REVISTA CIENTÍFICA

Vol. 17, No.2 (2021) mayo-agosto ISSN electrónico: 1683-8947

PROPUESTA DE PROCEDIMIENTO PARA LA CONSTRUCCIÓN SEMIAUTOMÁTICA DE TESAUROS EN BIBLIOTECAS UNIVERSITARIAS

PROPOSED PROCEDURE FOR THE SEMI-AUTOMATIC CONSTRUCTION OF THESAURI IN UNIVERSITY LIBRARIES

Andisleydis Acosta Bravo, Universidad Central “Marta Abreu” de Las Villas
aabravo@uclv.cu
https://orcid.org/0000-0001-8829-5197

Asleni Díaz Jiménez, Universidad Central “Marta Abreu” de Las Villas
https://orcid.org/0000-0002-8073-0040
adiazj@uclv.cu

Amed Abel Leiva Mederos, Universidad Central “Marta Abreu” de Las Villas
amed@uclv.edu.cu
https://orcid.org/0000-0002-9144-5018

Recibido: 11 de diciembre de 2020
Revisado: 24 de marzo de 2021
Aprobado: 9 de abril de 2021

Cómo citar: Acosta Bravo, A.; Díaz Jiménez,A.; & Leiva Mederos, A.A. (2021). Propuesta de procedimiento para la construcción semiautomática de tesauros en bibliotecas universitarias. *Bibliotecas. Anales de Investigacion;17* (1),

RESUMEN:

Objetivo Se proponen procedimientos de índole teóricos en la indización, y a la vez introduciéndose en parcelas de las Ciencias de la Computación para resolver problemas de la Red TIC del Proyecto VLIR desde una óptica multidisciplinar, dada las características imponen el reto de intentar desarrollar cualquier proceso normativo en el terreno de la indización para un entorno complejo y pluridisciplinar,

siendo la primera vez que se asume un trabajo en Ciencias de la Información en la Universidad Central "Marta Abreu" de Las Villas (UCLV). **Diseño/Metodología/Enfoque s**e utilizaron las técnicas de origen matemático: la Ley de Zip, TF-IDF, N-grams y Stop World Elimination, aportando un enfoque mixto predominantemente cuantitativo. La descripción sirve de guía para la construcción de léxicos especializados, al incluir los mecanismos de construcción basados en las reglas exigidas a nivel internacional. **Resultados/Discusión**: Desde el diagnóstico de los Procesos de Indización en las Bibliotecas asociadas a la Red TIC del Proyecto VLIR, se aplicaron métodos y técnicas en la recopilación de información permitieron obtener resultados relacionados a la experiencia en los procesos de indización y construcción de tesauros en las diversas universidades del país. **Conclusiones:** La creación de un procedimiento que proporcione la transformación de los vocabularios controlados en un lenguaje interoperable, facilita la indización y la recuperación eficiente de la información. **Originalidad/Valor**: El impacto social de uso estriba en que al contener datos estandarizados en formato SKOS, las plataformas que usa y desarrolla la red podrán interpretar con otras plataformas con fines similares dando visibilidad a la ciencia de la Red TIC.

PALABRAS CLAVE: lenguajes controlados; lenguaje interoperable; modelos computacionales; procedimiento para la automatización de tesauros; bibliotecas.

ABSTRACT:

Objective Theoretical procedures are proposed in indexing, and at the same time being introduced in plots of Computer Science to solve problems of the ICT Network of the VLIR Project from a multidisciplinary perspective, given the characteristics they impose the challenge of trying to develop any normative process in the field of indexing for a complex and multidisciplinary environment, being the first time that a job in Information Sciences is assumed at the Central University "Marta Abreu" of Las Villas (UCLV). Design / Methodology / Approach the techniques of mathematical origin were used: Zip's Law, TF-IDF, N-grams and Stop World Elimination, providing a predominantly quantitative mixed approach. The description serves as a guide for the construction of specialized lexicons, by including the construction mechanisms based on the internationally required rules. Results / Discussion: From the diagnosis of the Indexing Processes in the Libraries associated to the TIC Network of the VLIR Project, methods and techniques were applied in the collection of information that allowed obtaining results related to the experience in the indexing and thesaurus construction processes. in the various universities of the country. Conclusions: The creation of a procedure that provides the transformation of controlled vocabularies into an interoperable language, facilitates indexing and efficient information retrieval. Originality / Value: The social impact of use is that by containing standardized data in SKOS format, the platforms used and developed by the network will be able to interpret with other platforms with similar purposes, giving visibility to the science of the ICT Network.

KEYWORDS: controlled languages; interoperable language; computer models; procedure for the automation of thesaurus; libraries.

INTRODUCCIÓN:

En los últimos años se ha evidenciado un crecimiento de la producción documental en el ambiente digital. Lo anterior supone el empleo de nuevas técnicas para el procesamiento que satisfagan necesidades de los

usuarios y sistemas de información. Ante estos cambios se hace necesario, que tanto el emisor como el receptor de la información, logren entenderse en un entorno, en el que cada vez existe mayor auge del documento electrónico, lo que conlleva al aumento del ruido, que entorpece el proceso de comunicación. Para dar respuesta a esta necesidad surgen los lenguajes documentales, (Martín, 2009) los define como "sistema artificial de signos utilizados en las operaciones de indización que permite la representación de contenido documental para su posterior recuperación, sirviendo como medio para la interrogación, recuperación y difusión de información pertinente para el usuario." Con base en lo expuesto se tiene a la indización como un proceso comunicacional.

Los tesauros y otros lenguajes controlados impresos en formatos duros y gestados por bases de datos, son el punto de mira de la mayoría de las investigaciones encaminadas a estudiar su uso en la Web 3.0. Para ello se han evolucionado de grandes sistemas de etiquetado, hasta sistemas que sirven de soporte a la Web 3.0 mediante el uso de tecnologías de última generación, entre las que se encuentra *Extensible Markup Language*[1] (*XML*), *Resource Description Framework*[2] *(RDF), Web Ontology Language*[3] *(OWL)* y *Simple Knowledge Organization System*[4] *(SKOS).*

SKOS permite aprovechar la estructura base y el contenido de los esquemas conceptuales. De estos últimos se valen las listas de encabezamiento de materia, taxonomías, folksonomías, esquemas de clasificación, terminologías y tesauros para la descripción de los contenidos orientado a mejorar los sistemas de organización del conocimiento. Un tesauro que se represente mediante *SKOS* mantiene sus propiedades, al contar con un lenguaje de alto nivel y exigencia lógica como *RDF* se pueden vitalizar mejores lenguajes e interconectar los desarrollados y especificados por *SKOS.*

Los cambios que han operado en la concepción de la web y de los sistemas de información documental, hacen necesaria la búsqueda de metodologías eficientes para que los sistemas soportados en lenguajes controlados, evolucionen a sistemas cada vez más eficientes. Dichos procedimientos para el ámbito de la web semántica han sido descritos por varios autores como: (Pastor, 1992), (Moreiro y otros, 1999), (Pastor, 2009), (Fernández, 2010) y (Cavieres y otros, 2010), sin embargo, ninguno es eficiente para una red de bibliotecas, dado que sus aportaciones solo han sido usadas en determinados dominios o en determinados contextos y en una red de bibliotecas abunda la pluralidad de metodologías y formas de construcción de léxicos. Por ello constituye un reto resolver este problema para una red de bibliotecas universitarias como las de la Red TIC del Proyecto VLIR en Cuba.

Las primeras aproximaciones a estos trabajos surgen en la década del 60 en los Estados Unidos (EE.UU.) en actividades de las Ciencias de la Computación, conducidas a la extracción de palabras clave relacionadas con la actividad de recuperación de información. Estos aportes han evolucionado hasta hoy

[1] Lenguaje de Marcado Extensible: es un meta-lenguaje que permite definir lenguajes de marcas desarrollado por el *World Wide Web Consortium* o W3C (Consorcio Mundial de la red).
[2] Marco de Descripción de Recursos: lenguaje de descripción de la W3C, para la descripción de vocabularios que permite no solo ordenar los datos, sino que mediante la representación de las propiedades y sentencias ofrece una mejor visión semántica de los datos.
[3] Lenguaje de Ontología Web: lenguaje de marcado para publicar y compartir datos usando ontologías en la *World Wide Web* (Red Mundial).
[4] Sistema simple de organización del conocimiento: iniciativa de la *W3C* en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales.

con las transformaciones de las estructuras de marcado para ontologías, traducidas en diferentes esquemas de organización de la información. Por ello, procesos de construcción automática mediante técnicas incrementales son ineficientes hasta para la construcción de léxicos controlados. Su alto apego a la lógica y su poca sensibilidad a los contextos y los dominios los hace inoperantes en algunos contextos.

La Red TIC del Proyecto VLIR está compuesta por diversas bibliotecas universitarias con diferencias en recursos humanos, tecnológicos y sus usuarios. En dichas bibliotecas existen dificultades para la recuperación de la información y para la interoperabilidad semántica de sus contenidos y/o documentos. No todas las bibliotecas indizan por igual, se cometen errores y falta de consistencia en los procesos de indización. La inadecuada ejecución de estos procesos dificulta la eficiencia de la recuperación de información, lo que influye en el desempeño de las plataformas usadas para gestionar la información en la red.

Estas características imponen un reto al intentar desarrollar cualquier proceso normativo en el terreno de la indización para un entorno complejo y pluridisciplinar, siendo la primera vez que se asume un trabajo en Ciencias de la Información en la Universidad Central "Marta Abreu" de Las Villas (UCLV) desde la lingüística computacional para resolver problemas de la recuperación de la información. Aportando procedimientos de índole teóricos en la indización, y a la vez introduciéndose en parcelas de las Ciencias de la Computación para resolver problemas de la Red TIC del Proyecto VLIR desde una óptica multidisciplinar.

Antecedentes:

Se realizó una búsqueda a nivel nacional e internacional sobre investigaciones vinculadas a la materia que se trata en el trabajo. Esta exploración arrojó resultados favorables a nivel internacional, sin embargo, no existen trabajos de este tipo en el país:

- "Generación automática de tesauros. Propuesta de un método lingüístico-estadístico" (Moreiro et al., 1999). Se persigue diseñar y construir una plataforma de gestión de repositorios (tesauro de software) capaz de almacenar, procesar, gestionar y recuperar cualquier tipo de documento, sin importar su presentación, soporte y forma de acceso, para lograrlo se apoyan del tesauro de descriptores.
- "Proyecto de indexado automático para documentos en el campo de la física de altas energías" (Montejo, 2001). Dirigido por el Laboratorio Europeo de Física de Partículas (CERN) en Ginebra (Suíza) y encaminado a desarrollar un sistema automático de indexado por asignación. El indexado por asignación consiste en la selección de palabras clave dentro de un léxico controlado (en este caso un tesauro) que describan y resuman los conceptos más importantes tratados en un texto dado. El sistema propone palabras clave según el tesauro del laboratorio alemán *DESY* (Deutsche Elektronen-Synchrotron) a partir de artículos completos en inglés relacionados con Física de Altas Energías.
- "*Building a web thesaurus from web link structure*" (Creación de un diccionario de sinónimos web a partir de la estructura del enlace web) (Chen et al., 2003). Se propone un enfoque novedoso para la construcción automática de un tesauro específico de dominio desde la Web, utilizando

información de estructura de enlace. El enfoque propuesto es capaz de identificar nuevos términos y reflejar la última relación entre los términos a medida que la Web evoluciona.

- "Los Tesauros en la Web Semántica: *SKOS* y la norma ISO 25964" (Pastor, 2015). Una lectura crítica de los procedimientos para la automatización de tesauros atendiendo a la aplicabilidad de la Norma ISO 25964 en SKOS.
- "*Building thesaurus-based knowledge graph based on schema layer*" o "Elaboración de un gráfico de conocimiento basado en tesauros apoyado en la capa de esquema" (Qiao et al., 2017). Usa técnicas de Big Data a partir de la función MapReduce[5] para construir tesauros, las herramientas que sustentan ese desarrollo son los procesos de construcción de grafos a partir de reglas y heurísticas.

MÉTODOLOGÍA:

Se utilizaron para la discriminación de términos, las técnicas de selección de Rasgos, utilizadas en la Construcción de Tesauros:

1. **Stop world elimination:** esta técnica consiste en eliminar aquellos vocablos sin significado autónomo también llamados frecuentemente palabras de parada (stop words) (Salton y McGillm, 1986). Generalmente este tipo de palabras tienen altos niveles de aparición en textos ya que son las que comúnmente sirven de enlaces y como complementos preposicionales. Las palabras vacías tienen una alta frecuencia de aparición y son difíciles de discriminar (Rijsbergen, 1979), (Sahami y otros, 1988), por tanto, en la construcción de tesauros se aconseja su eliminación. En el trabajo posibilitó eliminar aquellos vocablos sin significado autónomo.
2. **Umbral de frecuencia de términos y Ley de Zpif:** es una técnica muy simple que se encarga de seleccionar las palabras que más se repiten en un dominio lingüístico o un corpus documental. Se conoce que sus postulados están asociados a las actividades computacionales realizadas por (Lunh, 1958) y (Lanquillón, 2002) calculando la frecuencia de ocurrencias de términos para determinar la representatividad de estos en una comunidad, de esta manera los vocablos que presenten menos frecuencia de aparición en un corpus pueden ser eliminados por ser poco representativos. (Sahami y otros, 1988) ha estudiado la ley de (Zipf, 1949) observando que es posible seleccionar con 50 %, de efectividad si se suprimen los términos que tienen o muy alta o muy baja frecuencia de aparición, todo esto está en dependencia del umbral que se delimite o sea la cantidad de términos que se desean obtener. En el trabajo se utilizó para obtener la frecuencia de términos en un corpus de una especialidad.
3. **Método N-grams:** este método se ha utilizado para construir expresiones compuestas a partir del tamaño de las cadenas de caracteres de longitud fija. Con N-Grams se realiza un filtrado de término muy similar a los que se realizan con la ley de Zipf y el TDF-IDF, de esta manera se calcula la frecuencia sobre cadenas de caracteres de una longitud antes preestablecida. En el trabajo facilitó obtener frases sustantivadas y frases compuestas en los tesauros.
4. **Frecuencia inversa de documento y TF-IDF:** esta medida ha sido utilizada en los trabajos de (González y otros,1999) y (Gil, 2017), el valor discriminante de un término se obtiene no de

[5] Es un modelo de programación para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras y al commodity computing (informática de productos básicos).

aquellos que poseen alta frecuencia sino de los que tienen frecuencia inversa o sea los términos menos frecuentes. En el trabajo permitió conocer la calidad de los términos en el corpus.

RESULTADOS Y DISCUSIÓN:

Diagnóstico de los Procesos de Indización en las Bibliotecas asociadas a la Red TIC del Proyecto VLIR

Los métodos y técnicas aplicados en la recopilación de información permitieron obtener resultados relacionados a la experiencia en los procesos de indización y construcción de tesauros en diversas universidades del país: Universidad Central "Marta Abreu" de las Villas (UCLV), Universidad de Camagüey (UC), Universidad de Holguín (UHO), Universidad de Pinar del Río (UPR), Universidad de Ciencias Informáticas (UCI) y La Biblioteca Nacional José Martí (BNJM), Universidad de la Habana (OH).

Se utilizó como técnica la entrevista estructurada, compuesta por 7 preguntas, de ellas 5 cerradas y 2 abiertas con el fin de obtener la información que requiere el estudio. Fue aplicada en el contexto del Congreso INFO-2018 por el Dr. Amed Leiva Mederos a un total de 6 especialistas. Para la medición de los procesos de indización se realizó un cuestionario con un total de 9 interrogantes, de ellas 8 cerradas y 1 abierta. Mediante este cuestionario se conoció la verdadera situación de la indización en bibliotecas universitarias.

La muestra escogida para realizar la entrevista y el cuestionario es no probabilística y de muestreo por redes (bola de nieve). Se localizaron los investigadores que son parte de la Red TIC y que se dedican a los procesos de indización. Se inició con los que realizan estas labores en la UCLV y estos introdujeron a otros miembros internos y externos de la red cuya experticia en la construcción de tesauros y en los procesos de indización fuera válida, obteniendo un total de 15 expertos.

El análisis mostró que de los centros laborales del país con experiencias en el proceso de indización y construcción de tesauros alcanza la condición de más destacados la UCLV con un 33%, la UPR con un 20% y la UC con un 13%. Para la recogida de información se tuvo en cuenta además de los centros laborales de información de importancia en el país, la categoría docente y científica de los encuestados. De un total de 15 encuestados la mayoría ostenta la categoría científica de Doctores y Másteres, de manera que representan cada uno el 40%. Durante el procesamiento de la información se detectó un 13.33% en los Licenciados en Filología mientras que los Técnicos solo alcanzan el 6.67%.

Todos los encuestados coinciden en que la construcción de tesauros es una actividad importante que se ha quedado rezagada en la práctica bibliotecaria en nuestro contexto, por ausencia de políticas en la red de bibliotecas del Ministerio de Educación Superior (MES), abandono de buenas prácticas en la década de los 90, falta de conocimiento en los especialistas con los que hoy cuentan las bibliotecas, por lagunas en su formación y con ello el aumento de los "paraprofesionales" en las plantillas, el uso exclusivo de las palabras clave del autor y el bajo salario que se atribuye a su trabajo. Evidenciando la gravedad de la situación, debido a los contenidos que presentan los servicios como repositorios y revistas digitales en la inclusión en sistemas de indexación o bases de datos bibliográficos que como requisitos de acceso miden la indización de calidad, lo que se traduce en el uso de tesauros en dicho proceso.

El lenguaje que utilizan los especialistas en información para realizar el proceso de indización puede ser libre o controlado. El lenguaje más utilizado por los encuestados es el controlado representado con un 71%, siendo el más tradicional y respondiendo al paradigma de indización centrado en el documento mientras que el libre representa el 29%.

En la construcción de tesauros como organización de orden léxico se tiene que la estructura principal es la más utilizada. El 100 % de los encuestados (15) solo llega a construir la estructura principal del léxico y en solo en 2 encuestados (13%) usan en la construcción de tesauros los índices auxiliares. Es evidente que la elaboración de los léxicos no es completa pues indicadores como el tipo de relaciones no es utilizado por ninguno, lo que evidencia que se obtengan pocos valores en el empleo de las relaciones jerárquicas, relaciones de equivalencia y las relaciones asociativas.

Dentro de la estructura principal todos los encuestados 15 (100%) coinciden en que la estructura que más utilizan es la alfabética, mientras que la jerárquica y la disciplinar son las menos empleadas con un 13%, la facética y la alfabética arbórea no son usadas por ninguno. Esto evidencia que se tienen concebidas muy pocas dimensiones de orden de los léxicos lo que puede afectar su uso y la organización del conocimiento en función de los usuarios.

También pudo observarse que los índices auxiliares, cuando se construyen la mayoría se hacen de forma alfabética, no se utilizan la construcción de índices de tipo KWIC y KWOC con una riqueza temática superior al alfabético, pero más costoso de construir. Las normas utilizadas para realizar los procesos de indización y construcción de léxicos son la Norma ISO y la Norma Cubana, esta última se encuentra desactualizada, pero es la más empleada por los especialistas encuestados, mientras que la Norma UNE y la ANZI no son relevantes en su uso.

El paso que consideran imprescindible para construir el léxico los encuestados es: la construcción (100%), utilizado por todos los especialistas, por su parte la planificación y la compilación representan el 50%, mostrando 25% cada uno, mientras que la introducción, difusión y actualización no son empleados. Se evidencia que los pasos para realizar los léxicos no siguen el ciclo completo regido en las Normas.

Para medir la calidad de la indización existen diversos métodos, los más representativos son la exhaustividad y la perspectiva de usuario, que son empleados por todos los encuestados, también se utilizan medianamente la precisión y la corrección, los demás métodos no son usados (consistencia extrínseca e intrínseca y otras medidas de recuperación). Es decir, los métodos más novedosos insertados en el contexto digital no son muy utilizados en las bibliotecas universitarias, esto demuestra que la indización es vista desde una perspectiva aislacionista más cercana al paradigma físico.

Ninguna de las técnicas: Métodos Tradicionales de trabajo Manual, Clustering, Redes Neuronales, Métodos Bayesianos, TF-IDF, MapReduce, VSM y Bag of word (bolsa de palabra), Semántica Latente, N-Grams, Basado en Datos de Referencias Bibliográficas, Traducción Automática, Cálculos de la Calidad de los Términos son usadas por los encuestados para la discriminación de términos o para la gestión del orden del tesauro. No existe una formación desde el pregrado que incluya las Minería de Texto y sus técnicas en la actividad informacional y de ahí que no se empleen estos procederes.

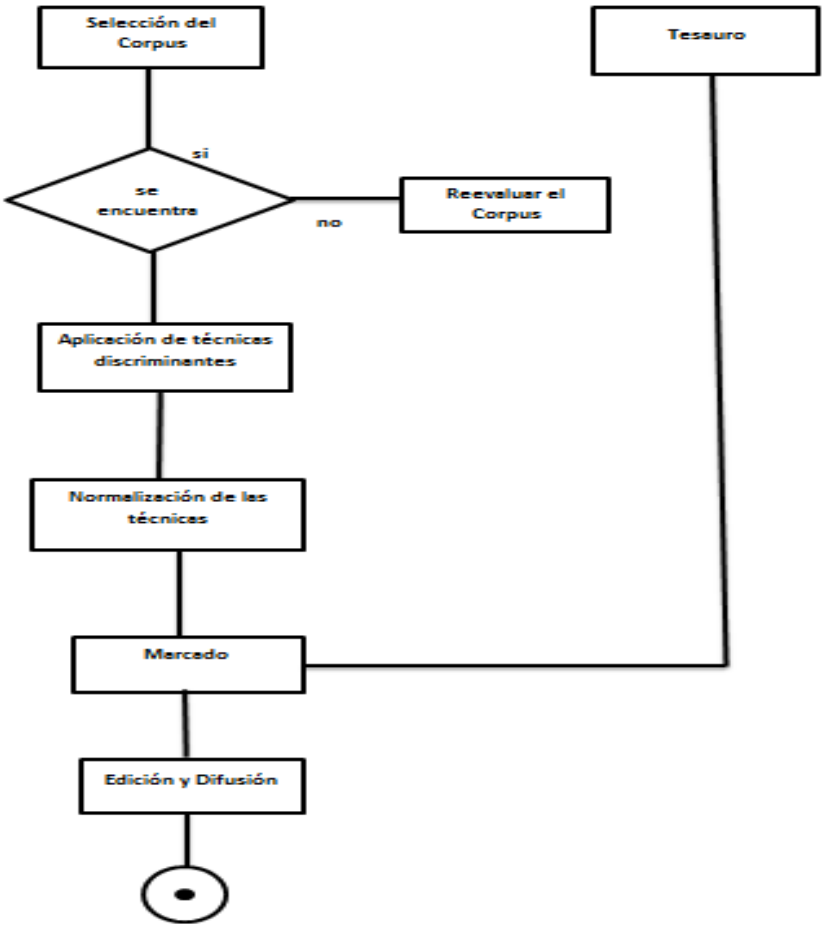
Procedimento para transformar en tesauros los vocabularios controlados de las bibliotecas asociadas a la Red TIC del Proyecto VLIR en Cuba

Procedimiento para la construcción de Tesauros

Tipo de Procedimiento: el tipo de procedimiento es semiautomático porque hay intervención humana, imprescindible para la iniciación del proceso global de construcción de un tesauro, es decir, existe mediación humana sin dejar todo el proceso al instrumento de inteligencia artificial. Ofrece mejorías al proceso de recuperación de información proporcionando una riqueza de componentes lingüísticos, generando de forma automática representaciones de un dominio y por consiguiente utiliza técnicas de origen tanto informático, estadístico y de inteligencia artificial, sin dejar a un lado las pertenecientes a las Ciencias de la Información.

Para introducir cada de una de las etapas por la que atraviesa este apartado se presenta un diagrama de flujo (Fig. 1), medio imprescindible que representa el curso de un proceso desde que inicia hasta que termina y se considera útil para ubicar al lector sobre los contenidos abordados en el trabajo.

Figura. 1. Diagrama de flujo de las etapas



Fase 1. Selección del Corpus

La primera fase se inicia empleando técnicas de recuperación de información desarrolladas en el ámbito de la lingüística de corpus, disciplina que ha mostrado su utilidad en proyectos a gran escala como las dos fases de *ACQUILEX*[6] (Adquisición de conocimiento léxico para sistemas de lenguaje natural) o, a menor escala dentro del mundo de los lenguajes documentales, los que pretenden generar clasificaciones, tesauros u ontologías.

Un corpus, para (Bowker y Pearson, 2002), puede ser descrito como "una gran colección de textos auténticos que han sido compilados de forma electrónica de acuerdo a una serie de criterios previos". Por lo demás, la lingüística de corpus queda definida como "el estudio del lenguaje basado en ejemplos extraídos a partir del uso del idioma en la "vida real" (Mcenery y Wilson, 1996). Biber y sus colaboradores describen los estudios de corpus de manera más concreta, y explican que estos consisten en análisis empíricos de una amplia colección de textos reales, para los que el investigador hace uso extensivo de material informático y aplica técnicas de análisis tanto cuantitativas como cualitativas. Con todas estas características, la lingüística de corpus atraviesa, hoy por hoy, un momento de creciente popularidad dentro del ámbito de estudio de la lexicografía (Biber y otros, 1998).

Existe una gama, cada vez más variada de corpus (Bowker y Pearson, 2002):

1. Corpus genéricos frente a corpus especializados
2. Corpus escritos frente a orales
3. Corpus monolingües frente a multilingües
4. Corpus sincrónicos frente a diacrónicos
5. Corpus abiertos frente a cerrados
6. Corpus de aprendices
7. Corpus impresos frente a los electrónicos

Todos estos corpus han permitido, además, la realización de diversos análisis. Con el corpus se ha profundizado en los ámbitos de la gramática, la semántica, la pragmática, el análisis discursivo, la lexicografía, la sociolingüística, la lingüística histórica, la traducción y la documentación, por medio de la representación del conocimiento.

En estos campos, el análisis de corpus es puramente descriptivo (frente a las aproximaciones de introspección propias de la gramática generativa) y se afana por identificar las asociaciones que rigen el funcionamiento de los diversos componentes textuales.

La lingüística de corpus orientada a la investigación textual intenta identificar y analizar patrones de uso (estructuras y rasgos lingüísticos) y correlacionarlos con variables extralingüísticas que puedan determinarlos. Para ello emplea técnicas de análisis cuantitativas y cualitativas: dentro de las primeras, pueden mencionarse programas de concordancias (ocurrencias de determinado tipo en un corpus), estadísticas (extensión oracional, número de palabras, etc.), búsquedas diversas (listas de palabras,

[6] El Proyecto Acquilex fue financiado por la Comisión Europea bajo la iniciativa de Investigación Básica.

índices, etc.). Ilustraremos el modo de trabajo de la lingüística de corpus con un modelo desarrollado en los últimos años por Douglas Biber, Finnegan, Susan Conrad, entre otros. Estos autores han propuesto un método para estudiar la variación textual –o registros– llamado análisis multidimensional: se basa en córpora que representan el rango total de los esquemas mayores de concurrencia en una lengua (Biber y otros, 1998).

Uno de los pilares fundamentales referente a la creación de un corpus son los criterios que deben guiar su diseño para que sea realmente representativo de la lengua que, valga la redundancia, representa. ¿Qué variedades de uso de la lengua debe incluir? ¿En qué proporción? ¿Cuál debe ser el tamaño de un corpus para que, realmente, represente una lengua o, mejor dicho, el uso que sus hablantes hacen de ella?

Este tipo de consideraciones son las que deben guiar los criterios de recopilación de los textos incluidos en el corpus. Aunque la literatura sobre este campo es extensa, la realidad es que hasta la fecha, casi todos los corpus se han diseñado con criterios internos al proyecto en cuestión, y sólo en determinados casos (*British National Corpus*[7], *Birmingham Collection of English Text*[8], Corpus CUMBRE o el Corpus *ARTHUS*[9]) se han hecho públicos los criterios de selección de los textos incluidos en el corpus (Pérez, 2002).

Representatividad, estandarización y tipología de los corpus han sido tres de los temas más debatidos entre la comunidad científica, con opiniones diversas recogidas en varios artículos y propuestas, algunas de ellas hechas en el seno de importantes proyectos europeos (Atkins, 1992), (Eagles, 1994), (Eagles, 1996a), (Eagles, 1996b), (Biber y otros, 1998). En *EAGLES*, por ejemplo, Sinclair define unos criterios mínimos que deben cumplirse para que un conjunto de textos en formato electrónico pueda ser considerado un corpus (cantidad, calidad, simplicidad y documentación), y clasifica los diferentes tipos de corpus que pueden existir, para así diferenciarlos de las colecciones de textos o los archivos (archives), ya que estos últimos no cumplen alguna de ellas:

1. El corpus debe ser tan grande como sea posible, teniendo en cuenta las posibilidades tecnológicas de la época.
2. Debe contener ejemplos de gran cantidad de tipos, con el fin de contemplar todas las posibles representaciones.
3. Debe ser una clasificación intermedia entre géneros.
4. Los ejemplos (cada fichero) debe tener un tamaño representativo.
5. Debe tener una fuente clara y válida.

La mayoría de los corpus usados por la comunidad científica no se ajustan a alguna de estas recomendaciones, aunque existen ya muchos proyectos que siguen las líneas de trabajo marcadas por *EAGLES.* Las dos primeras recomendaciones hechas por *EAGLES* recogen la polémica suscitada hace unos años a la que muchos se referían como calidad vs. cantidad, es decir, aquellos que daban más importancia al hecho de que el corpus fuera representativo y equilibrado y aquellos que, además, destacaban la importancia de que el corpus fuera lo más cuantioso posible (Pérez, 2002).

[7] http://www.natcorp.ox.ac.uk/
[8] http://www.english.bham.ac.uk/research/language/corpus.shtml
[9] http://adesse.uvigo.es/data/corpus.php

Por razones de tiempo no se pudo profundizar en esta polémica, aunque ambas argumentaciones tienen parte de razón, ninguna postura debe ser llevada a extremos. Se ha enfatizado en la representatividad del corpus, independientemente de qué textos o partes de textos han de incluirse o excluirse y los criterios que deben guiar la composición y el diseño del corpus, pero la representatividad sigue siendo un concepto bastante vago. Los estudiosos no parecen ponerse de acuerdo en cuáles son los rasgos (o los tipos de textos) que representan una lengua, ni qué proporción o que variables (número de lectores/ oyentes, amplitud geográfica de distribución, etc.) deben guiar la inclusión o exclusión de textos.

La selección del Corpus es la etapa más compleja del desarrollo de un tesauro. En esta etapa ocurre la formación del corpus que ha de servir de sustento al tesauro. Es recomendable seleccionar un corpus si y solo si es necesario construir un tesauro desde el inicio por no existir léxicos de referencia.

Esta fase se divide en dos sub-fases básicas: la evaluación y el etiquetado.

1. La evaluación de la fuente de información de donde se va a extraer el corpus en forma manual o mediante herramientas. Las evaluaciones de las fuentes llevan el análisis de los aspectos siguientes:
    a) **Autoridad:** este aspecto se relaciona con la identificación correcta del autor del documento donde se van a extraer los términos. Es importante revisar la calidad de: título, reputación, experiencia y currículum. Hay que valorar conocimiento y experiencia del autor o autores, además la reputación de la organización de la que se deriva la información, siendo un indicador potente de la fiabilidad y credibilidad de la información.
    b) **Ámbito temático:** especialización en relación con la temática que trata el tesauro ejemplo: bioquímica, genética, microbiología, biofísica y biología molecular, etc.
    c) **Audiencia a la que va dirigida la fuente:** es importante saber si el tesauro se dirige a: investigadores, estudiantes, académicos y científicos.
    d) **Actualidad:** otro aspecto importante es la actualidad de la información y la frecuencia o regularidad de su actualización en determinadas áreas temáticas. Esto propicia actualización y cubrimiento léxico.
    e) **Exhaustividad y cobertura:** la construcción de un léxico incluye la valoración del nivel de profundidad y exhaustividad con que se cubre el tema en cuestión, lo que influye en la calidad de los candidatos terminológicos.
    f) **Objetividad, falta de sesgo:** relacionada con el propósito con el que se ha escrito la información, debe tenerse en cuenta a la hora de interpretar y usar la información. Internet se ha convertido en una herramienta muy importante de publicidad y *marketing*, por lo que es aconsejable preguntarse por qué el autor ha puesto esa información en la red. Datos sobre la afiliación o autoría del recurso nos alertarán sobre un posible sesgo o un punto de vista parcial de la información dada.
    g) **Información primaria o secundaria:** hay que saber distinguir si es una fuente de información original, que contribuye a algo único sobre el tema, o apenas contiene información y sólo incluye enlaces.
    h) **Documentación de apoyo:** todo concepto que se coloque en el tesauro lleva una nota de alcance y definición; por tanto, toda fuente de la que se tome información tendrá asociados las referencias y las fuentes bibliográficas que le han servido de constituyente.
    i) **Calidad de la escritura y sintaxis:** que el texto esté bien escrito, con una buena gramática, sin faltas de ortografía, es importante para que el proceso etiquetado y extracción de términos sea

correcto. Los errores de sintaxis y gramaticales pueden indicar una falta de cuidado en la redacción del texto, y son elementos importantes para los procesos de construcción del léxico.

j) **Metadatos:** se necesita información estructurada. La descripción de la información se refleja en los metadatos y estos son el sistema principal para realizar los intercambios con otros sistemas.

2. El etiquetado es el proceso que obliga la colocación de etiquetas xml en el texto si se desea obtener términos de cada elemento retórico del texto. En el caso del sistema de léxico que se presenta usa el etiquetado automático mediante una herramienta computacional.

Fase 2 Técnicas Discriminantes (AntConc)

Control del Léxico: Para la construcción de un tesauro es imprescindible el proceso de identificación y adquisición de los componentes más representativos de un dominio. Este proceso es integrado por una serie de subprocesos: análisis léxico (AL), tratamiento de palabras vacías (TPV), tratamiento de términos flexionados (TTF), tratamiento de palabras compuestas (TPC) y filtrado de términos (FT). Cada uno de ellos ha sido tratado por Moreiro (1999) y teniendo en cuenta la praxis internacional son redefinidos.

Una de las herramientas que permiten este proceso es el *AntConc*. Los pasos que se deben seguir son:

1. Abrir el programa, en el interfaz de *Google*, escribir *AntConc* y seleccionar la versión correspondiente a *Windows: AntConc 3.2.1*.
2. Seleccionar el corpus (debe estar bien diseñado y compilado para evitar falsos o erróneos análisis de los resultados) y exportar a un documento en formato *.txt* para ser visualizado en el programa.
3. Abrir el documento *.txt* y remitirse a la opción *Start* para generar la lista de palabras que aparecen ordenadas por frecuencia.
4. Eliminar las palabras vacías del corpus documental, denominado *stopword list*, dígase artículos definidos e indefinidos, numerales, adverbios, así como palabras de contenido muy general. Para ello lo primero es crearse un archivo en texto plano *(.txt)* en el Bloc de Notas con todas las palabras que se desea que no aparezcan en el listado, deben estar separadas por coma o por saltos de párrafos como fue el caso del usado en el trabajo. Desde la pantalla principal de *AntConc*, remitirse a *Tool Preferences*, en el cuadro de <*Category*> desplegado en el panel izquierdo superior de la ventana, seleccionar la opción <*Word List*>, dentro del cuadro de sus preferencias conocido como <*Word List Preferences*> en la categoría <*Other Options*>, marcar la casilla <*Treat all data as lower case*>, seguido a esto remitirse a <*Word List Range Options*> y al lado de la opción <*Add words from file*> clicar <*Open*> para recuperar el fichero *.txt* que contiene la lista de exclusión, finalizar con la opción <*Apply*> y generar nuevamente el listado de palabras y observar los resultados.
5. Ordenar los resultados: en la opción *Sort* seleccionar el orden que se desea asignar (frecuencia, alfabéticamente, rango, entre otros), en el trabajo se organizaron alfabéticamente.
6. Establecer en la opción *Clusters/N-Grams* los tamaños de los *N-grams* y *Cluster* que utiliza un algoritmo *k-means* y dar en la opción *Start* para cargar los cambios ejecutados. En el caso presentado se empleó un trigrama de nivel de palabra porque mientras mayor es el tamaño del *N-grams* y *Cluster* mayor significado se obtiene.
7. Guardar los listados: ir a la opción *File*, en su menú desplegable elegir *Save Output Text File*. Los datos se guardan con vista a recuperarse sin tener que cargar nuevamente los textos.

Las técnicas aplicadas por Moreiro en investigaciones no son eficientes, sin embargo, algunas se utilizaron en el trabajo. Para la organización de los conceptos se procedió a usar la ley de *Zipf*, con el objetivo de conocer la frecuencia de palabras que existen en el corpus, luego se aplicó la técnica *stop word elimination* para suprimir las palabras vacías. Se empleó un *TF-IDF* con el fin de obtener los términos de mejor calidad lingüística, a partir de la frecuencia inversa de las palabras en los documentos, esto arroja los términos candidatos. La discriminación de los términos termina por el proceso de obtener cluster para analizar la pertenencia de los términos a diversos grupos, para ello se utilizó el algoritmo *K*-means, que permite agrupar los conceptos en *cluster* con un umbral definido, en este caso 3.

Se trabajó con el método de *N-Gramas* donde se obtuvieron trigramas para secuenciar caracteres en las estructuras lingüísticas y obtener mayores cuotas de significado, dividiendo la población de términos en grupos homogéneos en función de las características morfo- lingüísticas. Otras técnicas que constituyen heurísticas o algoritmos de *Stemming* no son utilizadas por no usarse en estos procesos corpus de contrastes o corpus previamente etiquetados de alto niveles de especialización.

Se propone evaluar el agrupamiento usando medidas de precisión y recobrado:

- Precisión; evalúa si los términos pertenecen o no al cluster que los agrupa, *a* constituye los términos agrupados en el cluster y *b* los términos que no pertenecen al cluster.

$$Precisión = \frac{a}{a + b}$$

Ecuación 27. Precisión

- Recobrado o *recall*; se evalúa mediante el cociente de los elementos que corresponden al cluster y los falsos negativos, es decir los elementos que fueron separados en el agrupamiento pero que pragmáticamente pertenecen a él. Donde *a* constituye los términos agrupados en el cluster y *c* los términos que de alguna forma han sido excluidos de este pero que indiscutiblemente corresponden a él.

$$Recall = \frac{a}{a + c}$$

Ecuación 28. Recobrado

Fase 3. Normalización de las Técnicas

Esta fase está precedida por la selección de conceptos. Concretamente se propone el uso de algunas reglas de construcción de términos para tener debidamente normalizados los candidatos terminológicos para el proceso de edición. Aquí se realiza la normalización de las entradas, formas gramaticales y se seleccionan los tipos de descriptores, además se describen algunas regularidades lingüísticas que la práctica de nuestro idioma así lo aconseja. Se recomienda usar la regla UNE.

Fase 4. Marcado

Subfase 1. Organización y adopción del esquema

En esta fase se describe la ontología del sistema. La ontología en *SKOS* realiza un conjunto de tareas de inferencia semántica de las que se deduce que el conocimiento está correcto (captura intuiciones) que se declaran a continuación:

a) C incluye D w.r.t. K ssi para *cada modelo* I de K, CI ⊆ DI esto indica que Conocimiento es mínimamente redundante (no hay sinónimos)
b) C es equivalente a D w.r.t. K ssi para *cada modelo* I de K, CI = DI Conocimiento es significativo(clases pueden tener instancias) C es *satisfiable* w.r.t. K ssi existe *algún modelo* I de K s.t. CI ≠ ∅; Consulta conocimiento x es una instancia de C w.r.t. K ssi para *cada modelo* I de K, xI ∈ CI <x,y> es una instancia de R w.r.t. K ssi para, *cada modelo* I de K, (xI,yI) ∈ RI
c) Consistencia del KB Un KB K es consistente ssi existe *algún modelo* I de K

La base de conocimiento descrita presenta una semántica donde una interpretación I satisface (modela) un axioma A (I ⊨ A):

- I ⊨ C ⊑ D ssi CI ⊆ DI
- I ⊨ C ≡ D ssi CI = DI
- I ⊨ R ⊑ S ssi RI ⊆ SI
- I ⊨ R ≡ S ssi RI = SI
- I ⊨ R+ ⊑ R ssi (RI)+ μ RI
- I ⊨ x ∈ D ssi xI ∈ DI
- I ⊨ <x,y> ∈ R ssi (xI,yI) ∈ RI
    - I satisface un Tbox T (I ⊨ T) ssi I satisface cada axioma A en T
    - I satisface un Abox A (I ⊨ A) ssi I satisface cada axioma A en A
    - I satisface un KB K (I ⊨ K) ssi I satisface ambos T y A

Los principios de la lógica descriptiva gestada en la ontología son una familia de formalismos basados en *KDF* (Knowledge *Description Formalism*-Lógica descriptiva). Los lenguajes particulares se caracterizan por ser un conjunto de constructores para construir conceptos y roles complejos a partir de otros más simples – Conjuntos de axiomas para dar aserciones acerca de conceptos, roles e individuos. Los constructores incluyen booleanos and (⊓), or (⊔), not (¬), y –; además utilizan restricciones acerca de "sucesores" en los roles usando ∃ y ∀.

Se declaran los axiomas que se utilizan en la ontología (tabla I):

Tabla I. Constructores para construir conceptos

Axioma	Sintaxis	Ejemplo
SubClassOf	$C_1 \sqsubseteq C_2$	$Ordered \sqsubseteq orderedColection$
equivalentClass	$C_1 \equiv C_2$	$Término \equiv Concept$
DisjointWith	$C_1 \sqsubseteq \dashv C_2$	$Collection \sqsubseteq \dashv ConceptsCheme$
InverseOf	$P_1 \equiv P_2^{-}$	$has\ broader \equiv has\ narrower^{-}$
transitiveProperty	$P^{+} \sqsubseteq P$	$has\ semantic\ relation^{+} \sqsubseteq related$

Posteriormente se explica la concepción de la base de conocimiento DL K = <T, donde la Taxonomía A> – T (Tbox) es un conjunto de axiomas de la forma:

a) C ⊑ D (inclusión de concepto)
b) C ≡ D (equivalencia de concepto)
c) R ⊑ S (inclusión de rol)
d) R ≡ S (equivalencia de rol)

A (Abox) es un conjunto de axiomas de la forma:

a) x ∈ D (instanciación de concepto)
b) <x,y> ∈ R (instanciación de rol)

Dos tipos de axiomas en Tbox: – *"Definitions".*

a) C ⊑ D o C ≡ D donde C es un nombre de concepto

Definiciones pueden ser cíclicas – Axiomas Generales de Inclusión (GCIs).

a) C ⊑ D donde C es un concepto arbitrario.

El uso fundamental de un tesauro es la recuperación de información, su aplicación es buscar por conceptos, los que constituyen el núcleo primordial de *SKOS*, los términos por los que se encuentran representados los conceptos pueden ser simples o compuestos.

Subfase 2. Marcado del tesauro

Conceptos

Los conceptos en *SKOS* son los términos que integran un tesauro. Se distinguen por las categorías de significado y su interacción con los objetos y las propiedades correspondientes de los términos utilizados para etiquetarlos. Los conceptos en *SKOS* se describen con la clase *skos:Concept*, que permite incluir términos para la indización de un recurso. La gestión o introducción de un concepto lleva dos pasos básicos:

a) elaborar o volver a utilizar un Identificador Uniforme de Recurso *(URI)* que posibilite distinguir específicamente e indubitable al concepto.
b) declarar una estructura *RDF* mediante la propiedad *rdf:type*, que exteriorice que el recurso reconocido a través de dicha *URI* es del tipo *skos:Concept*. Introduciendo los conceptos en *SKOS*.

Cada uno de los conceptos que se representan en el tesauro antes de llegar a ser visualizado, conllevan un proceso de análisis íntegro donde se detectan los que resultan útiles para recuperar información de determinado dominio, cada uno, se representa mediante términos de modo que para cada uno se escoge una de las posibles representaciones como el término preferente. Para alguna de estas acciones se hacen uso en *SKOS* de las etiquetas.

Etiquetas

Los términos o conceptos se utilizan a menudo en los tesauros para unidades de lenguaje natural con diferente carga semántica, esto explica que los términos se construyen a partir del uso de la lengua. Para ello en la construcción de lenguajes de indización se utilizan las siguientes etiquetas:

skos:prefLabel: utilizada para indicar preferencia en el léxico, ejemplo si usted tiene el término Archivología entonces preferirá usar Archivística.

skos:altLabel: se utiliza para asociar sinónimo a un concepto. Un ejemplo también puede ser cuando se relaciona un concepto en inglés con otro en español.

skos:hiddenLabel: etiqueta léxica oculta que permite la asociación de un recurso escrito de forma correcta a un recurso con variante ortográfica, se utiliza fundamentalmente en los procesos de indexación en los sistemas de recuperación de la información. Gracias a esta etiqueta quedan disponibles términos comprensibles para diversos robots que realizan búsqueda a texto completo e indización automática.

Cuando se trabaja con un tesauro se manejan varios tipos de relaciones semánticas entre los conceptos que lo integran, entiéndase como la relación que se establece entre un par de conceptos cuando el alcance de uno de ellos queda completamente dentro del alcance del otro. Las relaciones semánticas se establecen bajo tres tipos de categorías: equivalencia, asociación y jerarquía.

La relación de equivalencia está asociada a los problemas generados por errores de sinonimia y polisemia cuando se seleccionan los términos preferentes que identifican a cada concepto. Se distinguen entre preferentes y no-preferentes, utilizando las relaciones *USE* (usar o véase) y *UF* (usado por). *SKOS* sigue un patrón similar donde de igual manera las etiquetas preferentes se distinguen de las no-preferentes (alternativas), para ello se vale del empleo de las etiquetas *prefLabel* y *altLabel* ejemplificadas inicialmente.

La relación jerárquica propone que deberían existir grados o niveles de superordenación y subordinación donde el concepto superordinado represente una clase o un todo mientras que los subordinados se refieran a sus miembros o partes.

La Norma *UNE* es una copia fiel de la Norma Internacional ISO 25964-1 del 2014 que propone etiquetas con dos tipos de relaciones y plantea su uso recíproco: cuando se tiene un concepto X establecido como genérico de otro Y, entonces Y se convierte en un concepto específico de X (propiedad inversa).

SKOS proporciona dos propiedades estándar para representar estos términos, que permiten, la representación de los vínculos jerárquicos, como la relación entre un género y sus especies más específicas, o, en consonancia a su interpretación, la relación entre un todo y sus partes.

Un concepto X en este caso Archivos, que tiene como término genérico Y, para el caso, Archivos Cerrados atendiendo a la nota de significado. Siempre que esto ocurra se encuentra en presencia de una propiedad inversa, a lo que corresponde decir que para el ejemplo que se encuentra analizando en cuestión, Y conocido como Archivos Cerrados se convierte como acción alterna en Término Específico de X Archivos y el razonador del programa realiza una inferencia, aplicando la propiedad inversa sin necesidad del llenado de los campos de forma manual.

La relación asociativa, incluye las asociaciones entre pares de conceptos que, aunque no se encuentran relacionadas jerárquicamente, lo están semántica o conceptualmente, hasta el punto que es necesario hacer explícito en el tesauro el enlace que existe entre ellos porque pueden sugerir términos adicionales o alternativos, que son enriquecedores para el proceso de indización, las estrategias de búsqueda y el grado de conocimiento (UNE, 2014).

SKOS para denotar una relación de asociación entre dos conceptos utiliza la etiqueta *skos:related*. Las relaciones definitorias o aclaratorias, aunque no se incluyeron al inicio dentro de los tres tipos de categorías también se destacan en un tesauro, entiéndase por esta relación a la descripción de la aplicación de un descriptor. Se emplea para representarlo la notación NA (Nota de Alcance). *SKOS* también trabaja bajo ese precepto y establece que "los conceptos a veces tienen que definirse con mayor precisión utilizando documentación ("informal") que sea legible para las personas, como notas de alcance o definiciones*"* (Pastor y Martínez, 2010).

El Manual de *SKOS* es muy específico respecto a la propiedad *skos:note,* y la provee para representar este tipo de relación, se explota con fines de incorporar información de carácter general, sin embargo se especializa en otras como:

a) *skos:scopeNote:* proporciona información, posiblemente parcial, sobre el significado de un concepto, sobre todo como una indicación de cómo se limita su uso en los procesos de indización (Pastor y Martínez, 2010). Para reproducir cada una de las variantes de *skos:note* se tomó el concepto Archivos.
b) *skos:definition*: se utiliza para ofrecer una información detallada de la acepción de un concepto.
c) *skos:example*: se emplea para ejemplificar la función de un concepto.
d) *skos:historyNote:* se maneja para reflejar transformaciones de peso en la acepción o morfología de un concepto.

También contiene dos especializaciones de skos:note que resultan eficaces para los gestores o editores de sistemas de organización: *skos:editorialNote* y *skos:changeNote*.

a) *skos:editorialNote*: se explota como una especialización de configuración, mediante advertencias de tareas que quedan por elaborar o indicaciones para modificaciones editoriales que pudiesen ocurrir a corto o largo plazo.
b) *skos:changeNote:* se utiliza para notificar al editor o gestor de cambios muy específicos efectuados a un concepto como consecuencia de modificaciones en su administración y mantenimiento.

Es imprescindible tener en cierto nivel los conceptos seleccionados con sus respectivas etiquetas y asociaciones relacionales, hasta aquí pueden utilizarse como entidades independientes, sin embargo, una vez que se obtiene esto es necesario asociarlos a un vocabulario con un nivel de recopilación cuidadoso, en este caso un tesauro y más específico DOCUTES (Tesauro de Ciencias de la Documentación), al realizar esta operación de asociar por ejemplo el concepto Archivos al tesauro DOCUTES al activar el razonador este realiza inferencias y por defecto todos los términos que se asocian al concepto automáticamente lo hacen al tesauro, este proceso se denomina esquemas de conceptos. Para realizar esta

operación SKOS proporciona la propiedad *skos:ConceptScheme* específicamente mediante la propiedad *skos:inScheme*.

Hasta el momento no solo se seleccionan los conceptos y establecen etiquetas y relaciones, sino que además se puede referenciar su contexto de procedencia, pero qué sucedería si los conceptos no pertenecieran al mismo esquema, si se encuentran en esta parte del proceso desde diversos contextos puede que los términos guardaran un significado similar. Para ello *SKOS* ofrece la posibilidad de mapear los esquemas de conceptos mediante el uso de dos propiedades *skos:exactMatch* y *skos:closeMatch*.

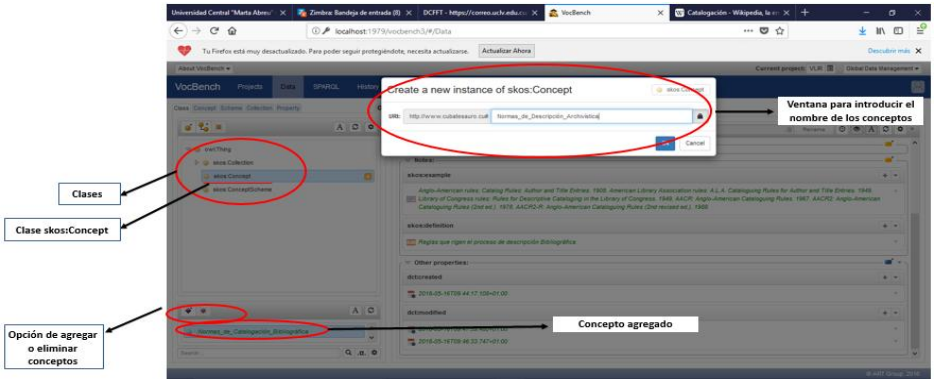
Otro aspecto importante además de poder esquematizar conceptos y mapearlos es establecer el orden de los mismos dentro de la colección, para realizarlo *SKOS* facilita la clase *skos:OrderedCollection* junto a la propiedad *skos:memberList*, primero hay que declarar que el concepto forma parte de una lista de miembros y luego al guardar los cambios apreciar cómo se van organizando los conceptos.

Fase 5. Edición y Difusión

La edición y difusión del tesauro se realiza mediante *VocBench*. El sistema gestiona en línea para todas las entidades de la red un tesauro único, editado y aprobado por sus bibliotecarios. El primer paso para instaurar el *VocBench* es el establecimiento de los roles de edición. Para ello se propone lo siguiente: 2 administradores de Sistema (asignan roles y dan permisos), 2 expertos en *SKOS* (modifican la ontología y gestionan los mecanismos de inteligencia artificial, 5 revisores (aprueban los términos que han de usarse en el sistema de indización de red), los editores (construyen las relaciones terminológica y proponen términos), finalmente el difusor o publicador es quien publica la ontología para que sirva a los sistemas de la red (*Dark Archieve*, ABCD, *Dsapce* y *Moodle*).

La primera interfaz es la de inicio donde el programa solicita las credenciales y en correspondencia a los roles establecidos otorga permisos específicos para el trabajo en el software. Los conceptos o términos pueden ser exportados directamente del Protegé mediante una ontología creada con antelación, para una mejor comprensión de sus potencialidades se presentan casos desde la introducción de los términos desde la etapa inicial (Fig. 2). El riesgo que se corre al trabajar desde el *VocBench* es que el trabajo del especialista al tratar el corpus es más riguroso pues no detecta errores que se cometen en la introducción de las familias de palabras con sus múltiples relaciones.

Figura 2. Introducción de un concepto



Las etiquetas son introducidas atendiendo a la preferencia del uso de un concepto con respecto a otro por tener asociado terminología con significado similar.

En el área de las notas se pueden agregar definiciones de los conceptos, ejemplos de diferentes conceptos, datos de carácter histórico, notas de edición, variación en el funcionamiento del término o transición que se deba ejecutar a corto o largo plazo. En la (Fig. 3 y 4) se ejemplifica la opción de añadir la definición de un concepto.

Figura 3. Introducción de una definición dentro del área de notas

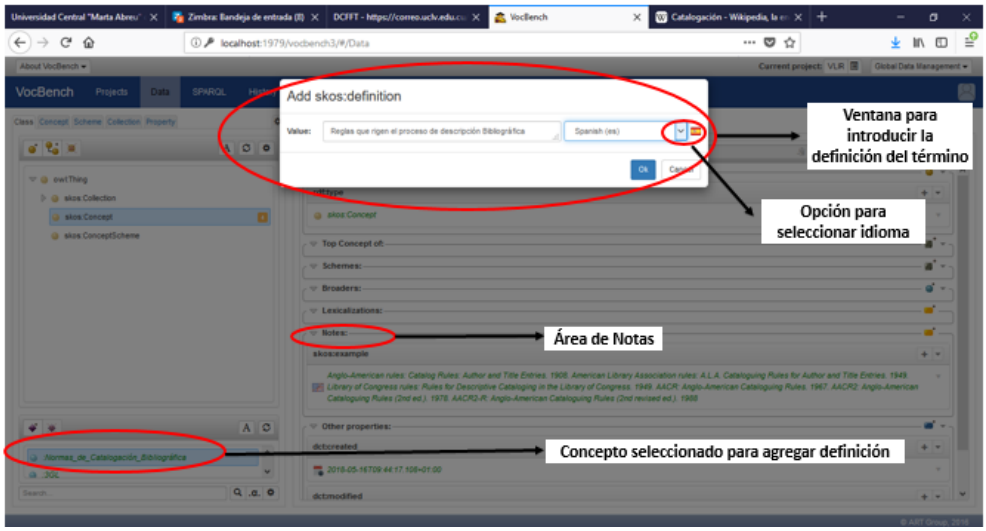
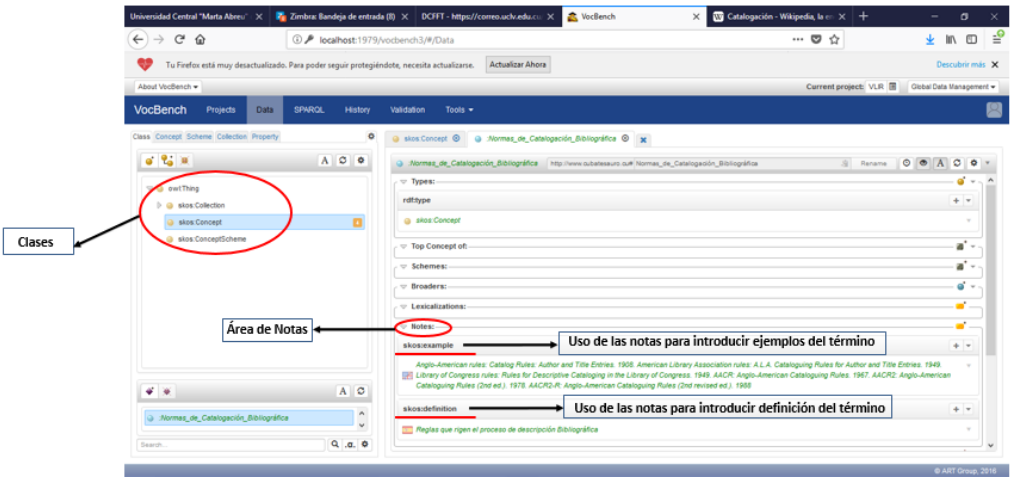
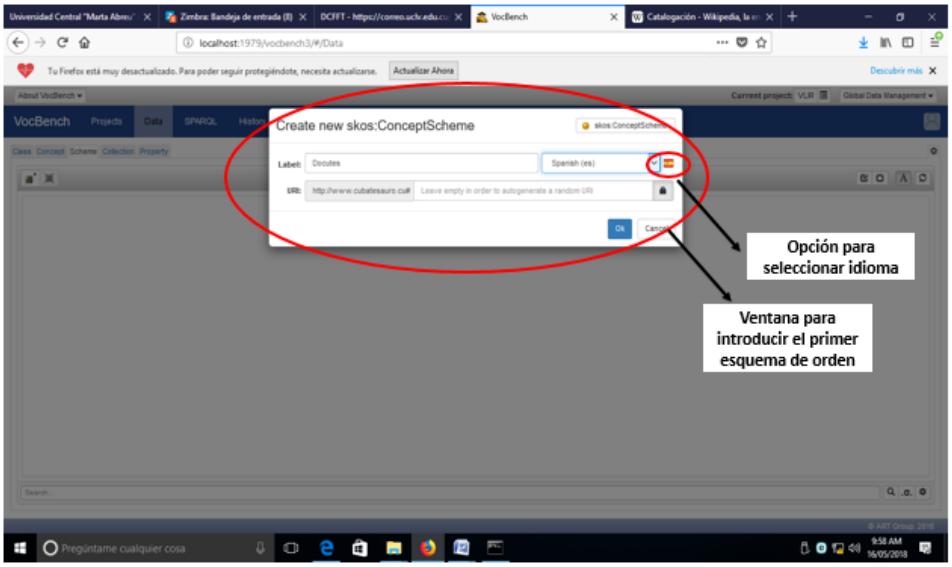


Figura 4. Definición y ejemplos del término dentro del área de notas



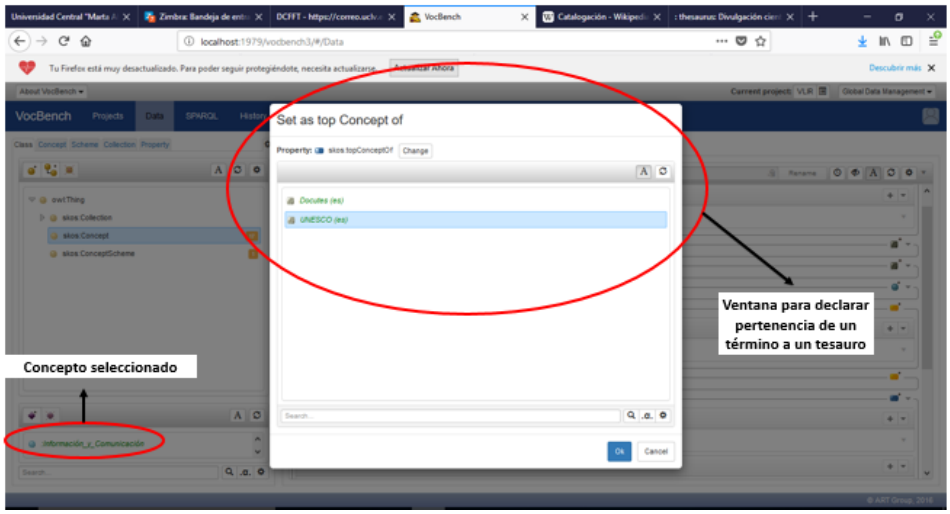
Para definir que un concepto pertenece a un esquema de orden ontológico específico, primero se deben crear los esquemas, esto es posible mediante la propiedad *skos:ConceptScheme* y para reflejar esta aplicación se introduce el primer esquema con el que se trabajó: DOCUTES (Fig. 5).

Figura 5. Creación del primer esquema de orden ontológico



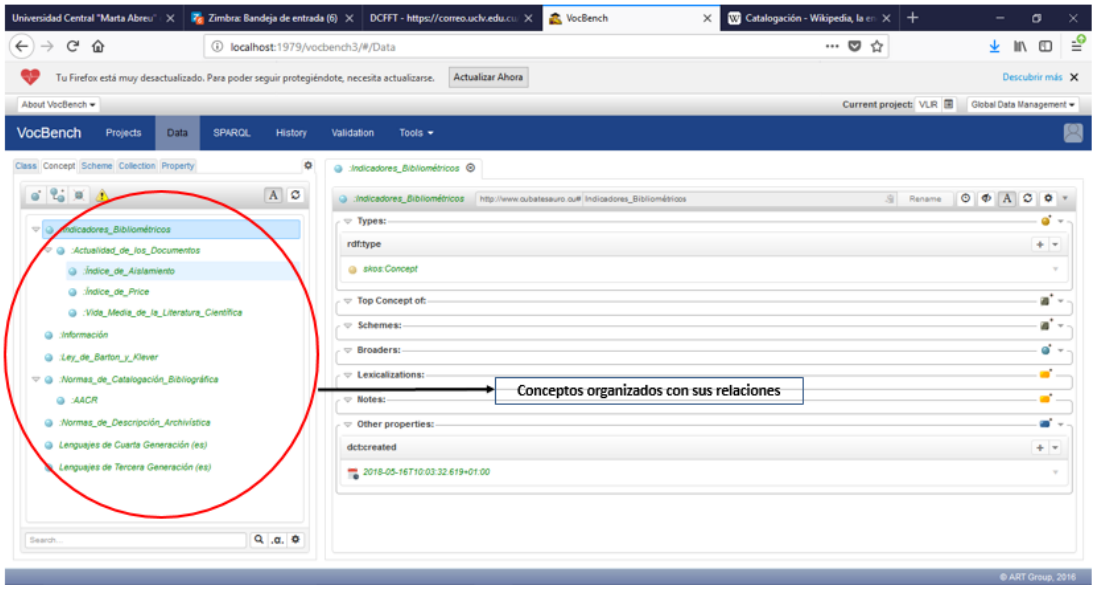
Al crearse los esquemas de orden DOCUTES y UNESCO se deben declarar los términos que pertenecen a cada uno, ante la posibilidad de que, aunque estuviesen en tesauros diferentes posean un significado similar. Luego que se tienen los conceptos registrados con el tesauro del que proceden, se puede establecer relación de similitud o cercanía (propiedad *skos:closeMatch*) entre los términos que lo integran. Para ello, se selecciona uno de los conceptos que se van a relacionar, en la ventana que aparece para ejecutar las propiedades se escoge la de cercanía; seguido a ello el programa se desplaza automáticamente hacia otra ventana donde se elige finalmente el término con el que guarda este tipo de relación el concepto seleccionado (Fig. 6).

Figura 6. Declaración de pertenencia de un término a un tesauro



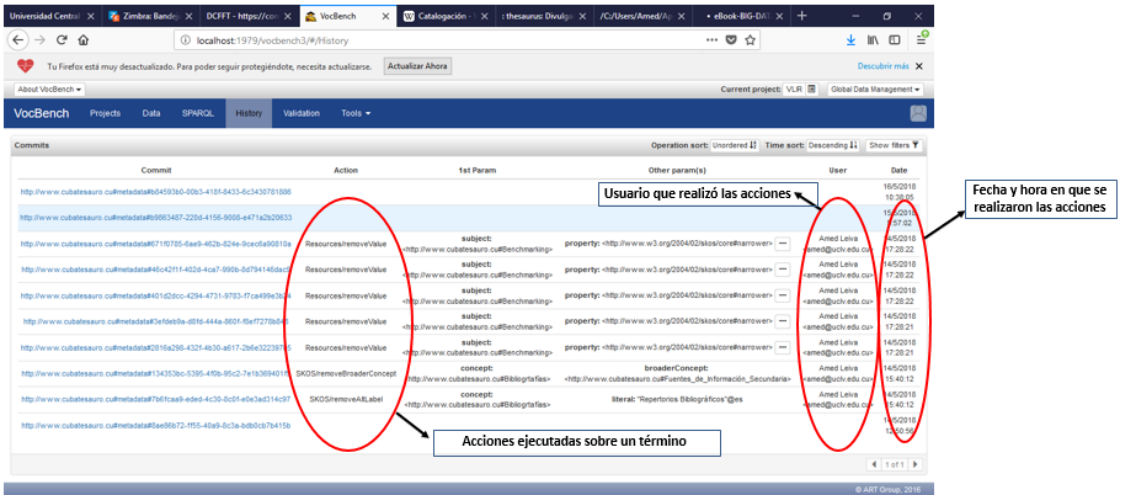
En la (Fig. 7) se visualiza la organización de algunos de los conceptos introducidos hasta el momento, estableciendo previamente los conceptos y sus relaciones, etiquetas y notas, estando esquematizados y mapeados.

Figura 7. Conceptos organizados con sus relaciones



Cuando se trabaja con los términos se puede visualizar el historial de las acciones realizadas sobre los mismos (Fig. 8) y una vez culminado todo el proceso de introducción de los términos con sus respectivas modificaciones proceder a la fase de validación o rechazo.

Figura 8. Historial de las acciones ejecutadas sobre los términos



Finalizando se pueden exportar los datos (Fig. 9) y observar el fichero xml con la visualización total de la ontología (Fig. 10)

Figura 9. Ventana para descargar la ontología

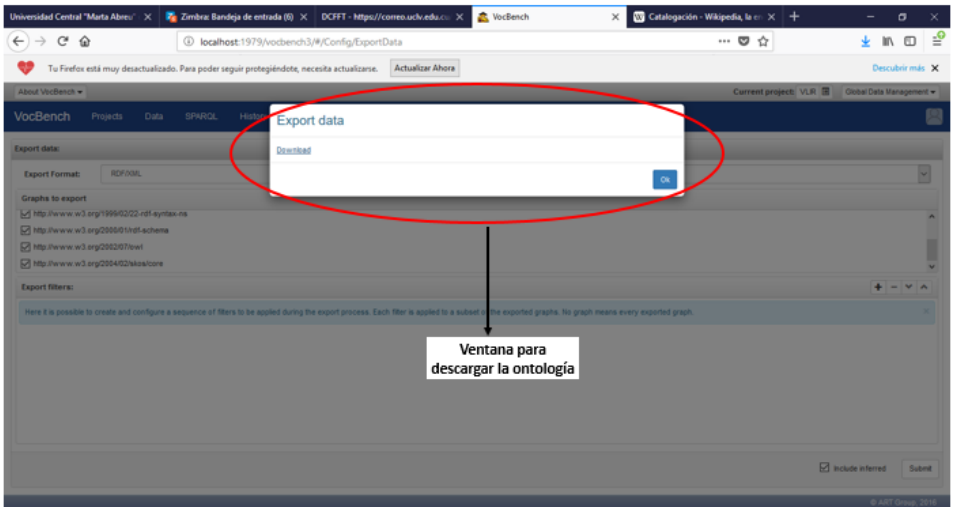
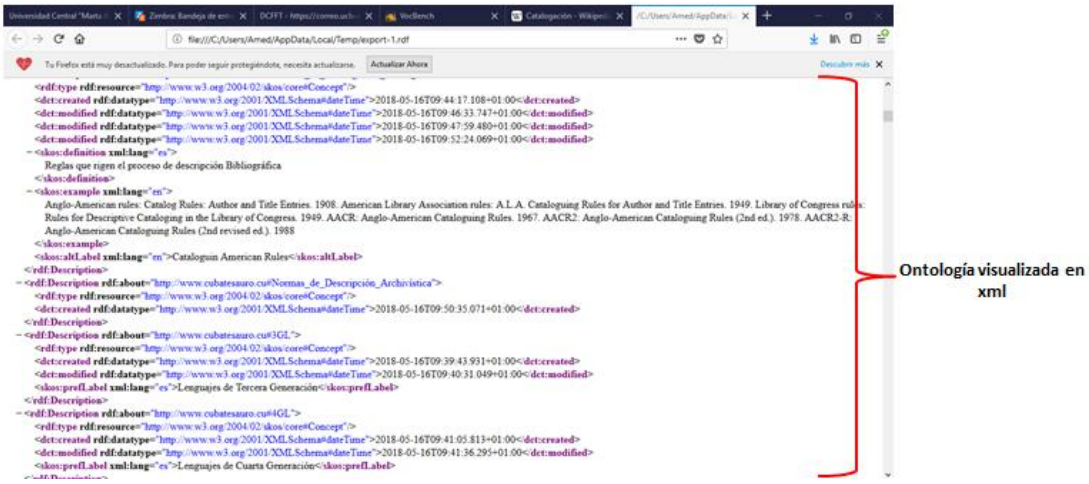


Figura 10. Visualización de la ontología en xml



El procedimiento para la construcción semiautomática de tesauros en las bibliotecas universitarias de la Red TIC del Proyecto VLIR en Cuba, permite transformar los vocabularios que hoy se utilizan en los procesos de indización, garantizando consistencia y normalización en los procesos de indización. Además, facilita una recuperación eficiente de la información en las bibliotecas universitarias de dicho proyecto. Los mecanismos de control e inferencia semántica asociados a SKOS, garantizan que los errores de sinonimia, polisemia, exhaustividad y corrección léxica sean eliminados en aras de garantizar la búsqueda y recuperación de la información de forma efectiva, sin ruido y con altos valores de precisión y recobrado. Desde el punto de vista metodológico sirve de guía para la construcción de léxicos especializados, al incluir los mecanismos de construcción basados en las reglas exigidas a nivel internacional. El impacto social de uso estriba en que al contener datos estandarizados en formato SKOS, las plataformas que usa y desarrolla la red podrán interpretar con otras plataformas con fines similares dando visibilidad a la ciencia de la Red TIC.

CONCLUSIONES:

1. Los referentes teórico conceptuales y metodológicos sobre la elaboración de tesauros han sido descritos desde la Ciencias de la Computación en su vertiente recuperacionista desde la década del 40, permeados de un paradigma multidisciplinar, utilizando metodologías provenientes de la lingüística, la semiótica textual, las Ciencias de la Información y la lingüística de Corpus.
2. Los procedimientos de construcción automática de tesauros se valen de técnicas mixtas para lograr su efectividad en la elaboración de léxicos, dichas técnicas han evolucionado desde simples análisis estadísticos hasta modelos de Big Data, lo que evidencia el culmen de técnicas y procedimientos que sustentan esta actividad.
3. Las técnicas discriminantes utilizadas para medir la calidad terminológica permiten la obtención de clusters de palabras candidatas a incluir en el tesauro, además de determinar la riqueza léxica del corpus.
4. El procedimiento cuenta con 5 fases que integran técnicas computacionales de manera sinérgica y holística procedentes del terreno de la minería de texto, aplicando reglas de construcción ontológica y de marcado, normando la estructura lingüística y difundiendo los contenidos estructurados en el sistema de conocimientos, lo que asegura su operatividad sistémica.
5. El procedimiento para la construcción de tesauros en SKOS para las bibliotecas asociadas a la Red TIC del Proyecto VLIR en Cuba, contiene los elementos necesarios para el desarrollo de la actividad de indización, solucionando en alguna medida las problemáticas existentes en cuanto a la interoperabilidad semántica amparada en el manejo de la Norma UNE.

REFERENCIA BIBLIOGRÁFICAS

Atkins, B. (1992). Tools for Computer-aided Corpus Lexicography: The Hector Project II. *Akadémiai Kiadó*, 41(1-4), 5-71.

Biber, D.; Conrad, S.; & Reppen, R. (1998). Corpus linguistics investigating language structure and use. *Tesol Quaterly*, 32(4), 789-790. https://doi.org/10.2307/3588017

Bowker, L.; Pearson, J. (2002). *Working with specialized language.* Routledge, Londres: *Language* & Literature.

Carras, C. (2004). *Tesauros y Ontologías.* 3 de abril de 2018 http://personales.upv.es/ccarrasc/doc/2003-004/tesaurosonto/principal.html#Presentacion

Cavieres, A.; Fredes, S.; Ramírez, A. (2010). Tesauros y Web Semántica: Diseño metodológico para estructurar contenidos Web mediante SKOS-Core. *Serie Bibliotecología y Gestión de Información, (57), 1-64.*

Chen, Z.; Liu, S.; Wenyin, L.; Pu, G.; Ma, W. Y. (2003). Building a Web Thesaurus from Web Link Structure. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* ACM, 48-55. https://doi.org/10.1145/860435.860447

Codina, L.; Pedraza, R. (2011). Tesauros y ontologías en sistemas de información Documental. *El Profesional de la Información*, 20(5), 555-563. https://doi.org/10.3145/epi.2011.sep.10.

EAGLES (1994). "Corpus Typology: a framework for classification". Informe Interno N. 2.1. En: Sinclair, John M., EAGLES Document 080294, pp. 1-18. Corpus Linguistics Group: Universidad de Birmingham,UK.

EAGLES, (1996). Preliminary Recommendations on Corpus Typology". 26 de marzo 2018 http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

EAGLES. (1996). Text Corpora Working Group Reading Guide". 26 de marzo 2018 http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html

Fernández, A. M. (2010). *La Construcción de tesauros académicos: un modelo general y un método inductivo con aplicación al "e-learning".* [Tesis de doctorado inédita]. Madrid, España: Universidad Complutense de Madrid.

Gil, I. (2017). SISA--Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules versus TF-IDF Rules. *Knowledge Organization. 44(3),139-163.*

González, J. A.; Díaz, I.; Lloréns, J., Morato, J.; Velasco, A. (1999). 5 de diciembre de 2017. Generación automática de tesauros. Propuesta de un método lingüístico-estadístico. *Revista Ciencias de la Información.* http://cinfo.idict.cu/index.php/cinfo/article/download/274/273

Lamarca, M. J. (2013). *Hipertexto, el nuevo concepto de documento en la cultura de la imagen.* [Tesis Doctoral inédita]. Madrid, España: Universidad Complutense de Madrid.

Lanquillón, C. (2002). Enhancing Text Classification to Improve Information Filtering. *Künstliche Intelligenz*, (2), 37-58.

Lunh, H. (1958). The Automatic creation of Literature abstracts. *Journal of Research of Development,* 2(2), 159-165. https://doi.org/159-165. 10.1147/rd.22.0159

Martín, C. (2009). Temas de Biblioteconomía: Lenguajes documentales Principales tipos de clasificación Encabezamientos de materia, descriptores y tesauros. 6 de mayo 2018 http://eprints.rclis.org/14817/1/lendoc.pdf

Mcenenry, A.; Wilson, A. (1996). *Corpus Linguistics*. Edinburgh, Scotland: Edinburgh University Press.

Montejo, A. (2001). Proyecto de indexado automático para documentos en el campo de la Física de Altas Energías. *Procesamiento del Lenguaje Natural*, (27), 295-296.

Moreiro, J. A.; Díaz, I.; Lloréns, J.; Morato, J.; Velasco, M. (1999). Generación automática de tesauros. Propuesta de un método lingüístico-estadístico. *Ciencia de la información*, 30(4), 51-60.

Organización Internacional de Normalización, (2014). *Información y documentación. Tesauros e interoperabilidad con otros vocabularios. Parte 1: Tesauros para la recuperación de la información.* Madrid, España: Asociación Española de Normalización y Certificación (AENOR).

Pastor, J. A. (1992). *Diseño, desarrollo e implementación de un sistema gestor de automatización de tesauros.* [Tesis de licenciatura inédita]. Murcia, España: Universidad de Murcia.

Pastor, J. A. (2009). *Diseño de un sistema colaborativo para la creación y gestión de tesauros en Internet basado en SKOS.* [Tesis de doctorado inédita]. Murcia, España: Universidad de Murcia.

Pastor, J. A. (2015). Los Tesauros en la Web Semántica: SKOS y la norma ISO 25964. Disponible en: https://digitum.um.es/xmlui/bitstream/10201/46824/1/norma-iso-web-semantica.pdf [Fecha de consulta 01/04/2018].

Pastor, J. A.; Martínez, F. J. (2010). Manual de SKOS (Sistema para la organización del conocimiento simple). *Anales de Documentación*, 13, 285-320.

Pérez, M. C. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de Lingüística del Español*, (18), 0-0.

Qiao, B.; Fang, K.; Chen, Y.; Zhu, X. (2017). Building thesaurus-based knowledge graph based on schema layer. *Cluster Computing*, 20(1), 81-91.

Ramírez, G.; Torres, A. (2017). *Mejora de las búsquedas en CMS donde predomina el contenido no estructurado.* [Tesis de Licenciatura inédita]. Buenos Aires, Argentina: Universidad Nacional de La Plata.

Rijsbergen, C. J. (1979). *Information Retrieval.* London, England: Butterworths-Heinemann.

Ruckhaus, E.; (2005). Lógicas Descriptivas y Ontologías. 10 de abril de 2018 https://ldc.usb.ve/~ruckhaus/materias/ci7453/clase51.pdf

Sahami, M.; Dumais, S.; Heckerman, D.; Hovitz, E. (1988) 5 de abril de 2018 A Bayessian approach to filtering junk a-mail. Disponible en: http://robotics.stanford.edu/users/sahami/papersdir/

Salton, G.; M. J. McGillm (1986). *Introduction to modern information retrieval.* Nueva York, United States: McGraw-Hill.

Stein, B. (1987). Methdologies for Documentary process. *Journal of Documentation*, 3(13).

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort. Reading.* Massachusetts, United States: Digital Library of India.