

Forma, significado y estructura de los sistemas de selección del conocimiento

(ISKO99: Lyon, Francia, Octubre 21-22, 1999)

Professor Michael BUCKLAND
School of Information Management and Systems.
University of California, Berkeley, Ca USA

Traducción del francés:
José Luis DEL RÍO SADORNIL

«Una bibliografía contemporánea preocupada por la claridad, ha lanzado esta breve definición: un documento es una prueba en apoyo de un hecho» (Suzanne Briet, 1951, 7).

Señoras, señores, vuestra invitación a este Congreso del Capítulo Francés de ISKO99 es un gran honor para mí, que os agradezco mucho. Tengo muy gratos recuerdos de mis estancias en Francia en mi juventud y debo reconocer la influencia en mi vida profesional de las obras de algunos documentalistas francófonos, sobre todo de Suzanne Briet y Paul Otlet. Se me ha invitado a dar una charla sobre los orígenes, el desarrollo histórico, el estado actual y el futuro de la indización. En lugar de tal discurso que, evidentemente, sería demasiado ambicioso por mi parte, he querido presentar algunas ideas sobre los sistemas de selección. ¿Qué fenómenos interesan a los investigadores en sistemas de organización del conocimiento? ¿Qué se puede decir sobre la estructura de tales sistemas? ¿Cómo pueden los usuarios de tales sistemas comprender lo que los datos, los documentos y los metadocumentos significan? ¿Cómo podemos juntar bajo una sola perspectiva la tecnología y el significado?. En pocas palabras, ¿cómo debemos caracterizar los sistemas de organización del conocimiento?. Las observaciones que voy a presentar aquí provienen de ideas desarrolladas con mis colegas de Berkeley.

Mi título es *«Forma, Significado y Estructura de los Sistemas de Selección del Conocimiento»*.

La palabra *«Forma»* se refiere a los fenómenos de interés o de utilidad (*information-as-thing*).

La palabra *«Significado»* es susceptible de designar, unas veces, el hacer (la significación como proceso – *information-as-process*), y otras, el estado (lo que quiere decir – *information-as-knowledge*).

Y utilizo *«Estructura»* para referirme al ingenio, el sistema y el saber hacer del documentalista.

1. ¿DE QUÉ NOS VAMOS A OCUPAR?

Los miembros de ISKO se interesan por la estructura del conocimiento. Sin embargo, cuando utilizamos no importa qué clase de tecnología para desarrollar sistemas operacionales, no nos preocupamos directamente más de los conceptos abstractos, que de los datos, del texto o de otros temas concretos. La tecnología es necesariamente material. En consecuencia, nos ocupamos de documentos, pero de documentos no importa cuál sea su forma. Los documentos no son solamente consecuencias del texto.

Hablar de «documento» de este modo, no es nada original. En 1937, el Instituto Internacional de Cooperación Intelectual, una organización creada por la Sociedad de Naciones, colaboró con la Unión Francesa de Organismos de Documentación (UFOD), a fin de definir los términos técnicos «documento» y lo que comprende:

Documento: todo soporte de conocimiento, fijado materialmente, susceptible de ser utilizado para consulta, estudio o prueba. Ejemplos: manuscritos, impresos, representaciones gráficas o figuradas, objetos de colecciones, etc. (Anón. 1937: 234).

2. LA INDIZACIÓN

Suzanne Briet (1894-1989), bibliotecaria, documentalista e historiadora, adelantó ya el concepto de «documento» en el 1951 en su interesante manifiesto *Qu'est-ce que la documentation?*. En él afirma desde el principio que «un documento es una prueba en apoyo de un hecho» (Briet, 1951, 7). Y enseguida explica que un documento es:

«...todo índice concreto o simbólico, conservado o registrado, con el fin de representar, de reconstruir o de probar un fenómeno ya físico, ya intelectual» (Briet, 1951, 7).

En consecuencia, no se puede considerar que la profesión de documentalista (*information manager*) está en ocuparse de textos, sino más bien, de toda especie de prueba, de testimonio, de evidencia, y de esta prueba que son los documentos, de forma concreta y no abstracta. Observemos que Briet ha empleado la palabra «índice». A mi parecer, la palabra «índice» quiere decir que un objeto no llega a ser una prueba (un documento) más que si se ha incluido este objeto en la relación de las otras pruebas (los otros documentos). Es decir, que los documentos deben ser recogidos de forma indizada, los unos con los otros.

Una aproximación más contemporánea sería decir que el sentido está consruuido por el espectador. Que todo objeto podría, en ciertas situaciones, ser prueba y ser un documento. Porque todo objeto es, en potencia, un significante y todo objeto concreto puede ser un documento. Esto es posible, pero tam-

bién poco verosímil. A pesar de todo, nosotros nos quedamos con dos suposiciones de Briet: Que todo objeto puede ser un documento; y que la esencia de la documentación está en incorporar voluntariamente estos objetos en las relaciones de indización. Estas relaciones son, a buen seguro, de un interés muy particular para los miembros de la ISKO.

3. ¿QUÉ HACER CON LOS DOCUMENTOS?

Además de ser creados, los documentos son seleccionados, representados y utilizados.

3.1. Selección

Reunimos documentos en las colecciones y recuperamos documentos de nuestras colecciones. Habitualmente se cree que estos son procesos diferentes. Pero el desarrollo de las colecciones y la recuperación de documentos de una colección son, a la vez, procesos de selección. En uno y otro caso, uno o varios documentos han recibido un rango privilegiado de cara a los demás documentos. En inglés se habla de «*information retrieval systems*» y de «*search engines*». Ciertamente, se busca y se extrae, pero hay también una elección. Personalmente, me gusta más la terminología de los años 1930: «mecanismos de selección».

3.2. Representación

Creamos sistemas abreviados, fichas y otras representaciones descriptivas de documentos. Estas representaciones pueden servir como sustituto del documento original. Creamos metadatos (*metadata*) que describen los datos (*data*). De hecho, hacemos las representaciones bibliográficas. Y hay una continuidad entre la reducida entrada de un índice y la versión completa de un documento, que comprende una descripción bibliográfica extensa. En el fondo, derivamos nuevas representaciones, a partir de documentos existentes.

3.3. Empleo

Es difícil predecir el empleo de los documentos. Ordinariamente, no sabemos quién va a utilizar éste o aquel documento. Puede ser que no lo utilice nadie. Normalmente, nosotros no sabemos si un documento ha sido leído cuidadosamente o examinado superficialmente, o bien, si se ha examinado el documento entero o consultado tan sólo una parte del mismo. Sobre todo, ignoramos las consecuencias intelectuales o prácticas de este examen del docu-

mento. La consulta de los metadatos se efectúa como una consulta de datos (de un documento), salvo que los metadatos sean muy breves y que hay menos base para captar la intención del autor (el indizador).

4. ESTRUCTURA

Veamos ejemplos de selección y de representación.

4.1. La indización automática

Con los documentos numéricos se pueden utilizar sistemas diversos de indización automática por logicales. El sistema KWIC ordena cada palabra del documento (con los términos que la acompañan) en una lista alfabética. Cada línea de un índice KWIC describe una pequeña parte del documento. La totalidad de las líneas KWIC derivadas del documento, constituye una representación mecánica del documento entero. Los sistemas en forma vectorial (por ejemplo el SMART) crean representaciones mecánicas más complejas.

4.2. Un catálogo de biblioteca

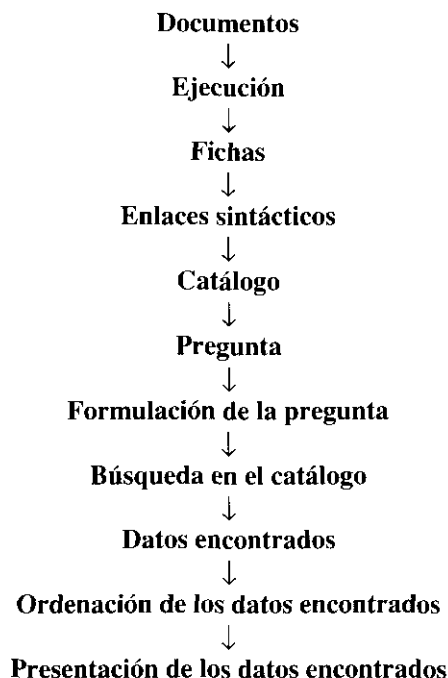
Pensemos que la creación de fichas bibliográficas sea para un fichero, o bien, como datos para un catálogo en línea. Se reúnen datos bibliográficos que representan un libro. Algunos datos son derivados del documento mismo: el título, el nombre del autor, el año de edición. Otros datos son derivados de otras fuentes, por ejemplo, una bibliografía nacional, OCLC, o un tesoro. Una ficha de catálogo es una representación de un libro, siguiendo las técnicas y las reglas convencionales con respecto a la forma, al contenido y al origen del libro (ISBD, AACRII, LCSH, LCC).

Sin embargo, los metadatos creados por el documentalista pueden ser modificados, verse normalizados, por relaciones o enlaces sintácticos impuestos por el editor del catálogo: USE, o UP (usar por), etc., y la armonización de los nombres personales.

En la misma línea, el usuario de la biblioteca quiere hacer una pregunta. Es probable que su demanda no se corresponda exactamente con la forma de las entradas del catálogo, por lo que deberá darse una versión distinta de la pregunta, con el fin de encontrar en él su correspondencia del modo más preciso.

Es preciso, por tanto, que la terminología de la pregunta concuerde con la terminología del catálogo. En la medida en que se dé esta correspondencia, se obtendrá el conjunto de documentos pertinentes. Además, los catálogos automatizados recogen generalmente los resultados obtenidos en orden alfabético antes de presentar los datos.

La figura siguiente muestra este proceso:



Podremos observar que el procedimiento contiene una serie de operaciones y que cada operación produce un nuevo conjunto de objetos, a partir de otro conjunto de objetos precedentes. Observemos que podemos mirar el catálogo entero como un documento complejo, que representa la colección entera. Al mismo tiempo, si nosotros lo consideramos a un nivel menos cercano, podremos observar el catálogo como un conjunto de fichas individuales (ellas mismas como pequeños documentos), cada una representando un libro en el conjunto de libros que constituyen la colección. Cada índice (autor, título, palabra clave, clasificación) es un subconjunto del catálogo.

Se encuentra que cada etapa produce un nuevo conjunto de objetos a partir de un conjunto precedente, por medio de algunas operaciones y que no hay dos categorías de operaciones.

1. Una categoría de operaciones la forman (ordenan, incluyen) los miembros de un conjunto. Comprendemos en esta categoría un orden total (*strictly ordered set*), de los conjuntos débilmente ordenados (*weakly ordered set*) —por encima del orden binario en dos subconjuntos: de los datos encontrados y de los datos no encontrados— y también la combinación de dos conjuntos para formar un subconjunto.

2. La otra categoría de operaciones comprende las transformaciones que modifican los miembros de un conjunto. Las fichas que se derivan, o bien las líneas KWIC o representaciones vectoriales, o de los documentos originales, constituyen esta segunda categoría.

Estas dos categorías de operaciones son también las dos clases de actividades que ya hemos observado antes, en la sección «¿Qué hacer con los documentos?»: la selección y la creación de representaciones o versiones, de documentos. Si se analiza el sistema bibliográfico, en la búsqueda de información y de filtración, se encuentra siempre esta estructura: *una cadena de operaciones sobre conjuntos de datos, que producen siempre un conjunto nuevo, ya (re)ordenado, ya transformado, sin excepción*. Por lo menos, en las investigaciones del Dr. Christian Plaunt y más, ninguno de los dos hemos encontrado la mínima excepción en esto. Parece que esta fórmula caracteriza todo sistema operacional de organización y nosotros no hemos encontrado excepción alguna en ello. Se ve que esta fórmula es característica de cualquier sistema operacional en la organización del conocimiento. (Buckland & Plaunt, 1994; Plaunt 199?).

5. COEXISTEN DIFERENTES VOCABULARIOS

Todo sistema de selección del conocimiento incluye múltiples vocabularios. Tal sucede en los casos primarios, por ejemplo, cuando un texto inédito es recorrido y cuestionado por una pregunta inédita, se dan, al menos, dos vocabularios:

1. **El vocabulario del autor del documento**, o bien, el vocabulario de varios autores; y
2. **el vocabulario del investigador.**

En los sistemas operacionales actuales, se encuentra, de ordinario, demasiados vocabularios simultáneos. En un catálogo de biblioteca se encontrará hasta tres vocabularios:

- el vocabulario de indización del documentalista, que modifica o suple el vocabulario del autor;
- las relaciones de sustitución (Use, UP, usa por, etc.), para armonizar o corregir los vocabularios de los documentalistas; y
- el vocabulario del investigador tal como lo formula en su pregunta.

En una palabra, siempre hay múltiples vocabularios en juego. Esperar que todos estos vocabularios sean idénticos y se armonicen alguna vez es desgraciadamente inútil.

Si se observan varios sistemas de selección, se ve enseguida la abundancia de vocabularios diferentes. He aquí un ejemplo: Yo mismo he querido buscar libros y artículos sobre la «*Coastal pollution*» (la polución en las costas del mar) en MELVYL, el catálogo en línea de la Universidad de California, y MEDLINE. Ni el uno ni el otro utilizan la frase «*Coastal pollution*» y una búsqueda booleana con «*coastal*» y «*pollution*» no ha dado resultado alguno, a pesar de que los documentos pertinentes existían en los dos sistemas.

En el catálogo MELVYL, utilizando el *Library of Congress Subject Headings*: se buscó por *marine pollution* y salió *coastal zone management, water-pollution, petroleum industry and trade, beach erosion, coasts, barrier islands; coastal changes, etc.*

Pero en MEDLINE, utilizando *MeSH*, se buscó por *seawater* y salió: *water pollution, bacteria, water microbiology, air pollution, environmental monitoring, bathing beaches, environmental pollution, etc.*

Observen la variedad de los vocabularios y lo poco que las dos listas tienen en común. Estos términos de indización son reales, ciertos, justificables, pero ¿quién podría imaginarse, al menos, la mitad?. Aquí nosotros hemos encontrado tres lenguajes diferentes: LCSH, MeSH y el mío.

6. CORRESPONDENCIA Y ASOCIACIONES ENTRE VOCABULARIOS

Es precisamente por esta multiplicidad de vocabularios por lo que se da siempre la posibilidad de una incompatibilidad, consecuencia del tránsito entre vocabularios y de una disonancia de sentidos. Un investigador puede emplear el término A y un autor ha empleado el término B. Ambos pueden referirse a la misma cosa, pero con sinónimos. Y sin embargo, es posible que los dos hayan empleado el término A para indicar diferentes sentidos, con homógrafos.

Los vocabularios intermediarios (sea el del documentalista, una pregunta formulada o la estructura syndética) pueden ser considerados como un intento de normalizar el empleo de los términos, a fin de rectificar todas las discordancias. El índice del documentalista rectifica el título dado por el autor en la representación del tema de un documento, por medio de un vocabulario estandarizado. Los investigadores experimentados saben cómo modificar sus preguntas o las de los demás de una manera tal, que responda el sistema convenientemente. Se dan tantas re-representaciones del sentido de un vocabulario, como transiciones de un vocabulario a otro. Cada una de estas re-representaciones ofrece una oportunidad para rectificar las disonancias entre el investigador y el documento, y además, ofrece también la posibilidad de que aparezcan aún más disonancias dispares. En un ejemplo sobre el artículo «*Conflicto Vietnamita*» el vocabulario del catalogador puede diferir a la vez del vocabulario del autor y del investigador. En este caso, una referencia cruzada, como «*Guerra del Vietnam*», Usar «*Conflicto Vietnamita*» corregiría la disonancia.

Alternativamente, un buen intermediario de búsqueda (humano o informático) podría orientar lo bastante, para demandar un cambio de terminología y la adaptación al vocabulario del sistema.

7. UNA DEFINICIÓN DE «VOCABULARIO»

Hemos hablado del vocabulario como si fuera un lenguaje ordinario. Pero si creemos que el concepto vocabulario es importante para los sistemas de selección del conocimiento, nos hará falta una definición técnica, precisa y suficiente, para la función del documentalista.

El *Oxford English Dictionary* (1989, vol. 19, 721) ofrece cuatro definiciones de Vocabulario:

1. un conjunto o lista de palabras con explicaciones breves de su sentido;
2. la extensión del lenguaje de una persona, clase, profesión o cosa;
3. la totalidad o suma de palabras que componen una lengua; y
4. figurativamente, un conjunto de formas artísticas o estilísticas, técnicas, movimientos, etc., a disposición de una persona particular, etc.

La noción fundamental es que «vocabulario» denota una enumeración de diferentes formas de expresar un sentido, el repertorio de las formas representativas. Es decir, que el repertorio de términos de indización es el vocabulario del documentalista.

8. LOS METADATOS COMO LENGUAJE

En el sistema de indización, los términos son, más de una vez, adaptaciones más o menos artificiales, del lenguaje cotidiano; o emplean una notación artificial, por ejemplo, «330», que en la *Clasificación Decimal de Dewey* significa *Ciencias Económicas*. Son sistemas para codificar el conocimiento. Evidentemente, cada uno es una clase de lenguaje. Describir es una actividad del lenguaje. Se ha reconocido después de mucho tiempo que los sistemas de indización son lenguajes. Se habla hoy de «metadatos», pero antes que de metadatos, se hablaba de «lenguajes documentales», «lenguajes de indización», o bien, de «metalenguajes» (Cfr. Coyaud, M., 1966).

Se puede, pues, utilizar la palabra «vocabulario» para indicar el repertorio de no importa qué lenguaje documental: los términos de un tesoro, los nombres de una clasificación; los valores de una categorización. En el marco de la documentación, se puede emplear el término «vocabulario» para indicar el repertorio de no importa qué campo MARC, o cualquiera otra forma de conjunto de metadatos. Es un concepto poderoso, porque -como decía- toda clase de sistemas de selección del conocimiento comprende una cadena de operaciones sobre los

conjuntos de datos, que produce siempre un conjunto nuevo, Así, el repertorio de cada conjunto sucesivo es el vocabulario de dicho conjunto. Y este concepto de vocabulario ha venido a ser un concepto enteramente fundamental.

9. LOS LENGUAJES HUMANOS TIENDEN A SER IMPRECISOS

Los vocabularios tienden a ser imprecisos, por dos principales razones:

1. Puede darse una falta de afinidad. Los términos son poco conocidos: ¿Quién sabe que para encontrar documentos referentes a los automóviles es necesario buscar por «TL 205» en la clasificación de la *Library Congress Classification*; por «180/280» en la clasificación de patentes de la *U.S. Patent Office*; y por «3711» en la *Standard Industrial Classification*?. Evidentemente, un índice o diccionario que recoja nuestro lenguaje cotidiano en cada lenguaje documental sería utilísimo. (Buckland con otros, 1999 [D-Lib]). Pero un índice es también necesario cuando un lenguaje «natural» es utilizado para la clasificación. En las estadísticas oficiales del comercio internacional de los Estados Unidos, no se encuentra *comercio* en «*automóviles*», porque no existen en su indización de lenguaje natural. Se encuentran datos estadísticos si se busca por «*cars*», pero estos datos se refieren ¡a los vagones del ferrocarril!. Es necesario buscar *automóviles* en «*Passenger motor vehicles, spark ignition engine*», un término de indización bastante descriptivo, pero desusado.

2. Por otra parte, la terminología queda imprecisa, porque el empleo de las palabras es dinámico. La lengua es algo vivo y lo que una palabra significa, puede cambiar con las circunstancias.

Mi parecer

Si el análisis que he presentado es correcto, las consecuencias son numerosas e importantes. Los sistemas de organización del conocimiento están en la base de los sistemas de lenguajes y vocabularios, lo que significaría que siempre serían imprecisos.

Las teorías formales de la información que utilizan la lógica, la entropía, el cálculo de la incertidumbre, y que son apreciados y prestigiosos en el campo de las ciencias de la información, deben quedar siempre incompletas; como útiles, puede ser, pero como inacabadas.

La mayor parte de las búsquedas de «digital libraries» se ocupan de cuestiones de infraestructura, y no de problemas fundamentales para los sistemas de organización del conocimiento.

Estos problemas fundamentales conciernen al lenguaje, la representación, la explicación semiótica de «...todo índice concreto o simbólico, *es conservado o registrado, con fines de representar, de reconstruir o de probar un fenómeno, ya físico, ya intelectual*».

BIBLIOGRAFÍA

- ANÓN (1937): «La terminologie de la Documentation». *Coopération Intellectuelle*, 77, 228-240.
- BUCKLAND, M. K. (1999): «The Landscape of Information Science: The American Society for Information Science at 62». *Journal of the American Society of Information Science*.
- BUCKLAND, M. K. (1999): «Vocabulary as a Central Concept in Library and Information Science». In *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23-26 May, 1999)*. Ed. By T. Arpanac et al. Zagreb: Lokve, pp. 3-12.
- BUCKLAND, M. K. (1997): «What is a “document”?» *Journal of the American Society for Information Science*, 48, n.º 9: 8.º4-8.º9. Also *Document Numérique (Paris)* 2, n.º 2 (1998): 221-230.
- BUCKLAND, M. K. (1999): Mapping Entry Vocabulary to Unifamiliar Metadata Vocabularies *D-Lib Magazine* 5 (1) January 1999. Online at.
- BUCKLAND, M. K. & C. PLAUNT. (1994). On the Construction of Selection Systems. *Library Hi Tech* 12:4:15-28.
- BUCKLAND, M. K. (1966): *Introduction à l'étude des langages documentaires*. Paris: Klincksieck.
- PLAUNT, C. (1997) *A Functional Model of Information Retrieval Systems and Processes*. Ph. D. dissertation, School of Information Management & Systems, University of California, Berkeley.