Biblios
ISSN 1562-4730 (online)

Experimentos de extracción terminológica orientados a la construcción de un tesauro de bibliotecologia

Adriana Suárez-Sánchez

Instituto de Investigaciones Bibliotecológicas y de la Información. Universidad Nacional Autónoma de México, México

ORIGINAL

Resumo

Objetivo. Evaluar dos métodos de obtención terminológica (extracción terminológica manual y extracción terminológica automatizada) como técnicas viables para la obtención de términos que puedan ser incorporados como descriptores de un tesauro de bibliotecología

Metodo. La metodología empleada fue exploratoria-cuantitativa y se fundamentó en dos experimentos de extracción terminológica: (1) extracción manual y (2) extracción automatizada. El proceso de extracción terminológica manual fue llevado a cabo por un profesional con formación académica multidisciplinaria mientras que en la extracción terminológica automatizada se empleó el programa WordStat. Tanto en el proceso de extracción manual como automatizada se partió del mismo corpus, formado por 283,585 palabras que corresponden a 59 artículos de la especialidad publicados en la revista Investigación Bibliotecológica durante los años 2019 y 2020.

Resultados. Los resultados muestran que: la extracción terminológica manual implicó una cantidad considerable de tiempo humano de dedicación pero el 82% de los términos resultaron útiles y fueron establecidos como descritores viables para el tesauro. En comparación la extracción terminológica automatizada fue un proceso que implicó menor tiempo, pero sólo el 12% de los términos fueron establecidos como descriptores viables para el tesauro.

Conclusiones. Se encontró que cada una de las técnicas de obtención terminológica resultó útil, pero presentaron diferencias. Mientras la extracción manual implicó un nivel alto de recursos humanos y tiempo, sus resultados se observaron excelentes. En contraste, la extracción automatizada requirió menor inversión humana y tiempo, pero la cantidad de términos útiles también fue menor. Se concluye que la experimentación con diversas técnicas de extracción terminológica es importante, asociada a la base terminológica que constituye el pilar de todo vocabulario controlado.

Palavras-chave

Extracción terminológica automatizada; Extracción terminológica manual; Terminología; Tesauros.

Experiment of terminology extraction oriented to the construction of a library science thesaurus

Abstract

Objective. The objective of this artcicle is to evaluate two terminology extraction techniques: manual terminology extraction and automated terminology extraction, to assess the effectiveness of each process in obtaining useful terms for the construction of a library thesaurus.

Method. The methodology used was exploratory-quantitative and was based on two terminology extraction experiments: (1) manual extraction and (2) automated extraction. The manual terminology extraction process was carried out by a professional with multidisciplinary academic training, while the automated terminology extraction process was carried out using WordStat program. Both, manual and automated extraction processes were based on the same corpus, consisting of 283,585 words corresponding to 59 articles about library and information science that were published in the journal Investigación Bibliotecológica during the years 2019 and 2020.

Results. The results show that: manual terminology extraction provided excellent results, 82% of the terms were useful and were established as viable descriptors for the thesaurus. In comparison, automated extraction was a time-consuming process, but only 12% of the terms proved useful and were established as viable descriptors for the thesaurus.

Conclusions. It was found that each of the terminology retrieval techniques was useful but presented differences. While manual extraction required a high investment of human resources and time, its results also showed high effectiveness. In contrast, automated extraction required less human investment and was fast in time, but its results in this experiment were

less accurate and useful. It is concluded that experimentation with various terminology extraction techniques is important, associated with the terminology base that is the cornerstone of any controlled vocabulary.

Keywords

Automatic term extraction; Manual terminology extraction; Terminology; Thesaurus.

1 Introducción

El origen del término tesauro proviene de la palabra griega θησαυρος" y en la antiguedad la denominación implicaba la noción de tesoro. Posteriormente, entre los siglos XVII y XIX, la función y estructura de los tesauros estuvo ligada a los diccionarios, como una compilación de palabras. En el periodo hubo dos obras que marcaron una diferencia entre el tesauro y el diccionario. La primera de ellas *English Synonymes Explained in Alphabetical Order with copious illustrations and examples drawn from the best writers* fue publicada en el año de 1816 por George Crabb. La segunda obra *Thesaurus Facilitate the Expression of Ideas and Assist in Literary Composition* fue publicada en 1852 por Peter Mark Roget. Como señala Gil Urdiciain (1998, p. 6) "la organización alfabética del Roget, con sus referencias de véase que reenvían los términos específicos a términos genéricos, la diferenciación de sinónimos y la organización jerárquica de sus entradas, permiten un control del vocabulario que hace posible considerarlo no sólo precursor de los tesauros documentales, sino que incluso podría ser utilizado en la actualidad como herramienta de control terminológico en recuperación de información."

Al paso del tiempo, la palabra tesauro tomó otro significado y poco a poco se integró en la disciplina bibliotecaria y otras áreas que organizan recursos de información desde una perspectiva temática. En el contexto bibliotecario, los tesauros son un tipo de lenguaje documental con amplia tradición (NAUMIS PEÑA, 2007). Desde la década de los sesentas, periodo en el que tuvieron gran auge, han sido implementados en el tratamiento temático de recursos informativos de áreas de conocimiento especializadas: medicina, derecho, literatura, física, historia, arquitectura, arte, etc. En la actualidad, los tesauros se han instaurado como instrumentos altamente útiles asociados a tres funciones preeminentes de las bibliotecas y otros centros documentales: (1) representación, (2) organización y (3) recuperación temática de los recursos en un sistema.

En 1986, la norma ISO-2788-1986 definía el tesauro como:

> Un instrumento de control terminológico que traduce a un lenguaje sistémico o documental el lenguaje empleado en los documentos y por los usuarios […] es un sistema dinámico de términos relacionado semántica y jerárquicamente, que se aplica a un campo específico del conocimiento (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 1986, p. 18)

El estándar más actual sobre tesauros ISO-25964-1 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011, p.12) señala que un tesauro es:

> Un vocabulario controlado y estructurado en el que los conceptos son representados por términos, organizados de manera que las relaciones entre los conceptos son explicitadas y los términos preferidos son acompañados por vínculos de entradas para sinónimos y cuasi sinónimos. El propósito de un tesauro es guiar tanto al indexador como al buscador para que seleccionen el mismo término preferido o la combinación de términos preferidos para representar un tema determinado.

En la bibliotecología, la construcción de tesauros y las técnicas asociadas a su desarrollo han sido actividades relevantes (CURRAS, 1998; LANCASTER, 2002; NAUMIS PEÑA, 2007; CHU, 2010). Actualmente, los tesauros son sistemas para la organización del conocimiento (HODGE, 2000) aplicables a la organización temática de colecciones físicas e instrumentos que se han insertado en el ciberespacio, donde predomina una necesidad constante de vocabularios que permitan la explicitación terminológica de áreas de conocimiento y la organización temática de los recursos de información. Refieron a Laguens García (2006, p. 106), quien señala que:

> Los lenguajes documentales, sean clasificaciones terminológicas o tesauros, siguen siendo instrumentos indispensables para estructurar la información y el conocimiento en sistemas organizados de almacenamiento y difusión de documentos, sobre todo si tenemos presente que en la catalogación actual se tiende a conceder una mayor relevancia a los descriptores como puntos de acceso al registro bibliográfico.

Desde la perspectiva de su construcción, los tesauros son sistemas fundamentados en diversas etapas de desarrollo: determinar sus objetivos, definir sus características, establecer recursos de vocabulario, asentar los términos, instaurar relaciones entre los términos, evaluar el producto final, etc., (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011). Entre las etapas mencionadas, la determinación terminológica vinculada a las denominaciones conceptuales del área de conocimiento tratada es fundamental (LANCASTER, 2002); en vista de que la concisión de un modelo de conceptos materializado en términos es la base de todo tesauro. Hjørland (2007) señala que, en gran medida, la base de la organización del conocimiento radica en el mapeo de conceptos y su correspondiente concreción en etiquetas lingüísticas. Bräscher (2014) coincide con Hjørland cuando afirma que los sistemas para la organización del conocimiento se fundamentan en conceptos que para su explicitación emplean estructuras lingüísticas.

Para la concreción de la base conceptual y terminológica que constituye el pilar de un tesauro, se cuenta con diversas técnicas de asentamiento terminológico, entre las cuales es posible mencionar (VIVALDI y RODRÍGUEZ, 2001; SUÁREZ-SÁNCHEZ, 2018, p. 68):

- Reúso de terminología previa
- Extracción terminológica manual
- Extracción terminológica automatizada
- Encuestas terminológicas a expertos del área de conocimiento
- Encuestas a usuarios del área de conocimiento

Entre las múltiples técnicas, cada una tiene particularidades en cuanto a proceso y resultados. A partir de ello, el presente trabajo tiene por objetivo:

- ❖ Evaluar dos métodos de obtención terminológica (extracción terminológica manual y extracción terminológica automatizada) como técnicas viables para la obtención de términos que puedan ser incorporados como descriptores de un tesauro de bibliotecología.

El artículo pretende contribuir en la reflexión y explicitación de métodos para la obtención de terminología que constituya la base de vocabularios controlados y mostrar un caso práctico del proceso de obtención de términos a partir de extracción manual y extracción automatizada. La experimentación de diversas técnicas de extracción terminológica para la construcción de tesauros y otros vocabularios controlados es un asunto preeminente, puesto que en la bibliotecología son necesarios sistemas para la organización del conocimiento que apoyen la indización y clasificación de recursos informativos en la biblioteca física, las bibliotecas digitales, los repositorios de información, las bases de datos y la web misma (CHU, 2010; GOLUB, TUDHOPE, ZENG y ŽUMER, 2014).

En este orden de ideas, el presente artículo evalua dos técnicas de extracción terminológica (manual y automatizada), orientadas a la construcción de un tesauro de bibliotecología. Si bien el estudio no considera todas las técnicas existentes para cumplir con esta finalidad y tampoco es totalitario -en el sentido de que no contempla un corpus absolutamente representativo del dominio-, sus alcances permiten: 1) reflexionar sobre dos métodos que pueden ser aplicados en la construcción de la base terminológica que constituye el pilar de todo vocabulario controlado y 2) es un artículo útil para profesionales de la información que se enfrentan a la tarea de construir un tesauro y se encuentran en la etapa de selección de los procesos y herramientas idóneos para lograr su objetivo. La justificación de llevar a cabo una investigación de tal naturaleza radica en la necesidad de contar con casos prácticos que orienten a los bibliotecarios y otros profesionales de la información en la construcción de vocabularios controlados.

2 Revisión de literatura

En la sección que se presenta a continuación se ofrece una breve revisión teórica sobre la extracción terminológica manual y la extracción terminológica automatizada. De manera general se retoman aspectos como definición, perspectivas desde el enfoque linguístico y documental y el proceso que las conforma. Es importante tener en cuenta que todo tarea de extracción terminológica debe considerar una amplia revisión teórica del tema que oriente las actividades a realizar. Así, en la medida que tengamos claras las etapas que forman cada proceso, enfocaremos adecuadamente los recursos humanos, económicos y tecnológicos de nuestro proyecto.

2.1 Extracción terminológica manual

La extracción terminológica manual (ETM) es una técnica antigua para el establecimiento de compendios terminológicos y ha sido ampliamente utilizada en la construcción de vocabularios controlados que incluyen desde diccionarios generales hasta vocabularios de especialidad (ARNTZ y PITCH, 1995). Al principio, fue asunto exclusivo de la terminología, pero poco a poco se vinculó con las ciencias de la información, en tanto que "la terminología presenta una serie de similitudes sorprendentes con las ciencias de la información" (SAGER, 1993, p. 25) ya que ambas son disciplinas pragmáticas enfocadas a construir entramados comunicativos en los que la terminología juega un papel importante.

La determinación de un sistema terminológico mediante extracción manual contempla las siguientes actividades (AUGER y ROUSSEAU, 2003):

- Elección del dominio
- Delimitación del campo de trabajo
- Medios de exploración del campo de trabajo
- Establecimiento de corpus
- Confección de la nomenclatura
- Tratamiento de la nomenclatura
- Clasificación de las unidades y presentación del léxico
- Normalización

La elección del dominio y delimitación del campo de trabajo dependen del enfoque del tesauro. Hasta hace algún tiempo la división de las ciencias era el precepto dominante en la construcción de vocabularios de especialidad (KUMBHAR, 2012), pero en los tiempos recientes han surgido nuevos campos de trabajo terminológico (BARITÉ, 2009). En consecuencia, los conceptos de "disciplina" y "área de conocimiento" conviven con la noción de "dominio", entendido como "cualquier grupo que es útil para la construcción de un sistema para la organización del conocimiento" (SMIRAGLIA, 2015, p. 86). Bajo esta denominación, es posible construir tesauros de áreas de conocimiento como "mecatrónica", colecciones como "museo de culturas populares", conjuntos de objetos como "alfarería mexicana" o catálogos de productos como "vinos de tienda Alianza", etc.

La ETM se "basa firmemente en un corpus de documentos" (SAGER, 1993, p. 189). El corpus "debe ser representativo del dominio a estudiar y de sus subdominios, en conformidad con el plan de trabajo previamente establecido" (AUGER y ROSSEAU, 2003, p. 38). La diversidad de las fuentes posibilita cubrir el conjunto de nociones y su autoridad garantiza la extracción de terminología especializada. Algunos criterios que guían la selección del corpus son: pertinencia en relación con el dominio, naturaleza del documento (didáctica, académica, publicitaria), nivel de especialización, lengua de expresión, origen geográfico, autoridad del autor (especialista, técnico, estudiante) y fecha de producción (AUGER y ROSSEAU, 2003).

La producción literaria representativa de un dominio es la que provee los términos más relevantes y acordes al lenguaje tanto de los autores principales como de los usuarios (CABRÉ, 1999). En el trabajo terminológico es conveniente contar con la visión experta de profesionales del dominio que evalúen la solidez literaria y representativa del corpus, puesto que "el desarrollo de vocabularios controlados, sostenidos en la legitimación que da la literatura misma, implica crear estructuras conceptuales con alto poder de representatividad, ricas en

expresiones y relaciones semánticas aceptadas y aptas para la búsqueda y el intercambio de información en diferentes contextos (BARITÉ, 2009, p. 22).

En la ETM la habilidad humana para detectar los términos es esencial. Éstos se perciben como símbolos lingüísticos tras los cuales subyacen conceptos que posibilitan la descripción, la clasificación y la previsión de los objetos cognoscibles (ABBAGNANO, 1963). La confección de la nomenclatura coloca al profesional frente a un conjunto de interrogantes: ¿qué términos hay que seleccionar? ¿Cómo reconocer las unidades terminológicas? ¿Cómo determinar el sentido exacto de los términos y las relaciones de sinonimia, cuasi-sinonimia, términos genéricos, términos específicos, etc.? (AUGER y ROSSEAU, 2003). Para facilitar la tarea, es importante que el proceso de extracción terminológica manual sea realizado por profesionales, tanto expertos en terminología como especialistas del dominio (SAGER, 1993).

Las palabras son "unidades léxicas, compuestas de uno o más fonemas, a las que corresponde un significado" (LUNA TRAILL, VIGUERAS ÁVILA y BAEZ PINAL, 2005, p. 169). Y, en comparación, los términos poseen tres rasgos diferenciadores (CABRÉ, 2009, p.10):

1) Un componente cognitivo, la percepción y categorización de la realidad por parte de las especialidades, los términos vinculan la representación de dicha categorización de la realidad.
2) Un componente lingüístico, por cuanto las unidades terminológicas son signos lingüísticos, pertenecen a las lenguas naturales, forman parte de sus gramáticas y se describen a través de las mismas propiedades, estructuras y condiciones que describen las unidades lingüísticas,
3) Un componente social, los términos sirven para la comunicación de los expertos, formar nuevos expertos y para divulgar el conocimiento especializado; aunado a ello, identifican grupos socio-profesionales.

No todas las palabras constituyen unidades terminológicas. Para su selección, un término, debe ser una unidad lingüística con significado particular y delimitado, cuya función es el establecimiento de una relación entre una realidad (real o abstracta) y su denominación (CINTRA, TÁLAMO, LARA Y KOBASHI, 2002, p.40). Los términos simbolizan conceptos del dominio; de manera indirecta, cuando seleccionamos un término operamos con conceptos (GUINCHAT Y MENOU, 1983). El primer criterio para la elección de términos es su pertenencia al dominio establecido y, el segundo, su idoneidad (léxica, sintáctica y semántica) en relación con el tesauro al que se agregará (LANCASTER, 2002). Los conceptos, que usualmente se concretan en términos para la construcción de vocabularios controlados, representan cosas y sus partes físicas, materiales, actividades o procesos, eventos o sucesos, propiedades de personas, objetos, materiales o acciones, disciplinas o campos temáticos, unidades de medida, tipos de gente u organizaciones (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011, p.18).

Dado que el término es un símbolo convencional en cierto dominio de saber (CABRÉ, 2009), su función es asociar una realidad (concreta, abstracta o imaginaria) con un signo lingüístico que le ha sido designado por consenso. Así pues, la selección terminológica se enfoca en la recopilación de categorías léxicas denominativas que pueden ser:

- Unidades simples: formadas por una sola palabra: "indización", "metadatos".
- Unidades compuestas: formadas por dos o más palabras que no se pueden disociar sin cambiar el sentido significativo: "análisis documental", "comportamiento informativo"
- Unidades sintagmáticas: formadas por varias palabras dependientes de un núcleo sustantivo: "sistemas para la organización del conocimiento".

La recopilación terminológica en un corpus genera un sistema de términos. A lo largo de su trabajo, los encargados del proyecto seleccionan unidades idóneas que forman parte de la terminología del dominio. En la elección se consideran aspectos como pertenencia al dominio, garantía literaria, garantía de uso, garantía lógica, garantía estructural y adecuación terminológica. En esta etapa también se considera el establecimiento asociativo de sinonimia o envíos terminológicos. El propósito es establecer con precisión un banco terminológico que será la base del tesauro (LANCASTER, 2002).

Por último, la presentación de la terminología puede concretarse en una base de datos sencilla (tablas, listados) o bien en una base compleja que contemple el término, el concepto, la sinonimia y relaciones genéricas y/o partitivas, según el vocabulario controlado que se intenta construir. Para una lista temática, por ejemplo, resulta suficiente con la presentación de la terminología en un listado, mientras que los sistemas más complejos como tesauros implican la identificación del término y el establecimiento de relaciones con otros términos, lo que conlleva el uso de sistemas robustos para su codificación, generalmente programas especializados.

2.2 Extracción terminológica automatizada

La extracción terminológica automatizada (ETA) hunde sus raíces en el procesamiento del lenguaje natural. Un área de la ciencia y la tecnología "vinculada con la catalogación, la categorización, la clasificación y la búsqueda de grandes cúmulos de información, particularmente en forma textual" (STRZALKOWSKI, 1999. p. xiii) que emergió con gran fuerza en la década de los sesentas, de la mano de la propuesta de Hans Peter Luhn respecto a las palabras en contexto (Keywords in Context-KWIC) para determinar las unidades léxicas relevantes en un documento con fines de análisis terminológico o indización automatizada.

La ETA se fundamenta en el trabajo realizado por los extractores terminológicos (ET), programas informáticos con la capacidad para proponer candidatos a términos a partir del tratamiento automático de un corpus de textos especializados (ESTOPÁ, 2009). Los ET datan de la década de los sesenta y tienen por objetivo el tratamiento lingüístico/ terminológico de grandes cantidades de datos en poco tiempo a partir de criterios sistemáticos.

Para la detección de los términos, los extractores terminológicos disponen de tres principios (CHUNG, 2003):

1) Principios estadísticos: comparan el número de ocurrencias de una palabra en un corpus técnico en comparación con el mismo corpus o un corpus de consulta. La aproximación estadística basada en el uso de frecuencia y rango de la forma de las palabras es la técnica de operación más común en los programas de extracción terminológica automatizada.
2) Principios lingüísticos: consideran aspectos lingüísticos como selección de estructuras nominales, sintaxis terminológica, derivaciones, flexiones y lematización. Su incorporación en los programas de extracción terminológica asume conocimiento experto: cómputo, matemáticas y lingüística.
3) Principios híbridos: engloban ambas posibilidades, es decir, tanto principios estadísticos como lingüísticos.

De manera general, la ETA considera las siguientes etapas (CHUNG, 2003; BENAVENT y PARRILLA, 2006; LUO, XIE, CHEN y YE, 2018):

- Establecimiento del corpus
- Preparación del corpus
- Criterios de exclusión de no-términos
- Análisis en el extractor terminológico
- Depuración de la propuesta generada por el extractor terminológico

El establecimiento del corpus considera el tamaño y representatividad de la muestra. La calidad de los términos que se obtengan, igual que la extracción manual, depende de la representatividad de los documentos seleccionados. El corpus deberá (SINCLAIR, 1991, p. 18):

1) Ser lo suficiente amplio para generar un conglomerado de ocurrencia de palabras y permitir descripciones numéricas que constituyen la base de los sistemas estadísticos. Sinclair recomienda un corpus mínimo de 90000 -100000 palabras.
2) Formarse de textos completos que constituyan obras de conocimiento integras.
3) Coincidir idiomáticamente con el lenguaje del tesauro que se está construyendo para evitar problemas de traducción.

La preparación del corpus contempla varios aspectos antes de ser ingresado al extractor terminológico, entre ellos están: formato de los archivos de ingreso, eliminación de imágenes, tablas o gráficos, trabajo de depuración y limpieza, listas de paro, etc. (LÓPEZ MATEO y OLMO CAZEVIEILLE, 2017). Las posibilidades de los extractores automáticos en relación con tales aspectos son diversas, WordSmith, enfocado al trabajo léxico, admite solamente corpus en formato TXT y guardados en algún dispositivo de almacenamiento; en comparación, WordStat, para el trabajo de minería textual, posibilita la carga de archivos en formato diverso y de fuentes tanto almacenadas en algún dispositivo como recursos de la web (Facebook, Twitter, Survey Monkey, entre otros).

En la extracción terminológica automatizada la exclusión de no-términos es una etapa crucial. Los no-términos son unidades léxicas con nula posibilidad de representar conceptos del dominio. En su descarte se construyen listas de paro que incorporan unidades léxicas vacías, tales formas están asociadas con categorías gramaticales del lenguaje natural en el que se desarrolla la extracción automatizada (CHUNG, 2003). En el caso de la lengua española, una lista de paro deberá incluir: artículos, preposiciones, conjunciones, interjecciones, verboides, entre otras unidades denominadas palabras vacías. En este rubro, Gil Leiva (2010, p. 200) identificó una lista de palabras vacías del español actual (con 273 elementos) que pueden ser retomada para procesos de extracción automatizada. Es importante destacar que algunos extractores automatizados, como WordStat incluyen listas de paro precargadas con posibilidades de inclusión de más elementos, mientras que en otros extractores, Wordsmith por ejemplo, es necesario incorporarlas.

Una vez que el corpus ha sido establecido y las listas de paro se han integrado, la extracción terminológica es relativamente sencilla. En pocos minutos un extractor terminológico es capaz de llevar a cabo la tarea para la que ha sido desarrollado ya sea mediante principios estadísticos, lingüísticos o híbridos (ESTOPÁ, 2009). Actualmente, en el mercado se comercializan diversos extractores terminológicos como SDL Multiterm, Standalone Terminology Extraction Tools, Simple Extractor y Term Suite. Por demás, algunos grupos de investigación -cuya materia prima de trabajo son los términos- han desarrollado programas propios como LEXTER y TermExt que responden a necesidades específicas e incluyen patrones léxicos y/o sintácticos del dominio, corpus comparativos y bases terminológicas previas.

Por último, es importante mencionar que los extractores terminológicos no generan un listado de términos del dominio sino un listado de candidatos a términos. Posteriormente, la validación de tal listado de términos candidatos implica el factor humano que debe escoger, entre los candidatos, cuáles estructuras constituyen verdaderamente términos del dominio. Al final, como menciona Chung (2003, p. 230), “el estudio de los términos técnicos requiere conocimiento experto de especialistas en la materia temática” para su validación e incorporación en un tesauro o vocabulario controlado.

3 Metodologia

La metodología de estudio aplicada en la investigación fue exploratoria-cuantitativa. Por una parte, realizamos un acercamiento exploratorio a un tema de análisis (extracción terminológica para un tesauro documental) y, por otra, evaluamos de modo cuantitativo dos procesos de extracción terminológica (extracción terminológica manual y extracción terminológica automatizada). Consideramos las siguientes variables para evaluar cada una de las técnicas de extracción terminológica:

a) Resultados del proceso: medido en el número de términos que pueden funcionar como descriptores en el tesauro de bibliotecología que nos interesa construir.

b) Recursos humanos: medido en el número de horas de tiempo humano dedicadas al desarrollo de cada proceso de extracción terminológica.

c) Recursos tecnológicos: medido en el inventario de elementos tecnológicos inherentes al desarrollo del proceso.

d) Recursos económicos: medido en el número de aspectos monetarios asociados al proceso.

3.1 Selección del corpus

El corpus que se procesó tanto en la extracción terminológica manual como en la automatizada fue el mismo. Consistió en los volúmenes 33 y 34 de la revista *Investigación Bibliotecológica: bibliotecología, archivonomía e información* que:

> *Es una revista científica mexicana editada por el Instituto de Investigaciones Bibliotecológicas y de la Información de la Universidad Nacional Autónoma de México […] La RIB tiene como propósito publicar resultados científicos originales e inéditos del quehacer investigativo en las ciencias bibliotecológica y de la información, derivados de investigaciones originales realizadas en México y en otras partes del mundo* (INSTITUTO DE INVESTIGACIONES BIBLIOTECOLÓGICAS Y DE LA INFORMACIÓN, 2021).

Se seleccionó tal publicación porque, entre las revistas mexicanas de bibliotecología, es relevante, su publicación es constante, se distribuye en libre acceso a texto completo, mantiene relación con el ámbito académico y se encuentra indizada en Scopus, Latindex, Dialnet y otras bases de datos. Dado que *Investigación Bibliotecológica: bibliotecología, archivonomía e información* incluye textos en español, inglés y portugués, los textos en idioma diferente al español fueron excluidos porque una característica esencial del estudio fue obtener términos para la construcción de un tesauro monolingüe en lengua española y la recomendación indica que es preferible trabajar con corpus en el idioma del instrumento que se desea construir. En total se seleccionaron 59 artículos que constituyeron una muestra de 283,585 palabras contenidas en los títulos, resúmenes, palabras clave y el texto completo del documento. Como se señaló en la introducción, en este experimento se trabajó con un corpus prueba, con características reducidas y focalizadas. En el futuro, cuando se trabaje de manera formal en el tesauro bibliotecológico que se intenta construir, se contemplarán diversas fuentes documentales con rasgos sincrónicos y geográficos más amplios, como recomienda la teoría.

3.2 Preparación de corpus y listas de paro

a) Extracción terminológica manual*:* se descargaron los artículos en PDF-Formato de Documento Portátil, mediante el programa *Adobe Acrobat Reader DC* y se compendiaron en una carpeta.

b) Extracción terminológica automatizada*:* el corpus fue descargado en una carpeta en formato PDF para su posterior ingesta en el extractor terminológico. Originalmente el corpus se ingresó en formato PDF a *WordStat*, en vista de que el software permite la ingesta de archivos en formatos diversos (PDF, Word, Excel, etc.), pero al revisar el corpus se detectaron graves problemas de reconocimiento textual. Por esta razón, el corpus tuvo que ser transformado previamente a formato TXT-Texto sin formato con el programa *Adobe Acrobat* e ingresado en este último formato a *WordStat*.

En cuanto a las listas de paro, el extractor *WordStat* tiene compendios establecidos de manera previa para diversas lenguas (inglés, español, francés, alemán, entre otros), por lo que sólo se seleccionó la preferencia de idioma y automáticamente las listas de paro se agregaron al proceso.

3.3. La extracción terminológica

a) Extracción terminológica manual*:* el proceso fue realizado por un profesional con formación multidisciplinaria, integrada por formación académica en bibliotecología y lingüística, particularmente sobre terminología. La extracción terminológica manual fue desarrollada de manera periódica, cubriendo artículo por artículo en sesiones de trabajo de aproximadamente dos horas.La metodología seguida por el profesional consistió en: lectura detallada de cada uno de los artículos, marcado de los términos conforme los iba detectando en el proceso de lectura y eliminación de términos repetidos al finalizar el trabajo en cada documento. Durante el proceso se identificó que algunas veces el

profesional tenía duda sobre si cierta unidad léxica constituía un término del dominio. Ante tales casos, el experto rotulaba el término dudoso y, al finalizar la revisión del artículo, consultaba fuentes de información (diccionarios, glosarios y bases de datos de bibliotecología) para discernir sobre la validez o el descarte del término.

b) Extracción terminológica automatizada*:* el proceso lo realizó un profesional de la información, bibliotecólogo, mediante la aplicación del programa WordStat, una herramienta de análisis de contenido y minería de textos que entre sus funciones incluye un módulo para la extracción de términos técnicos y frases comunes a partir de corpus textuales (PROVALIS RESEARCH, 2021. La metodología seguida por el profesional en el proceso de extracción terminológica automatizada consistió en: la compilación del corpus que se procesaría en archivos TXT, la selección de una lista de paro para idioma español (WordStat integra listas de paro definidas que se pueden seleccionar), adjuntar el corpus en la función de procesamiento del programa (Project>Import Archives), procesar el corpus para obtener términos, mediante el uso de WordStat, específicamente en las funciones que se señalan en el apartado de generación de listados de candidatos a términos que aparece a continuación.

3.4 Generación de listados de candidatos a términos

a) Extracción manual: el profesional con formación multidisciplinaria marcó los candidatos a términos en los archivos digitales, formato PDF, y los compiló en un archivo *Excel*.

b) Extracción automatizada: el profesional de la información hizo las operaciones técnicas adecuadas para que el programa *WordStat* generará listados de candidatos a términos y luego los exportó a *Excel*. En cuanto al uso de WordStat, el programa se empleó en combinación con *QDA Miner* y las funciones utilizadas fueron: listas de exclusión de términos (Exclusión list), extracción automática de tópicos (Topics>Extraction) y extracción de frases para términos compuestos (Frases>Extraction).

3.5 Validación de términos

a) Extracción terminológica manual: una vez que el experto concretó el listado de candidatos a términos, un grupo de bibliotecarios expertos en el dominio (2 profesionales con estudios de Licenciatura en bibliotecología, 3 personas con estudios de Maestría en bibliotecología y estudios de la información y 3 Doctores en bibliotecología y estudios de la información) revisaron el listado y emprendieron el proceso de validación que derivó en la selección de los términos ideales para ser incluidos en el tesauro, considerando sus aspectos sintácticos, léxicos y semánticos. Dichos profesionales revisaron y validaron el listado porque son quienes conforman el equipo base que está a cargo del tesauro que se intenta desarrollar.

b) Extracción terminológica automática: a partir del listado de términos candidatos generado por el programa WordStat, el mismo grupo de bibliotecarios (8 profesionales que constituye el equipo base) revisó la propuesta e, igual que en la extracción manual, seleccionaron los términos adecuados para el tesauro.

4 Resultados

4.1 Extracción terminológica manual

4.1.1 Candidatos a términos

A partir del proceso manual se obtuvieron los siguientes resultados (Figura 1):

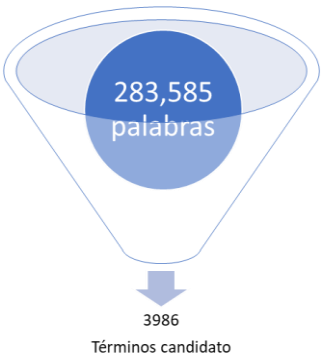

Figura 1. Términos candidato en extracción manual
Fuente: elaboración propia, 2021.

Del corpus formado por 283,585 palabras se recuperaron 3,986 términos, lo que supone un 1.4% de unidades léxicas que son términos. Para obtener tales datos se consideró el total de palabras del corpus sobre el número de términos presentes. Se detectó que los términos presentes en el corpus trabajado son reducidos, si consideramos la relación: 100% palabras-1.4% términos. Pese a ello, desde la teoría terminológica no existe una media que indique cuándo un corpus presenta densidad terminológica alta o baja; sólo los estudios de densidad terminológica comparativa pueden evidenciar tal aspecto. Tal tema abre una brecha de investigación en la que se trabaje la densidad terminológica en corpus bibliotecológicos, por ejemplo, comparando una revista contra otra, un autor frente a otro y así.

4.1.2 Características

Los candidatos a términos extraídos manualmente se categorizaron en tres grupos: términos simples, términos compuestos y unidades sintagmaticas. Como se advirtió en la revisión de literatura, los términos pueden categorizarse según el número de morfemas que los componen: desde los términos simples, formados por una unidad léxica nominal; los términos compuestos, derivados de unidades léxicas nominales unidas por preposiciones o adyacencia y las unidades terminológicas sintagmáticas que se componen de una cadena nominal extensa.

Algunos ejemplos de términos simples derivados del proceso fueron:

Bibliotecología
Colecciones
Datos
Información
Tesis

Los términos compuestos incorporaron casos como:

Web semántica
Vocabularios abiertos
Unidades de información
Sistemas de clasificación
Registros bibliográficos

Un último grupo, correspondió a unidades terminológicas sintagmáticas:

Sistemas para la organización del conocimiento
Listas de encabezamientos de materias
Sistemas de recuperación de información
Requisitos funcionales para registros bibliográficos
Protocolo de transferencia de hipertexto

Se encontró que los candidatos a términos representan mayoritariamente cosas, actividades o procesos, propiedades de personas y campos temáticos. En la identificación de tales categorías se retomó el estándar *ISO25964-1: Information and documentation-Thesauri and interoperability with other vocabularies-Part 1: Thesauri for information retrieval* (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011), en donde se señala que los términos de un tesauro pueden fragmentarse en categorías exclusivas, fundamentadas en rasgos compartidos: cosas, actividades, materiales, eventos, campos temáticos, tipos de personas, lugares y eventos (Figura 2):

Figura 2. Términos y referentes en extracción manual
Fuente: elaboración propia, 2021.

4.1.3 De términos candidatos a términos aceptados

Una vez que se contó con una lista de términos candidatos, generada por el especialista que realizó la extracción terminológica manual, siguió la etapa de validación. Un grupo de bibliotecólogos revisó el listado de "términos candidato" y aprobó su traslado al estatus "término aceptado", es decir, un término que puede funcionar como un descriptor del tesauro de bibliotecología que se pretende construir. La validación de los términos se llevó a cabo en sesiones de trabajo conjuntas en las que participó todo el equipo base, encargado de la construcción del tesauro (8 profesionales). La aprobación de los términos implicó la revisión del listado término por término y, en este proceso, la opinión de todos los expertos fue considerada para llegar a un consenso mayoritario sobre la aprobación o descarte de cada término candidato.

En el traslado de los términos del estatus "término candidato" a "término aceptado" se consideraron los siguientes criterios:

a) Garantía literaria: definida por la aparición del término en la literatura de la especialidad y sus líneas de investigación.
b) Adecuación terminológica: definida por la pertinencia léxica, sintáctica y semántica del término en la literatura de la especialidad.

Entre los 3,986 términos candidatos, 3,267 (81%) cumplieron con las características antes señaladas, fueron aprobados por todo el equipo base en la validación terminológica y obtuvieron el estatus de términos aceptados (Figura 3):

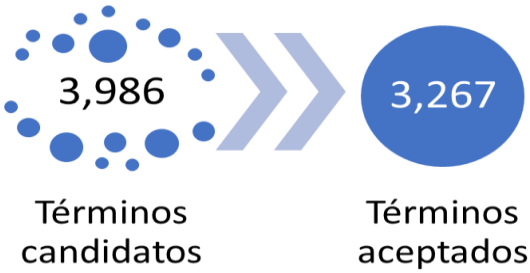

Figura 3: De términos candidatos a términos aceptados en extracción manual
Fuente: elaboración propia, 2021.

Un porcentaje mínimo de términos (19%) no obtuvieron el estatus "término aceptado". Los casos más frecuentes estuvieron asociados con:

a) Formas en singular: ante las formas singulares como "ontologia" o "artículo" se dio prioridad a la forma plural, dado que se trata de objetos contables (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011: 27).
b) Términos que son instancias (entidades a nivel nombre propio) (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2010): tecnologías, bases de datos, fuentes de información, proyectos o sitios web, por ejemplo, "Bibliografía Nacional Británica", "Tesauro UNESCO" o "Photomerge".
c) Términos con pertenencia dubitativa al dominio: más adecuados para tesauros de otras áreas de conocimiento como "Objetos de estudio", "Aprendizaje significativo" o "Apoyo a tareas".

A partir de los resultados obtenidos, es posible afirmar que la extracción manual de terminología, fundamentada en la selección de fuentes literarias del dominio, es altamente efectiva en el establecimiento de términos para la construcción de un tesauro de la disciplina bibliotecaria. Tal afirmación se fundamenta en la relación términos candidatos>>términos aceptados, en donde el 81% de las unidades léxicas seleccionadas por el profesional de la información fueron útiles como descriptores del tesauro que se desarrolla.

4.1.4 Ponderación de variables costo/beneficio

La ponderación de las variables costo/beneficio en la ETM quedaron establecidas como sigue:

1. *Resultados del proceso*: obtuvimos 3.267 términos que nos servirán para integrar al tesauro de bibliotecología.
2. *Recursos humanos*: el procesamiento de las 283,585 palabras del corpus, la selección de los términos del corpus y la construcción de los listados de términos candidatos se llevo a cabo en un tiempo de 158 horas de dedicación exclusiva de un profesional con formación multidisciplinaria.
2. *Recursos tecnológicos*: se emplearon los programas: *Adobe Acrobat Reader DC* y *Microsoft Excel*.
3. *Recursos económicos*: el factor crucial en la extracción manual radicó en los aspectos monetarios asociados al nivel de especialización del profesional que seleccionó los términos. Se encontró una estrecha relación como sigue: a mayor especialización profesional, mayor costo económico.

4.2 Extracción terminológica automatizada

4.2.1 Candidatos a términos

A partir del proceso automatizado se obtuvieron los siguientes resultados (Figura 4):

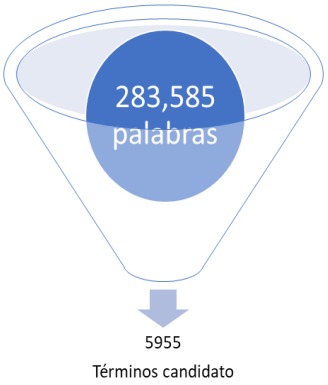

Figura 4. Términos candidatos en extracción automatizada
Fuente: elaboración propia, 2021.

Del corpus formado por 283,585 palabras se recuperaron 5,955 términos candidatos, arrojados por el programa WordStat tras el proceso de extracción terminológica, lo que supone un 2.1% de unidades léxicas que son términos. La tecnología redujo el tiempo de procesamiento textual a media hora de conversión de archivos PDF a TXT y cinco minutos destinados para la extracción del listado de términos candidatos en comparación con las 158 horas empleadas por un profesional con formación multidisciplinaria que se necesitaron en la extracción terminológica manual. Se identificó que el proceso de extracción automatizada representa una significativa economía de tiempo en comparación con la extracción terminológica manual.

4.2.2 Características

Las posibilidades de WordStat, programa empleado para el proceso automatizado, permiten obtener términos simples y términos compuestos formados por dos, tres, cuatro, cinco y más palabras, según los parámetros establecidos en las opciones de configuración de preprocesado y posprocesado que el programa ofrece. En el experimento el rango consideró obtener términos simples y compuestos (con extensión de hasta 10 palabras).

Los términos simples derivados de la ETA fueron 3,454. En comparación, los términos compuestos fueron 2,501 y su longitud máxima fue de 6 palabras (Sistemas para la organización del conocimiento). Los resultados reafirman el trabajo de Benavent y Parilla (2006) quienes proponen que la longitud de procesado de términos no exceda más allá de 6 o 7 palabras, con el objetivo de reducir ruido en la recuperación de candidatos.

Entre los términos simples encontramos:

Datos
Información
DOI
Internet
Web

Los términos compuestos incluyen candidatos como:

Datos bibliográficos
Producción científica
Datos enlazados
Recuperación de información
Competencias digitales

Igual que en la extracción manual, se observan referentes conceptuales asociados a cosas, actividades o procesos, propiedades de personas y campos temáticos. Tenemos entonces que, independientemente del modo de extracción terminológica, las entidades nominales de un corpus siempre estarán en relación con tal tipo de referentes, enunciados por el estándar *ISO25964-1: Information and documentation-Thesauri and interoperability*

with other vocabularies-Part 1: Thesauri for information retrieval (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011) (Figura 5):

Figura 5. Términos y referentes en extracción automatizada
Fuente: elaboración propia, 2021.

4.2.3 De términos candidatos a términos aceptados

El programa WordStat generó una lista de términos candidatos posibles de exportar a Excel. Después, el grupo de bibliotecarios que forma el equipo base (8 profesionales) revisó la propuesta y aplicaron los mismos criterios que en la extracción manual para el traslado del estatus término candidato a término aceptado.

Se obtuvieron los siguientes resultados (Figura 6):

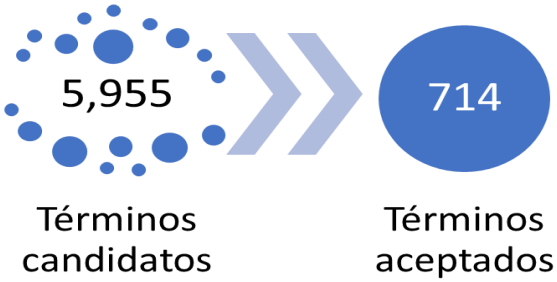

Figura 6: De términos candidatos a términos aceptados en extracción automatizada
Fuente: elaboración propia, 2020

En la extracción terminológica automatizada, sólo el 12% de los términos candidatos, esto es 714, fueron trasladados al estatus términos aceptados y se detectó que un número considerable de los términos candidatos presentaban los siguientes problemas:

a) Errores de caracteres textuales que no fueron corregidos en el proceso de limpieza del corpus y palabras registradas con cortes textuales, por ejemplo, “atri”- “butos” o “Catalo”-“gación”.
b) Selección de palabras generales como términos, por ejemplo, “proceso”, “resultados”, “figura”, “México”, “nivel”, “científica”
c) Apellidos y nombres personales tomados como términos: su alta ocurrencia los llevó al estatus de términos, por ejemplo, “Sánchez”, “María”, “García”.
d) Términos generales: que por su alta frecuencia en el corpus sobrepasaron a los términos del dominio, por ejemplo: “educación”, “profesores”, “trabajo social”.

Los resultados coinciden con lo señalado por Benavent y Parrilla cuando señalan que, si bien los extractores terminológicos automatizados agilizan el proceso de obtención terminológica, ocasionalmente presentan inconvenientes como exceso de ruido y unidades que no deberían aparecer en el listado de candidatos (BENAVENT y PARRILLA, 2006). La extracción terminológica automática es muy rápida en tiempo, pero no iguala los resultados, en número de términos útiles, de la extracción terminológica manual.

4.2.4 Ponderación de variables costo beneficio

1. *Resultados del proceso*: obtuvimos 714 términos que nos servirán para integrar al tesauro de bibliotecología.
2. *Recursos humanos*: el proceso automatizado requirió el trabajo de un profesional con conocimientos para operar el programa Wordstat desde la ingesta del corpus hasta la exportación del listado de términos candidatos. Además, también se requirió trabajo humano en la transformación del corpus de formato PDF a TXT, debido a los problemas de reconocimiento textual que evidenció *Wordstat*. En total fueron requeridas 6 horas de trabajo humano. Dichos recursos humanos también se caracterizaron por ser altamente especializados en el manejo del software, donde fue necesario que el profesional supiera usar todas las funciones de WordStat: subir los archivos, activar listas de paro, sacar términos simples, términos compuestos, exportar listados e Excel, etc.
2. *Recursos tecnológicos*: se emplearon los siguientes programas: *Adobe Acrobat Reader DC*, *Abobe Acrobat*, el paquete *WordStat con base QDA Miner* y *Microsoft Excel*.
3. *Recursos económicos*: el factor crucial en la extracción terminológica automatizada radicó en dos aspectos: factor humano experto para llevar a cabo el proceso de extracción con el programa establecido y recursos tecnológicos de costo considerable, especialmente el programa *WordStat con base QDA Miner*.

5 Conclusiones

A partir del objetivo principal del estudio que fue “evaluar dos técnicas de extracción terminológica: extracción terminológica manual y extracción terminológica automatizada, para analizar la efectividad de cada proceso en la obtención de términos útiles para la construcción de un tesauro de bibliotecología” es posible mencionar que mediante ambos procesos es posible obtener términos que pueden funcionar como descriptores de un tesauro de bibliotecología; sin embargo, cada método evidencia características y resultados diferentes:

1. La extracción terminológica manual: es una técnica muy antigua, cuenta con un procedimiento bien establecido que autores como Curras (1998), Lancaster (2002) y Naumis Peña (2007) señalan detalladamente. Aunado a ello, el estándar ISO-25964-1 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011) ofrece indicadores fundamentales para desarrollar el proceso. Tras el experimento, descubrimos que la extracción de términos mediante este método contempla una inversión alta de tiempo y recursos humanos para su desarrollo, pero los resultados obtenidos -en número de términos útiles- son muy buenos. Consideramos que esta opción es útil si en el proyecto para la construcción del tesauro se cuenta con recursos humanos especializados y tiempo suficiente en la etapa de selección de los términos que serán incorporados en la herramienta.

2. La extracción terminológica automatizada: es una técnica que combina el talento humano con las herramientas tecnológicas. Tras el experimento, descubrimos que es una opción rápida para la obtención de términos de un dominio determinado, pero el número de términos útiles que aportó fue bajo en comparación con el número de términos obtenidos mediante extracción manual: en proporción de 3267 términos derivados del proceso manual frente a 714 derivados del proceso automatizado. Tal hecho permite dilucidar que las tecnologías de la información son una excelente ayuda en muchas actividades, sin embargo, en procesos de detección terminológica aún presentan deficiencias que deberán ser analizadas en trabajos posteriores. Pese a ello, consideramos que este proceso es recomendable cuando el proyecto de construcción del tesauro contempla corpus

muy extensos que son difíciles o imposible de procesar de forma humana y cuando se cuenta con poco tiempo para el procesamiento del compendio documental.

A partir de esta investigación, evidenciamos algunas líneas de investigación futuras:

1. Primero, la necesidad de emprender estudios comparativos sobre la efectividad de diversos programas de extracción terminológica automatizada, por ejemplo, contrastar la potencialidad de extractores automatizados como WordStat, WordSmith, Voyant Tools, AntConc y otros programas tanto de fuente abierta como aquellos disponibles en el mercado.

2. Segundo, es necesario el estudio de las otras técnicas de obtención terminológica no tratadas en este artículo (reutilización de vocabularios previos, entrevistas a expertos del dominio y entrevistas a usuarios del dominio).

Se concluye que la construcción de tesauros documentales sobre diversos campos de conocimiento ha sido una tarea fundamental de la bibliotecología y en nuestros días los tesauros siguen siendo de interés para la disciplina debido al surgimiento constante de nuevos dominios de conocimiento que requieren ser ordenados y la imperante necesidad de organizar recursos de información, desde una perspectiva temática que contemple profundida terminológica y relaciones semánticas. Los tesauros se hacen necesarios tanto en la indización de los recursos de información en la biblioteca física como en contextos bibliotecológicos digitales (bibliotecas, repositorios de información y bases de datos) donde, incluso, están siendo colocados en interfaces que los usuarios consultan para encontrar los descriptores que representan adecuadamente sus necesidades de información. Dentro de este orden de ideas, nuestro estudio coincide con lo enunciado por Chung (2003) y Benavent y Parilla (2006): la construcción de tesauros es una actividad con tendencia a la continuidad y la concreción de la base terminológica que los sustenta se observa como un área de trabajo vigente que deberá enriquecerse continuamente.

Referencias

ABBAGNANO, Nicola. Diccionario de filosofía. México: Fondo de Cultura Económica, 1963.

ARNTZ, Reiner y PICHT, Heribert. Introducción a la terminología. Madrid: Fundación Germán Sánchez Ruipérez, 1995.

AUGER, Pierre y ROSSEAU, Lois Jean. Metodología de la investigación terminológica. Málaga: Universidad de Málaga, 2003.

BARITÉ, Mario. Garantía literaria y normas para construcción de vocabularios controlados: aspectos epistemológicos y metodológicos. Scire: Representación y organización del conocimiento, v. 15, n. 2, pp. 13-24, 2009.

BENAVENT, Paloma y PARRILLA, Sara. Análisis de la extracción automática de términos con el programa informático ExtraTerm. Fòrum de Recerca, n. 12, pp. 1-10, 2006.

BRÄSCHER, Marisa. Semantic relations in knowledge organization systems. Knowledge Organization, v. 41, n.2, pp. 175-180, 2014.

CABRÉ, María Teresa. La teoría comunicativa de la terminología, una aproximación lingüística a los términos. Dans Revue française de linguistique appliquée, v. 14, pp. 9 -15, 2009.

CABRÉ, María Teresa. La terminología: representación y comunicación elementos para una teoría de base comunicativa y otros artículos. Barcelona: Universitad Pompeu Fabra, 1999.

CINTRA, Anna, TÁLAMO, María, LARA, Matilda y KOBASHI, Nair. Para entender as linguagens documentárias. São Paulo: Polis, 2002.

CURRAS, Emilia. Tesauros: manual de construcción y uso. Madrid: Kaher II, 1998.

CHU, Heting. Information representation and retrieval in the digital age. Medford, New Jersey: Information Today, 2010.

CHUNG, Teresa. A corpus comparison approach for terminology extractión. International Journal of Theoretical and Applied Issues in Specialized Communication, v. 9, pp. 221- 246, 2003.

ESTOPÁ, Rosa. Los extractores de terminología logros y escollos. En ALCINA CAUDET, María Amparo, (coord.). Terminología y sociedad del conocimiento. España: Bern: Peter Lang, 2009, pp. 117-146,

GIL URDICIAIN, Blanca. Orígenes y evolución de los tesauros en España. Revista general de Información y Documentación, v.8, n.1, pp.63-110, 1998.

GIL LEIVA, Isidoro. La automatización de la indización, propuesta teórico- metodológica: aplicación al área de biblioteconomía y documentación. Murcia Universidad de Murcia, 2010.

GOLUB, Koraljka; TUDHOPE, Douglas.; ZENG, Marcia y ŽUMER, Maja. Terminogy Registries for knowledge organizatión systems: functionality, use, and attributes. Journal of the association for information science and technology, v. 65. n.9, pp. 1901-1016, 2014.

GUINCHAT, Claire y MENOU, Michel. Introducción general a las ciencias y técnicas de la información y de la documentación. París: UNESCO, 1983.

HJØRLAND, Birger. Semantics and knowledge organization. Annual Review of Information Science and Technology, v. 41, n.1, pp. 367-405, 2007.

HODGE, Gail. Systems of knowledge for digital libraries: beyond traditional authority files. Washington: Council on Library and Information Resources, 2000.

INSTITUTO DE INVESTIGACIONES BIBLIOTECOLÓGICAS Y DE LA INFORMACIÓN. Investigación Bibliotecológica [en línea]. Disponible en: http://rev-ib.unam.mx/ib/index.php/ib (Recuperado el 16 de junio 2021).

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. Functional requeriments for subject autority data (FRSAD). A conceptual model. Washington: IFLA, 2010.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO. 25964. Information and documentation-Thesauri and interoperability with other vocabularies-Part 1: Thesauri for information retrieval. Ginebra, Suiza: ISO, 2011.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO:2788-1986.Documentation-Guidelines for the establishment and development of monolingual thesauri. Ginebra, Suiza: ISO, 1986.

KUMBHAR, Rajendra. Library classification in the 21 century. Oxford: Chandos Publishing, 2012.

LAGUENS GARCÍA, José Luis. Tesauros y lenguajes controlados en Internet. Anales De Documentación, v. 9, n.9, pp. 105-121, 2006.

LANCASTER, Frederich. W. El control del vocabulario en la recuperación de la información. Valencia: Universidad de Valencia, 2002.

LÓPEZ MATEO, Coral y OLMO CAZEVIEILLE, Françoise. Metodología para la extracción e identificación de candidatos a términos en el ámbito de la bioquímica. Terminàlia, n.16, pp. 18-28, 2017.

LUNA TRAIL, Elizabeth, VIGUERAS ÁVILA, Alejandra y BAEZ PINAL, Gloria. Diccionario básico de lingüística. México: Universidad Nacional Autónoma de México, 2005.

LUO, Zhiwei, XIE, Rong, CHEN, Wen y YE, Zatao. Automatic domain terminology extraction and its evaluation for domain knowledge graph construction. Web Intelligence, v. 16, n.3, pp. 173-185, 2018.

NAUMIS PEÑA, Catalina. Los tesauros documentales y su aplicación en la información impresa, digital y multimedia. Buenos Aires: Alfagrama, 2007.

PROVALIS RESEARCH. WordStat: software de análisis de contenido y minería de textos [en línea]. Disponible en: https://provalisresearch.com/es/products/software-de-analisis-de-contenido/ (Recuperado el 16 de junio 2021).

SAGER, Juan. Curso práctico sobre el procesamiento de la terminología. Madrid: Fundación Germán Sánchez Ruipérez: Pirámide, 1993.

SINCLAIR, John y SINCLAIR; Les. Corpus, concordance, collocation. Oxford: Oxford University Press, 1991.

SMIRAGLIA, Richard. Domain analysis for knowledge organization. Nueva York: Chandos, 2015.

STRZALKOWSKI, Tomek. Natural language information retrieval. Kluwer Academic, 1999.

SUÁREZ SANCHEZ, Adriana. Ontologías: fundamentos y aplicaciones, una aproximación desde la perspectiva bibliotecológica. Ciudad de México: Universidad Nacional Autónoma de México, 2018.

VIVALDI, Jorge y RODRÍGUEZ, Horacio. Improving term extraction by commbining differents techniques. Terminology, v. 7, n. 1, pp. 31-48, 2001.

Datos del autor

Adriana Suárez-Sánchez

Licenciada en Lingüística, Maestra en Bibliotecología y Doctora en Bibliotecología y Estudios de la Información. Su área de especialización es la organización temática de la información en contextos digitales, mediante sistemas en red (tesauros, mapas tópicos, folksonomías, taxonomías digitales, ontologías y anillos semánticos). Actualmente labora como Investigadora Asociada en el Instituto de Investigaciones Bibliotecológicas y de la Información de la Universidad Nacional Autónoma de México. Es profesora de asignatura de la materia de Indización en el Colegio de Bibliotecología de la Facultad de Filosofía y Letras de la UNAM y docente de la materia Fundamentos de la Organización Documental en el sistema de Prerrequistos del Posgrado en Bibliotecología y Estudios de la Información de la UNAM. Su línea de investigación actual es la sistematización y organización de la información desde una perspectiva temática. Ha publicado varios trabajos en revistas nacionales e internacionales sobre: sistemas para la organización del conocimiento, taxonomías, ontologías y folksonomías.

asuarez@iibi.unam.mx

Received-Recibido-Recibido: 2021-04-14

Accepted-Aceptado-Aceitado: 2022-12-09