por establecer relaciones de carácter múltiple etre esas cosas, modelo que parece tener más sentido em sistemas electrónicos contemporáneos, y que se puede exportar com facilidad al mundo de los documentos analógicos.

Sob essa mesma ideia, Maria Guercio (2002) explorou o ambiente digital no âmbito de um projeto de desenvolvimento de esquemas de classificação, com o foco em facilitar a interoperabilidade entre os setores do organismo produtor. A autora entende que para alcançar este objetivo, o único meio é a classificação através de uma organização lógica e funcional, qualquer que seja o suporte documental.

Já Maria Mata Caravaca (2017, p. 31), ao afirmar que a construção de esquemas de classificação é praticamente inexplorada no campo arquivístico, carecendo, portanto, de padronização metodológica, reforça que

> Records classification schemes, in which hierarchies and associative relationships can be (pre-) established, are fundamental to effectively manage digital records and constitute organized archives.

Ainda para Caracava (2017) os documentos de um produtor devem ser classificados por classes hierárquicas (do geral ao específico), identificados por códigos. Além disso, é importante ressaltar que a autora atenta para o fato de que, apesar das soluções e avanços dos sistemas informatizados, a classificação continua sendo considerada uma função arquivística essencial no ambiente digital, afirmando que tanto os documentos eletrônicos quanto os analógicos necessitam de um modelo de organização com o objetivo de manterem as relações e contextualização dos documentos.

As relações entre os documentos precisam ser estáveis, e não estabelecidas de forma aleatória, de modo a contribuir para a manutenção do vínculo arquivístico, garantir a prova documental e o significado do documento ao longo do tempo e do espaço. A classificação funcional pode oferecer a garantia para esta estabilidade, principalmente ao representar não apenas as funções e atividades, como também seus documentos correlatos.

4. Espécie e tipo documental

Como discutimos, a classificação dos documentos de arquivo determina o lugar ocupado pelas séries documentais (2) no contexto de produção documental. Sendo assim, percebe-se a organicidade, na medida em que sua materialização no plano de classificação espelha as atividades e funções desenvolvidas pelo órgão. Nessa perspectiva, pode-se entender como o tipo documental foi produzido, por quê e para quê.

Sendo um campo de estudo, "a tipologia documental é a ampliação da Diplomática na direção da gênese documental", conforme coloca Bellotto (2000). Ao trazer a importância do tipo documental para a contextualização nas atribuições, competências, funções e atividades do órgão produtor, é possível enxergar sua relevância nos diversos estudos da área arquivística, uma vez que é a tipologia que define o arquivo, como ressalta Rodrigues (2005, p. 22):

> Reflexo e produto material da ação desenvolvida no processo administrativo, a especificidade do arquivo vem comprovada pela tipologia documental produzida. Se o acesso aos documentos é uma questão que deve ser analisada do ponto de vista das políticas de arquivos, por outro, é possível estudá-la do

ponto de vista da metodologia usada para gestão de documentos.

Conceitos	*Definições*	*Fontes*
Série documental	"Conjunto de documentos produzidos por um mesmo produtor no desenvolvimento de uma mesma função e cuja atuação administrativa foi plasmada em um mesmo tipo documental"	Martín-Palomino y Benito La Torre Merino, 2000, p. 21-22.
	"Unidade por excelência sobre a qual gira a totalidade dos processos documentais na realização de inventários, estabelecimento de prazos de retenção, unidade de referência na descrição documental e controle físico de entrada e saída de documentos nos arquivos"	Sierra Escobar, 2004, p. 51, tradução nossa
	"Conjunto de documentos produzido de maneira contínua no tempo como resultado de uma mesma atividade e regulada por uma norma de procedimento"	Cruz Mundet, 2011, p. 326, tradução nossa
	"Sequência de unidades de um mesmo tipo documental"	Camargo e Bellotto, 1996
Espécie	"A configuração que um documento assume de acordo com a disposição e a natureza das informações nele contidas, obedecendo a fórmulas convencionadas, estabelecidas pelo direito administrativo ou notarial"	Bellotto, 2008
Tipologia documental	"(...) é a ampliação da diplomática em direção da gênese documental, perseguindo a contextualização nas atribuições, competências, funções e atividades da entidade geradora/acumuladora". "(...) o objeto da tipologia é a lógica orgânica dos conjuntos documentais. Utiliza-se a mesma construção diplomática para assinalar o registro do que se quer dispor ou do que já foi cumprido com a mesma função. Por isso mesmo, a tipologia pode ser chamada de diplomática arquivística ou, melhor ainda, de diplomática contemporânea"	Bellotto, 2008
Tipo documental	"Configuração que assume uma espécie documental de acordo com a atividade que a gerou"	Camargo e Bellotto, 1996

Tabela I.

Neste sentido, "o tipo [documental] é a configuração da espécie documental de acordo com a atividade que a gerou" (Camargo; Bellotto, 1996). Se o tipo documental corresponde a uma atividade administrativa, acaba por assumir sua coletividade dentro da estrutura organizacional correspondente, de forma que sua denominação se encontrará sempre ligada à espécie relacionada a esta atividade. A espécie é, portanto, a "configuração que um documento assume de acordo com a disposição e a natureza das informações nele contidas", obedecendo a fórmulas convencionadas, estabelecidas pelo direito administrativo ou notarial (Bellotto, 2008).

Cumpre destacar que esta abordagem sobre tipo documental que vinculamos à classificação funcional é proposta por Heloísa Bellotto (2018) quando avança nas reflexões de arquivistas espanhóis (Heredia Herrera, 1991, 2007; Duplá Del Moral, 1997; Fugueras, 2003) sobre a espécie/tipo documental, isto é, uma perspectiva da tipologia documental para a Arquivologia pautada na Diplomática, a chamada Diplomática contemporânea.

Bellotto vai além ao propor o tipo documental como uma extensão da espécie com a atividade que gerou o documento. Deste modo, nosso entendimento a respeito do tipo documental está em consonância ao colocado por esta autora, afinal, o tipo documental é o encontro da espécie carregada da função que produziu o documento.

Após a determinação dos tipos documentais, de acordo com os princípios explicitados, podem-se estabelecer as séries documentais, já que as mesmas refletem o conjunto de tipos que retratam a mesma atividade. Para que haja a normalização da produção documental, é fundamental que a classificação seja realizada previamente a esta produção, pois como coloca Heredia Herrera (1999), ela permite normalizar a criação de espécies documentais sem possibilitar que se gerem ou se reproduzam documentos desnecessários à organização, impulsionando a gestão de documentos.

Os tipos documentais formam as séries documentais próprias de cada órgão produtor porque possuem igual modo de produção, de tramitação e de resolução final do procedimento que lhe deu origem no contexto das atribuições (competências, funções, atividades e tarefas) desempenhadas por um órgão administrativo. A partir do reconhecimento do tipo, se forma a série documental, definida "como a sequência de unidades de um mesmo tipo documental" (Camargo & Bellotto, 1996). Portanto, a denominação da série documental obedece à fórmula do tipo: espécie + atividade (verbo + objeto da ação), sob a qual incide os critérios de Classificação, avaliação, descrição e planejamento de produção de documentos (Rodrigues, 2008).

Estas reflexões tornam-se fundamentais para o entendimento dos conceitos de série, espécie e tipo documental e para uma aplicação na elaboração de plano de classificação funcional vinculado ao tipo documental.

Bellotto refere-se ao estudo da gênese documental para a identificação de documentos e destaca que "[...] os estudos da Diplomática e tipologia levam a entender o documento desde o seu nascedouro, a compreender o porquê e o como ele é estruturado no momento de sua produção" (Bellotto, 2006, p.45). A autora reforça que a identificação dos documentos deve ser compreendida à luz do contexto em que foi produzido, não sendo possível dissociar a diagramação e a construção material do documento do seu contexto jurídico-administrativo de gênese, produção e aplicação.

5. Relações entre função e tipo documental

Ao relacionar o aporte teórico da classificação funcional e sua importância para a gestão de documentos, com o entendimento de alguns autores como Henttonen e Kettunen (2011) e Mokhtar et al.(2016) a respeito da classificação funcional vinculada ao documento, foi possível identificar que o tipo documental explícito no plano de classificação torna o vínculo com suas funções e atividades mais claros.

O tipo documental - considerado como um patamar representado no plano de classificação funcional - está relacionado à perspectiva de Heloísa Bellotto, a qual reconhece que a classificação funcional reflete as ações no documento, sendo diferente das demais perspectivas apresentadas no campo dos arquivos.

É através do estudo do contexto de produção documental que será possível estabelecer o vínculo arquivístico, ligação entre as funções e atividades de um órgão produtor e os documentos que ele produz. Neste sentido, é na produção do documento que podemos identificar as unidades documentais produzidas, de modo que, após a fixação das séries documentais, são facilitados os trabalhos de classificação dos documentos institucionais.

A partir da elaboração de um plano de classificação funcional vinculado ao tipo documental, há a possiblidade dos produtores dos documentos reconhecerem seus documentos com mais objetividade e relacionarem tais documentos às suas atividades e funções desempenhadas no âmbito da competência maior do órgão produtor.

Maria Guercio (2001) reafirma o pensamento de Giorgio Cencetti sobre o fato dos documentos e suas relações serem recíprocas e persistentes, ou seja, entende que o vínculo arquivístico pressupõe tanto a imparcialidade dos documentos (o fato dos documentos serem acumulados como instrumentos essenciais de atividades práticas e para fins de disposição e utilização), quanto a autenticidade dos documentos (a necessidade real de autodocumentação do produtor, organizando os documentos para garantir sua fiabilidade).

Reforçamos que a classificação funcional vinculada ao tipo documental, atende melhor à manutenção do contexto de produção ao longo do tempo, fundamentalmente em ambientes digitais, visto que não apenas as ações institucionais estarão representadas no instrumento, mas também seus documentos.

Se o documento é produto de uma ação, nasce para registrá-la a ponto de servir como prova desta, é assim que deve ser a classificação: através das atividades e funções que lhes deram origem. Ao classificar os documentos observando as atividades e funções do órgão produtor, compreende-se a organicidade, ficando claras as ações que geram os documentos.

Apresentamos os exemplos abaixo de espécies e tipos documentais preconizados por Heloísa Bellotto de modo que podemos melhor visualizar melhor sua perspectiva.

Espécie	*Tipo Documental*
Livro	Livro de registro de entrega das amostras
Formulário	Formulário de controle de aquisição de instrumentos de calibração
Formulário	Formulário de fiscalização diária de contrato de terceirização de dietas hospitalares
Ata	Ata de reunião do Conselho Deliberativo
Prescrição	Prescrição de nutrição enteral
Prescrição	Prescrição de medicamentos controlados
Relatório	Relatório de atividades semestral para o Comitê de Ética em Pesquisa

Tabela II. Espécies e tipos documentais (elaboração das autoras com base em Bellotto, 2018)

Se o tipo documental corresponde a uma atividade administrativa, acaba por assumir sua coletividade dentro da estrutura organizacional correspondente, de forma que sua denominação se encontrará sempre ligada à espécie relacionada a esta atividade (Bellotto, 2018, p. 449, tradução nossa):

> A espécie documental, quando acompanhada de uma atividade, é um tipo [documental]. (...) A espécie é como uma fórmula vazia que se torna um tipo

> quando, no momento da gênese documental, adicionamos a atividade, lhe agregamos algo e lhe damos vida. A atividade seria como a razão funcional, seria a espécie em funcionamento.

Atribuir maior importância ao procedimento administrativo é fundamental para que o contexto de produção documental seja preservado no momento de classificar um documento de acordo com a função e atividade que o gerou. Nesse sentido, Ana Célia Rodrigues (2018, p.437) atenta para o fato de que foi Bellotto que deu um novo enfoque, a partir de 1982 no Brasil, a respeito do referencial teórico sobre a tipologia documental, trazendo para os estudos de diplomática a diferença entre espécie documental (objeto da Diplomática) e o tipo documental (objeto da Arquivística, da tipologia documental), algo ainda inovador na área.

6. Considerações finais

Após a análise feita nesta revisão de literatura, por que então consideramos importante identificar os metadados presentes no tipo documental? São diversas as vantagens de se identificar os elementos do tipo documental, a saber:

- Possibilidade do acesso à informação de uma maneira mais ampla, devido ao conhecimento de diversos metadados presentes no tipo documental, atendendo ao preconizado na LAI (Lei de Acesso à Informação, Lei nº 12.527, de 18 de novembro de 2011);
- Classificação dos dados de forma mais consistente quanto às informações pessoais, no que tange à LGPD (Lei Geral de Proteção de Dados, Lei nº 13.709, de 14 de agosto de 2018), e consequentemente, maior proteção dos dados presentes nos tipos documentais;
- Maior rigor no que se refere à classificação da natureza do assunto, ou seja, se o tipo documental traz informação ostensiva ou sigilosa.

A classificação dos documentos de arquivo pela metodologia funcional determina o lugar ocupado pelos tipos documentais no contexto de produção, isto é, a organicidade. Na medida em que a materialização do tipo documental é evidenciada no plano de classificação, é possível enxergar as atividades e funções desenvolvidas pelo órgão produtor.

A classificação funcional aliada à representação do tipo documental no plano de classificação, demonstrando o contexto de produção de documentos, pode ser apropriada pela organização de arquivos na medida em que facilita, inclusive, o acesso mais efetivo aos arquivos. As funções e atividades do produtor, levantadas pela metodologia funcional e registradas até o nível do tipo documental, favorecem a visualização do contexto tanto dos documentos produzidos quanto do produtor como um todo. Além do acesso aos documentos ser facilitado, todo o processo de avaliação documental e tomada de decisões é melhor realizado. Ademais, torna-se possível uma gestão do conteúdo informacional dos tipos documentais, favorecendo o controle de dados e informações pessoais e sigilosas.

Portanto, a partir das análises desta investigação, consideramos que a classificação funcional vinculada ao do tipo documental corrobora tanto para a manutenção do vínculo arquivístico quanto para o reconhecimento dos documentos por seu produtor, sejam documentos em papel ou num ambiente digital.

Notas

(1) A obra “Arquivos Modernos: princípios e técnicas” de Theodore Roosevelt Schellenberg foi publicada originalmente no ano de 1956, sendo a 1ª edição publicada no Brasil em 1973. Neste trabalho, optou-se por usar a 6ª edição de 2006, traduzida por Nilza Teixeira Soares.

(2) De acordo com Camargo e Bellotto (1996), série é a sequência de unidades de um mesmo tipo documental.

Referências

Bak, Greg (2010). La clasificación de documentos electrónicos: documentando relaciones entre documentos. // Tabula: Revista de Archivos de Castilla y León/Asociación de Archiveros de Castilla y León, Salamanca.13, 59-77.

Bellotto, Heloísa Liberalli (2000). Como fazer análise diplomática e análise tipológica em arquivística; reconhecendo e utilizando o documento de arquivo. São Paulo: Associação de Arquivistas de São Paulo; Arquivo do Estado. (Projeto Como Fazer)

Bellotto, Heloísa Liberalli (2008). Diplomática e tipologia documental em arquivos. 2 ed. Brasília, DF: Briquet de Lemos/ Livros.

Bellotto, Heloísa Liberalli (2018). Concepto de especie documental como antecedente al tipo en la teoria archivística. // Boletín ANABAD. LXVIII, Núm. 3-4, jul/dez. Madrid, 446-455.

Bellotto, Heloísa Liberalli (2022). O entendimento da espécie e do tipo documentais na teoria e na prática arquivísticas. // Officina: Revista da Associação de Arquivistas de São Paulo, São Paulo. 1:1, 09-16.

Brasil. Conselho Nacional de Arquivos (2001). Classificação, temporalidade e destinação de documentos de arquivo relativos às atividades-meio da administração pública. Rio de Janeiro.

Camargo, Ana Maria de Almeida; Bellotto, Heloísa Liberalli (Coord.) (1996). Dicionário de terminologia arquivística. São Paulo: Associação dos Arquivistas Brasileiros, Núcleo Regional de São Paulo; Secretaria de Estado da Cultura. 142 p.

Cruz Mundet, José Ramón (2011). Diccionario de Archivística: (con equivalencias en inglés, francés, alemán, portugués, catalán, euskera y gallego). Madrid: Alianza, 363 págs. ISBN: 978-84-206-5285-6.

Delgado Gómez, Alejandro (2010). Sistemas de clasificación en múltiples dimensiones: la experiencia del Archivo Municipal de Cartagena. // Tabula: Revista de Archivos de Castilla y León. 13, 125-136.

Dicionário brasileiro de terminologia arquivística (2005). Rio de Janeiro, Arquivo Nacional.

Duplá Del Moral, Ana (1997). Manual de archivos de oficina para gestores. Madrid: Marcial Pons Ediciones AS.

Foscarini, Fiorella (2010). La clasificación de documentos basada en funciones: comparación de la teoria e y la práctica. Tabula: Revista de Archivos de Castilla y León. // Asociación de Archiveros de Castilla y León, Salamanca.13, 41-57.

Foscarini, Fiorella (2010). Records Classification and Functions: An Archival Perspective. Knowledge Organization, 33(4) 188-198. 38 references.

Fugueras, Ramon Alberch I (2003). Los archivos entre la memoria historica y la sociedad del conocimiento.

Guercio, Maria (2001). Principles, Methods, and Instruments for the Creation, Preservation, and Use of Archival Records in the Digital Environment. // The American Archivist. 64, 238-269.

Guercio, Maria (2002). Records classification and content management: old functions and new requirements in the legislations and standards for electronic record-keeping systems. European Archives News – Insar, Barcelona, 432-442.

Guercio, Maria; Carloni, Cecilia (2015). The research archives in the digital environment. JLIS.it. 6:1. https://doi.org/10.4403/jlis.it-10989.

Henttonen, Pekka; Kettunnen, Kimmo (2011), Functional classification of records and organizational structure. // Records Management Journal. 21:2, 86-103. http://dx.doi.org/10.1108/09565691111152035

Heredia Herrera, Antonia (1991). Archivistica. Estudios básicos. Sevilla.

Heredia Herrera, Antonia (1995). La Norma ISAD (G) y su terminología. Análisis y alternativas, Madrid, ANABAD.

Heredia Herrera, Antonia (1999). La identificación y la valoración documentales en la Gestión Administrativa de las Instituciones Públicas. http://www.anabad.org/boletinpdf/pdf/XLIX(1999)_1_19.pdf (2022-01-28).

Heredia Herrera, Antonia (2007). En torno al tipo documental. Arquivo & Administração. 6:2. http://hdl.handle.net/20.500.11959/brapci/51509 (2022-01-31).

Heredia Herrera, Antonia (2013). Manual de archivística básica: gestión y sistemas. México: Benemérita Universidad Autónoma de Puebla.

International Standard ISO 15489-1 (2001). Information and documentation – records management Part 1: General.

International Organization for Standardization (2007), ISO 23081-2. Information and Documentation. Records Management Processes. Metadata for Records. Part 2. Conceptual and Implementation Issues, ISO, Bon.

La Torre Merino, José Luiz y Martín-Palomino y Benito, Mercedes (2000). Metodología para la identificación y valoración de fondos documentales. Madrid: Ministerio de Educación, Cultura y Deportes; S.G. de Información y Publicaciones. (Escuela Iberoamericana de Archivos: experiencias y materiales).

Mata Caravaca, Maria (2017). Elements and Relationships within a records classification scheme. // JLIS.it 8:2, 19-33. http://doi.org/10.4403/jlis.it-12374.

Mokhtar, Umi Asma. Yusof, Zawiyah M. Ahmad, Kamsuriah. Jambari, Dian Indrayani (2016). Development of function-based classification model for electronic records. // International Journal of Information Management. 36: 626–634.

Orr, Stuart A (2009). Functions-Based Classification of Records: Is it Functional?. UK: Northumbia University.

Pinto, Fernanda Bouth (2017). Plano de classificação por assunto ou funcional: análise de metodologias e equivalências para classificação de documentos de arquivo no Instituto Nacional de Infectologia Evandro Chagas. 183 f. Dissertação (Mestrado em Gestão de Documentos e Arquivos) - Programa de Pós-Graduação em Gestão de Documentos e Arquivos, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro.

Rodrigues, Ana Célia (2005). Tipologia documental como parâmetro de classificação e avaliação em arquivos municipais. // Cadernos de Estudos Municipais. Universidade do Minho (Portugal). 17/18, 11-46.

Rodrigues, Ana Célia (2008). Diplomática contemporânea como fundamento metodológico da identificação de tipologia documental em arquivos. São Paulo: Universidade de São Paulo. (Tese de Doutorado). http://www.teses.usp.br (9009-10-15).

Rodrigues, Ana Célia (2018). Tipología documental: diálogos entre la archivística y la diplomática para la construcción del método de identificación del documento de archivo. Boletín ANABAD. 68:3-4.

Schellenberg, T. R. (1980). Documentos públicos e privados: arranjo e descrição. Rio de Janeiro: FGV.

Schellenberg, T. R. (2006). Arquivos modernos: princípios e técnicas. 6.ed. Rio de Janeiro: Ed. da FGV.

Sierra Escobar, Luis Fernando (2004). Como identificar y denominar uma serie documental: propuesta metodológica. // Biblios. 5:20.

Sousa, Renato Tarciso Barbosa de (2005). Classificação em Arquivística: trajetória e apropriação de um conceito. São Paulo: Tese (Doutorado) – Escola de Comunicações e Artes – Universidade de São Paulo.

Sousa, Renato Tarciso Barbosa de (2022). A classificação funcional de documentos de arquivo é uma abstração intelectual ou um instrumento prático? // Acervo. 35:2, 1-21, 5 abr.

Yusof, Zawiyah M. and Mokhtar, Umi Asma. (2015) Records and Information Management: The Requirement for Functional Classification. Open Journal of Social Sciences. 3, 215-218. http://dx.doi.org/10.4236/jss.2015.33032.

Enviado: 2022-03-29. Segunda versión: 2022-09-16.
Aceptado: 2022-10-27.

La educación de posgrado desde la gestión del conocimiento: estudio de caso en la Universidad de las Ciencias Informáticas de Cuba

Postgraduate Education from knowledge management: case study at the University of Informatics Sciences of Cuba

Eylin HERNÁNDEZ-LUQUE, Vivian ESTRADA-SENTÍ, Miguel Ángel HERNÁNDEZ-DE LA ROSA

Universidad de las Ciencias Informáticas de Cuba, Cuba, eluque2005@gmail.com

Resumen

La creación, apropiación y socialización del conocimiento son claves en las actividades investigativas y científicas que se desarrollan en la Educación de Posgrado. El objetivo de este artículo es identificar los aspectos que permitan evaluar la capacidad, las condiciones tecnológicas y el saber del posgrado, desde la comprensión del proceso de gestión del conocimiento, su importancia, intención, los beneficios, obstáculos y resultados que tiene su aplicación. El análisis documental permitió una caracterización de la Universidad de Ciencias Informáticas de Cuba en cuanto a la aplicación de la gestión del conocimiento en los estudios de posgrado y propone transponer la propuesta de Nonaka & Takeuchi (1998) para potenciarla. Se aplicaron técnicas y procedimientos cuantitativos para las pruebas estadísticas Kaiser-Meyer-Olkin (KMO), la prueba de esfericidad de Bartlett y el test de Alfa de Cronbach, para comprobar la calidad científica de los cuestionarios en términos de fiabilidad y validez de constructo con el software científico de analítica predictiva IBM SPSS v23. Arrojando una buena relación entre las variables, validez de constructo y fiabilidad de los cuestionarios. La muestra estuvo compuesta por 145 estudiantes de posgrado de programas académicos de la Universidad de las Ciencias Informáticas. Se propone generalizar los cuestionarios para áreas de investigación que evalúen la aplicación y necesidad de gestionar conocimiento.

Palabras clave: Educación de posgrado. Gestión del conocimiento. Socialización del conocimiento. Educación superior. Universidad de las Ciencias Informáticas de Cuba. Cuba.

Abstract

The creation, appropriation and socialization of knowledge are key in the research and scientific activities that are developed in Postgraduate Education. The objective of this article is to identify the aspects that allow to evaluate the capacity, the technological conditions and the knowledge of the postgraduate, from the understanding of the knowledge management process, its importance, intention, the benefits, obstacles and results that its application has. The documentary analysis allowed a characterization of the University of Informatics Sciences of Cuba in terms of the application of knowledge management in postgraduate studies and proposes to transpose the proposal of Nonaka & Takeuchi (1998) to enhance it. Quantitative techniques and procedures were applied for the Kaiser-Meyer-Olkin (KMO) statistical tests, Bartlett's sphericity test and Cronbach's Alpha test, to verify the scientific quality of the questionnaires in terms of reliability and construct validity with scientific predictive analytics software IBM SPSS v23. Showing a good relationship between the variables, construct validity and reliability of the questionnaires. The sample consisted of 145 postgraduate students from academic programs at the University of Informatics Sciences. It is proposed to generalize the questionnaires for research areas that assess the application and need to manage knowledge.

Keywords: Postgraduate education. Knowledge management. Socialization of knowledge. Higher education. University of Informatics Sciences of Cuba. Cuba.

1. Introducción: gestión del conocimiento y educación de posgrado

1.1. Conocimiento y gestión del conocimiento

El conocimiento para Guzón (2018) es socialmente relevante, porque está al servicio de la solución de problemas, potencialmente volcado a la innovación, y es capaz de favorecer la creación de competencias para la asimilación y creación de tecnologías y saberes. El proceso de apropiación social del conocimiento es aquel en el que las personas acceden a los beneficios del conocimiento y participan en actividades de producción, transferencia, evaluación, adaptación, aplicación del conocimiento; y se tiene la capacidad social de usar el conocimiento.

Desde su concepción, Nonaka & Takeuchi (1995) refieren que el conocimiento es un recurso clave para el logro de ventajas competitivas, basado en

los sujetos que se forman para dar solución a problemas de carácter social e individual. A criterio de Greiner (2007) el conocimiento tácito es aquel que reside en cada persona, por lo cual, no es posible separarlo de quién lo posee para distribuirlo a otros o almacenarlo en medios físicos; y, por el contrario, el conocimiento explícito puede ser codificado, extraído, almacenado, distribuido, difundido o divulgado.

Teniendo en cuenta la sistematización, el análisis de la bibliografía y los estudios realizados en el desarrollo de la tesis de maestría de esta investigadora (Hernández-Luque et al., 2018), se asume para el desarrollo de esta investigación que el conocimiento es un recurso que tiene cada persona, que se reconstruye, se transforma continuamente, se puede socializar y tiene como base el uso de información para solucionar problemas y estimular la obtención de resultados. Tiene carácter social, porque se apropia en la ejecución de una tarea práctica y en la relación con los demás. Es resultado de una actividad que genera necesidad hacia la obtención del mayor provecho posible; por sí solo no existe, es inherente a las personas; no constituye valor hasta que no se manifieste como resultado. No es estático en el pensamiento, sino dinámico: se desarrolla y se transforma cada vez que el sujeto tiene un nuevo intercambio con el objeto y con los sujetos, junto a los cuales aprende.

El análisis anterior muestra la multiplicidad de definiciones existentes y tipos de conocimientos, lo que hace apropiado y pertinente entender el conocimiento desde la perspectiva de su gestión.

A pesar de que varios autores apuntan que no hay consenso en torno a una definición única de la gestión del conocimiento (GC) (Hislop et al., 2018; Yee et al., 2019), se puede afirmar que busca integrar de manera intencional a las personas, los procesos y las tecnologías, con el objeto de construir e implementar la infraestructura intelectual de la organización (Laal, 2011). Al mismo tiempo, precisan que en el marco referencial de la GC se distinguen las actividades de asimilar, aplicar, compartir, definir, identificar, capturar, organizar la información y el conocimiento.

Teniendo en cuenta los fundamentos de las normativas de la GC —ISO 30401:2018, ISO 9001:2015 (Organización Internacional de Normalización, 2015, 2018; Ricardo, 2021; Simeón-Negrín, 2004)—, se destaca que las organizaciones deben potencializar el conocimiento que ellas tienen y que les permite describir y modelar de manera coherente integrada y articulada todos los componentes de la GC (derivado de las personas, de los procesos y procedimientos documentados y de los procedentes de la organización y su entorno). Es una necesidad conocer lo que está pasando con los nuevos conocimientos, con los cambios de la tecnología, la valoración de las expectativas, los riesgos, las oportunidades de mejoramiento y fortalecimiento, así como la identificación de la pérdida de conocimiento, en las organizaciones. Por lo que, se necesita la implementación de proceso de adquisición, transferencia, retención, manejo de conocimiento. Es criterio de la ISO 30401 (2018) que lo más importante es la promoción de la cultura para la GC en las organizaciones, que no es más que generar espacios y condiciones de trabajo colaborativo y de promoción de la transferencia de conocimiento.

Por otro lado, los autores comparten el criterio de que la GC típicamente distingue entre dos tipos de conocimiento —tácito y explícito— y asume que la GC presenta dos estrategias principales: de codificación y de socialización (Davenport & Holsapple, 2011; Greiner et al., 2007; Nonaka & Konno, 1998; Rodríguez-Montoya & Zerpa-García, 2019). Coincidiendo con Greiner (2007) la estrategia de codificación intenta recolectar el conocimiento producido por los sujetos para categorizarlo, almacenarlo digitalmente, actualizarlo y hacerlo conocido y accesible de manera explícita; mientras que la estrategia de socialización busca implantar los mecanismos necesarios para activar y mantener comunidades de saberes en las cuales los sujetos establezcan vinculaciones de carácter social entre ellos con el propósito de promover y fomentar la comunicación interpersonal y el intercambio de conocimientos, dado que la socialización para el intercambio directo del conocimiento tácito es crítica para la creación de conocimiento.

Se dirige la atención a la espiral evolutiva de conversión del conocimiento y procesos de autotrascendencia de Nonaka & Konno (1998). En esta espiral, las combinaciones de interacciones entre el conocimiento explícito y tácito conducen a cuatro patrones posibles de conversión: socialización, externalización, combinación e internalización.

A partir de la sistematización realizada y el análisis de la bibliografía a (Alavi & Leidner, 2013; Drucker, 1995; Foray & Lundvall, 1998; ISO 30401, 2018; Ikujiro Nonaka & Takeuchi, 1999; Pérez, 2020; Polanyi, 1962; Raneda-Guirriman et al., 2017; Rodríguez-Montoya & Zerpa-García, 2019) se identificó que la GC permite la creación, validación, presentación, distribución y aplicación del conocimiento; y se caracterizan las interrelaciones de conversión del conocimiento: de tácito a explícito y de explícito a tácito, destacando como de especial significación la disposición de los usuarios a compartir conocimiento. El análisis también permite inferir que la capacidad para

aprender es esencial para mantener, renovar y compartir los conocimientos, así como la necesidad de incluir el aprendizaje y su papel en la creación del nuevo conocimiento.

En estos análisis se pudo constatar la existencia de modelos de GC para trabajar las relaciones del conocimiento tácito y del conocimiento explícito y las relaciones derivadas de estas, entre los que se pueden señalar el modelo Socialización, Externalización, Combinación e Internalización (SECI) que aborda el enfoque de la espiral del conocimiento tácito y explícito (Nonaka & Takeuchi, 1995), así como el modelo de Capital intelectual (CI) en el cual se combinan los enfoques de las teorías cognitivas de la organización, del capital intelectual y la economía basada en el conocimiento (Gómez-Bayona et al., 2019). Otro modelo analizado es el modelo de CI B&O (Barboza & Ochoa, 2016) desarrollado a partir de la combinación de modelos GC de Nonaka y Takeuchi (1995), Choo (1998), CI de Onge (1996) y Bueno (2002). En su propuesta se analiza el capital intelectual como centro principal de la organización con apoyo de los componentes tecnológicos.

Se toma como base el modelo matricial de conversión o transferencia del conocimiento estudiado y referenciado por Rodríguez-Montoya y Zerpa-García (2019), basado en Davenport y Holsapple (2011), Greiner et al. (2007), Nonaka y Konno (1998) e Ikujiro Nonaka y Takeuchi (1999), en el que los conocimientos pasan por un proceso que los transforma de tácitos (contenidos en los sistemas de información, en las bases de datos y en las personas) a explícitos (capturados y almacenados en un formato reutilizable que permite realizar búsquedas) y otra vez en tácitos, lo cual permite que otras personas de la organización puedan aprenderlos y utilizarlos.

Se considera que la Educación de Posgrado debe potenciar la GC mediante acciones dirigidas en cuanto a la transformación del conocimiento (Figura 1):

- Tácito-Tácito: mediante la organización de tutorías, charlas y conferencias.
- Tácito-Explícito: mediante la publicación de artículos, informes, videos, notas técnicas, reportes de desarrollo, libros o fórum digitales que dejan evidencias.
- Explícito-Tácito: se ejecuta cuando el sujeto (el proceso destino) utiliza el conocimiento explícito para generar valor y adquirir experiencia (aprender haciendo). En este subproceso existe un procedimiento independiente para los eventos científicos y publicaciones y las defensas de tesis de maestría y doctorado por su envergadura y por poseer todas las formas de transformación implícitas.
- Explícito-Explícito: se organiza y evidencia a partir de aquellos trabajos explícitos que recogen un conjunto de trabajos anteriores y aunque no generan nuevo conocimiento lo ordenan con un fin que aporta beneficio a la organización como son los documentos de procesos, resoluciones, planes de organización, estrategias de desarrollo, resúmenes de búsquedas bibliográficas y monitoreo tecnológico y competencial.

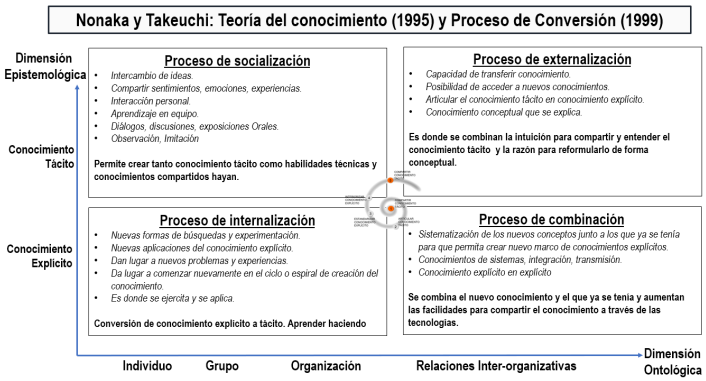

Figura 1. Modelo matricial de conversión y creación del conocimiento (a partir de Nonaka y Tekeuchi, 1995, 1999)

1.2. Los retos de la educación de posgrado

Varios autores, entre los que se destacan, Streck (2015), Luna-Nemecio (2019) y Martínez Aguilera (2018) refieren que la esencia de la educación está en la trilogía del saber ser, saber hacer y saber pensar. Además, señalan que se profundiza más en la formación, el aprendizaje, la evaluación, la investigación, la GC, así también cómo promover el aprendizaje permanente para hacer frente a las altas exigencias y al carácter fluctuante del mercado laboral (Amber & Domingo, 2016; Becker et al., 2017; Ortega-Carbajal et al., 2015; Shujahat et al., 2017; UNESCO, 1995)

La Educación Superior tiene como uno de sus retos asumir el liderazgo social en la creación de conocimientos para abordar retos mundiales, así como el aprendizaje a distancia y el uso de las TIC para ofrecer oportunidades de ampliar el acceso a la educación de calidad (George & Salado, 2019). Además, debe formar a las personas para que sean capaces de transformar la información en conocimiento; y constituye una condición clave tener la capacidad de innovación y transformación de los procesos propios para lograr mayores resultados del aprendizaje.

En este sentido Brew y Saunders (2020) argumentan la necesidad de definir un nuevo tipo de Educación Superior en la que los estudiantes y académicos trabajen progresivamente hacia el desarrollo de comunidades de práctica inclusivas de construcción de conocimiento, donde se promuevan enfoques de la enseñanza "basados en la investigación" y "orientados a la investigación", en los que los estudiantes realicen investigaciones e indagaciones y desarrollen las habilidades y técnicas asociadas.

En el presente siglo XXI, la Educación de Posgrado se inserta en el complejo y contradictorio panorama mundial. Según Sallán (2015) debe considerarse como un contexto de gestión, de realización personal y de promoción del cambio social, orientado siempre a la mejora. Silva (2017) añade que la sociedad basada en el conocimiento es considerada como heredera de la revolución industrial y como la fase más avanzada de la globalización, que ha traído consigo la construcción de un nuevo enfoque o modelo. Se trata de la socioformación, que tiene como finalidad, no solo la formación integral de la persona, sino de cómo ésta contribuye a la sociedad, y considera que el saber es el recurso de mayor cuantía. Por ello, demanda nuevas formas de enseñar y de aprender, siendo necesario un activo y consciente proceso de aprendizaje, es decir, la autogestión de conocimientos (Ministerio de Educación Superior, 2019).

1.3. La gestión del conocimiento como respuesta

Estos enfoques permiten a los autores del presente trabajo compartir el criterio obtenido de la sistematización realizada sobre las investigaciones de (George & Salado, 2019; González & Jover, 2020; Luna-Nemecio et al., 2019; Páez-Suárez et al., 2021; Ponjuán-Dante, 2015; Rodríguez-Montoya & Zerpa-García, 2019; Salazar-Gomez & Tobon, 2018; Sentí et al., 2016; Sentí & Cárdenas, 2010; Simeón-Negrín, 2004; Soto-Balbón & Barrios-Fernández, 2006) sobre la necesidad de conocer las formas de expresión de la GC. Se concluye que en la sociedad basada en el conocimiento, el saber es un recurso esencial, donde tienen que existir nuevos enfoques de enseñar y aprender, fomentando la autogestión del conocimiento. Es precisamente en la Educación de Posgrado en las universidades, donde se contribuye a la formación integral acorde a la evolución de las TIC para enfrentar la sociedad cambiante, fortaleciendo la investigación, la especialización de conocimientos desde la actividad investigativa y la actualización de contenidos, por lo que es una necesidad fortalecer las acciones destinadas a la formación y entrenamiento investigativo y científico.

Desde esta perspectiva y teniendo como base la sistematización, el análisis de la bibliografía y los estudios realizados (Hernández-Luque et al., 2021), es opinión de la autora que en la Educación de Posgrado en las universidades deben potenciarse los programas sustentándolos en las tecnologías y en correspondencia con las exigencias de la sociedad basada en el conocimiento, donde los procesos de innovación, investigación y desarrollo sean protagónicos. Debe acerse teniendo en cuenta las actividades de la GC, el rendimiento de los recursos, el uso de las TIC, la apropiación, creación y socialización del conocimiento, así como la disponibilidad y uso de plataformas y herramientas que fortalezcan la integración y comprensión de buenas prácticas, de modo que:

- Es necesario incorporar herramientas didácticas basadas en plataformas digitales, que potencien la actividad virtual, creando las condiciones para el aprendizaje significativo, para apropiarse del conocimiento y compartir las buenas prácticas adquiridas que le indiquen al posgrado lo que conoce, las condiciones que tiene y la capacidad que presenta para aplicar la GC (Figura 2).

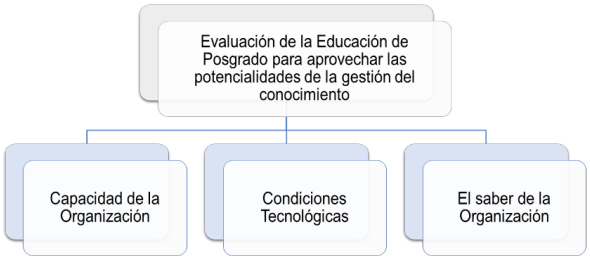

Figura 2. Evaluación de la Educación de Posgrado para aprovechar las potencialidades de la GC (Hernández-Luque et al., 2021)

- Se necesita profundizar en la teoría más específica sobre la evaluación de la GC y sobre instrumentos utilizados para medir procesos iguales o semejantes, creando condiciones para fortalecer el aprendizaje, para conocer y comprender el proceso de la GC, su importancia, su intención, los beneficios que aporta su aplicación, así como los resultados que se pueden obtener y los obstáculos que enfrenta su aplicación, lo que evidencia la necesidad de conocer los requerimientos que se deben cumplir para compartir las buenas prácticas adquiridas y socializar la información y el conocimiento que se adquiere en la Educación de Posgrado (Figura 3).

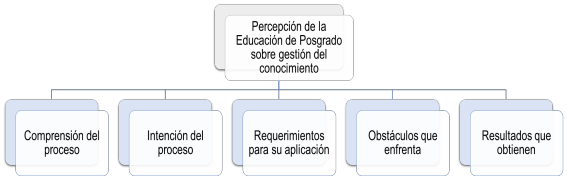

Figura 3. Percepción de la Educación de Posgrado sobre la GC (Luque et al., 2021)

A partir del estudio realizado los autores del artículo asumen que la actividad de posgrado constituye un área de resultados claves en las universidades, que es esencial para la sostenibilidad del desarrollo y socialización del conocimiento. Además, coincide que la transferencia de la tecnología, se considera un factor esencial para la producción y la socialización del conocimiento, que se apropia en todos los espacios de posgrado, pero este camino exige tener en cuenta las características y necesidades profesionales en el contexto universitario.

En las resoluciones cubanas (GOC-2020-381-EX27; GOC-2019-776-O65; GOC-2018-57-EX13) se establecen los procesos de continuidad y evaluación de los estudios de posgrado, así como los reglamentos para Educación de Posgrado en Cuba y los centros autorizados a superación profesional de posgrado.

La Universidad de Ciencias Informáticas de Cuba (UCI) es una universidad con un modelo flexible de centro docente-productor que le permite formar profesionales altamente calificados. Desde la Dirección de Educación de Posgrado en la universidad se establecen los indicadores, se fijan metas, objetivos, se ofrecen oportunidades de superación a través de cursos, pasantías, entrenamientos, diplomados, maestrías, doctorados, constituyendo la formación investigativa un elemento esencial en todos los programas. Además, desarrollan importantes eventos científicos que se aprovechan para el intercambio de conocimientos. La Educación de Posgrado, se fortalece especialmente con la superación profesional y tiene como objetivo la especialización, la reorientación y la actualización permanente de los graduados universitarios para el mejor desempeño de las actividades de formación académica (UCI, 2021). Desde al análisis documental se evidencia la necesidad de identificar los aspectos que permitan evaluar la capacidad, las condiciones tecnológicas y el saber del posgrado, desde la comprensión del proceso de gestión del conocimiento, su importancia, intención, los beneficios, obstáculos y resultados que tiene su aplicación.

Desde esta perspectiva, es opinión de los autores que en la Educación de Posgrado en las universidades se desarrollen programas asertivos adaptados a las necesidades actuales y en correspondencia con las exigencias de la sociedad basada en el conocimiento, donde los procesos de innovación, investigación y desarrollo sean protagónicos, de modo que es necesario incorporar herramientas didácticas basadas en plataformas digitales, que potencien la actividad virtual, creando las condiciones para el aprendizaje, para apropiarse del conocimiento y compartir las buenas prácticas adquiridas que indiquen a los agentes del posgrado lo que conoce, las condiciones que tiene y la capacidad que presenta para aplicar gestión del conocimiento.

A partir del estudio realizado se asume que la actividad de posgrado constituye un área de resultado clave en las universidades, que es esencial para la sostenibilidad del desarrollo. La transferencia de la tecnología se considera un factor esencial para la socialización del conocimiento que se apropia en todos los espacios del posgrado, pero este camino exige tener en cuenta las características y necesidades profesionales en el contexto universitario. Esta es la línea que se investiga y en la que se enmarca el estudio que se realiza.

2. Objetivo

La Educación de Posgrado en la Universidad de las Ciencias Informáticas (UCI) está estructurada en estrecha relación con la política científica, las líneas y proyectos de investigación, desarrollo e innovación de la institución, logrando una integración coherente entre la investigación y el posgrado. Los programas que ofrece amplían la cultura científica y brindan conocimientos avanzados fundamentalmente en el área de la informática y ramas afines, no obstante, aun cuando se realizan ingentes esfuerzos en el cumplimiento de la formación de doctores, no se cumple con el plan previsto: no siempre se estructura bien la pirámide de investigación a partir del tutor, aspirante, tema de maestría, investigación científica y los proyectos de investigación.

Para lograr ventajas competitivas sostenibles, la Educación de Posgrado en la UCI necesita fortalecer la aplicación de la GC para identificar, generar, compartir y apropiarse del conocimiento desde las tecnologías. Es por ello que se debe perfeccionar la Educación de Posgrado desde una mirada de la GC en lo académico, lo investigativo y lo formativo, viéndolo desde el contenido, desde la concepción y desde la didáctica. En las actividades de posgrado los temas infotecnológicos tienen que ser contenido, donde las personas se formen y trabajen con gestores bibliográficos, para que aprendan a reconocer ba-

ses de datos de impacto, los algoritmos esenciales para realizar búsquedas bibliográficas efectivas en revistas indexadas, arbitradas.

La UCI cuenta con las herramientas y plataformas tecnológicas para realizar la formación permanente en la Educación de Posgrado, y debe continuar dirigiendo su esencia al proceso formativo, donde la tecnología es el medio y no el fin. Los autores del presente trabajo consideran que resulta más apropiado que en la Educación de Posgrado se perfeccione no solo los contenidos, sino también los elementos didácticos y metodológicos para potenciar la virtualidad, donde siempre se garantice que el recurso humano se dote de las herramientas de gestión del conocimiento para su mejor desempeño.

A pesar de los resultados positivos que se tienen en la Educación de Posgrado en la UCI, se considera que aún hay debilidades en el método de trabajo y en la aplicación de los materiales bibliográficos relacionados con la GC en la Educación de Posgrado en Cuba, en particular para la formación posgraduada en la UCI.

De ahí que el objetivo se concreta en diagnosticar la comprensión y aplicación de la gestión del conocimiento desde la Educación de Posgrado, además de registrar las necesidades percibidas de los estudiantes del posgrado académico en la UCI.

3. Método

La muestra fue definida para este estudio, siguiendo el criterio de (Hernández Sampieri et al., 2014), mediante un muestreo aleatorio estratificado; y la integran los estudiantes que cursan la Maestría en Gestión de Proyectos Informáticos (MGPI-5ta Ed), la Maestría en Calidad de Software (MCSw-4ta Ed) y la Maestría de Informática Avanzada (MIAv-2da Ed), por ser los programas valorados de excelencia en esta universidad por la Junta de Acreditación Nacional (JAN), que en total reúnen a 107 maestrantes. Además, del Programa de Doctorado en Informática también valorado de excelencia por la JAN, que agrupa 38 doctorandos, para un total 145 personas.

Para la recogida de datos se elaboraron sendos cuestionarios (Hernández-Luque et al., 2021; Hernández Luque et al., 2021). Luego, se procedió a su evaluación por expertos, mediante un Análisis Factorial Exploratorio con el objetivo de evaluar la validez de constructos y fiabilidad de los instrumentos con el software SPSS v.23. Una vez validados se aplicaron a la muestra determinada. Además, para la gestión y emisión de reportes de los instrumentos aplicados se utilizó la herramienta LimeSurvey v1.52. Posteriormente se aplicaron diferentes pruebas estadísticas tales como la prueba Kaiser-Meyer-Olkin, la prueba de esfericidad de Bartlett y el test de Alfa de Cronbach a cada una de las tres escalas elaboradas, para comprobar su calidad científica en términos de validez y fiabilidad.

4. Resultados y discusión

Los expertos evaluaron por cada ítem (i) la comprensión, referido al grado en el que cada ítem expresa de manera concreta su enunciado, (ii) la factibilidad del ítem, en el cual expresa el grado en el que el ítem puede ser contestado, (iii) la pertinencia, se refiere al grado con el que el ítem realmente mide la comprensión de los estudiantes en cuanto a la gestión del conocimiento en la Educación de Posgrado sustentado en las TIC. Para su aplicación se empleó una escala tipo Likert que va desde 1 (muy en desacuerdo), 2 (algo en desacuerdo), 3 (ni de acuerdo ni en desacuerdo), 4 (algo de acuerdo) y 5 (muy de acuerdo).

4.1. Comprensión, intención, requerimientos, obstáculos y resultados

Se analizaron 35 ítems sobre estos aspectos.

Antes de aplicar el AFE se empleó el test de Kaiser-Meyer-Olkin (KMO) y la prueba de esfericidad de Bartlett como supuestos estadísticos. El resultado mostró un coeficiente KMO = ,748 que implica una buena relación entre variables (Kaiser, 1974). En tanto, la prueba de esfericidad de Bartlett ofrece un p=0,000 lo que justifica que se puede realizar el análisis factorial. Resultados que muestran la factibilidad de aplicar el AFE. De modo que, las escalas elaboradas son válidas y fiables.

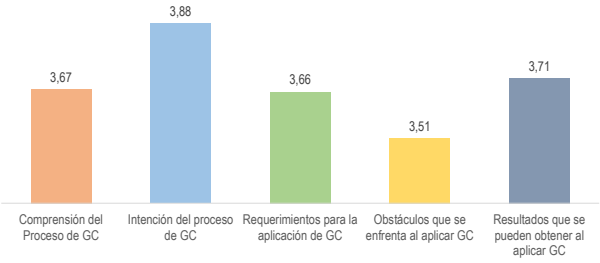

Figura 4. Comparación de las escalas de percepción en la Educación de Posgrado sobre la GC

Las valoraciones globales (Figura 4) sintetizan las necesidades percibidas por los estudiantes de posgrado en cuanto a la comprensión, intención, requerimientos, obstáculos que se enfrentan y resultados que se pueden obtener al aplicar la gestión del conocimiento desde las actividades investigativas y científicas del posgrado.

Se constata en todos los indicadores valoraciones superiores a las medias de las escalas, por lo que se confirma que se realizan acciones para GC en la Educación de Posgrado. Tras esta constatación, se necesita comprender la importancia, diseñar estrategias para socializar, crear espacios y utilizar herramientas GC, integrar conocimiento y buenas prácticas.

Los resultados detallados de cada ítem y su dimensión (Figura 5) destacan con una media mayor a 4,0 la necesidad de interconectar con sustento en las TIC el conocimiento explícito que se genera con el conocimiento tácito que no se socializa, así como la de formar mayor cantidad de especialistas en gestión del conocimiento.

Figura 5. Comprensión del proceso de la GC

Figura 6. Intención del proceso de la GC

Además, entre las acciones que se desarrollan en las actividades de posgrado se encuentran los chequeos, las reuniones, los despachos, los talleres, que se planifican y organizan en planes de trabajo y con una media de 4,94 (Figura 6). Se considera que gestionar grandes volúmenes de datos y reproducir información en estos espacios, son las actividades principales para lograr una adecuada GC. De modo que, se percibe que no se logra comprender que —a través de las acciones que promuevan la cultura científica y la competencia investigativa— se logrará aprender, compartir experiencias, evitar repetir los errores y duplicar esfuerzos, en todas las actividades científicas e investigativas que se desarrollan en el posgrado.

Se evidencia con una media de 4,83 (Figura 7) la necesidad de identificar qué, cómo, para qué y por qué se necesita aplicar la GC en la Educación de Posgrado, lo cual permitirá que se realicen actividades para socializar las buenas prácticas, aprender sobre los referentes teórico-metodológicos de la GC y así garantizar mejor control de los recursos y medios disponibles.

Con una media de 3,28 se plantea la necesidad de impulsar una cultura organizacional basada en el conocimiento, que permita la identificación, desarrollo y formación de los involucrados en el proceso de posgrado, para aumentar el índice de satisfacción, motivación y retención de estos, y para establecer un crecimiento de la base de conocimiento e índice de impacto en el desarrollo de las actividades de posgrado.

Figura 7. Requerimientos para aplicar la GC

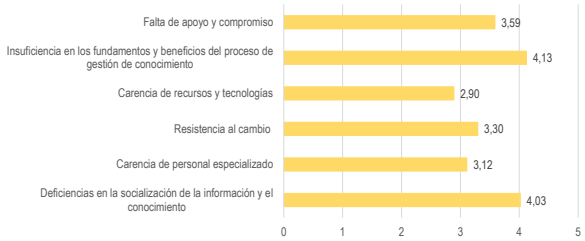

Figura 8. Obstáculos para aplicar la GC

Se corrobora (Figura 8) con una media de 4,03 en las preguntas abiertas que hay un alto índice de fluctuación de personal, lo que dificulta la estabilidad en los resultados y el seguimiento de las actividades, así como el logro de los objetivos. Además, se coincide con una media de 4,30 en cuanto a que hay un alto dinamismo en las actividades que se desarrollan en la universidad, lo que produce movimientos de actividades dificultando el cumplimiento de las planificaciones de investigación y posgrado. Es por ello que se necesita que la Educación de Posgrado sea capaz de recibir y procesar información, de aprender siempre de lo aprendido, de crear conocimientos a partir de la información procesada y de utilizarla de manera eficaz para la toma de decisiones.

Se evidencia con medias entre 4 y 5 (Figura 9) que los estudiantes de posgrado conocen los resultados que se pueden alcanzar al lograr una

mejor estimulación del aprendizaje y de las experiencias adquiridas por los maestrantes y doctorandos dentro y fuera de su entorno laboral, lo que contribuirá a elevar la transformación de la articulación entre docencia, producción, investigación y extensión desde el uso intensivo e intencionado de las TIC para fortalecer este proceso de la GC.

Figura 9. Resultados que se obtienen al aplicar la GC

Es importante resaltar que hay una representación con una media de 3,93 por parte de las mujeres y de 3,5 por parte de los hombres, que evidencia que la mujer comprende mejor el proceso de la GC, su intención, así como los requerimientos que se necesitan para aplicar la GC y los obstáculos que se deben enfrentar para lograr mejores resultados.

Manifiestan además las mujeres que la GC ha sido considerada por muchos en la Educación de Posgrado en la UCI, pero comprendida y valorada por pocos, siendo el principal requisito el de identificar, preservar, documentar y socializar el conocimiento del que se apropia mediante el acceso a fuentes documentales y en los intercambios de experiencias, lo que conlleva a convertir el conocimiento tácito a explícito. El éxito no está en quien sabe más, sino en los que hacen mejor uso de lo que saben; y ese es el mayor reto de la Educación de Posgrado en la UCI.

4.2. Capacidad, condiciones tecnológicas y saber de posgrado

Se analizaron 15 ítems sobre estas dimensiones. Igualmente, antes de aplicar el AFE se empleó el test de KMO y la prueba de esfericidad de Bartlett como supuestos estadísticos. El resultado mostró un coeficiente KMO = ,675 que implica una buena relación entre variables (Kaiser, 1974). En tanto, la prueba de esfericidad de Bartlett ofrece un p=0,000, lo que justifica que se puede realizar el análisis factorial. Resultados que muestran la factibilidad de aplicar el AFE. De modo que, las escalas elaboradas son válidas y fiables.

Al evaluar las medias sobre la capacidad, las condiciones y el saber de la Educación de Posgrado para aplicar la GC desde las actividades investigativas y científicas de posgrado (Figura 10), se evidencia que en el posgrado se cuenta con las condiciones tecnológicas y la capacidad para aplicar la GC; sin embargo, existen limitaciones en integrar y socializar conocimientos y buenas prácticas.

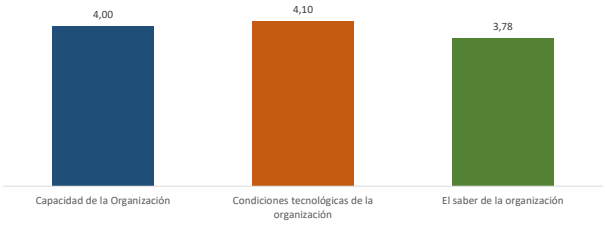

Figura 10. Comparación de las escalas sobre la capacidad, las condiciones y el saber de la Educación de Posgrado para aplicar la GC

Los resultados detallados de cada ítem y su dimensión (Figura 11) reflejan con una media entre 3,97 y 4,13 que la Educación de Posgrado tiene la capacidad y las condiciones tecnológicas para aplicar la GC.

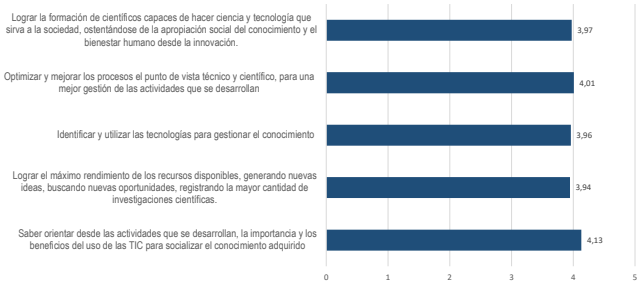

Figura 11. Capacidad de la Educación de Posgrado para aplicar la GC

Por ello, los sujetos están de acuerdo en que puede orientar desde las actividades que se desarrollan sobre la importancia y los beneficios del uso de las TIC para socializar el conocimiento adquirido, así como para lograr el máximo rendimiento de los recursos disponibles, generando nuevas ideas, buscando nuevas oportunidades, registrando la mayor cantidad de investigaciones científicas. Además, la GC ofrece la posibilidad de formar profesionales capaces de hacer ciencia y tecnología que sirva a la sociedad, a partir de la apropiación social del conocimiento y el bienestar humano desde la innovación.

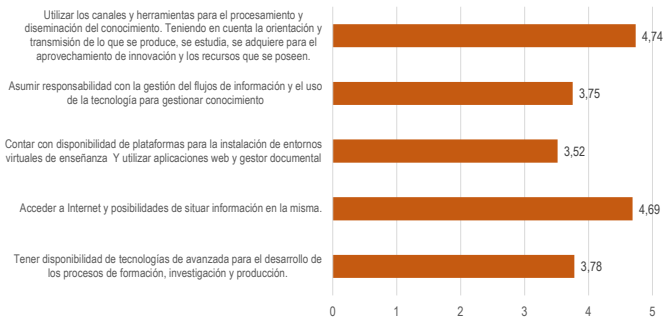

Figura 12. Condiciones tecnológicas para aplicar la GC

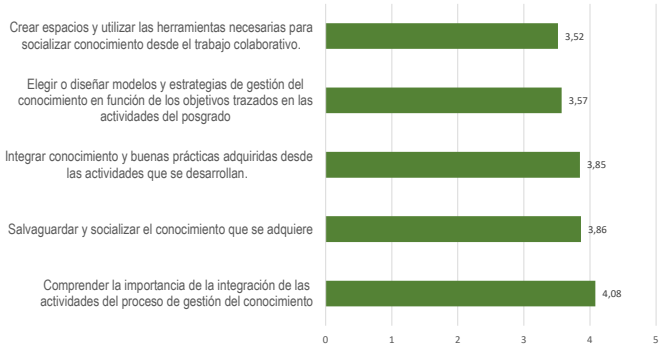

Figura 13. El saber de la Educación de Posgrado para aplicar la GC

Por otro lado, con una media entre 3,52 y 4,74 (Figura 12) evalúan la necesidad de utilizar los canales y herramientas para el procesamiento y diseminación del conocimiento, lo que conlleva a tener en cuenta la orientación y transmisión de lo que se produce y se estudia para el aprovechamiento de la innovación y los recursos que se poseen. Además, los sujetos piensan que se necesita lograr una adecuada interrelación entre el aporte de las tecnologías, el capital intelectual y la cultura organizacional para que se logre socializar el conocimiento y las buenas experiencias adquiridas en las actividades de posgrado.

Con una media de 3,52 (Figura 13) se evidencia que es necesario lograr identificar ¿qué se necesita saber? y ¿quién sabe qué?, para fortalecer y desarrollar habilidades que se puedan aprender en todo contexto y momento. Por ello es importante y significativo lograr adquirir, representar, socializar y administrar el conocimiento que se apropia desde la práctica.

El diagnóstico muestra que las actividades de posgrado constituyen una vía fundamental para mejorar las actividades investigativas y científicas fundamentadas claramente en la UCI, porque potencia el constante perfeccionamiento y sus aportes prácticos influyen positivamente en los procesos que transforman su objeto social.

No obstante, se evidencian limitaciones en los referentes teórico-metodológicos de la GC, lo cual refleja las debilidades para socializar, utilizar el conocimiento que se genera en las actividades científicas e investigativas de posgrado. De este modo, se constata la importancia de un adecuado tratamiento a la gestión del conocimiento para aprovechar sus potencialidades en la educación de posgrado de la UCI.

5. Conclusiones y recomendaciones

La aportación principal de esta investigación es la formalización de una estrategia metodológica para la GC en la Educación de Posgrado que consta de cinco etapas (Figura 14) y la propuesta de un sistema de acciones organizado en cada una de estas etapas teniendo en cuenta el modelo Nonaka yTakeuchi (1995).

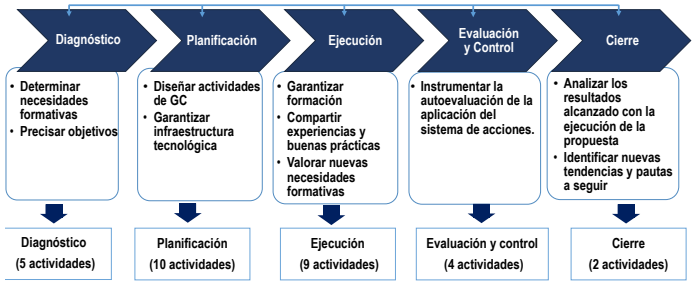

Figura 14. Etapas de la propuesta

De esta forma, se mapearon las relaciones que se establecen entre el sistema de acciones y las etapas de la propuesta con los procesos de socializaciones, externalización, internalización y combinación para fortalecer la apropiación, la aplicación y la socialización de los conocimientos en la Educación de Posgrado de la UCI en correspondencia con el avance científico-tecnológico actual.

Para el diseño del sistema de acciones (Figura 15) se recomienda que:

- el proceso de socialización tenga como propósito identificar y compartir conocimiento tácito entre las personas, incidiendo en la difusión del conocimiento individual; por tanto, las acciones deben lograr que el conocimiento individual sea trasmitido y recibido para generar nuevos conjuntos de conocimientos y experiencias;
- el proceso de externalización tenga como propósito la interacción colectiva de conocimiento; por tanto, las acciones deben estar dirigidas a generar, crear, innovar y compartir nuevas ideas para el desarrollo de conocimiento;

- el proceso de internalización tenga como propósito interiorizar el nuevo conocimiento explícito, por lo que las acciones deben retener e incorporar el nuevo conocimiento adquirido para su desarrollo y aplicabilidad en las actividades de posgrado;
- el proceso de combinación tiene como propósito la interacción colectiva, convirtiendo el conocimiento en nuevas formas explícitas, de modo que las acciones se potencian con el uso de las herramientas tecnológicas para la transformación y desarrollo del nuevo conocimiento.

Figura 15. Sistema de acciones desde modelo SECI

En definitiva, tras analizar los referentes teórico-metodológicos de la Educación de Posgrado y la gestión del conocimiento, se profundizó en las actividades de creación, apropiación y socialización del conocimiento en las actividades científicas y de investigación que se desarrollan, constatando la necesidad de fortalecer el valor del conocimiento y su gestión.

La caracterización de la gestión del conocimiento en la Educación de Posgrado que se obtuvo a través de los cuestionarios refleja la existencia de deficiencias en comprender y aplicar el proceso, en facilitar la gestión del contenido evitando el exceso de información, así como en aprovechar todas las potencialidades en la articulación de la información y el conocimiento que se genera con el aprendizaje de las actividades del posgrado.

Esta investigación también tiene una aportación teórica y metodológica, al identificar los aspectos que permitan evaluar la Educación de Posgrado para aprovechar las potencialidades de la gestión del conocimiento desde la capacidad, las condiciones tecnológicas y el saber de la organización. Sus principales elementos son la percepción sobre la gestión del conocimiento en la la Educación de Posgrado desde la comprensión e intención del proceso de GC, así como la identificación de los requerimientos y obstáculos y de los resultados que se encuentran al aplicarla. Su validación se convierte en una herramienta útil para futuras investigaciones.

El estudio presenta limitaciones y se recomienda el análisis desde la investigación científica aumentando el tamaño muestral para alcanzar una mayor generalización de resultados.

Referencias

Aguilera, G. del R. M.; González, E. B. O. (2018). Habilidades intelectuales específicas que favorecen el desarrollo de competencias para la investigación en la licenciatura en educación física. // Educando Para Educar. 33, 77-86.

Alavi, M.; Leidner, D. E. (2013). Review: knowledge management and knowledge management systems: conceptual foundations and research issues. // MIS Quarterly. 25:1, 107-136.

Amber, D.; Domingo, J. (2016). Desempleo y precariedad laboral en mayores de 45 años. Retos de la formación e implicaciones educativas. // Revista Iberoamericana de Educación. 73:1, 121-140.

Barboza, A.; Ochoa, I. (2016). Modelos de Gestión del Conocimiento O&B y Capital Intelectual B&O para Organizaciones. // REVECITEC Urbe. 6:1.

Becker, S. A.; Cummins, M.; Davis, A.; Freeman, A.; Giesinger, C. H.; Ananthanarayanan, V. (2017). The NMC Horizon Report: 2017 Higher Education Edition. Austin, Texas: The New Media Consortium.

Bernaza-Rodríguez, G. J. (2013). Construyendo ideas pedagógicas sobre el posgrado desde el enfoque histórico-cultural.

Brew, A.; Saunders, C. (2020). Making sense of research-based learning in teacher education. // Teaching and Teacher Education. 87, 102935.

Camporredondo, A. G. (2018). Desarrollo local en cuba: retos y perspectivas. // Desarrollo local en Cuba: retos y perspectivas. https://bit.ly/3uBvdMr

Davenport, D.; Holsapple, C. W. (2011). Knowledge Organizations. // Schwartz, G. (Editor), Encyclopedia of Knowledge Management. London: Idea Group Inc., 451-458.

Drucker, P. (1995). Dirección por excepción. México: Editorial Cecsa.

Foray, D.; Lundvall, B. (1998). The knowledge-based economy: from the economics of knowledge to the learning economy. // Neef, D.; et al.(Eds.). The Economic Impact of Knowledge, 115-121. Boston: Butterworth-Heinemann.

George, C. E.; Salado, L. I. (2019). Competencias investigativas con el uso de las TIC en estudiantes de doctorado. // Apertura, 11:1, 40-55. https://bit.ly/3LdfuML

Gómez-Bayona, L.; Londoño-Montoya, E.; Mora-González, B. (2019). Modelos de capital intelectual a nivel empresarial y su aporte en la creación de valor. // Revista CEA. https://bit.ly/3J86ffi

González, A. F.; Jover, J. N. (2020). Creación de capacidades y desarrollo local: el papel de los centros universitarios municipales.

Greiner, M. E.; Böhmann, T.; Krcmar, H. (2007). A strategy for knowledge management. // Journal of Knowledge Management. 11:6, 3-15.

Hernández-Luque, E.; Estrada-Sentí, V.; Keeling-Alvarez, M. (2018). Perspectivas y desafíos de la gestión del conocimiento y la competencia investigativa en la educación del posgrado. // UCE Ciencia Revista de Postgrado. 6:1.

Hernández-Luque, E.; Zulueta-Veliz, Y.; Hernández-de la Rosa, M. A. (2021). La gestión del conocimiento en el posgrado: un instrumento para su diagnóstico. // Atenas. 3:55, 86-99.

Hernández Gutiérrez, D.; Benítez Cárdenas, F.; Sánchez Hernández, Y.; Manzano Rivera, S. A. (2006). La nueva

universidad cubana y su contribución a la universalización del conocimiento.

Hislop, D.; Bosua, R.; Helms, R. (2018). Knowledge Management in Organizations: A Critical Introduction. Oxford University Press, Oxford.

Kaiser, H. F. (1974). An index of factorial simplicity. // Psychometrika. 39, 3136.

Laal, M. (2011). Knowledge management in higher education. // Procedia Computer Science. 3, 544-549.

Luna-Nemecio, J.; Tobón, S.; Juárez-Hernández, L. (2019). Socioformation and complexity: towards a new concept of sustainable social development. // Human Development and Socioformation. 1:2, 1-13.

Luque, E. H.; Sentí, V. E.; de la Rosa, M. A. H. (2021). Diseño y validación de un cuestionario para evaluar la gestión del conocimiento en la educación de posgrado. // Revista EDUSOL. 21, 29-43.

MES. (2013). Plan de estudios "D" Ingeniería en Ciencias Informáticas.

Ministerio de Educación Superior, M. (2019). Resolución No. 140/2019. Reglamento de la Educación de Posgrado de la República de Cuba. (GOC-2019-776-O65). Gaceta Oficial. 2019(138), 1442-1447.

Nonaka, I.; Konno, N. (1998). The concept of "ba": Building a foundation for knowledge creation. // California Management Review. 40:3, 40-54.

Nonaka, I.; Takeuchi, H. (1995). The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press.

Nonaka, Ikujiro; Takeuchi, H. (1999). La organización creadora de conocimiento. México: Ed. Castillo Hnos. Número de Registro 723.

Organización Internacional de Normalización (2015). ISO 9001:2015. Sistema de Gestión de Calidad. Conocimiento de La Organización. https://bit.ly/3rA450n

Organización Internacional de Normalización (2018). ISO 30401:2018. Knowledge Management Systems. https://bit.ly/3l0St5o

Ortega-Carbajal, M. F.; Hernández-Mosqueda, J. S.; Tobón-Tobón, S. (2015). Análisis documental de la gestión del conocimiento mediante la cartografía conceptual. // RaXimhai. 11:4, 141-160.

Páez-Suárez, V. (2020). La Didáctica de la Educación Superior ante los retos del siglo XXI (I. Bermúdez Lamadrid & S. Lima Montenegro (eds.); Editora Ed. Sello Editor Educación Cubana.

Páez-Suárez, V.; Soto-Saez, E. M.; Olivera-Romero, J. J. (2021). Fundamentos epistemológicos de la relación conocimiento, gestión del conocimiento y la labor educativa en la formación profesional.

Pérez-Zubillaga, Z. R. (2014). La formación continua en las TIC de los profesores de la UCCFD "Manuel Fajardo."

Pérez, M. G. V. (2020). Visibilidad de la producción de conocimiento: componente estratégico de la Gestión Universitaria. CyCL Controversias y Concurrencias Latinoamericanas. 11:20, 353-363.

Polanyi, M. (1962). Personal Knowledge.

Ponjuán-Dante, G. (2015). La gestión del conocimiento desde las ciencias de la información: responsabilidades y oportunidades. // Revista Cubana de Información en Ciencias de la Salud. 26:3, 206-216.

Raneda-Guirriman, C.; Rodríguez-Ponce, E.; Pedraja-Rejas, L. (2017). La gestión del conocimiento en instituciones de Educación Superior. // Revista de Pedagogía. 38:102, 13-30.

Ricardo, M. A. (2021). Knowledge Management and the 2030 Agenda for Sustainable Development in the United Nations Context. Ciencias Administrativas. // Revista Digital FCE-UNLP. 9:17, 80-84.

Rodríguez-Montoya, C.; Zerpa-García, C. E. (2019). Gestión del Conocimiento en Programas de Postgrado: un Modelo Prescriptivo. // Pixel-Bit, Revista de Medios y Educación. 55, 179-209.

Salazar-Gomez, E.; Tobon, S. (2018). Análisis documental del proceso de formación docente acorde con la sociedad del conocimiento. // Revista Espacios, 39:45, 17.

Sallán, D. R. (2015). Innovación, aprendizaje organizativo y gestión de conocimiento en las instituciones educativas. // Educación. 24:46. https://revistas.pucp.edu.pe/index.php/educacion/article/view/12245

Sentí, V. E.; Cárdenas, F. B. (2010). La gestión del conocimiento en la nueva universidad cubana. // Universidad y Sociedad. 2:2. https://bit.ly/3goRw34

Sentí, V. E.; Rodríguez, J. P. F.; Baquerizo, R. M. P. (2016). La socialización del conocimiento y el empleo del webquest en apoyo al aprendizaje PED-066. Universidad, 2016.

Shujahat, M.; José, M.; Hussain, S.; Nawaz, F.; Wang, M.; Umer, M. (2017). Translating the impact of knowledge management processes into knowledge-based innovation : The neglected and mediating role of knowledge-worker productivity. // Journal of Business Research. November, 0-1.

Silva, H. (2017). Globalización y Sociedad del Conocimiento. // Investigaciones en Educación. 17:2, 45-56.

Simeón-Negrín, R. E. (2004). Bases para la introducción de la gestión del conocimiento en Cuba. "Cuba posee una verdadera riqueza de conocimientos." // Ciencia, Innovación y Desarrollo. Revista de Información Cientifica y Tecnológica. 9:2, 6-8.

Soto-Balbón, M. A.; Barrios-Fernández, N. M. (2006). Gestión del conocimiento. Parte I. Revisión crítica del estado del arte. // ACIMED. 14(2).

Streck, D. R.; Redin, E.; Zitkoski, J. J. (2015). Diccionario. Paulo Freire. 2da edición traducida al castellano. Lima: CEAAL.

UCI. (2021). Universidad de las Ciencias Informáticas. https://bit.ly/3LgkksE

UNESCO. (1995). Declaración mundial sobre la educación superior en el siglo XXI : Visión y Acción. // Educación Superior y Sociedad. 9:2, 97-113.

Yee, Y. .; Tan, C. .; Thurasamy, R. (2019). Back to basics: building a knowledge management system. Strategic Direction. 35:2, 1-3.

Enviado: 2022-02-24. Segunda versión: 2022-09-12.
Aceptado: 2022-11-10.

Diretrizes para a compatibilização de SOCs com vistas a uma recuperação inteligente da informação

Directrices para la compatibilidad del SOC con miras a la recuperación inteligente de la información

Guidelines for KOS compatibility towards intelligent information retrieval

Nilson Theobald BARBOSA (1), Maria Luiza de Almeida CAMPOS (2)

(1) Universidade Federal do Rio de Janeiro - Cidade Universitária, Rio de Janeiro/RJ, nilson@tbarbosa.org
(2) Universidade Federal Fluminense – Niterói/RJ, Universidade Federal da Bahia – Salvador/BA, marialuizalmeida@gmail.com

Resumen

Se presenta un enfoque para la compatibilización de vocabularios heterogéneos con el fin de permitir la recuperación inteligente de información en diferentes bases de datos, asegurando que los vocabularios originales se mantengan sin cambios. Este estudio se caracteriza por un enfoque cualitativo que supone un desarrollo interpretativo de los datos recogidos a partir de la investigación bibliográfica y documental. Como producto de esta investigación se presenta un conjunto de directrices que, apoyadas en métodos, técnicas y algoritmos computacionales, apuntan a la posibilidad de crear procesos automatizados de compatibilidad semántica de vocabularios que conduzcan a un proceso inteligente de recuperación de información distribuida en diferentes SOCs.

Palabras clave: Sistemas de organización del conocimiento. Interoperabilidad semántica. Lenguaje intermedio. Coordenadas semánticas. Recuperación de información.

Abstract

An approach to the compatibilization of heterogeneous vocabularies is presented aiming to allow the intelligent retrieval of information in different databases, ensuring that the original vocabularies are kept without change. This study is characterized by a qualitative approach that supposes an interpretative development of data collected from bibliographic and documentary research. As a product of this research, a set of guidelines is presented that, supported by methods, techniques, and computational algorithms, points to the possibility of creating automated processes of semantic compatibilization of vocabularies that may lead to an intelligent process of retrieval of information distributed in different KOS.

Keywords: Knowledge organization systems. Semantic interoperability. Intermediate language. Semantic coordinates. Information retrieval.

1. Introdução

Já há algum tempo vivemos no mundo dos dados digitais, com a evolução das tecnologias criadas há aproximadamente meio século que dariam início a mais uma revolução em nossa história e permitiram o surgimento da Internet. Nos tempos atuais parece que nossa capacidade de criar e produzir dados ultrapassa de longe nossa capacidade de gerenciar e permitir que estes dados, além de estarem acessíveis, sejam compreendidos, enfim que façam sentido e sejam fonte de informações para quem delas necessita. Gantz e Reinsel (2010) mostravam em um relatório que entre 2010 e 2020 o total de registros digitais criados e replicados pelo mundo teriam uma evolução para inconcebíveis 35 trilhões de gigabytes. Esta previsão foi confirmada em 2020 e temos hoje uma perspectiva de atingirmos perto de 150 zettabytes em 2024, considerando o volume de dados criados e capturados em todo o mundo (Statista, 2021).

Todo o avanço tecnológico e computacional que contribui com a geração deste volume de dados e cria redes com maior desempenho, sejam as redes de fibra ótica, as redes sem fio ou a vindoura rede 5G, e que usa repositórios “infinitos” e fornece computação de alto poder de desempenho na palma da mão tem, apesar disso, um limitador de forte impacto para a utilização plena de toda esta possível informação. Uma miríade de repositórios de dados e seus conteúdos com todos os tipos de dados se multiplicam exponencialmente. Estes repositórios, seus dados e suas linguagens de indexação heterogêneas em todos os seus níveis, seja dentro das organizações seja na Web como um todo, ainda não são capazes de oferecer aos seus usuários uma recuperação plena e semântica da informação que está disponível.

Para que possamos ter pleno usufruto deste capital de conhecimento, um forte limitador nos persegue, aliás desde antes da criação da Internet, que é a incapacidade da humanidade de resolver seus problemas de divisão culturais e linguísticos

e fundamentalmente de resolver a incompatibilidade de seus sistemas de classificação. Pierre Lévy chama este processo envolvendo um enorme crescimento computacional, difusão de dados, produção de registros e consumo crescente de informações digitais de 'memória digital participativa', afirmando que esta memória está apenas em vias de constituição, a despeito de todo avanço tecnológico. Temos como desafio a automatização das operações cognitivas de análise e interconexão das informações que supostamente estão disponíveis. Não sabemos, ainda, como transformar de forma sistemática e efetiva este oceano de dados em conhecimento e menos ainda como transformar o meio digital em observatório reflexivo de nossas inteligências coletivas (Lévy, 2014).

A interoperabilidade física entre computadores e entre bases de dados é uma questão já bem resolvida pela tecnologia, a compatibilização entre termos de mesma grafia em diferentes vocabulários já é bem realizada pelo alto desempenho computacional disponível e facilitada pela conectividade física, mas transpor a barreira semântica, em que precisamos compatibilizar conceitos, tenham eles a mesma expressão verbal ou não, é uma questão ainda a ser resolvida.

Apesar de parecer tentador criar sistemas de indexação e vocabulários unificados para resolver este problema, em um ambiente aberto e não controlado nem sempre é possível recorrer a esta solução, e precisamos partir para soluções que possibilitem recuperar informações de bases indexadas por sistemas heterogêneos sem fazer alterações nestas bases ou em seus vocabulários através do estabelecimento de correspondências e mapeamentos de conceitos, e não simplesmente de termos verbais, entre estes vocabulários.

Portanto, esta é, em síntese, a questão que perseguimos aqui. A possibilidade de compatibilização semântica automatizada de diferentes sistemas de organização do conhecimento sem a alteração de seus ambientes originais, com a utilização de metalinguagens, que parecem ser capazes de fornecer as bases teóricas para o desempenho desta tarefa. Estas metalinguagens devem ser formadas como uma linguagem intermediária entre os diferentes vocabulários fonte possibilitando diferentes atores navegarem por esta linguagem e, de forma contextual e semântica, recuperarem a informação pretendida, não mais com base em simples comparações de cadeias de caracteres, mas sim em seu significado.

A seguir colocaremos o problema brevemente relacionado à criação de um ambiente para uma recuperação inteligente da informação, para depois apresentarmos as diretrizes que consideramos poder auxiliar na elaboração deste ambiente. Apresentaremos em sequência uma síntese das técnicas que podem ser utilizadas para o estabelecimento de equivalências e proximidades semânticas entre os conceitos e por fim as considerações finais.

2. Um espaço para a recuperação inteligente da informação: as linguagens intermediárias e as coordenadas semânticas

Após a disseminação da computação, da Internet e da Web, estamos diante de muitas iniciativas que procuram resolver a disparidade e heterogeneidade entre os sistemas de informação. Soergel (1972, 1974) já abordava esta questão e apresentava discussão que endereça estes problemas de compatibilidade. Uma solução apresentada, para enfrentar a abundância de tesauros e a contínua criação de novos, seria a criação de um Tesauro Fonte Universal, armazenado em computador, onde os elementos de todos os tesauros existentes poderiam ser coletados, assim como a indicação de todas as suas relações. Apesar de este tesauro universal poder ser usado como uma fonte de informação para descritores existentes e relações entre seus conceitos, e para criação de novos sistemas utilizando os conceitos existentes, estabelecer este tesauro universal seria um grande empreendimento e sua realização poderia apenas criar mecanismos de compatibilidade únicos para os sistemas que originem seus elementos da fonte comum (Dahlberg, 1981).

Como um caminho de compreensão e solução do problema, nosso olhar se volta inicialmente para a norma internacional que trata de interoperabilidade entre vocabulários e que oferece linhas de atuação para obter este fim. Observamos que a norma ISO 25964 parte 2 elenca diferentes modelos estruturais para a realização de mapeamentos entre vocabulários, a saber, Unidade estrutural, Ligação direta e Estrutura central, e estes modelos apresentados reforçam a visão do caminho para um vocabulário único, uma vez que no primeiro caso não chegamos sequer a ter mapeamentos, no segundo caso temos a proposta dos mapeamentos um-a-um, unidirecionais, custosos e difíceis de implantar, e no terceiro caso temos a utilização ou criação de um vocabulário central, que além de também ser usado para indexação é usado como um possível comutador entre outros vocabulários menores ou mais específicos. Portanto, se tomarmos por base o documento padrão que se propõe a normatizar os processos de interoperabilidade entre sistemas de organização do conhecimento somos levados a estabelecer processos de compatibilização que levam à criação de vocabulários únicos.

Trilhando um caminho diferente, nosso interesse se volta para uma abordagem de criação de linguagens intermediárias como uma estratégia para a compatibilização de vocabulários, por considerar que as necessidades de compatibilização de linguagens para o momento atual, com vistas a uma Web Semântica devem ter por base este arcabouço teórico da Ciência da Informação.

As discussões que embasam a proposta de métodos de léxicos intermediários para a compatibilização de vocabulários remontam aos trabalhos publicados por Hammond e Rosenborg (1962), Newman (1965) e Henderson et al.(1966), tendo a questão da compatibilidade e conversibilidade recebido especial atenção no relatório da UNESCO de 1971. Neste relatório temos uma definição de compatibilidade como sendo "uma qualidade de sistemas cujos produtos podem ser intercambiados, apesar de suas diferenças de notação, estrutura, suportes físicos etc., sem qualquer mecanismo especial de conversão" (Unesco, 1971). Além disso, conversão é definida como "o processo de transformar registros de informação, com respeito à codificação, estrutura de dados etc., de modo a fazê-los intercambiáveis entre dois ou mais sistemas usando diferentes convenções" (Unesco, 1971, p.147).

Uma importante contribuição foi também desenvolvida por J. C. Gardin (1967, 1973) e pelo seu grupo de trabalho na França, definindo que um léxico intermediário é destinado a acessar documentos indexados em termos de uma linguagem de indexação para outra sem que haja a perda de informação. Ele implica no mapeamento de dois ou mais vocabulários para uma linguagem intermediária ou neutra.

As investigações empíricas apresentadas por Wellisch (1972), Agraev et al.(1974), Smith (1974), Svenonius (1975) e Wersig (1975) apresentam estudos que definem a natureza das linguagens de indexação comparadas, a metodologia para comparação de linguagens de indexação e a estrutura dos elementos das linguagens de indexação mais adequadas para intercambiamento, sendo de especial interesse os estudos de Horsnell (1975) sobre a criação de um "léxico intermediário" (Dahlberg, 1981).

Considerando as abordagens seminais estudadas, a definição de linguagem intermediária vista e aceita aqui será como mostrada por Dahlberg (1981) e Neville (1970, 1972), baseada em uma codificação de conceitos, que permite o estabelecimento de uma equivalência conceitual de descritores de diferentes linguagens (Campos et al., 2009) e, de forma compatível com esta definição, a premissa aqui assumida é a importância e a necessidade de se realizar qualquer processo de compatibilização, alinhamento e mapeamento de vocabulários sem que os vocabulários originais sejam alterados ou tenham suas características modificadas, considerando as grandes dificuldades, especialmente administrativas e políticas, de conseguir realizar estas tarefas.

Um dos modos de realizar a construção deste dispositivo é a utilização do método da matriz de compatibilização conceitual de Dahlberg (1981). Partindo de seu método analítico-sintético, Dahlberg propõe a construção de uma matriz representativa da compatibilidade conceitual entre sistemas ordenados. Esta matriz é um mapeamento da potencialidade semântica das linguagens a serem compatibilizadas e, a partir daí, pode fornecer os resultados da análise de compatibilidade entre estas linguagens sob os pontos de vista sintático, estrutural e semântico.

Nesse sentido, também o método de Neville (1972) chamado de reconciliação de tesauros, tem por base o mesmo princípio de construção de léxicos intermediários apresentados por Natacha Gardin (1969) e Coates (1970), pressupondo que a compatibilização dos sistemas de organização do conhecimento deve considerar não só a sintaxe dos termos descritores, mas também os seus conteúdos conceituais, isto é, suas significações, expressas pelas suas definições (compatibilidade semântica). Este método prevê a elaboração de uma linguagem intermediária, baseada na codificação numérica de conceitos (onde cada conceito poderia ser identificado por um código numérico), possibilitando (i) estabelecer equivalência conceitual entre termos descritores de diferentes linguagens e, (ii) realizar a conversão automática de termos equivalentes e de termos específicos para genéricos (Bocatto e Torquetti, 2012).

Também nesta direção, consideramos de grande interesse para nosso trabalho discutir o conceito de "Sistema de Coordenadas Semânticas", apresentado por Pierre Lévy (Lévy, 2014, p. 312, 2019, p. 27), cuja formulação e procedimentos metodológicos permitem uma aproximação com o conceito de dispositivos de comutação, e consequentemente com os léxicos intermediários, colocando uma visão atual e centrada em procedimentos computacionais e automáticos para esta implementação.

Pierre Lévy vem se dedicando a explicitar uma construção teórica, que ele denomina de IEML – Metalinguagem da Economia da Informação, e argumenta que a sua principal hipótese para propor tal metalinguagem é a de que ainda não inventamos sistemas simbólicos que se encaixam no novo meio digital. Ao propor esta construção, ele apresenta a IEML principalmente como uma

linguagem artificial que se traduz automaticamente em línguas naturais (Lévy, 2014).

Neste sentido, na construção teórica proposta por Lévy, nos interessa sobremaneira o conceito de linguagem ponte e de sistema de coordenadas semânticas apresentadas na IEML. Como linguagem ponte podemos entender uma linguagem artificial ou natural usada como uma linguagem intermediária para tradução entre muitas línguas diferentes – para traduzir entre qualquer par de idiomas A e B, uma função traduz A para a linguagem ponte L, depois de L para B. Como sistema de coordenadas semânticas é importante entender que sua função seria de permitir um sistema de endereçamento que possibilite computar as relações e distanciamentos semânticos existentes entre as linguagens (Lévy, 2014).

Entendemos que a totalidade da proposta de Lévy apresenta proposições de difícil implantação dada a sua complexidade e extensão, mas por outro lado, apresenta caminhos metodológicos para a construção de dispositivos que permitam comutação entre diferentes ontologias que particularmente interessam em nossa pesquisa. Uma destas propostas é o estabelecimento de identificadores únicos, chamados de Uniform Semantic Locators (USL). As operações calculáveis realizadas nos conjuntos de sequências que são os USL são ao mesmo tempo operações realizadas sobre o sentido que estes conjuntos representam (ou seja, os conceitos). “A principal ideia a ser retida é a de que um caminho qualquer no espaço hipertextual das conexões entre USL pode ser representado por uma função e a de que essa função pode ter uma pertinência semântica” (Lévy, 2014, p.478).

Dessa forma, conforme Lévy, a esfera semântica e a linguagem IEML funcionam como um sistema de codificação do sentido concebido para tornar automaticamente calculáveis operações sobre os conceitos e sobre suas operações semânticas, e tudo isso repousa, na prática, sobre a existência de um conjunto de circuitos semânticos matriciais funcionando como convenção de tradução dos textos IEML para os circuitos semânticos selecionados em línguas naturais e vice-versa.

Compreendemos que a proposta de Lévy é uma criação desta convenção de tradução que avance para todas as coisas existentes, representando um sistema de comutação universal, mas defendemos também que esta proposta pode ser analisada sob o ponto de vista da criação de dispositivos cujo funcionamento é coerente com as propostas de construção de linguagens intermediárias, ou seja, dispositivos de comutação, apresentadas pela Ciência da Informação para a recuperação da informação em ambientes heterogêneos.

Para isso recuperamos a definição de que a unidade básica da IEML, o USL, não se limita a descrever um conceito, mas pode ser usada para intermediar consultas em uma base dados. Neste ambiente com uma coleção de identificadores únicos (ou USL), é possível calcular os mais semelhantes a outros USL, ou seja, os mais representativos da coleção, e aqueles que têm menos em comum com outros membros da coleção, ou seja, a maioria dos dispositivos da coleção.

A proposta de Lévy nos leva a compreender que os conceitos e suas representações dentro de um sistema de comutação para recuperação da informação entre várias linguagens podem ser recuperados através dos métodos de mapeamento que sejam voltados para gerar linguagens intermediárias e podemos ver que em sua proposta é necessário que estes identificadores únicos apresentem descritores que representem seu significado semântico, o que Dahlberg em seu trabalho procurou descrever através do Registro do Conceito (Dahlberg, 1981).

Portanto, os processos computacionais de alinhamento e mapeamento que são capazes de compatibilizar SOCs e seus termos podem ser utilizados para a criação de uma “esfera semântica”, não universal e global, mas que atenda à recuperação da informação em ambientes com diversos sistemas de indexação. A definição dos “USL”, baseadas no Registro de Conceito de Dahlberg, para estes ambientes, pode levar a um processo de recuperação da informação que seja capaz de fornecer a um usuário interessado em utilizar este ambiente de multivocabulários o significado de cada conceito, sua representação em cada linguagem utilizada, e a medida de distância semântica de cada conceito para outros conceitos, iguais ou semelhantes. É este objetivo que perseguimos na formulação das diretrizes que apresentamos a seguir.

3. Diretrizes para elaboração de Linguagens Intermediárias entre SOCs

No contexto deste artigo, apoiados na proposta de Nurcan et al.(1999), vamos destacar alguns aspectos na apresentação das Diretrizes (DIR01 a DIR06), ou seja, iremos denominá-las como um objetivo a ser atingido, logo após apresentaremos uma descrição apontando a sua finalidade e importância e por último apresentaremos instruções visando alcançar os propósitos esperados enunciados na denominação da Diretriz.

3.1. DIR 01: Desenvolvimento e manutenção dos sistemas de organização do conhecimento por profissionais especializados

Num momento em que, tanto os novos vocabulários criados, como a imensa quantidade de sistemas já existentes, precisam participar de processos automatizados de compatibilização de informações com vistas à sua recuperação entre sistemas heterogêneos, a questão da utilização de padrões sólidos é questão essencial para que este fim seja alcançado.

Nesse sentido, a construção destes sistemas de organização do conhecimento, talvez hoje mais do que nunca precisem ser desenvolvidos por profissionais altamente qualificados para sua construção, pois, caso contrário, todos os esforços na criação de técnicas modernas, algoritmos de alta eficiência e computadores e redes de alto desempenho, não serão capazes de avançar na tarefa de interoperar sistema heterogêneos. De fato, o desenvolvimento de Sistemas de Organização de Conhecimento deve ser um trabalho conjunto entre aqueles que dominam os processos de classificação de domínio, os especialistas do domínio, e profissionais de TI.

É importante notar que esta recomendação não se aplica apenas no momento da criação dos sistemas, mas deve ser seguida por toda sua vida, em seus processos de manutenção e atualização, e da mesma forma se aplica tanto a taxonomias e tesauros, quanto a construção de ontologias, pois o conhecimento referente ao tratamento e organização da informação é essencial para ser combinado com a utilização das novas tecnologias e novas linguagens de representação de sistemas de organização do conhecimento.

3.2. DIR 02: Utilização de linguagens de representação padrão, compatíveis e abertas para construção dos sistemas de organização do conhecimento

Proporcionar aos sistemas de organização do conhecimento participantes capacidade de serem lidos e interpretados por agentes de software de forma compatível com representações padrão definidas pelas tecnologias da web semântica, em especial aquelas defendidas pelo consórcio W3C.

As tecnologias ligadas à web semântica apresentam múltiplas opções para representação de dados e temos diversas possibilidades para representação de sistemas de organização do conhecimento dentre RDF, RDF-S, OWL, SKOS, entre outras.

A análise do modelo SKOS nos mostra que este formato, pela sua difusão e pela facilidade de conversão entre diferentes modelos, inclusive a partir de vocabulários e tesauros que possuem apenas um modelo de dados descritivo, nos permite dizer que este pode ser considerado um modelo preferencial para a representação de sistemas de organização do conhecimento em geral, e em especial, tesauros e taxonomias.

Outro aspecto importante é a recomendação por este modelo de representação assumido pelo World Wide Web Consortium, sob a justificativa que sua difusão mundial e a sua representação em RDF proporciona um padrão facilmente interoperável entre diferentes instituições.

No caso de SOCs ainda não representados em SKOS é uma boa prática realizar sua conversão (utilizaremos aqui como exemplo um tesauro por oferecer mais elementos para consideração, mas o procedimento se aplica a diferentes tipos de SOC) para o modelo SKOS. Ao transferir toda sua base de conhecimento e suas relações estruturais podemos utilizar alguns procedimentos, baseados nas recomendações e padronizações do W3C, que objetivam converter o SOC em questão para uma codificação SKOS/RDF.

Estes procedimentos se iniciam com a análise do tesauro em questão de forma a verificar as suas relações padronizadas (tais como, TG, TGP, TE, TEP e TA) e as não-padronizadas. Como resultado deste passo teremos um catálogo de todos os itens de dados e todas as restrições, como uma lista de todas as características do tesauro.

Em seguida, todas estas características devem ser mapeadas para o formato SKOS RDF. Nesse momento se define como cada item de dados será representado no esquema SKOS, gerando uma tabela de cada item do tesauro para o item SKOS que o represente.

Por fim, um especialista em tecnologia da informação deve ser capaz de realizar a função de elaboração de um programa de conversão que gere o SOC em seu formato SKOS RDF. Nesse processo não há interferência nos dados originais e os dois vocabulários têm as mesmas informações, apenas representados de forma diferente, mas facilitando sobremaneira a utilização de agentes para compatibilização, pela sua representação em uma linguagem de dados padronizada, voltada para ser usada por programas de computador.

Como resultado deste processo teremos SOCs capazes de participar de um processo de compatibilização, mesmo organizados em diferentes instituições ou departamentos, permitindo a ação de agentes de software que realizem sua leitura e criação dos objetos registros de conceito.

3.3. DIR 03: Utilização de definições para os conceitos nos SOC

Possibilitar um ganho na identificação do significado semântico de cada conceito a ser compatibilizado, a partir da explicitação de suas características ou propriedades, sob a forma de definição conceitual, que possa ser extraída e tratada por agentes de software.

As notas de escopo/aplicação, onde as definições de um conceito são apresentadas em tesauros, são úteis para um processo de recuperação ou de compatibilização manual. Ao se propor uma situação de compatibilização semântica por processos automáticos elas ganham uma nova importância, uma vez que a utilização de técnicas e tecnologias baseadas em algoritmos computacionais podem ser aplicados e estas informações serem comparadas entre diferentes conceitos.

Diversas técnicas podem ser usadas para a comparação dos textos livres e não estruturados usados para compor as definições (que se encontram nas notas de escopo/aplicação), que por sua vez podem estabelecer parâmetros para determinar a aproximação semântica e a similaridade entre dois conceitos, adicionalmente às informações estruturais extraídas do SOC. Uma destas técnicas é a análise de distribuição semântica, que determina esta similaridade como resultado da similaridade da distribuição linguística (Boleda, 2020). Outra técnica atual é a de word embedding, que agrupa na verdade um conjunto de técnicas para mapear de forma sintática e semântica um texto em linguagem natural, com a utilização de meios estatísticos. Como resultado, palavras de um texto são levadas para um espaço vetorial e podem ser comparadas com palavras de mesmo conteúdo semântico em outro texto, a partir da criação de um embedding space, que represente semanticamente as palavras determinantes do sentido de cada um dos textos (Goldberg, 2017).

O emprego de tais técnicas nas notas de escopo vai no sentido de estabelecer uma representação semântica destas notas, através da extração de frases e palavras que representem seu significado e permitir que esta informação seja utilizada em conjunto com a expressão verbal originalmente comparada, com as comparações que utilizam a estrutura e taxonomia do SOC e com as comparações que utilizam os termos associados. A utilização das técnicas que traduzem a interpretação semântica das notas de escopo pode elevar sua utilização de simples texto de apoio para usuários humanos dos SOC para importantes meios de estabelecer conexão semântica entre conceitos.

Desta forma, nosso objetivo aqui é determinar que a inclusão de notas de escopo para o maior número possível de conceitos nos vocabulários representa esforço decisivo para que o processo de automação da compatibilização e correspondência destes conceitos ocorra de forma mais eficiente e precisa possível. Assim, as informações geradas da nota de escopo através das técnicas de distribuição semântica e word embedding deveriam ser incluídas ao registro de conceito (ver diretriz DIR 05 a seguir) para cada conceito analisado e assim este parâmetro ser também utilizado para a definição da similaridade e distância semântica entre os conceitos.

Além disso, normalmente este campo é utilizado de forma livre e não estruturada, de forma coerente com o que é apontado pela norma ISO 25964, mas a inclusão na norma de uma proposta de futura padronização deste campo, transformando-o em um campo estruturado com informações que pudessem, de forma imediata e automática, participarem do processo de compatibilização, traria ganho significativo para a compatibilização dos SOC mais comumente utilizados, tais como os tesauros.

Assim, um atributo comum, como as notas de escopo, já largamente utilizado para o aumento das capacidades semânticas dos conceitos de um SOC pode ser trazido para compor o sistema proposto aqui, de coordenadas entre vocabulários, aumentando as possibilidades de mapeamentos com alto grau de acerto e acurácia.

3.4. DIR 04: Estabelecimento de identificadores únicos para os registros de conceito

Garantir que a representação dos conceitos, em especial os registros de conceito, que forem gerados de forma automática possam ser identificados de forma única, mesmo em um ambiente aberto na Internet.

Para Dahlberg (1981), um dos requisitos para a criação do registro do conceito, abordado anteriormente, é a utilização da notação utilizada pelo SOC para representar este conceito e possibilitar que participe do processo de compatibilização através da criação de suas matrizes de compatibilidade semântica, sendo de grande importância e um requisito ao processo de compatibilização.

Se num ambiente manual esta identificação notacional dos conceitos já era considerada importante, num ambiente de criação de um espaço semântico gerado de forma automática por agentes de software, passa a ser essencial e condição básica para sua implementação.

Ao abordar e propor a criação da esfera semântica e o espaço de coordenadas semânticas que a compõe, Pierre Levy (2014) também ressalta a

necessidade imperativa de estabelecer um sistema de identificação e endereçamento dos conceitos. Para isto, o autor justifica esta posição afirmando que, com relação ao meio digital, a única certeza que temos é que sua história acaba de começar, ao estabelecer um vetor de crescimento do processo de codificação digital que vivemos em nossa história recente.

Neste sentido, por volta de 1995, com a Web, temos as conexões entre os dados, e a sua identificação realizada através dos Uniform Resource Locators (URL). Para a concretização da sua proposta de esfera semântica, Levy, propõe a criação dos Uniform Semantic Locators (USL), cuja função é a identificação e endereçamento dos conceitos, com a utilização da Information Economy Meta Language (IEML) (Lévy, 2014).

Esta proposta claramente não se coloca como uma substituição das camadas anteriores, e não prescinde delas, da mesma forma que a camada da web não substituiu as suas camadas prévias, pois, considerando as tecnologias atuais, será necessário endereçar dados no meio digital, em seus diversos níveis, usando protocolos de internet e URLs. Nesse caso, se acrescenta uma nova camada de codificação, que permitirá interpretar e utilizar conceitos melhor do que se faz com dados da web e suas URL.

A formação desta esfera semântica completa e unifica, segundo Levy, a ação dos autômatos processadores de símbolos, por sua vez interconectados pela internet, com o conjunto dos dados interconectados pela web. A introdução desta nova camada de endereçamento possibilita a interconexão dos dados, criando uma forma de sinergia diferente daquela que se conhece hoje. O sistema de endereçamento virtual proposto pela IEML define que cada USL distinto codifica um conceito distinto. Como cada conceito pode ser traduzido em línguas naturais, essa identificação na metalinguagem funciona como uma linguagem pivô entre as línguas e os sistemas simbólicos naturais.

Definimos aqui a tomada deste caminho, onde cada conceito precisa ter sua codificação única, mas é necessário fazer sua representação sem que tenhamos ainda disponível um sistema de codificação global, como proposto por Levy na IEML, mas ainda assim estabelecer meios para que os conceitos que formarão nosso espaço semântico possam ser representados por identificadores únicos.

Desta forma, como não temos disponível uma codificação global, e os simples URL que ligam páginas na web, não servem para nosso propósito, recorremos à utilização de um Uniform Resource Identifier (URI), que precisa se apoiar em uma construção do tipo URL, mas pode ser capaz de estabelecer códigos de identificação únicos para objetos na Web.

Para a formação do URI de cada conceito, contendo suas informações, podemos, portanto, partir das notações já definidas em cada SOC, ou estabelecer uma notação sequencial caso um determinado sistema não a possua. Desta forma, o SRI responsável por realizar as buscas descentralizadas pode ser capaz de funcionar utilizando todo o espaço semântico disponível (ver diretriz DIR 05, a seguir) de forma inteligente.

Desta forma, os URI dos registros de conceito gerados deverão ser constituídas minimamente pelos seguintes campos: a) protocolo utilizado (http ou https) e domínio da instituição ou setor detentor e responsável pelo vocabulário de origem; b) nome do vocabulário; e c) identificador único extraído do vocabulário, ou numeração sequencial gerada durante o processo de geração dos registros de conceito.

Como resultado deste processo temos a capacidade de estabelecer, utilizando as camadas disponíveis do endereçamento digital, um identificador único para os conceitos que farão parte de nosso espaço semântico, e que permitirá suas interligações semânticas propostas a seguir.

3.5. DIR 05: Elaboração do espaço semântico a partir da extração dos registros de conceito e seu mapeamento semântico

Elaborar um espaço semântico a partir da extração das informações que farão parte da identidade de cada conceito, da criação dos registros de conceito, e da realização de um mapeamento semântico entre os conceitos de diferentes SOC, que possam vir a ser utilizados em um sistema de recuperação inteligente.

Esta diretriz foi construída a partir dos passos seguidos pelos experimentos de compatibilização realizados em Barbosa (2021), onde pudemos reproduzir caminhos a serem seguidos por um agente de software ao realizar um processo de compatibilização e correspondência.

Para a composição dos registros de conceito, de forma a atender aos processos de compatibilização, propomos a inclusão básica dos seguintes campos, que serão extraídos dos sistemas de organização do conhecimento: (i) expressão verbal do conceito; (ii) notação do conceito extraída do sistema de organização do conhecimento ou gerada de forma automática e sequencial; (iii) termo genérico maior, ou seja, o nível mais abrangente de sua escala hierárquica; (iv) termo genérico imediato; (v) termos específicos; (vi) termos associados; e (vii) definição extraída de sua nota de escopo.

Os passos propostos são especialmente voltados para um processo cujo objetivo não é simplesmente estabelecer um apontamento, ou um mapeamento de um determinado conceito para outro numa situação um para um, e de mesma forma não tem como objetivo a fusão, junção ou integração de vocabulários, mas sim criar objetos denominados registros de conceito que, de forma digital, armazenem as informações semânticas para um determinado conceito em um determinado SOC e, além disso, estabeleçam relações de apontamento para conceitos de outros SOC que possam ser relacionados semanticamente, visando possibilitar interoperação em espaços semânticos. Conforme mostrado em Barbosa (2021), estas relações de apontamento e mapeamento propostas aqui não se limitam a identificar equivalências em conceitos com expressões verbais iguais ou similares, mas identificar equivalência conceitual mesmo com diferentes formas verbais e, de mesma forma, serem capazes de identificar que determinados conceitos com mesma expressão verbal podem ter significados semânticos completamente diferentes e por isso não podem ser mutuamente mapeados.

O processo que recomendamos aqui tem por objetivo a utilização das técnicas e algoritmos computacionais descritos em Barbosa (2021, p. 169), para que seja possível o projeto de agentes de software que não simplesmente estabeleçam ponteiros interligando conceitos similares semanticamente, mas que sejam capazes de determinar quanto um determinado conceito se relaciona a outro, ou seja, se apresenta total compatibilidade semântica, ou uma compatibilidade parcial, determinada num intervalo entre 0 e 1, ou seja,

0 < similaridade semântica calculada <= 1

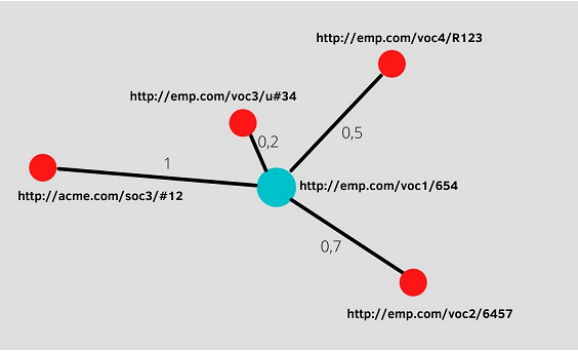

Figura 1. Correspondências semânticas a partir de um conceito (Barbosa, 2021)

Na figura 1 podemos ver que os conceitos foram identificados através de uma URI e, por exemplo, o conceito http://emp.com/voc1/654 foi mapeado para quatro outros conceitos de diferentes vocabulários, e suas medidas de similaridade semânticas para cada um deles foram explicitadas pelas suas relações de ligação. O espaço semântico a ser criado é um espaço multidimensional onde todos os conceitos extraídos se relacionam por estas distâncias semânticas.

Para a operacionalização deste processo, de acordo com as propostas de Neville (1972) e Dahlberg (1981), o primeiro passo é buscar pela identidade sintática entre as expressões verbais presentes nos vocabulários, assumindo como pressuposto sua possível identidade semântica. Como mostramos em nosso experimento e nas discussões teóricas sobre as possíveis e diversas técnicas usadas (Barbosa, 2021), esta identidade sintática se inicia pela total igualdade de caracteres, passando pela pesquisa de subcadeias de caracteres, plurais, formas verbais, e assemelhados.

Para isso, o primeiro passo a ser tomado, para cada expressão verbal de cada vocabulário a ser compatibilizado, é extrair de sua estrutura as informações que irão compor seu registro de conceito, tais como, a própria expressão verbal, sua notação no vocabulário, o termo genérico maior, o termo genérico, os termos específicos, os termos associados e sua nota de escopo.

Após este procedimento inicial, as operações se voltam para dois procedimentos básicos, ou seja,

1. verificar se as formas verbais que se equivalem sintaticamente representam conceitos que são semanticamente iguais ou semelhantes, ou se trata de polissemias, e
2. descobrir nos vocabulários participantes possíveis expressões verbais que, apesar de dessemelhantes sintaticamente, apresentam similaridade semântica que as tornem capazes de serem mapeadas dentro do espaço semântico construído.

Portanto, para chegar a estes propósitos, os passos a serem seguidos, relacionando e detalhando as ações a serem implementadas, são:

1. para cada um dos registros gerados, identificar registros de conceito nos outros vocabulários participantes que possuam a mesma expressão verbal;
2. para cada um dos registros gerados, identificar registros de conceitos nos outros vocabulários participantes que sejam representados por expressões verbais consideradas semelhantes pela aplicação exaustiva das técnicas de nível de elemento, tais como tokenização, lematização, identificação de plurais, ordem

dos termos invertida, extração de hifens e outras similares, listadas com mais detalhes na seção 4 deste trabalho.

3. para cada um dos registros gerados, identificar registros de conceitos mesmo com expressão verbal diferente nos outros vocabulários, mas que tenham semelhança em sua estrutura, usando as mesmas técnicas do item (2), em seus termos genéricos, termos específicos e termos associados;
4. a partir daí aplicar as técnicas de análise de taxonomia e grafos, que permitem identificar similaridades pela utilização da estrutura, validando ou não os conceitos de expressão verbal igual, semelhante, ou dessemelhantes descobertos - observar aqui os procedimentos realizados para os mapeamentos descritos em Barbosa (2021);
5. extrair o significado semântico principal das notas de escopo, através das técnicas de distribuição semântica e word embedding, de forma que permita incluir este resultado nos procedimentos de correspondência e no cálculo da distância semântica entre os conceitos;
6. para cada situação ocorrida anteriormente, estabelecer uma medida de similaridade semântica calculada que vetorize o grau de compatibilidade de cada termo com os termos descobertos que sejam possíveis de serem compatíveis;
7. armazenar as distâncias semânticas calculadas para cada conceito apontado, nos próprios registros de conceito, utilizando os identificadores únicos para representar os conceitos mapeados (figura 1).

Na figura 2 mostramos os sistemas de organização do conhecimento S1, S2, ..., Sn, que respectivamente são utilizados para indexar as bases B1, B2, ..., Bn, e que são percorridos pelo agente de software AS, responsável por executar as operações detalhadas acima e criar o espaço semântico ES.

O que apresentamos aqui, portanto, como caminho a ser seguido não é o estabelecimento de uma matriz de mapeamento booleano, onde a ligação entre conceitos existe ou não existe. O caminho apresentado é o estabelecimento de uma base de dados de registros de conceitos, representando um grupo de SOCs participantes que, ao ser gerada a partir dos caminhos propostos aqui vai ser utilizada por um sistema de recuperação inteligente da informação (diretriz 06), oferecendo uma visão semântica do conjunto de sistemas de organização do conhecimento compatibilizados e servindo de base para buscas interativas inteligentes entre os diversos SOCs por parte dos usuários do sistema, como veremos a seguir.

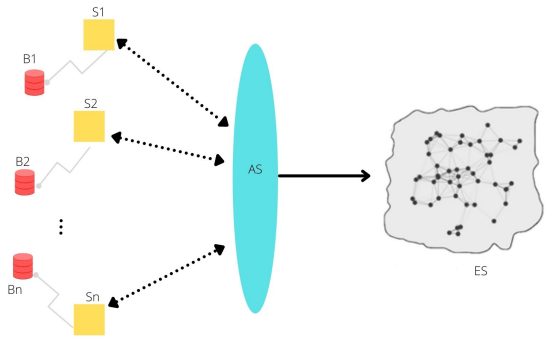

Figura 2. Sistema de criação do espaço semântico (Barbosa, 2021)

3.6. DIR 06: Estabelecimento de um espaço de recuperação inteligente da informação que trabalhe com os registros de conceitos e suas distâncias semânticas

Estabelecer um sistema de recuperação da informação que faça uso do espaço semântico gerado onde conceitos interconectados pelas similaridades semânticas calculadas possam oferecer buscas semânticas inteligentes em ambientes heterogêneos multivocabulários.

Para que seja possível fazer uso efetivo dos processos de compatibilização, correspondência e mapeamento de vocabulários e seus conceitos, estes processos devem ser voltados, como já defendemos anteriormente neste trabalho, para que seja possível a realização da recuperação inteligente da informação, onde se procura suplantar as barreiras impostas pela diversidade e heterogeneidade dos sistemas de organização da informação utilizados para indexar documentos e fornecer informações relevantes para os usuários.

Para isto, nossa proposta se cristaliza na construção de ambientes de recuperação da informação que façam uso do espaço semântico proposto anteriormente, resolvendo a heterogeneidade, atingindo uma interoperabilidade semântica entre vocabulários e adicionalmente trazendo para as mãos do usuário a possibilidade de interagir e definir os limites de compatibilidade que interessam aos seus propósitos.

Este sistema de recuperação da informação ao ser acionado por um usuário ao percorrer a hierarquia de um vocabulário, ou mesmo a partir de um termo livre, poderá ser capaz de identificar os conceitos que atendem àquela busca em outros vocabulários que participem do mesmo espaço semântico, oferecendo conceitos similares nos

outros vocabulários e sendo capaz de afirmar para o usuário, o quanto cada um daqueles conceitos é similar ao conceito pesquisado, em cada um dos vocabulários. Desta forma, o sistema de recuperação da informação pode iniciar sua busca em uma das estruturas taxonômicas de um dos vocabulários participantes e, a partir daí, oferecer os mapeamentos para todos os outros vocabulários armazenados em seu espaço conceitual, fornecendo ao usuário as identidades descobertas e oferecendo as possibilidades semânticas a partir dos registros de conceito e seus mapeamentos.

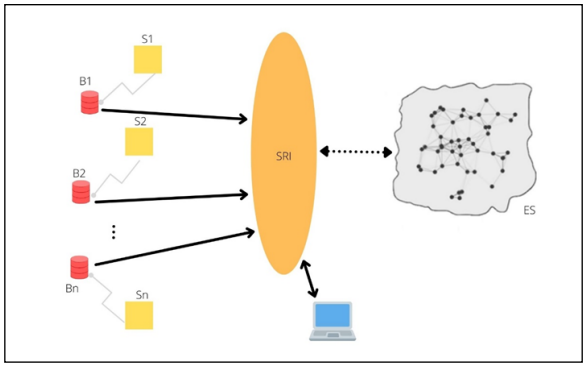

Figura 3. Sistema de recuperação da informação atuando no espaço semântico (Barbosa, 2021)

Um ambiente completo para representação deste sistema pode ser visto na figura 3, onde podemos ver os diversos participantes deste processo para a recuperação de informações distribuídas e indexadas por vocabulários heterogêneos.

Os diferentes passos para a criação e funcionamento deste sistema inteligente de recuperação são dados como:

1. O agente de software AS, programado com algoritmos que executam as técnicas demonstradas em nosso experimento (Barbosa, 2021) e em conformidade com a diretriz 5, executam continuamente seus códigos programáticos, extraindo as informações dos sistemas de organização do conhecimento S1, S2, ..., Sn, e criam e mantém atualizado o espaço semântico ES. Este processo é contínuo, pois uma vez gerado este espaço semântico, as atualizações e manutenções realizadas nos SOCs devem estar representados no espaço semântico. Esta base de dados comporta, portanto, os registros de conceitos extraídos dos SOC e seus interapontamentos, com suas distâncias semânticas calculadas. Em acordo com a diretriz 4, estes registros são identificados sob a forma de URI, garantindo a sua identificação única no sistema;
2. Um usuário interessado em recuperar informações das bases de dados indexadas pelos sistemas de organização do conhecimento do ambiente, faz acesso ao sistema de recuperação da informação SRI, que estabelece uma comunicação com o usuário, buscando no espaço semântico e mostrando os conceitos encontrados para cada SOC e suas distâncias semânticas, podendo oferecer ao usuário a escolha de um limite de similaridade que atenda sua busca. Consideramos que esta busca, conforme descrevemos e propusemos em nosso texto e reafirmamos aqui, não apresenta os mapeamentos matriz de duas dimensões, em linha-coluna, mas sim se apresentam em uma estrutura navegacional multidimensional, onde conceitos estão interligados e são apresentados como uma possibilidade de descobrimento de conhecimento por parte do usuário.
3. Uma vez que definidos os conceitos participantes da busca, em cada SOC participante, de forma automática ou com opcional interferência do usuário, o SRI, então extrai das bases de dados B1, B2, ..., Bn, os documentos indexados pelos termos representativos dos conceitos em cada base de dados. Reafirmando, portanto, que esta recuperação em cada base de dados pode ser feita com os termos sintáticos exatos buscados pelo usuário e confirmados pela similaridade semântica estabelecida pelos processos da diretriz 5, mas também pode ter expressão verbal diferente na recuperação dos documentos em cada base, uma vez que o conceito buscado pode ter esta representação diferente por diversos motivos em cada SOC participante, mas similaridade em seu significado semântico.
4. As escolhas dos limites semânticos estabelecidos e posteriormente validados pelo usuário em sua busca podem ficar armazenadas no sistema, de forma que componham um aprendizado para que futuras buscas multivocabulários possam utilizar este conhecimento acumulado.
5. A adição de novos sistemas de organização do conhecimento a um ambiente como este enriquece o espaço semântico gerado e permite um processo cada vez mais rico de recuperação de informações de forma inteligente, mesmo se tratando de bases de dados diferentes e heterogêneas.

Esta diretriz, portanto, se refere ao processo de construção de um sistema de recuperação da in-

formação que seja capaz de interpretar os objetos 'registros de conceito' armazenados no espaço semântico e fornecer ao usuário buscante uma informação recuperada com ótimos índices de precisão e revocação, superando a barreira da heterogeneidade e da simples igualdade sintática para sua solução.

Portando, como produto desta diretriz obtemos um ambiente de recuperação inteligente da informação que trabalha com os registros de conceitos, ou seja suas identidades semânticas, e suas distâncias semânticas. Este ambiente, composto centralmente pelo espaço semântico construído, é flexível em sua constituição, pois permite a inclusão de novos SOC, que indexem novas bases de dados e é aberto em sua concepção pois utiliza as tecnologias padronizadas e abertas da web semântica. Além, disso é um ambiente interativo, que propõe a participação do usuário ao permitir sua interferência no nível de compatibilidade pretendido.

4. Técnicas aplicadas

Consideramos que o caminho apontado pelas nossas diretrizes para compatibilização e correspondência de vocabulários heterogêneos, a partir de bases teóricas desenvolvidas por autores da Ciência da Informação, deve poder ser colocado em prática utilizando-se, entre outras, técnicas de manipulação de cadeias de caracteres e análises de hierarquias que sejam adequadas ao propósito desejado. Para isso recorremos a diferentes técnicas, algoritmos e sistemas já desenvolvidos pela Ciência da Computação para que seja possível implementar de forma prática nossa proposta teórica.

Inicialmente abordaremos algumas técnicas de correspondência que podem ser utilizadas por algoritmos e sistemas para identificar as possíveis correspondências entre termos de vocabulários. Conforme Euzenat e Shvaiko (2013) e Angermann e Ramzam (2017), e a partir da análise dos procedimentos apresentados em Achichi et al. (2017), Algergawy et al. (2018) e Algergawy et al. (2019), foi possível confirmar nove grandes grupos de tipos de técnicas de correspondência (cada um com diversas subcategorias e especificidades), em que cinco são voltadas para o nível de elemento, com valores literais, e quatro voltadas para o nível de estrutura usando uma estrutura "é-um".

As Técnicas de Nível de Elemento utilizam os valores literais dos conceitos, e/ou suas propriedades, para medir a similaridade semântica. Podemos preliminarmente citar cinco técnicas de nível de elemento: baseadas em recursos formais, baseadas em recursos informais, baseadas em strings (cadeias de caracteres), baseadas em linguagem e baseadas em restrições.

As técnicas baseadas em recursos formais se reportam e mapeiam a conhecimentos prévios fortemente estruturados. Estes recursos podem ser, por exemplo, taxonomias de mais alto nível, específicas para um domínio e padronizadas, como taxonomias que representam grupos de diferentes domínios ou taxonomias de mesmo domínio, mas mais gerais e abrangentes. As baseadas em recursos informais também usam a mesma técnica, mas podem se referir a recursos não padronizados, tais como diretórios de índices estruturados em nível superior às taxonomias a serem compatibilizadas. Nesses dois casos os elementos das taxonomias são apenas comparados e mapeados aos elementos das taxonomias globais.

Ainda no nível de elemento temos as técnicas baseadas em cadeias de caracteres que identificam correspondências com base na comparação e igualdade destas cadeias. Estas técnicas tratam de avaliar a comparação entre os termos e até de suas descrições (Cheatham e Hitzler, 2013). Esta similaridade pode ser calculada, de modo geral, de dois modos: a similaridade de nome e a similaridade de descrição.

A similaridade de nome mede quão similar uma palavra ou grupo de palavras é similar a outra ou a outras. Estas medidas podem ser feitas de múltiplas formas, conforme vemos a seguir. A distância Levenshtein define esta similaridade como o mínimo número de trocas necessário para transformar uma cadeia de caracteres em outra. Cada troca pode ser a transformação necessária para um caractere, seja a sua remoção, inserção ou substituição (Levenshtein, 1966). A distância de Bailey, denominada pelo autor como Euclidiana, mostra o comprimento da conexão necessária para combinar um ponto no espaço euclidiano com outro ponto. Por meio deste, cada caractere de uma sequência é atribuído a um ponto no espaço euclidiano (Bailey, 2004). Já o processo de distância de Hamming (Hamming, 1950; Tanenbaum, 2003) exige que as duas cadeias tenham o mesmo comprimento em número de caracteres e apresenta o número de caracteres diferentes em um mesmo índice posicional. A medida de distância de Lin (Kernighan e Lin, 1970), calcula a probabilidade de uma string ocorrer dentro de um termo. Por fim, Wu e Palmer (1994) classificam cada termo de acordo com a sua profundidade dentro de um corpo de texto usado para comparação.

A similaridade de descrição, por sua vez, leva em consideração termos compostos que devem ser comparados com outras sequências. As principais medidas de similaridade por descrição usadas

hoje são: a distância Jaccard, que representa a semelhança entre dois conjuntos de strings, a similaridade Cosine que considera as sequências como vetores para compará-las e a TF-IDF (Term Frequency–Inverse Document Frequency), que usa a importância de um termo, com base em sua ocorrência em um documento, como uma das bases de comparação (Jones, 1972; Tan et al. 2005).

As próximas técnicas do nível de elemento são as técnicas de correspondência baseadas em linguagem. Estas técnicas são normalmente usadas em conjunto com as técnicas baseadas em cadeias de caracteres e a comparação é geralmente apoiada pelo uso de conhecimento referente ao domínio, que permite analisar o contexto dos conceitos comparados. Podemos citar seis categorias predominantes nestas técnicas:

A Lematização e Morfologia agrupam diferentes formas de inflexão de uma palavra, de forma que elas possam ser analisadas como um único item, como por exemplo seu tipo mais comum, singulares e plurais. Tratam também de coisas tais como abreviaturas.

A Tokenização quebra um texto em palavras, frases, símbolos, ou outros "tokens" significantes. Um único token pode conter mais de uma palavra. Nesse caso o método é apelidado de N-Gram, onde N associa o número de palavras associado ao token.

Já a Eliminação reduz os tokens eliminando os elementos considerados supérfluos, por exemplo, stop-words.

Os métodos com léxicos ou tradutores são usados para traduzir entre idiomas. Normalmente são usados tradutores automáticos, tais como Microsoft Bing e Google Tradutor.

O método de similaridade é utilizado para analisar uma possível similaridade semântica entre conceitos, usando bases de dados, como, por exemplo, a WordNet.

Em seguida, temos a desambiguação de sentidos, que é utilizado para analisar o sentido da sentença no contexto considerado, ou seja, qual o token mais importante a ser considerado para comparação.

Por fim, para completar as técnicas de nível de elemento, temos as técnicas de correspondência baseadas em restrições, que analisam a estrutura interna do sistema de organização do conhecimento utilizado. Sempre agindo em conjunto com outros métodos, esta técnica pode avançar na superação da heterogeneidade conceitual. Podemos dividi-los em duas categorias. Consideramos inicialmente uma similaridade por tipo dos atributos, porque estes elementos descrevem os conceitos em um domínio. Nesse caso, dois conceitos de mesmo tipo, mas de nomes diferentes, que compartilhem a mesma descrição em diferentes atributos podem ser assumidos como similares semanticamente. Por exemplo, dois conceitos tais como "carro de passeio" e "automóvel" que estejam compartilhando atributos como número de portas, espaço na mala e número de bancos, podem ser assumidos como semanticamente similares. A outra categoria que temos é chamada de propriedades-chave, que são usadas para descrever os conceitos que pertencem, por exemplo, a uma taxonomia. Nesse caso, quando os conceitos dentro de uma taxonomia são estruturados de acordo com um determinado ponto de vista correspondente, as taxonomias como um todo podem ser assumidas como similares (Angermann e Ramzan, 2017).

Consideramos a seguir outro grupo de técnicas, aquelas baseadas em taxonomia e que se dedicam a explorar os subconceitos (especialização) e superconceitos (generalização) em uma taxonomia, também conhecidos com relações do tipo é-um. As taxonomias podem diferir no número total de conceitos e na quantidade de relações utilizadas. A análise de similaridade de dois conceitos, em diferentes estruturas, avalia seus subconceitos e seus conceitos superordenados e quanto menos diferentes estas estruturas forem, mais similar semanticamente eles serão. Estas técnicas, em adição às técnicas em nível de elemento, nos permitem criar algoritmos que estabeleçam as medidas semânticas propostas em nossas diretrizes.

Com as técnicas baseadas em grafos, uma taxonomia é considerada como um grafo identificado. Assim, as relações de paridade, ou irmandade, também são tomadas em consideração ao comparar conjuntos e subconjuntos e a distância entre cada um, usando técnicas matemáticas de análise de grafos, tais como, homomorfismo, similaridade de caminhos, similaridade de filhos e similaridade de folhas.

As próximas técnicas consideradas são bem interessantes e focam nas técnicas baseadas em instância. Neste caso a indicação de similaridade entre dois conceitos depende de suas instâncias. Esta similaridade, assim como as anteriores, também depende de dois conjuntos a serem comparados, pois define que conceitos similares devem ter instâncias similares. Apesar de nem todos os sistemas de organização do conhecimento apresentarem a ocorrência de instâncias em suas representações do conhecimento de um domínio, estes objetos, quando presentes, são

de grande utilidade na comparação e no estabelecimento de mapeamentos semânticos e medidas entre conceitos.

Por fim, de uso bastante limitado na literatura e de descrição bastante dispersa e pouco densa, temos as técnicas de correspondência baseadas em modelos que usam lógicas de descrição para superar a heterogeneidade da taxonomia. Solucionadores de satisfação determinam se existe uma interpretação que satisfaça um dado operador booleano, que pode ser verdadeiro ou falso e, Raciocínio de Lógicas de Descrição que é uma família de linguagens formais de representação do conhecimento. Um raciocínio é uma técnica que é capaz de inferir consequências lógicas de um conjunto de entidades (Angermann e Ramzan, 2017).

As técnicas e métodos mostrados acima são, na verdade, grupos e categorias de métodos, onde em cada categoria temos diversas variações e especificidades. Desde os mais complexos com técnicas de manipulação de grafos até aqueles de manipulação de cadeias de caracteres, são métodos não necessariamente criados para os propósitos de compatibilização, mas sim usados em diferentes aplicações e gerados por variados propósitos. A utilização destas técnicas pode e deve levar a serem combinadas de múltiplas formas ao serem aplicadas em um processo de compatibilização. Estas combinações de técnicas geram os diferentes algoritmos e procedimentos que podem ser utilizados sobre sistemas de organização do conhecimento para mapeamento e alinhamento de termos, visando a recuperação dos documentos e informações indexados por estes termos. Em suma, um algoritmo de correspondência usa uma estratégia peculiar consistindo em uma ou mais (geralmente mais de uma) técnicas de correspondência para superar a heterogeneidade de vocabulários. Estes algoritmos, conforme extraído dos relatórios recentes da OAEI (Ontology Alignment Evaluation Initiative), podem ser agrupados naqueles voltados para resolver quatro tipos de heterogeneidade, a saber, terminológica, conceitual, sintática e semiótica.

A heterogeneidade terminológica ocorre quando os descritores dos conceitos são diferentes, podendo ocorrer pelo uso de diferentes idiomas, por exemplo, ou diferentes sublinguagens técnicas, ou pelo uso de sinônimos. Em suma, pelo diferente uso do idioma.

A heterogeneidade conceitual ocorre quando duas taxonomias usam diferentes modelos, representando o domínio em questão com diferentes conceitos, por exemplo, dois conceitos semanticamente similares têm em uma taxonomia um número diferente de subconceitos em relação a outra taxonomia.

Já a heterogeneidade sintática com um viés estrutural, aqui neste caso, ocorre quando diferentes modelos de dados são utilizados para armazenar as taxonomias, por exemplo, uma armazenada em OWL e outra armazenada em RDF. Nesse caso, preliminarmente é necessário realizar uma tradução entre os formatos ou linguagens de representação.

A heterogeneidade semiótica surge quando pessoas fazem diferentes interpretações cognitivas dos conceitos, em especial, nas relações é-um. Por exemplo, quando um usuário de uma taxonomia não espera encontrar “Marcopolo” e “Ferrari” na mesma categoria de “Automóveis”, por exemplo, porque apesar de ambas serem marcas de tipos de veículos, uma serve como meio massivo de transporte de pessoas e o outro representa um modo de dirigir individual e esportivo.

Todas estas técnicas apresentadas aqui podem ser usadas para criar sistemas de correspondência, onde definimos que um sistema para correspondência de taxonomias e ontologias é um aplicativo ou um conjunto de aplicativos de software que tem por objetivo identificar e resolver diversos tipos de heterogeneidade na execução de uma operação de correspondência (Otero-Cerdeira et al., 2015). Ou seja, o que chamamos aqui de sistema de correspondência é um programa de computador desenvolvido para resolver um determinado problema de compatibilização entre ontologias específicas em um determinado contexto. Para isso o desenvolvedor utiliza bases de dados, ou data-sets, pré-determinados para aplicar diferentes algoritmos e gerar um mapeamento entre dois SOC em questão.

Portanto, a determinação da equivalência entre os conceitos e, mais importante, conforme descrito em nossas diretrizes, o estabelecimento das medidas de compatibilidade semântica entre dois conceitos pode ser alcançado com a aplicação dos métodos resumidamente descritos nesta seção, metodologicamente organizados em aplicativos de software escritos para este fim. Estes aplicativos de software, ao atuar sobre a nomenclatura dos conceitos, sobre sua posição relativa nas hierarquias taxonômicas, sobre a comparação de suas instâncias e sobre suas definições, quando presentes, são capazes de estabelecer as medidas de compatibilidade entre conceitos, conforme representado esquematicamente na figura 1. Desta forma a navegação neste espaço semântico criado será capaz de estabelecer uma compatibilização entre vocabulários heterogêneos de forma replicável, aberta e automática,

executada por máquinas, com vista a um processo de recuperação da informação inteligente e preciso.

5. Considerações finais

A linha de raciocínio que norteou o desenvolvimento desta pesquisa sempre foi o estudo e a compreensão das causas da heterogeneidade entre sistemas de organização do conhecimento e os caminhos para a sua interoperabilidade, de forma que seja possível estabelecer processos de recuperação inteligente da informação. Estes processos devem ter por objetivo permitir aos usuários recuperar informações de bases diversas indexadas por vocabulários heterogêneos, sem que seja necessário alterar estes vocabulários e sem que o usuário precise manualmente percorrer diferentes estruturas taxonômicas, dispendendo tempo precioso com informações imprecisas e com análise manual de múltiplas fontes e bases de dados. Em outras palavras, procuramos estudar e propor soluções que permitam que a capacidade de obter da Internet e da Web dados que respondam às necessidades dos indivíduos seja capaz de ser tão efetiva como o avanço do desempenho dos computadores e das redes.

Para isto este foi nosso foco principal: a recuperação da informação. Como diz Lévy, o cenário contemporâneo dificulta conseguir achar a informação que nos interessa e que tem mais valor, pois buscar e receber uma resposta avassaladora de informações não relevantes ou não significativas é quase tão ruim como buscar informações e conseguir achar pouco ou nenhum resultado.

Portanto, nossa pesquisa enfrenta na prática um problema que o mundo tem hoje, ao tentar produzir soluções para a recuperação da informação, não somente de dados de pesquisa, mas também bases bibliográficas, de livros e publicações científicas e dados em geral dispersos, armazenados e indexados de forma descentralizada. Isto nos levou a este trabalho, que se soma aos esforços de estudar e propor soluções de forma interdisciplinar que superem a heterogeneidade em todos os níveis dos dados digitais produzidos atualmente.

As propostas de Neville e Dahlberg têm destaque e importância reconhecidos na Ciência da Informação para a compreensão e solução do processo de compatibilização de vocabulários, mas tem uma grande dependência de atuação do ser humano. Assim, como seus próprios autores colocam, estes procedimentos não foram feitos diretamente para implementação em processos automatizados, mas como pudemos demonstrar em nosso experimento, tem um papel de grande importância com sua contribuição metodológica para dar forma e sentido à utilização de técnicas computacionais cujos propósitos se voltem para a compatibilização e correspondência de sistemas de organização do conhecimento.

Por fim, com base nos estudos e experimentos realizados e mostrados com detalhes em Barbosa (2021), fomos capazes de apontar um caminho com diretrizes que possam ser capazes de serem aplicadas a vocabulários de ambientes heterogêneos e gerar um espaço semântico. Possibilitando, assim, que a partir daí, possa ser utilizado por um SRI que permita a recuperação inteligente da informação nestes ambientes, de forma flexível, permitindo sua expansão, e interativa, que permita a participação do usuário de forma ativa nas suas buscas.

Portanto, nossa proposta foi apresentar aqui um caminho possível a ser implementado utilizando os recursos e a inteligência que já estão presentes nos sistemas de organização do conhecimento tradicionais que, aliados às técnicas e procedimentos computacionais que já temos disponíveis e às bases teóricas da Ciência da Informação, possa apontar uma solução para a heterogeneidade. Este caminho e diretrizes podem, por um lado, resolver problemas e compatibilidade com vocabulários já existentes, mas podem também serem usados para orientar e propor possibilidades de construção mais adequadas para SOC ainda a serem desenvolvidos, que utilizem novas tecnologias.

Referências

Achichi, M. et al.(2017). Results of the Ontology Alignment Evaluation Initiative 2017. Ontology Alignment Evaluation Initiative Conference. Viena.

Agraev, V. A. et al.(1974). Information retrieval system compatibility. // Automation Documentation and Mathematical Linguistics. 2, 29-37.

Algergawy, A. et al.(2018). Results of the Ontology Alignment Evaluation Initiative 2018. Ontology Alignment Evaluation Initiative Conference. Monterey.

Algergawy, A. et al.(2019). Results of the Ontology Alignment Evaluation Initiative 2019. Ontology Alignment Evaluation Initiative Conference. Auckland.

Angermann, H.; Ramzan, N. (2017). Taxonomy Matching Using Background Knowledge: Linked Data, Semantic Web and Heterogeneous Repositories. Springer International Publishing.

Bailey, D. (2004). An efficient euclidean distance transform. Proceedings of the 11th International Semantic Web Conference. Berlin, Germany: Springer. 394-408.

Barbosa, N. T. (2021). Para uma economia da informação semântica: a construção de ambientes semânticos para a recuperação inteligente da Informação. Rio de Janeiro: Universidade Federal Fluminense. Tese de doutorado.

Bocatto, V. R. C.; Torquetti, M. C. (2012). Interoperabilidade entre linguagens de indexação. // Informação & Informação. Londrina. 17:3, 76-101.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. // Annual Review of Linguistics. 6, 213-234.

Campos, M. L. A.; Campos, M. L. M.; Davila, A. M. R.; Gomes, H. E.; Campos, L. M.; Lira, L. (2009). Aspectos Metodológicos no Reúso de Ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos. // RECIIS. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde. 3, 64-75.

Cheatham, M.; Hitzler, P. (2013). String similarity metrics for ontology alignment. International Semantic Web Conference ISWC 2013. Heidelberg: Springer. 294–309.

Coates, E. J. (1970). Switching languages for indexing. // Journal of Documentation. London. 26:2, 102-110.

Dahlberg, I. (1981). Towards establishment of compatibility between indexing languages. // International Classification. 8:2, 88-91.

Euzenat, J.; Shvaiko, P. (2013). Ontology Matching, 2 ed. Heidelberg: Springer.

Gantz, J.; Reinsel D. (2010). The digital universe decade - are you ready? IDC White Paper. May.

Gardin, J. C. (1967). Recherches sur l'indexation automatique des documents scientifiques. // Revue d'informatique et de recherche opérationnelle. 1:6, 27-46.

Gardin, J. C. (1973). Document analysis and linguistic theory. // Journal of Documentation. London. 29:2, 137-68.

Gardin, N. (1969). Le lexique intermédiaire: un nouveau pas vers la cooperation internationale dans le domaine de l´information scientifique et technique. // Bulletin de l'UNESCO: à l' Intention des bibliothèques. Paris. 23:2, 66-71.

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies.

Hamming, R. W. (1950). Error detecting and error correcting codes. // Bell System Technical Journal. 29:2, 147-160.

Hammond, W; Rosenborg, S. (1962). Experimental study of convertibility between large technical indexing vocabularies. // Technical report IR-1. Datacontrol Corporation. Silver Spring. Ago.

Henderson, M. M. et al.(1966). Cooperation, Convertibility, Compatibility Among Information Systems: A Literature Review. // National Bureau of Standards. Jan.

Horsnell, V. (1975). The Intermediate Lexicon: an aid to international co-operation. // Aslib Proceedings. 27:2, 57-66.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. // Journal of Documentation. 28:11, 21.

Kernighan, B.; Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. // Bell System Technical Journal. Blackwell Publishing. 49, 291.

Levenshtein, W. (1966). Binary codes capable of correcting deletions, insertions and reversals. // Soviet Physics Doklady. Springer. 10, 707–710.

Lévy, P. (2014). A Esfera Semântica. Tomo 1: Computação, cognição, economia da informação. Editora Annablume.

Lévy, P. (2009). From Social Computing to Reflexive Collective intelligence: The IEML Research Program. CRC, FRSC, University of Ottawa.

Lévy, P. (2019). IEML - A metalinguagem da Economia da Informação - Livro Branco. Pré-print, não publicado.

Newman, S. M. ed. (1965). Information Systems compatibility. Washington: Spartan Books.

Nurcan, S. et al.(1999). Change process modeling using the EKD – Change Management Method. 7th European Conference on Information Systems, ECIS' 99. Copenhagen, Denmark. 513-529.

Otero-Cerdeira, L.; Rodríguez-Martínez, F. J.; Gómez-Rodriguez, A. (2015). Ontology matching: a literature review. // Expert Systems with Applications. 42, 949.

Smith, L. C. (1974). Systematic searching of abstracts and indexes in interdisciplinary areas. // American Society of Information Science. 25, 343-353.

Soergel, D. (1972). A universal source thesaurus as a classification generator. // Journal of the American Society for Information Science. 23:5, 299-305.

Soergel, D. (1974). Indexing languages and thesauri: Construction and maintenance. Los Angeles, CA: Melville Wiley Information Science Series.

Statista (2021). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. // Statista Research Department. https://www.statista.com/statistics/871513/worldwide-data-created/ (2022-03-01).

Svenonius, E. (1975). Translation between hierarchical structures: an exercise in abstract classification. Ordering systems for global information networks. 204-211.

Tan, P. N.; Steinbach, M.; Kumar, V. (2005). Introduction to data mining. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Tanenbaum, Andrew S. (2003). Redes de Computadores. 4ª Edição. Editora Elsevier.

Unesco (1971). Unisist study report on the feasibility of a world science information system. Paris.

Wellisch, H. (1972). A concordance between UDC and Thesaurus of engineering and scientific terms. Proceedings of the International Symposium UDC in Relation to Other Indexing Languages. Novi, Yugoslavia.

Wersig, G. (1975). Experiences in compatibility research in documentary languages: Ordering systems for global information networks. 423-430.

Wu, Z.; Palmer, M. (1994). Verbs semantics and lexical selection. 32nd Annual Meeting on Association for Computational Linguistics (ACM). New Mexico. 133-138.

Enviado: 2022-03-30. Segunda versión: 2022-09-26.
Aceptado: 2022-09-26.

Índice de autores

Author index

Índice de materias en español

Subject index in Spanish

Índice de materias en inglés

Subject index in English