

V Congreso del Capítulo Español de ISKO, (International Society for Knowledge Organization) y IV Coloquio Internacional de Ciencias de la Documentación
Tendencias de investigación en organización del conocimiento

Comunicación para la Sección Científica: La organización del conocimiento en el entorno digital. Título de la comunicación:

Modelo conceptual de objetos para la representación y rastreabilidad de documentos digitales

Autora: Mela Bosch
melabosch@europa.com
Universidad Nacional de La Plata

Resumen

Se comenta el modelo conceptual desde el paradigma de orientación a objetos. Se lo define como una especificación del dominio del problema a través de la representación mediante objetos. Respecto del problema de los documentos se parte metodológicamente de la premisa de que en los documentos en el mundo digital no estamos ante una unidad documental: esto es así porque SGML, (y XML) presentan un concepto de documento distribuido por lo que el material que constituye un documento se puede encontrar en diferentes archivos y documentos simultáneamente.

Considerando esto, el modelo conceptual parte de la definición de documento como un soporte de información que constituye una unidad estructurada y convalidada de uno o más objetos documentales.

Por ello se impone una representación reticular de los objetos documentales, en tanto que las formas más difundidas de representación documental con diferentes propósitos, tales como MARC, recurren a las estructuras jerárquicas.

Proponemos usar otros recursos de representación que dan más libertad para el reuso y la rastreabilidad, recurriendo a la asociación entre objetos y dentro de ésta a la agregación como un tipo de asociación con propiedades semánticas adicionales.

En este modelo se plantea, siguiendo a Blaha y Premerlani, un agrupamiento físico de objetos documentales (*physical aggregation*) y un agrupamiento catalográfico (*catalog aggregation*) para extender la semántica de los objetos documentales. El lenguaje gráfico de modelado que se usará es UML y en el futuro se lo aplicará a diferentes casos.

Palabras claves: Documentos digitales; Modelo conceptual de objetos; Paradigma de Orientación a Objetos

Conceptual object model for representation and traceability of digital documents.

The conceptual model of this work is based upon the object oriented paradigm. It is explained as an especification of the problem domain by the means of the representation through objects.

Regarding documents, the starting point is methodologically following the premise that digital documents are not a documental unit. This because SGML –and XML– present a concept of document in a distributed form. Therefore, the material that a document is made of can be found in several files and documents simultaneously. Each of these is independent from the rest and can be organized in a logical structure. This considered, the conceptual model begins with the definition of a document as information support that forms a structured unit of one or more documental objects.

This is why a reticular representation of documental objects imposes; the most known forms of documental representation for several purposes –such as MARC– tend to hierarchical structures.

Our proposal is to use other forms of representation that may provide further freedom for traceability and recycling, by the means of association between objects and within this the use of aggregation, which is a type of association with additional semantic properties.

In this model we suggest –following Blaha and Premerlani– a physical aggregation of documental objects and a catalog aggregation in order to extend the semantics of the documental objects. The graphic language for modeling that will be used is the UML and it will be applied to several cases.

Keywords: Conceptual object model; Digital documents; Object Oriented Paradigm

1. Un modelo conceptual para representar documentos digitales

1.1 La descripción de documentos: la abstracción de la obra física

La forma tradicional de descripción bibliográfica se basa en una transcripción de elementos presentes en la obra física. Se trata de un recorrido que tomó varios años, pero en este momento se cuenta con un estándar internacional, el ISDB (International Standard of Bibliographic Description), desarrollado y difundido por la International Federation of Library Associations, (IFLA).

Esta vasta experiencia tiene como base la abstracción de elementos presentes en la obra física a partir de un proceso de distinción, selección y ordenamiento.

El resultado, tal como existe en ISDB, es un lenguaje de descripción constituido por: 1. un contenido estipulado para los elementos, los cuales se agrupan en áreas de descripción; 2. un ordenamiento preciso de esos elementos; 3. una sintaxis, por medio de indicadores de separación que son signos de puntuación que tienen una articulación entre sí; 4. una referencia de la fuente de los datos y 5. una indicación de la lengua natural en que deben ser transcritos esos datos.

El enfoque que proponemos no es incompatible con el lenguaje de descripción de ISBD, es complementario como medio para llevar adelante la tarea de distinción y selección. Puede aplicarse a diversos corpus documentales, en especial los formados por documentos digitales.

El instrumento de representación es el modelo conceptual, en nuestro caso, este modelo se realiza dentro del Paradigma de Orientación a Objetos.

1.2. El modelo conceptual en el Paradigma de Orientación a Objetos

La palabra modelo, como indican Blaha y Premerlani (BLAHA, 1999) tiene dos dimensiones:

1. una vista abstracta de un sistema
2. una fase de desarrollo (análisis, diseño, implementación)

Desde esta segunda perspectiva, el modelo forma parte de la fase de análisis. El resultado es representado por medio del modelo conceptual, el cual asume la primera definición de modelo, en el sentido de una vista abstracta.

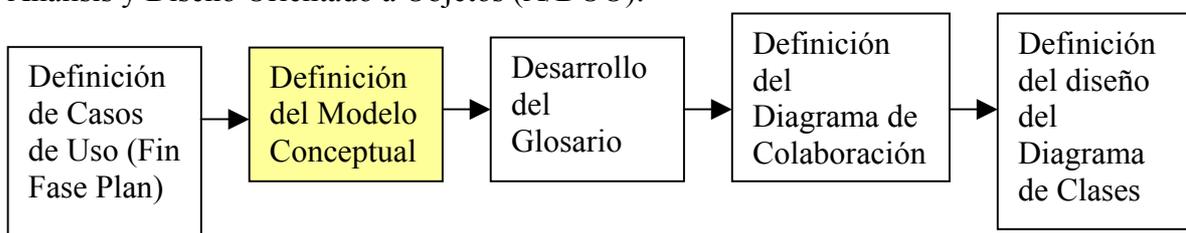
El modelo conceptual de objetos es, entonces, una identificación de conceptos, atributos y asociaciones, desde la perspectiva de la descomposición del dominio del problema. (Larman, 1998)

Según este autor el análisis orientado a objetos se propone una especificación del dominio del problema a través de la clasificación mediante objetos, tendiente a la comprensión de los términos usados en ese dominio. El resultado es representado por medio del modelo conceptual.

Si consideramos cualquier situación de análisis aplicando el Paradigma de Orientación a Objetos, el modelo conceptual se encuentra según Larman en lo que él denomina Fase de Análisis que sigue a la Fase de Plan y Elaboración en la que se establecen los casos de uso.

Esta Fase de Análisis comprende el Modelo Conceptual, con el establecimiento de asociaciones y atributos, el registro de los términos en el Glosario, los diagramas de secuencia y los de conducta o contratos.

Para Larman este modelo conceptual aparece de la siguiente manera en el esquema de Análisis y Diseño Orientado a Objetos (A/DOO):



Es importante diferenciar entre el concepto de modelo conceptual y el de objetos de software. El modelo conceptual representa el dominio del problema, es una abstracción del mundo real. Justamente, nos referimos a este autor porque consideramos valiosa su determinación de diferenciar el modelo conceptual del modelo de desarrollo de software, por lo que un mismo modelo conceptual puede servir para el desarrollo en diferentes plataformas de software.

Es algo que en la teoría se expresa claramente, pero que en la práctica no es sencillo. Se tiende a pensar el dominio del problema en términos de entidades software. Esto según este autor, simplifica la complejidad del dominio del problema y obstaculiza luego el establecimiento de clases que efectivamente den cuenta del dominio. Este obstáculo puede ser tan fuerte como observar el dominio del problema en términos de análisis de estructuras de datos, con lo cual se observará el dominio del problema en términos de funciones y procesos.

Esta problemática no es desdeñable en el ámbito de la representación de documentos. Es más, la delimitación de elementos en muchos casos, aparece vinculada con la descomposición necesaria para ser tratada con un software relacional, por ejemplo.

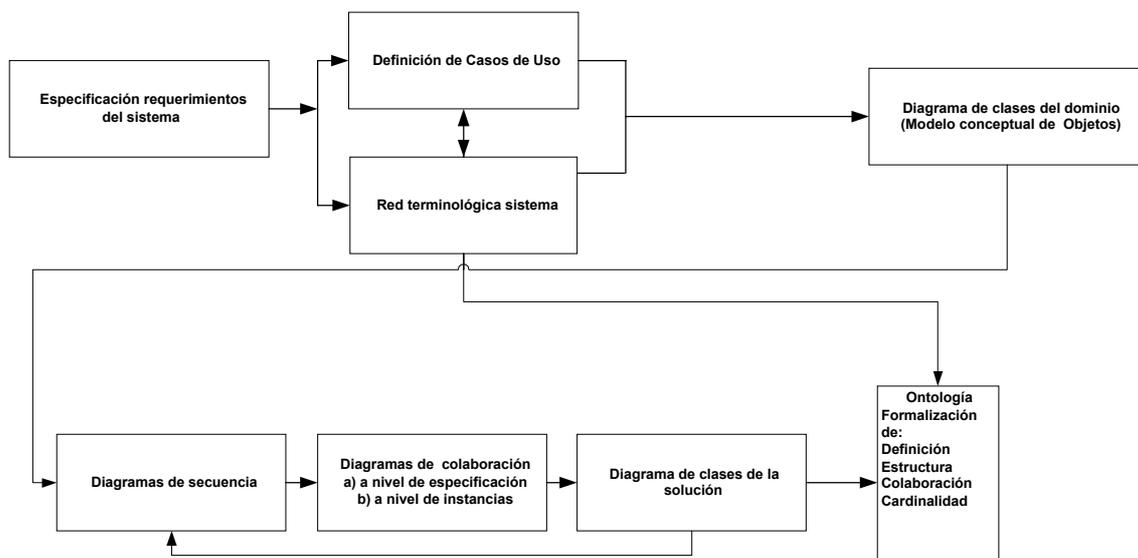
Para evitar esto el Modelo Conceptual según Larman muestra: conceptos, asociaciones entre conceptos y atributos de conceptos. Así cada concepto consta de un símbolo o término que lo representa unívocamente. Una definición intensional que dice qué es el concepto y cuales son sus atributos y una extensión que indica el conjunto de ejemplos a los que se aplica.

En síntesis, el proceso de construcción de conceptos puede ser visto como una descomposición del espacio del problema en unidades mínimas de comprensión.

La tarea más compleja es entonces la identificación y diferenciación de conceptos. Estas unidades mínimas se expresan por medio de términos, por lo que la presencia del glosario es fundamental.

Teniendo en cuenta todo esto, y considerando la importancia del aspecto terminológico como insumo para el Modelo Conceptual, y su valor para el establecimiento de la ontología del sistema, hemos desarrollado un esquema más amplio:

Esquema de Análisis y Diseño Orientado a Objetos (A/DOO):



2. Formas de representación de los documentos: Aspectos Metodológicos.

2.1 Estado del arte

En la AACR2 y en MARC las estructuras conceptuales presentes son de tipo jerárquico. En tal sentido, hace ya algunos años Frías (Frías, 1996) realizó una completa revisión de la representación en el modelo entidad/relación de las entidades presentes en un catálogo bibliográfico. Notó que las entidades bibliográficas están relacionadas con entidades no bibliográficas. E hizo notar además que se dan relaciones de tipo multidimensional. Indicó, también acertadamente, que las relaciones jerárquicas son de difícil representación en lógica de relaciones, tales como las relaciones verticales (todo parte) y las relaciones cronológicas (sucesiones y derivativas).

Esto en cuanto a los documentos en soportes impresos. A nosotros nos interesa ir más allá, al difuso campo entre la documentación, la archivística y la bibliotecología, y cubrir aspectos de los documentos en la vida de las organizaciones. Estos documentos son, cada vez en mayor medida, documentos digitales.

Así, nos encontramos con complejidades adicionales: 1. el emisor y responsable de los documentos y 2. no estamos ante una unidad documental.

Esto es así porque SGML, (y XML) presentan un concepto de documento distribuido por lo que el material que constituye un documento se puede encontrar en diferentes archivos y documentos simultáneamente. Cada uno de éstos son independientes entre sí y se organizan en una estructura lógica. (Bosch, 2001)

Esto transcurre en la vida virtual de las intranets, redes de empresas, entes administrativos, organizaciones en general. Aquí ante la dispersión de documentos y de las piezas de datos presentes en fragmentos de metadata se impone como requisito la rastreabilidad: quién dijo qué, cuándo, cómo y dónde. (Oliveros, 1996)

La rastreabilidad es además una de las propiedades requeridas para un modelo conceptual. (Karakostas, 1995)

Nuestro enfoque, finalmente, se orienta hacia el manejo de conocimiento en las organizaciones. (Bosch, 2002)

2.2. Definición de documento digital

En este contexto definimos a los documentos digitales como: un soporte de información que constituye una unidad estructurada y convalidada de uno o más objetos documentales.

Por lo que cada objeto documental debe poseer: 1. estructura constante; 2. información; 3. convalidación: alguien debe ser responsable por su estructura o la información.

Considerando estos tres items podemos considerar las siguientes categorías documentales: 1. **Predocumentos:** formularios, plantillas, etc.; son aquellos documentos de los que vulgarmente decimos que “están en blanco”. Si consideramos que para nosotros un documento es un soporte de información que tiene estructura y convalidación, cuando un documento tiene sólo estructura constante y convalidación, y no tiene más información esa, es un predocumento. La estructura constante es obvia, en cuanto a la convalidación, por ejemplo, al hacer un formulario alguien en la organización lo convalida definiendo que, como y cuando se va a usar ese formulario y no otro. 2. **Protodocumentos:** pueden tener información y estructura, no tienen convalidación. Ejemplo, cuando alguien escribe algo y se lo manda a otra persona; es un borrador, un mensaje de correo sin firma digital, etc. 3. **Documentos:** hay información, estructura y convalidación

Estas categorías documentales tienen una importancia fundamental tanto en la cadena de eventos de la vida de los documentos digitales, como en la representación estática de sus características. (Las ideas sobre la definición y categorías de documentos tienen como fuente el trabajo de Hugo García (García, 2002).

3. La base teórica para un modelo conceptual de objetos para representar documentos digitales

Dada la brevedad de la presente comunicación nos limitamos a dar las bases metodológicas y teóricas de nuestro desarrollo.

En este sentido en los apartados 1 y 2 presentamos el concepto de modelo conceptual y nuestra definición de documento.

En este último apartado presentaremos, a partir de la metodología indicada, las bases teóricas de nuestra propuesta: modelo conceptual de objetos para representar documentos digitales.

Tal como indicamos, para modelar la parte documental se tiende a hacer jerarquías. Sin embargo se pueden usar otros recursos. Proponemos discriminar dos conjuntos, el de la colección documental, para representar la totalidad de documentos y que referencia a cada documento individual, y el del catálogo documental que referencia a la totalidad, sin controlar cuántos, sino cuáles.

Tengamos en cuenta que además, según nuestra definición metodológica, un documento no es una unidad documental sino un conjunto de objetos documentales, que tienen a su vez diferentes categorías documentales. Esto nos lleva a plantearnos cómo puede darse la articulación entre la colección documental y el catálogo documental. Para modelizar esto se requiere de una consideración especial en cuanto a las formas de asociación entre objetos.

Proponemos una aplicación de lo que desarrollaron Blaha y Premerlani. (Blaha, 1998). Los autores plantean los diferentes tipos de asociación entre objetos, más allá del simple conocimiento, que sirven para dar más semántica al modelo conceptual.

Presentan las formas de asociación múltiple, con máximo y mínimo de multiplicidad, dentro de éstas las asociaciones ternarias, las asociaciones Exclusive-Or; las asociaciones cualificadas con roles que sirven para reducir o acotar los efectos de la multiplicidad.

Luego plantean el tema de la agregación. La agregación, para ellos, es una asociación con propiedades semánticas adicionales. La agregación se da entre un todo, llamado *assembly* (que traduciremos como agrupamiento) y sus partes, llamadas componentes. Lo que determina si una asociación es una agregación es la aplicación del test de is-part-of.

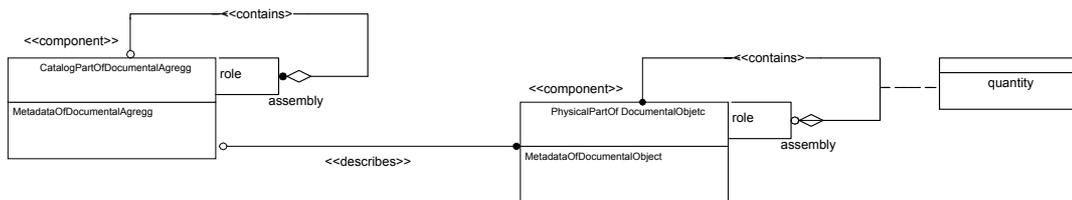
La agregación, según indican estos autores, es valiosa porque tiene características interesantes para nuestra representación conceptual de documentos, tales como la transitividad y la antisimetría. La transitividad permite la clausura transitiva del agrupamiento, por medio de una navegabilidad entre los nodos, en tanto que la antisimetría da libertad a los objetos del agrupamiento: en una generalización los objetos desaparecen cuando “muere” la superclase.

El aporte teórico fundamental que tomamos de estos autores para el desarrollo de un modelo conceptual de objetos documentales es la diferencia que ellos realizan entre lo que denomina *physical aggregation* y *cataloguing aggregation*. Nos permitimos usar los términos en inglés tal como aparecen en el original.

Physical aggregation, es aquella donde cada componente está dedicado al menos a un agrupamiento.

En la *cataloguing aggregation* tenemos una agregación donde cada componente es reusable a través de varios agrupamientos.

Dijimos que diferenciábamos dos conjuntos en nuestro modelo conceptual de documentos. El del catálogo documental, que utiliza la *cataloguing aggregation* y por tanto presenta los documentos con sus partes. En tanto la colección documental es necesaria para identificar el conjunto de objetos documentales que constituye un documento, y asimismo individualizar a cada uno de sus componentes para, de una manera rastreable, determinar además de cuál grupo u objeto documental exacto se trata. Para ello proponemos usar la *physical aggregation*.



En síntesis: utilizando estas bases teóricas en el modelo conceptual que desarrollamos, los diferentes objetos documentales van formando lo que se llama un *assembly*, o sea un conjunto agregado de objetos documentales: cada uno de estos agrupamientos es un documento.

La forma de poder expresar cada *assembly*, se tomó de Blaha y Premerlani, quienes, tal como reseñamos, proponen un dispositivo formado por otro objeto que se llama *cataloguing aggregation*, que es el catálogo de todos los objetos que forman la *assembly* y se relaciona con una *physical aggregation* que indica un aspecto concreto (va a dar el equivalente al Id de cada elemento de la *assembly*). La *assembly* construida se representa en la *catalog aggregation* y forma un conjunto de objetos documentales, por ejemplo: AgrupamientoDocumental1 (*assembly*1). Cualquier objeto integrante del AgrupamientoDocumental1 puede a su vez formar parte de otro agrupamiento (AgrupamientoDocumentalN (*assembly*N).

Cada objeto documental en la *cataloguing aggregation* que responde a un agrupamiento debe estar identificado en la *physical aggregation* y forma parte de esa única agrupación.

De manera que se establece una relación entre la *cataloguing aggregation* y la *physical aggregation*, o por otros objetos. Con lo que el problema de la multidimensionalidad y los objetos “no bibliográficos” aludidos por Frías (*ver ut supra*) pueden ser representados.

Los autores además proponen hacer una semántica extendida para la *physical aggregation* que permite: propagación de operaciones, propagación de valores por defecto, versiones, locking, entre otros aspectos. Sobre esta semántica trabajaremos en breve.

Para la *cataloguing aggregation* también proponen una semántica extendida para la descripción de los componentes, en lo que puede no haber comportamiento, solo datos que se describen y son reusados por la *physical aggregation*, o por otros objetos que conocen a la agregación. En este punto, para el modelo conceptual de documentos, la consideración de items de metadata es fundamental y hacia allí nos orientamos.

Para concluir, este desarrollo forma parte de un proyecto de investigación sobre tratamiento semántico de corpus documentales para estrategias comunicacionales en organizaciones, de la Facultad de Periodismo y Comunicación Social de la Universidad Nacional de La Plata, Argentina.

Referencias:

Blaha, M.; Premerlani, W. **Object Oriented and design for database applications**, Prentice Hall, 1998. Cap. 3. *Advanced Object modeling. 3.2. Link and association concepts*, p. 48 y ss.

Bosch, Mela. *La gestión del conocimiento en el medio digital. Viejos problemas de tratamiento de información y aspectos nuevos. Ciencias de la Información*: La Habana, v. 33, n.1, abril 2002. p. 35-43.

Bosch, Mela. *Documentos y lenguaje de marcado: conceptos, problemas y tendencias. El profesional de la información*: Barcelona, nov. , 2001, v. 10, n. 11, pp. 4-9.

- Frías, José A. *La estructura conceptual del registro bibliográfico: Una revisión*. **Scire. Representación y Organización del Conocimiento**: Zaragoza: Universidad de Zaragoza. 2: 2, Jul.-Dic. 1996. pp.53-77.
- García, Hugo. *La documentación científica*. En: Dai, Daniel (coord) **Pensar y hacer ciencia**. Buenos Aires:Hernandarias, 2002. Cap 13.
- Karakostas, V.; Loucopoulos, P. **System requirements engineering**. London: McGraw-Hill, 1995. p.76.
- Larman, Craig. **Applying UML and patterns**. New Jersey: Prentice Hall, 1998.
- Oliveros, Alejandro et al. **Integración de Escenarios con el Léxico Extendido del Lenguaje en la elicitación de requerimientos : Aplicación a un caso real**. Buenos Aires: Universidad de Belgrano, 1996.