TEMA 6. La indización. Los sistemas de indización

6.1. CONCEPTO Y PROCESO

La indización, como la elaboración de resúmenes, está estrechamente ligada al contenido informativo de los documentos. En el caso de la indización cobra una relevancia más visible la teoría de los conceptos, dado que el objeto de la indización es la identificación, selección y representación de estos.

Así, Hjørland (2001, 2017b) ha analizado las relaciones semánticas de una serie de conceptos próximos con los que se vincula directamente la indización como son: *subject*, *topic*, *theme*, *domain*, *field* y *content*. La indización está conectada especialmente al concepto de *aboutness* que Hjørland considera sinónimo de *subject* y del que destaca la dificultad de su identificación en un documento. Señala que el *subject* de un documento es ese "algo" que la indización y la recuperación tienen que ser capaces de identificar. Se relaciona, por tanto, con las preguntas a que un documento tiene que responder. Dado que un documento puede, en principio, proporcionar respuestas a infinidad de cuestiones, la indización deberá establecer prioridades basadas en los grupos de usuarios a los que se dirige. La mejor indización será la que logre la más acertada prognosis del uso futuro del documento.

Las teorías relativas a la indización giran en torno a la problemática de la subjetividad/objetividad, la interpretación del indizador y la discusión sobre si los *subjects* son inherentes a los documentos o, por el contrario, están vinculados a las necesidades de los usuarios y a la finalidad del servicio de información. Hjørland (2017b) explicita las diferentes posiciones al respecto clasificándolas en cuatro teorías de la indización: racionalismo, empirismo, historicismo y pragmatismo. Atenderemos a las que consideramos más fructíferas que son la primera y la última.

Desde un punto de vista racionalista un documento tendría un número determinado de conceptos esenciales que podrían ser identificados y descritos de manera neutral e independientemente de los distintos puntos de vista o necesidades de los usuarios. La idea de que habría una forma correcta de indizar cada documento estaría asociada con la creencia de que indizadores

expertos conseguirían una alta consistencia en la indización porque existirían ciertas reglas no escritas que guiarían el proceso de indización. Sin embargo, diversos estudios empíricos, como el de Soler y Gil (2011), han demostrado que no es así ya que en la indización rara vez se consigue un nivel de consistencia deseable.

La teoría pragmática enfatiza que la indización no puede ser neutral sino que se hace siempre con un objetivo, teniendo en cuenta las necesidades de los usuarios y del centro o sistema para el que se indiza. La asignación de una materia a un documento es un acto político que contribuye a facilitar ciertos usos de ese documento a expensas de otros posibles. A título de ejemplo de aproximaciones pragmáticas podemos señalar la epistemología feminista como una de esas aproximaciones críticas y a Hope Olson (2002) como su principal representante.

En este sentido, indica Moreiro (2002) que la indización puede ser exhaustiva o selectiva. En opinión de Hjørland (2018) la idea de una indización neutral debe ser abandonada y reemplazada por la de una indización orientada o *slanted indexing* como es denominada por Guimaraes (2017). Para estos autores la subjetividad sería inevitable e, incluso, necesaria ya que la indización no puede estar orientada únicamente al contenido sino que ha de estarlo, igualmente, a las necesidades y/o solicitudes de los usuarios.

Gil Leiva (2008) ha reflexionado sobre el concepto de indización y ha sistematizado las definiciones de los principales estudiosos, al igual que las etapas del proceso de indización establecidas. Recogemos a continuación dos ejemplos que se ubican respectivamente en los dos marcos teóricos mencionados. De un lado la definición clásica de Neet (1989) quien apunta que *"indizar es analizar los documentos y aislar, en la riqueza de la lengua natural empleado por los autores, todos los conceptos esenciales que deben ser retenidos con vista a búsquedas posteriores"*. De otro, la postura de Lancaster (2003) quien defiende que lo principal es dar a un ítem las etiquetas que permitan que sea recuperado por los miembros de la comunidad a la que se dirige el documento.

Señala Hjørland (2018) que el principal criterio de calidad de la indización se vincula con que todos los conceptos relevantes se extraigan de los documentos para poder responder a las consultas de los usuarios. No

obstante, hay que ir más allá. No se trata solo de qué porcentaje de preguntas pueden contestarse de forma satisfactoria sino de qué preguntas se contestan de forma más completa o satisfactoria y cuáles de forma menos completa o satisfactoria. Ello guarda relación con una indización más orientada al documento o al usuario, respectivamente.

No vamos a profundizar en la vinculación indización-recuperación ni en la discusión de la utilidad o no del ítem recuperado para el usuario. Sobre el concepto de relevancia se ha escrito mucho considerándola principalmente desde tres puntos de vista: desde el punto de vista del sistema (o relevancia algorítmica), desde el punto de vista del usuario (o relevancia cognitiva) y desde un punto de vista temático (o relevancia orientada al dominio) (Hjørland, 2010; Schamber, 1994).

Coincidimos sobre la imposibilidad, y no necesariamente conveniencia, de perseguir la neutralidad en la indización aunque en la docencia tratamos de que los estudiantes identifiquen los conceptos mejor desarrollados en el documento y tengan la capacidad de realizarlo al menos por su mayor representación en el texto. Conviene considerar que en el aula no tenemos unos usuarios de referencia a quienes dirigirnos. Tratamos, por tanto, de que sean capaces de identificar, seleccionar y nombrar los conceptos que serían imprescindibles y los que serían posibles en la indización de cada texto. Sin perder de vista que la finalidad es satisfacer a los usuarios realizamos, por tanto, una indización orientada al contenido del documento más que orientada a las posibles solicitudes de información (Hjørland, 2017b).

En otro lugar (Rodríguez Bravo, 1996a), ya reflexionamos sobre las etapas del proceso de indización tomando como punto de partida la norma UNE (1991) y las aportaciones de Rowley (1988). Para abundar en el proceso de la indización recomendamos, asimismo, el trabajo de Mai (2001).

La primera etapa sería la de *Reconocimiento del contenido documental*, en ella se procederá a la lectura del documento o en su caso al visionado o a la audición con las peculiaridades que estas distintas formas de acceso al documento comportan. En el caso de la lectura como forma de acceso al contenido resultan de interés los trabajos de Fujita orientados a la formación de los indizadores (Fujita y Rubi, 2006; Fujita, 2007, Redigola y Fujita, 2015).

La segunda etapa tendría como objetivo *la Identificación de las nociones principales*. A medida que realiza la lectura, el documentalista identifica los conceptos sobre los que trata el documento. Como señala Cunha (1989) en esta fase el indizador tratará de identificar la organización metodológica del discurso del autor a través de la segmentación del texto para, a continuación, aislar los conceptos traductores del contenido de esos segmentos. En este sentido habrá que considerar las propuestas explícitas e implícitas identificadas por el análisis del texto y que traducidas en términos de indización implicarán un mayor o menor grado de acierto en relación al contenido del texto o discurso que será transmitido a los potenciales consumidores.

Si el documento trata de varios temas diferentes habrá que subdividirlo en varias partes (unidades bibliográficas), pudiendo cada una de ellas ser considerada como una publicación independiente. El indizador tendrá que ponerse en el lugar de los usuarios potenciales del documento y determinar el contenido informativo de este mediante la identificación de la idea o de las dos o tres ideas que constituyen la razón esencial de que el documento haya sido publicado, pasando por alto todas las informaciones superfluas, marginales o imprecisas, (de modo que pueda evitarse luego la recuperación de documentos no pertinentes -ruido-), y detectando las informaciones implícitas (de modo que pueda soslayarse luego la no inclusión de documentos pertinentes -silencio-).

La tercera etapa sería la de *Selección de los términos de indización*. Una vez que se han identificado las nociones principales y previa extracción, es necesario ordenarlas basándonos en la observación de las relaciones entre las posibles palabras-clave, de recurrencia, equivalencia, oposición, paralelismo, simetría, inversión, etc. A continuación, se elegirán los términos que mejor representen estas nociones y se extraerán. En esta etapa, indica Lancaster (2003), el indizador debe hacerse varias preguntas acerca del documento: ¿De qué trata?, ¿Por qué se ha añadido a nuestra colección? Y ¿qué aspectos interesan a nuestros usuarios?

6.2. LOS SISTEMAS DE INDIZACIÓN

Los sistemas de indización pueden estar fundamentados en las palabras (indización por unitérminos), en los conceptos (indización por descriptores) o en los temas (indización por materias). Además los sistemas de indización pueden utilizar lenguaje libre (derivado del texto) o lenguaje controlado (asignado a partir de un lenguaje documental, un tesauro o una lista de encabezamientos). Igualmente pueden ser postcoordinados (unitérminos y descriptores) o precoordinados (materias).

Los lenguajes libres, fundados en el principio de postcoordinación, se componen de un vocabulario no predefinido que se va generando a partir de la realización de procesos de indización. De este tipo son las listas de descriptores libres y las listas de palabras clave. Los lenguajes libres no son propiamente lenguajes documentales puesto que para que reciban este nombre el vocabulario ha de estar controlado.

Son lenguajes controlados los demás tipos de lenguajes documentales: tesauros, listas de encabezamientos de materia y clasificaciones. Presentan un vocabulario previamente elaborado, y admiten un limitado número de modificaciones en el momento de su utilización.

Existe abundante literatura acerca de las ventajas e inconvenientes que conlleva el uso del lenguaje libre y del lenguaje controlado. Del análisis comparativo de ambos se suele concluir que uno neutraliza las deficiencias del otro, por ello, muchas bases de datos combinan la utilización de ambos en las distintas fases del tratamiento documental. Según Lancaster, los sistemas con lenguaje natural ofrecen una ventaja sobre los sistemas que utilizan un lenguaje controlado. El uso de un vocabulario ilimitado permite una gran especificidad en la recuperación; es más probable que el sistema con lenguaje libre de mejores resultados en comparación con los sistemas de lenguaje controlado, cuanto más específica tenga que ser la información.

Los vocabularios controlados también tienen ventajas. Un vocabulario controlado tiene tres funciones fundamentales: tiende a reducir las ambigüedades semánticas, a mejorar la consistencia en la representación de la materia y a facilitar la realización de búsquedas amplias. La primera función se consigue diferenciando los distintos significados de los homógrafos, la

segunda mediante el control de los sinónimos y cuasi-sinónimos, y la tercera estableciendo una estructura que una los términos relacionados semánticamente. Existe una relación entre los costes o esfuerzo en el input y el output de los sistemas de recuperación. En los sistemas con lenguaje controlado el coste y el esfuerzo se encuentran en la fase de entrada, mientras que en los sistemas con lenguaje libre los soporta la fase de salida, es decir, la búsqueda en la base de datos. Un usuario experimentado puede desarrollar una estrategia que compense la falta de control del vocabulario en la fase del input. En esencia, utilizará la estrategia de búsqueda para conseguir los mismos resultados que podría proporcionarle un vocabulario controlado. En conclusión, el vocabulario controlado es más práctico: proporciona al usuario un punto de búsqueda, en vez de dos o más, y reduce la posibilidad de que la búsqueda sea incompleta. Sin embargo, puede perderse alguna información.

Los sistemas de indización son principalmente tres: la indización por unitérminos o palabras clave utiliza el lenguaje natural, la indización por descriptores y por materias se fundamenta en el uso de lenguajes documentales –tesauros y listas de encabezamiento respectivamente-. La indización por materias es la correlación sucesiva de diferentes encabezamientos que expresan el tema o temas de un documento, es por tanto una indización precoordinada, es decir se produce la coordinación en el momento del almacenamiento y su principal ventaja es que prácticamente no da cabida a falsas combinaciones entre los términos, ya que cada cual ocupa su posición. Se utiliza prioritariamente en los catálogos de bibliotecas. En la indización por unitérminos o palabras clave y en la indización por descriptores la coordinación se produce en el momento de la recuperación y por ello se llaman sistemas de indización postcoordinados. El sistema de indización por descriptores evita la posible ambigüedad, porque no se basa ya en las palabras sino en los conceptos, y para evitar las falsas combinaciones precoordina los términos cuando es necesario. La utilización de la indización controlada por tesauros –indización por descriptores- es de uso general en bases de datos y centros de documentación. La indización por palabras clave es de uso habitual en los motores de búsqueda y complementariamente en bases de datos y catálogos de bibliotecas donde se recuperan los documentos

a partir de unitérminos del título, el resumen o los descriptores y encabezamientos de materia. El proceso de indización a la salida es el mismo que a la entrada: extracción de los conceptos de la demanda del usuario y luego traducción de estos conceptos a términos del lenguaje documental. Pero cuando se trata de la indización de las preguntas o cuestiones, en la indización postcoordinada, se añade una etapa suplementaria: la formulación de la pregunta bajo la forma de ecuación lógica. Como dice Neet (1989:121), con la indización postcoordinada se afirma una tendencia nueva: el esfuerzo de síntesis se ha desplazado hacia la búsqueda. El indizador se ocupa del análisis, el que busca de la síntesis. La búsqueda será combinatoria

EN RESUMEN:

Existen diversos sistemas de indización puesto que el proceso puede estar fundamentado en las palabras (indización por unitérminos), en los conceptos (indización por descriptores) o en los temas (indización por materias). Además, los sistemas de indización pueden utilizar lenguaje libre (derivado del texto) o lenguaje controlado (asignado a partir de un lenguaje documental, un tesauro o una lista de encabezamientos, por ejemplo). Igualmente, los SOC pueden ser postcoordinados (unitérminos y descriptores) o precoordinados (materias y clasificaciones). Los principales criterios para establecer la sistematización de los sistemas de indización aparecen recogidos en Rodríguez Bravo (2011a).

Veamos un ejemplo de los principales tipos de indización:

La **indización por unitérminos** se fundamenta en las palabras. Cada palabra con significado constituirá un término de indización. Se trata de una indización postcoordinada, la combinación de términos no se realiza en el momento de la indización sino cuando se procede a la búsqueda.

La **indización por conceptos** se basará generalmente en los descriptores y para ello se utilizará un Tesauro, es decir un vocabulario controlado. Se trata asimismo de una indización postcoordinada.

La **indización por materias** implica la coordinación de conceptos para conformar un tema en el momento de la indización. Es, por tanto un sistema

precoordinado y utiliza en su formulación, generalmente, un vocabulario controlado como son las Listas de Encabezamientos de Materia.

Ejemplo. Pensemos en un documento que trata de la "automatización de las bibliotecas universitarias en España".

Según los tres tipos de indización que hemos observado la solución sería:

Indización por unitérminos o palabras:

Automatización, Bibliotecas, España, Universidades (utilizo el orden alfabético pero el orden es indiferente. Eso sí se utilizarían sustantivos).

Indización por conceptos:

Automatización, Bibliotecas universitarias, España (utilizo el orden alfabético pero el orden es indiferente. Si el concepto es compuesto ya no será un unitérmino sino sintagmático. En este caso sustantivo+adjetivo).

Indización por temas o materias:

Bibliotecas universitarias-Automatización-España (aquí los conceptos que conforman el tema se introducen en un orden determinado y no es indiferente. Generalmente cuando hay dos conceptos temáticos el más específico va delante del más genérico y a continuación suelen ir las nociones que indican lugar, tiempo y forma).

La indización por materias es la más compleja habida cuenta de que no se pide solo que se identifiquen los conceptos principales del documento sino que se ordenen. Su ventaja es que ese mismo orden permite eliminar las falsas combinaciones en la recuperación.

Filosofía-Historia (historia de la filosofía)

Historia-Filosofía (filosofía de la historia)

El sistema que más problemas de ambigüedad produce es la indización por palabras. No siempre es fácil expresar conceptos con sustantivos unitérminos. Pensemos en "bibliotecas públicas" o "bibliotecas nacionales".

En las prácticas de esta asignatura vamos a tratar de realizar una indización en vocabulario libre, es decir, directamente derivada del texto, y basada en los conceptos, por tanto, postcoordinada. Es decir,

vamos a tratar de identificar y extraer los principales conceptos de los documentos.

No vamos a utilizar un tesauro, pero sí vamos a realizar una mínima normalización de esos conceptos. Si son conceptos contables irán en plural y si son incontables, abstractos, disciplinas, etc. En singular. Cuando haya opción utilizaremos el masculino.