
El tesoro, el tesoro documental y el tesoro conceptual

Es bueno, recordar aquí que la palabra tesoro usada en la acepción de tesoro documental como se explicará más adelante surge en la bibliotecología cuando se comienzan a procesar grandes volúmenes de información especializada por medio de computadoras.

Un tesoro es una herramienta documental utilizada en el ámbito de la indización y recuperación de información, que representa un dominio de conocimiento determinado mediante una estructuración conceptual. La estructura conceptual proporciona una organización semántica mediante la explicitación de las relaciones conceptuales y de la restricción del significado de los términos que los representan.

En cuanto a la primera vez que el *tesoro* fue utilizado como un tipo de instrumento lingüístico en la indización de los documentos, Chaumier presenta en su libro la opinión de Vickery y Lancaster. Según Vickery, fue Helen Brownson de la American National Science Foundation, la primera que lo empleó en la: "Dorking Conference on Classification Research" (1957), presentando la "aplicación de un tesoro mecanizado basado en redes de significados relacionados". En cambio, Lancaster le otorga el mérito a Hans Peter Luhn, quien lo empleó en un texto del mismo año 1957.¹

Hans Peter Luhn fue científico de la computación para IBM, creador del algoritmo de Luhn, del concepto de Diseminación Selectiva de Información, de los resúmenes automáticos basados en la frecuencia estadística de términos y del sistema KWIC (Key Words In Context). Este emigrante alemán en Estados Unidos, hijo de impresores trabajó en la *División de Investigación sobre Recuperación de Información en IBM*.² La trascendencia del trabajo de Luhn es difícil de opacar, pero en todo caso Helen Brownson adoptó y aplicó el concepto de tesoro con mucha celeridad.

La primera noticia sobre la palabra *tesoro* proviene del mundo helénico y se usaba para designar a las pequeñas capillas donde se guardaban los exvotos y las donaciones que frecuentemente eran muy ricas y valiosas, verdaderas joyas. Estas capillas estaban enclavadas en el recinto del santuario o témenos donde se encontraba el oráculo. El único tesoro restaurado que ha llegado a nuestros días pertenecía al templo de Apolo y la descripción bastante exacta de cómo fue el recinto sagrado se conoce gracias a las informaciones de Pausanias en el siglo II AC.

1 Jacques Chaumier (1986) *Análisis y lenguajes documentales: el tratamiento lingüístico de la información documental*. p. 11

2 Wikipedia: the free encyclopedia (Consultada el 12 de septiembre de 2006) http://en.wikipedia.org/wiki/Hans_Peter_Luhn

Posteriormente se usó para designar ciertos diccionarios, índices o repertorios de términos³, con la misma connotación de lugar donde se guarda lo más importante, que en este caso es menos corpóreo porque se trata del saber humano. A lo largo de la historia, el término tesoro está asociado a obras de estructuras muy variadas tanto diccionarios como repertorios terminológicos o vocabularios. De hecho, la voz *thesaurus* en latín fue, durante años, el término más utilizado para designar tanto un diccionario muy exhaustivo (de lenguas muertas) como un vocabulario que contuviera todas sus unidades léxicas:

En la actualidad, la voz de *tesoro* en el ámbito bibliotecológico se usa en el sentido de redes de significados relacionados, tal como lo planteó Helen Brownson en 1957, aunque sigue apareciendo en algunos diccionarios y enciclopedias usado con el valor de diccionario, catálogo o antología⁴. El *Oxford English Dictionary* define el *tesoro* como "una lista clasificada de términos, especialmente términos llave, en un campo en particular, para usar en la indización y recuperación de información"⁵

En 1969 Aitchison publica el *Tesoro facetado* que muestra las ventajas de indizar y clasificar con una correspondencia entre facetes y descriptores (Gilchrist, 2003: 9) En un principio los sistemas de indización y recuperación que utilizaban el tesoro trabajaban con tarjetas perforadas en lotes, pero luego se aplicó a sistemas de recuperación en línea, con éxito. La permanencia de la herramienta determinó la propuesta de normas para construirlos (*International Organization for Standardization 2788*, 1986). Incluso en el año 2000 Aitchison y Gilchrist publican un manual para elaborar tesoros. La National Information Standards Organization en el año 2005 publica las "Guidelines for the construction, format and management of monolingual controlled vocabularies". Estas últimas normas agregan las taxonomías como herramientas a ser desarrolladas con su apoyo y sobre todo involucra la organización de información electrónica.

En el campo de la literatura el tesoro es una lista de palabras con significados similares sinónimos, habitualmente acompañada por otra lista de antónimos. Un ejemplo sería un tesoro dedicado a un campo especializado, que contiene la jerga que se emplea en dicho campo del conocimiento. En el mundo de habla inglesa, es clásico el Tesoro de Roget's cuya función es, según su autor, además de ayudar al escritor a encontrar la

3 J. Martínez de Souza (1993) *Diccionario de bibliología y ciencias afines*. – p. 776

4 *El pequeño Larousse ilustrado* 1996 en color. -- México: Ediciones Larousse. -- 1792 p.

5 The Oxford English Dictionary. Vol 17, p. 924

El tesoro, el tesoro documental y el tesoro conceptual

palabra que exprese mejor su pensamiento, también estimular su intelecto y sugerirle palabras o ideas relacionadas.⁶ En cierto modo, en el tesoro que usan los documentalistas y en el Tesoro de Roget existen ciertos objetivos similares, pero como se explicará en el siguiente apartado, el primero difiere en el uso y en la presentación.

El término tesoro en español es muy usado en diccionarios, por esta razón se propone el uso del término tesoro documental como se indica en el nombre del trabajo que como término compuesto evita la polisemia que lo acompaña.

El tesoro ha ido evolucionando como una herramienta de recuperación de información inestimable en los procesos de organización del conocimiento. Las designaciones terminológicas usadas para definir esa evolución no reflejan las verdaderas diferencias. En este caso la informática que ha sido la disciplina que ha otorgado una funcionalidad diferente al tesoro ha impuesto su terminología, creando gran confusión. En general, se han difundido dos designaciones: tesauros terminológicos y tesauros conceptuales. Se plantearía así que con la primera designación los tesauros no trabajarían con conceptos. Nada más alejado de la realidad porque en la organización del conocimiento se ha buscado desde un principio, la representación del contenido documental mediante conceptos y no los términos presentes en los documentos. A los fines de este trabajo se utilizará el término tesauros documentales, en lugar de tesauros terminológicos.

No se deja de reconocer el beneficio que ha otorgado a los sistemas de información el manejo tecnológico del tesoro documental. Su desarrollo se ha ido adecuando paulatinamente a los cambios que las tecnologías de la información suponen, aunque su aplicación en la Web ha resultado hasta ahora poco fructífera. Muchos trabajos están estudiando la relación de la semántica y las tecnologías, principalmente las representaciones del espacio semántico.

El espacio semántico es el medio en el que se involucran mínimo dos personas, las cuales pueden tener diferentes percepciones ya sea de lengua, cultura, posición política, relación familiar, nivel educativo, entre otros, por lo que es un espacio de intercambio de valores y significados, que inciden en las conductas informacionales. Por lo tanto es muy difícil transmitir el mensaje con claridad y precisión debido a que en el proceso del emisor al receptor pueden existir muchas interferencias que hagan más complicada y un poco distorsionada la comunicación. Desde una perspectiva pragmática, en la actualidad encontramos múltiples modelos y

propuestas en que se aplican estos principios. Como muestra está la recuperación de información en Internet a partir de motores de búsqueda. En este caso, la inferencia en un sistema y un espacio semántico consistirían en vincular la información que se encuentra latente en la red y en las bases de datos con las necesidades de información del usuario de manera que se logre establecer la mayor precisión y relevancia en la búsqueda realizada. A esto se le denomina indización semántica latente (Erk, 2009: 106).

La indización persigue la estructuración conceptual de determinadas porciones de la realidad. Ahora bien, la existencia del tesoro promueve la propuesta de una estructura conceptual en un nivel de detalle que recuerde al indizador la posición que deberá asignar a un determinado descriptor documental en la disciplina, pero hasta que no fueron insertados en los sistemas de información como elementos informáticos validados, no ofrecía grandes posibilidades de interpretación y navegación por los descriptores de la disciplina a organizar al momento de la recuperación. La informática obliga a un desarrollo semántico de mayor profundidad puesto que proporcionan una descripción lógica y formal de la información que almacenan, y por lo tanto puede ser interpretada tanto por usuarios humanos como por programas informáticos.

La Web semántica supone una oportunidad para utilizar los tesauros en los procesos de descripción y recuperación de información, para muestra tenemos las múltiples iniciativas de representación de tesauros mediante RDF (Resource Description Framework), que han dado lugar al desarrollo de SKOS (Simple Knowledge Information System), un modelo para esquemas conceptuales aplicables en la Web semántica y que actualmente se encuentra en desarrollo por el World Wide Web Consortium (W3C). La organización del conocimiento más novedosa tiene por objeto combinar diferentes sistemas de organización del conocimiento en Internet mediante metadatos compartidos y ontologías formales que funcionan a partir de los principios de la lógica deductiva.

Es necesario puntualizar que el desarrollo de un vocabulario RDF/XML para representar tesauros supone explotar de manera innovadora el modelo canónico que nos proponen las normas: internacional, norteamericana, británica y la IFLA (ISO 2788:1986, la ANSI/NISO Z39.19, las BS 8723 y las IFLA Guidelines for Multilingual Thesauri) sobre construcción y mantenimiento de tesauros. Estas normas describen un tesoro como un conjunto de términos de diverso tipo entre los que se establecen relaciones semánticas, en un contexto que le otorga sentido. No obstante, la Web precisa de una visión en la que se comparta información, como por ejemplo que las jerarquías incluyan instancias, atributos y meronimias, así como una mayor riqueza en las asociaciones asociativas.

6 Roget's Thesaurus. p. xviii

El tesoro, el tesoro documental y el tesoro conceptual

Existen, por lo tanto, múltiples iniciativas dirigidas a la definición de vocabularios bajo esquemas de RDF para la representación de tesauros.

Lo anterior significa como ya se explicó: dos modelos básicos para la estructuración de tesauros: uno documental de relaciones en un espacio semántico (el clásico) y otros denominados tesauros conceptuales porque existe “una navegación directamente en el documento, en una red semántica conceptual por la que se ingresa de términos más genéricos de una faceta hacia los más específicos, e inversamente (navegación vertical), que a su vez permite la transición de una clase hacia otra y de un campo de la ciencia hacia otro mediante las relaciones asociativas (navegación horizontal) por nudos polijerárquicos basados en conceptos” (Deweze, 1998).

En este segundo tipo de tesauros que son de carácter automatizado en la construcción y en la recuperación de información se aprecia un incremento pragmático de las relaciones en especial las de asociación, donde ahora se reflejan también ideas combinadas que se vinculan entre sí con una red de relaciones semánticas de distinto tipo, principalmente jerárquicas (término), asociativas (término relacionado) y de equivalencia (término equivalente, término preferente) (Tudhope, Harith & Jones, 2001). En esta segunda alternativa se concibe al tesoro como un conjunto de conceptos entre los que se establecen relaciones semánticas similares a las que se establecen en los tesauros documentales. Los conceptos se vinculan con términos, a modo de etiquetas léxicas, entre los que pueden o no establecer relaciones léxicas que es la propuesta del modelo SKOS.

Los tesauros documentales posibilitan la representación de esquemas más compactos y cercanos al modelo tradicional, pero los tesauros basados en conceptos relacionados automáticamente permiten navegar en las estructuras de información, aunque en un principio su

comprensión no resulte tan intuitiva como los primeros. (Cfr. Jones, 1993). El cambio de paradigma en la modelización de tesauros supone un gran esfuerzo aunque dicho cambio se ve compensado con creces al optimizarse los procesos de elaboración, mantenimiento y productos derivados de la aplicación del tesoro en la indización de documentos (Pastor Sánchez, Martínez Méndez & Rodríguez Muñoz, 2009). La principal razón para este cambio se debe a que los tesauros documentales conllevan problemas de aplicación en procesos de indización cuando se realizan operaciones de cambio de términos preferentes y términos no preferentes. De esta manera, cuando un descriptor con el que se ha indizado un documento pasa a ser un no descriptor hay que realizar una reorganización de los índices que asocian dicho descriptor con los documentos.

La integración del tesoro conceptual permite no sólo definir las categorías conceptuales, sino que se indique la posición en la jerarquía y relaciones conceptuales y que se especifiquen sus propiedades y también toda la información que hereda de los conceptos jerárquicamente superiores. Se le puede además asociar una base de datos terminológica, con variantes y equivalentes lingüísticos para los términos relacionados a cada concepto. Además todos estos parámetros de información especificados (definición, categorización, jerarquía, propiedades y herencia) son codificados de manera formal y explícita, o sea que no solamente serán entendibles mediante la lectura por parte de un humano, sino que pueden ser interpretados o descodificados por un programa informático y por lo tanto "manipulables" automáticamente.

Catalina Naumis Peña
Ariel Antonio Morán Reyes