

La recuperación de la información: ¿Lenguaje natural vs. Lenguaje controlado?

Susana Soto, M.A., Ph.D.

Directora Sistema de Bibliotecas Universidad Abierta Interamericana

Lo primero que llamó mi atención cuando me invitaron para que hablara sobre el tema fue la sintaxis que utilizaron “*Lenguaje natural vs. Lenguaje controlado*”. Siempre en términos de opuestos presentados como irreconciliables.

Entonces releí uno de mis primeros artículos sobre Clasificación¹ que justamente describía el escenario de la Clasificación -entendida ésta en un sentido amplio- en términos de opuestos en conflicto.

Su lectura me llevó a recordar todos los planteos que se habían sucedido a lo largo del Siglo XX, no solamente en nuestro país sino también en el extranjero.

La clasificación sistemática vs. La clasificación alfabética

Este conflicto, que aun estaba vigente cuando estudiaba mis primeras materias de Bibliotecología, dividió a los bibliotecarios entre los defensores del catálogo sistemático y los defensores del catálogo alfabético de materia. Recitar las ventajas y desventajas de uno y otro catálogo era un clásico de los parciales y finales cuando empezaba a cursar mis primeras materias en la facultad.

Indización precoordinada vs. Indización poscoordinada

En los albores de la automatización, fue la batalla central. Posiblemente de todos los conflictos, éste haya sido el más vano. Un experimento realizado en la escuela de bibliotecología de Aberystwyth demostró que ambos sistemas si estaban bien aplicados y bien utilizados tenían capacidades de recuperación similares. Por otra parte, en un entorno manual, la indización poscoordinada era y es totalmente impráctica. Un catálogo en fichas, una bibliografía impresa necesitaban un buen sistema de indización precoordinada.

Lenguaje natural vs. Lenguaje controlado

El advenimiento de los tesauros trajo esta nueva discusión entre los defensores del lenguaje tal cual aparecía en el documento y los que abogaban por la normalización y sistematización de los descriptores.

Estos conflictos generaron pasiones conexas.

Los defensores del catálogo alfabético de materia abjuraron de los sistemas de clasificación, a los que relegaron en el mejor de los casos para la ubicación en el estante

por materia. Hasta que un buen día se encontraron que más de media docena de bibliografías nacionales indizaban según la Clasificación Decimal de Dewey.

Mientras tanto algunos amantes del catálogo sistemático se perdieron en el virtuosismo de la clasificación construyendo números de clase cada vez más largos y complejos, que volvían a su amado catálogo de materia altamente impopular no solo con los desamparados usuarios sino también con los colegas menos expertos que aun no habían alcanzado su mismo virtuosismo.

Los fieles de la indización precoordinada con lenguaje natural se aferraron a sus listas de encabezamiento de materia argumentando que era lo más fácil y simple para el usuario. Sin embargo, las cosas no iban a quedar simples para el usuario. En la batalla por la supervivencia, algunas listas empezaron a agrupar los encabezamientos por categorías temáticas y a dibujar árboles de relaciones, hasta que finalmente solo están disponibles como tesauros en - línea y por Internet, siempre para que sea más fácil y rápido encontrar el encabezamiento correcto.

Por su parte los tempranos defensores de la indización poscoordinada se lanzaron a la desenfadada pasión de construir tesauros. En nuestro medio esta pasión tuvo efectos muy peculiares, en una década los tesauros se adueñaron de la práctica profesional, al punto tal que algunos colegas consideraron un desmedro no indizar con tesoro. Cuantos más tesauros se usaban para clasificar un documento mejor. Otros dejando de lado la gestión total de la biblioteca se abocaron a la construcción del tesoro propio. Jamás se dignaron a examinar la relación costo-beneficio. Más aún, eliminaron de la práctica profesional a las clasificaciones sistemáticas como entelequias del pasado. Esto último fue un fenómeno típico de nuestro país, que no coincidió con el resurgimiento de tales clasificaciones en el escenario del control bibliográfico internacional.

Entre conflicto y conflicto se desgranaron una a una las décadas del siglo XX y hacia sus finales otras preocupaciones como la automatización e implementación de los OPAC absorbió totalmente a los responsables de Procesos Técnicos.

Cuando todo parecía aquietarse, la irrupción de Internet desató un nuevo conflicto justo a tiempo para empezar otro milenio:

Indización humana vs. Indización automática

¿Pueden los robots reemplazar el análisis temático hecho por expertos?

¿Es inútil seguir estudiando clasificación y todavía más inútil clasificar para nuestros catálogos en línea?

Ahora están en tela de juicio todos los sistemas de clasificación sistemáticos o no, pre o pos coordinados, listas de encabezamientos o tesauros. Está en tela de juicio una parte substancial de nuestro quehacer profesional.

Si miramos con desenfado los conflictos de la clasificación en el siglo XX alcanzaremos la suficiente objetividad para sacar conclusiones que esclarezcan la controversia actual.

En primer lugar todos estos conflictos han olvidado lo que una colega mexicana llamó la infodiversidad para indicar no solo que la información es compleja, sino que además al igual que en el caso de los organismos vivos, esa complejidad es esencial. Si se altera la biodiversidad de un ecosistema se pone en peligro la supervivencia de todos sus miembros. De la misma manera si se elimina la diversidad y la complejidad de la información deja de ser información.

Si la información es esencial y necesariamente compleja no puede esperarse que el acceso temático a la información sea uno solo.

La única manera de administrar la diversidad es con alternativas. Los conflictos que enumeré antes fueron y son alternativas del acceso a la información que pueden complementarse unas con otras.

El desarrollo del catálogo alfabético de materia en las bibliotecas norteamericanas fue de la mano con la popularidad de una clasificación sistemática, Dewey, que se usaba para ubicar la colección en acceso libre al estante.

Cuando un usuario tenía que buscar algo sobre un tema desconocido y se cansaba de seguir los “véase” y los “véase además” de gaveta en gaveta, dejaba el fichero y se iba a pasear entre las estanterías prolijamente arregladas en clases de lo general a lo particular. Encontraba una organización de la información que lo orientaba en un área poco conocida mucho más rápido que a través del lenguaje natural del catálogo.

En cambio, los catálogos sistemáticos de materia florecieron en las bibliotecas europeas sobre todo antes de la Segunda Guerra Mundial, cuando el acceso libre al estante era anatema. Si el estante no era una alternativa, la única posibilidad de ofrecer una organización sistemática era a través del catálogo. En el entorno manual era prácticamente imposible generar los dos tipos de catálogo de materia. Cuando en la década del 70 aparecieron en el Reino Unido, los primeros programas de automatización y catalogación cooperativa todos incluyeron encabezamientos de materia y números de clase para que cada biblioteca organizara el acceso temático como quisiera.

La automatización de los catálogos nos dio más alternativas de acceso a la información, en teoría podemos buscar por cada campo de la base de datos, pero fundamentalmente enriqueció el acceso temático porque permitió combinar sistemas diferentes de clasificación en una misma base.

La pre y pos coordinación también deben interpretarse como mutuamente complementarias y no como enemigas.

Ya dijimos que en el entorno manual, los sistemas poscoordinados eran imprácticos. Todos esos sistemas que estudiábamos en Documentación allá por los años 70, los Uniterm, las fichas con muesca, eran de relativa utilidad en una biblioteca

especializada pequeña, pero con ellos no se podían administrar grandes colecciones de documentos especializadas o no.

El advenimiento de las bases de datos con acceso en línea fue lo que dio sentido a los lenguajes de indización poscoordinados. Las búsquedas booleanas para ser efectivas necesitan de descriptores simples con bajo nivel de precoordinación.

Sin embargo esto no iba a durar mucho tiempo, el crecimiento imparable de las grandes bases de datos bibliográficas como Medline, Agris, Eric y todas las demás que siguieron nos ofrecieron acceso a millones de registros antes de terminar los 80. Si Medline se aferrara a la teoría pura de construcción de tesauros y no hiciera precoordinación en los MeSH no terminaríamos más de combinar descriptores con operadores booleanos y menos aun de verificar los resultados de las búsquedas.

Por último el entorno electrónico también permite superar la disputa entre el lenguaje natural libre y el lenguaje natural controlado, porque una base de datos bibliográficas bien diseñada y bien administrada tiene que permitir ambos accesos, es decir palabras claves en título, resumen, etc. y descriptores o encabezamientos de materia.

Un número de clase, un encabezamiento de materia, un descriptor, palabras del lenguaje natural, todas son alternativas válidas para recuperar información. Su efectividad y eficacia dependen de su correcta aplicación y de su acertada elección en el momento de la búsqueda. Algo que casi siempre se soslayó cuando se defendía una u otra alternativa.

Hace más o menos un mes tuve que preparar una clase de entrenamiento de un producto electrónico para médicos y utilicé como hilo conductor de la sesión una búsqueda que uno de ellos me había pedido. El tema era "*Criterios de admisión a las escuelas de medicina*"

La clase consistió en mostrarles como buscar en Medline, CINAHL, AMED y en una base de texto completo. Las tres primeras son bases de datos bibliográficas. La única manera de recuperar un conjunto razonable de documentos en Medline es utilizando un descriptor precoordinado, en 3.000.000 de registros no puedo poscoordinar conceptos tan amplios como "admisión", "criterios" y "escuelas". Se utilizaron descriptores precoordinados en las tres bases de datos, en las tres eran diferentes, aunque querían decir lo mismo, las tres arrojaron resultados diferentes que se superponían parcialmente.

Cuando llegó el turno de la base datos de texto completo, la búsqueda temática es solamente por lenguaje natural libre. En este caso la búsqueda arrojó resultados mucho más altos. Si en Medline habíamos encontrado alrededor de 74 citas, en la base de texto completo encontramos casi 200 citas. La cantidad de registros no relevantes a simple vista en este último grupo era alta, pero aun así había citas que no se habían encontrado en ninguna de las bases anteriores.

¿Cuál es la mejor? Todas las bases son buenas, lo ideal es explorar las cuatro bases de datos y comparar resultados. Pero lo importante es que el usuario tenga las alternativas y elija de acuerdo con su tiempo, sus ganas y sus necesidades. Lo importante

es que nosotros conozcamos las alternativas y se las sepamos ofrecer y explicar a los usuarios.

Si en vez de plantearnos pares en oposición, nos manejamos con alternativas que se complementan estaremos excelentemente preparados para ganar posiciones en la batalla por la indización de la web.

Los bibliotecarios sabemos por experiencia que un solo sistema no cubre todas las necesidades de acceso a la información, mucho antes de soñar con una base datos, construimos catálogos e índices con triple acceso: autor, título y materia ficha por ficha.

También sabemos que el lenguaje natural es ambiguo y cambia de acuerdo con el contexto y el punto de vista de la especialidad. Conocemos las ventajas y desventajas de la organización jerárquica y la búsqueda al azar.

Estamos entrenados para describir documentos por características extrínsecas e intrínsecas.

Tenemos experiencia en encontrar libros raros y revistas poco conocidas y en identificar exitosamente pedidos de compra a partir de datos incompletos e incorrectos.

En suma estamos perfectamente equipados para catalogar la web y no porque seamos mejores o peores que los robots de búsqueda, sino porque los buscadores dan ciertas alternativas utilizando diferentes formas de indización automática y nosotros podemos ofrecer otras alternativas de recuperación utilizando “indización asignada”.

Ni nuestros sistemas, ni las máquinas buscadoras van a poder abarcar la totalidad de la web. La inmensidad de los recursos disponibles en Internet es tal, la diversidad de razones por la que se busca es tal, que hay trabajo de sobra para bibliotecarios e informáticos.

No voy a analizar cada tipo de buscadorⁱⁱ, pero sí voy a referirme brevemente a los problemas que plantean los buscadores y las soluciones que podemos aportar los bibliotecarios.

En primer lugar Internet es caótica, pero esto es parte de su esencia. Internet derivó de un proyecto del Pentágono para sobrevivir al caos que ocurriría tras un ataque nuclear. El caos se maneja con caos. Si Internet fuera totalmente estructurada y ordenada dejaría de ser lo que es y no hubiera traído ningún cambio a nuestras vidas.

Segundo los buscadores tienen habilidades limitadas para presentarnos los resultados. Cuanto más funcionan en automático más multitud de datos recogen y nos lanzan. No siempre nos dejan limitar la búsqueda para acotar los resultados, Yahoo fue uno de los primeros que empezó a categorizar las páginas web y esto se hace a mano.

Tercero hay gran cantidad de páginas web sin indizar por ningún buscador, las páginas dinámicas resultado de una búsqueda en una base de datos, los catálogos, las páginas sin vínculos, los documentos en formato PDFⁱⁱⁱ.

Cuarto, los buscadores no son claros a la hora de explicar cómo indizan, qué indizan y hasta dónde indizan. En consecuencia nunca está claro si no encuentro algo porque no está, porque es una limitación del buscador, o si es realmente porque no existe.

Quinto, las búsquedas simples, aquellas que puedo reducir a un concepto expresado en un término son las que tienen mejores posibilidades, pero aun así los resultados son azarosos. Utilicé toda la semana un mismo buscador para localizar los sitios web de editoriales jurídicas. En algunos casos encontré la editorial al tope de la lista sin demoras, en otros casos recuperé tal cantidad de citas que tuve que abandonar. En un caso en vez de darme el sitio web de la editorial, el buscador había levantado un vínculo de la página web de una universidad, con lo cual para llegar a la editorial tuve que entrar por la página de la universidad con la consiguiente pérdida de tiempo. Y por último ayer a la noche encontré el sitio web de una importante editorial jurídica viajando en subte. Todo el vagón estaba tapizado con propaganda de la editorial, y ahí bien grande aparecía el sitio web, pero el buscador no lo tenía.

¿Qué podemos ofrecer nosotros como contrapartida?

Podemos catalogar los recursos de la web desde el punto de vista, necesidades e intereses de nuestros usuarios utilizando nuestras tradicionales herramientas de catalogación y clasificación.

Podemos aplicar nuestros criterios de selección y desarrollo de colecciones a la exploración sistemática de la web para levantar aquellos recursos que nuestros usuarios necesitan.

Podemos aplicar nuestros conocimientos de Referencia para evaluar, describir y explicar los recursos disponibles en Internet.

Así como un día diseñamos nuestras propias bases de datos para automatizar los catálogos, ahora podemos diseñar nuestras propias bases de datos para sistematizar los recursos de Internet.

Y cuando lo hagamos no dejemos ninguna alternativa afuera. Podemos establecer estándares para la construcción de edificios, el amoblamiento, el desarrollo de colecciones y hasta para la descripción catalográfica, pero nunca vamos a construir un sistema de indización o de clasificación único que funcione como un auténtico standard, es decir que sirva de la misma manera en cualquier situación.

ⁱ Soto, Susana. Los conflictos de la Clasificación. (En: *Boletín Bibliotecológico de La Plata*, 1981/82, (2) pp. 7-9)

ⁱⁱ Denenberg, Ray. Structuring and Indexing the Internet. [copia bajada de Internet sin datos del URL, 21/10/1999 15:34]

ⁱⁱⁱ Pedley, Paul. The Invisible Web. (En: *The Library Association Record*, 2000 v102 (11) pp.628-33)