

Metadatos y Tesauros: aplicación de XML/RDF a los sistemas de organización del conocimiento en Intranets.

Eva M^a Méndez Rodríguez
Universidad Carlos III de Madrid.
Dpto. Biblioteconomía y Documentación

Resumen:

La gestión del contenido semántico de los datos se está convirtiendo en un aspecto estratégico para las organizaciones que están asumiendo las tecnologías web emergentes. Las nuevas tecnologías para el acceso y el intercambio de información aumentan la visibilidad de la información corporativa y generan grandes expectativas para encontrar, entender y compartir el contenido informativo. En esta comunicación se trata la adecuación de estructurar la información a través de metadatos en el contexto de las Intranets, así como de la adecuación de RDF/XML para constituir tesauros que permitan optimizar la recuperación de información y mantener la consistencia en entornos de información distribuidos.

Palabras clave: Internet/Intranet / Metadatos / Tesauros / XML/RDF / Recuperación de información

1. Introducción.

Las intranets corporativas, es decir, la utilización de la arquitectura web para gestionar la información internamente, se han convertido en una solución óptima para compartir grandes cantidades de datos en el contexto concreto y finito de una organización. El entorno informativo de las intranets representa uno de los ámbitos más fértiles y poco explotados donde el profesional de la información puede aportar un valor añadido en la gestión del conocimiento integral. El fenómeno Internet de que "cualquiera puede producir información" ha trascendido al mundo de las intranets dificultando asimismo la recuperación de información relevante, que demanda una mejor gestión del contenido. El potencial de la tecnología para la producción y difusión de conocimiento hace imperativo el desarrollo de una estrategia de integración de la información en las organizaciones.

Estamos en la Era de la información, donde la gestión eficaz de contenidos se está convirtiendo en una verdadera industria. En la mayor parte de las organizaciones existen distintos servidores web con contenidos valiosos y heterogéneos para la toma de decisiones, accesibles a través de su Intranet que requieren, no sólo una gestión eficaz del contenido, sino también del contexto. El desarrollo de la tecnología web y la incorporación de la semántica a la información legible por máquina, proponen la solución para una recuperación más relevante de información almacenada en la intranet: por un lado XML (eXtensible Markup Language) plantea un horizonte para la gestión

del contenido, mientras que los esquemas de metadatos ofrecen el fundamento para la gestión del contexto. Sin embargo, la utilización de los mismos elementos de metadatos no garantiza que éstos sean compatibles, por ello se precisa además una indización por conceptos basada en bases de conocimiento u ontologías. En este sentido, es fundamental el papel de los vocabularios controlados y listas de autoridad para determinar el contenido de los metadatos y, por ello, la necesidad de redefinir el concepto de herramientas terminológicas en un contexto de información distribuido.

La mayoría de los estudios que se realizan sobre el valor de las herramientas terminológicas en el nuevo contexto de la información distribuida [MILLER,2000. ROSA, 1999, etc] se centran en la utilización de sistemas de organización del conocimiento en Internet. Sin embargo, señalaré algunas reflexiones importantes que me han llevado a circunscribir, en esta comunicación, la revalorización de los tesauros, ante el problema de la recuperación de información en texto completo, al entorno de la información web corporativa:

- a) Las intranets se desarrollan según los mismos estándares que Internet (HTML, XML, etc.) y como Internet, es normalmente un conjunto de recursos descentralizados. Sin embargo, las intranets suponen entornos finitos —o al menos previsibles— de información, además de tener una mayor homogeneidad temática y una complejidad de tipos de información controlable. Estas características hacen de que la Intranet pueda asumir con más facilidad el reto de la organización y recuperación de la información.
- b) Por otra parte, los sistemas de recuperación de información en Internet de propósito general (tipo Altavista, Northernlight, etc.) se basan en la extracción automática de la información y carecen de técnicas de gestión del conocimiento y por tanto no pueden dar una respuesta precisa a una pregunta concreta sobre el contenido semántico de los documentos, y por ello recuperan tanto ruido. Sin embargo, todos los sistemas de recuperación de información de calidad en la Red —los denominados *subject-gateways*, que prefiero llamar "sistemas de recuperación de información de organización bibliotecaria"— que centran sus esfuerzos en la selección, descripción y organización, de recursos de un área temática. Sólo en contextos muy concretos de recuperación de información en Internet se utilizan normas de valor semántico como vocabularios o tesauros tanto para describir el

contenido de los documentos como para realizar las búsquedas. V.gr. el sistema SOSIG (*Social Science Information Gateway*) <<http://www.sosig.ac.uk/>> utiliza un tesoro derivado del *HASSET thesaurus*, desarrollado por el *Data Archive* en la Universidad de Essex partiendo del Tesoro de la UNESCO.

- c) Mientras que Internet es un entorno no finito, multilingüe, y heterogéneo, una Intranet es en sí misma un sistema de información temático, una *subject-gateway* de visibilidad limitada, finita, más homogénea y tipificable, y normalmente mono/bilingüe. Por ello parece ser un entorno informativo proclive para basar la recuperación de información en sistemas de organización del conocimiento como tesauros y clasificaciones, que normalicen los atributos de los metadatos descriptivos aplicables.

2. Acceso consistente a la información en una Intranet: Metadatos e Intranets

Una Intranet convencional está formada por diversos servidores web, servidores de ficheros, depósitos de datos especializados y cientos o miles de documentos; sin embargo, a pesar de que el contenido está físicamente más accesible, esto no quiere decir que esté más organizado. Sin una organización concreta y coherente, los usuarios de una Intranet sólo pueden buscar un número limitado de recursos. Es pues fundamental organizar, catalogar y describir la información disponible de tal forma que se pueda especificar el contenido y el contexto de la información, el propósito de la misma, indicar las relaciones entre los distintos datos, establecer quién es el autor/creador/propietario de la información, especificar la validez de la información, etc., esto es, asignar metadatos al conocimiento almacenado en la Intranet.

Los metadatos son información documentada a través de herramientas de tecnologías de la información que mejoran la comprensión, tanto técnica como comercial, de los datos y de los procesos relacionados con ellos [SEINER, 2000]. Esta definición es sin duda mucho más elocuente que la de "datos sobre los datos". Los metadatos, en el contexto de una Intranet o de un Datawarehouse, son también información estructurada sobre la información distribuida, datos asociados a los objetos de información que proporcionan un conocimiento más completo sobre dónde encontrar una información y cómo esta puede o no ser útil según sus características. Todas las

empresas tienen metadatos: las bases de datos, los modelos de información, informes, etc. todos los componentes de un sistema de información están constituidos sobre metadatos.

En el contexto de Internet están surgiendo distintos formatos de metainformación algunos de propósito general como el Dublin Core (DC) y otros de propósito específico y temático como p. ej. el CDWA (*Categories for the Description of Works of Art*) para la documentación sobre arte y patrimonio, etc. En el contexto de una Intranet corporativa, sin embargo, no existe tanta literatura sobre esquemas predefinidos de metadatos, se puede desarrollar un modelo *ad hoc* a la temática de la empresa a través de una DTD (*Document Type Definition*) de XML o SGML, o bien utilizar un formato que pueda interpretar el motor de búsqueda.

Kelly Doran [DORAN, 1999] señala tres tipos de información que debe contener el esquema de metadatos aplicado en su proyecto de Intranet de la empresa Weyerhaeuser, en el que podemos fijarnos para definir qué tipo de metainformación precisa una Intranet:

- *Información bibliográfica básica*, de tal forma que se consignen los siguientes elementos: Título, autor, tipo de fichero (que en el caso que describe Doran, será por defecto HTML), tamaño del fichero, fecha de creación y última actualización.
- *Información contable y sobre la gestión de documentos*, donde se incluirán los siguientes elementos de metadatos: contacto, email del contacto, responsable, la organización que financia el recurso, código de confidencialidad de la información, fecha de caducidad u obsolescencia de la página.
- *Información descriptiva del contenido del recurso*, donde deben figurar los siguientes elementos: descriptores controlados de materia, materias añadidas (no controladas), categoría a la que pertenece (seleccionada de una lista controlada) y categorías añadidas (para incluir nuevas categorías no incluidas en la lista).

Darlene Fichter [FICHTER, 1999] habla de dos tipos de metadatos en Intranets: por un lado, la información bibliográfica y la de gestión, que denomina metadatos

administrativos o factuales, y por otro, los metadatos estrictamente descriptivos. Esta autora abre una lanza a favor de la metainformación factual, argumentado que suelen pasarse por alto y sin embargo aportan datos rápidos e importantes sobre los documentos (fecha de creación, de modificación, el tamaño del fichero, el idioma en el que está escrito, el creador/autor, etc.) que puede ser de gran utilidad para la búsqueda, recuperación e intercambio de información en entornos corporativos; además señala la facilidad de asignación de estos metadatos, en algunos casos incluso automática en las aplicaciones de creación de documentos web, frente a la complejidad de los metadatos descriptivos y la competencia documental que debe tener el personal de la Intranet para desarrollar tesauros y asignar descriptores controlados. De la complejidad de los metadatos descriptivos y del reto que suponen para su incorporación en una Intranet profundizaremos en el apartado siguiente, en el que relacionaremos las estructuras de metainformación con los sistemas de organización del conocimiento.

Independientemente de su tipología, la asignación de metadatos y la destreza de una empresa en la gestión de datos, información y conocimiento determinará el éxito de una compañía.

3. Metadatos y sistemas de organización del conocimiento: continente y contenido.

El problema intelectual de la caracterización del contenido de los documentos, ha constituido siempre una rémora en el trabajo documental. No obstante, asegurar la consistencia y el control del vocabulario han sido temas de gran preocupación para los profesionales de la información y retoman su importancia en los modelos de metadatos aplicados a Intranets para coadyuvar a la consistencia de los resultados en la recuperación.

La asignación de metadatos administrativos en una Intranet es casi automática y de gran utilidad, sin embargo la combinación de atributos relativos a los metadatos descriptivos tiene mayor complejidad, como adelantábamos antes. Para que la adopción de un esquema de metadatos sea operativa no basta sólo con la coherencia de los elementos o la normalización de la estructura (continente), sino que es preciso también un control terminológico y normas de valor semántico que gobiernen la formulación de

su contenido, esto es listas de autoridad, vocabularios controlados, clasificaciones y tesauros.

3.1. Metadatos y vocabularios controlados

Crear sistemas de organización del conocimiento para toda la Red sería imposible porque implicaría la creación de un megatesauro o una clasificación general, por ejemplo, que abarcase todo el conocimiento humano, todas las lenguas de producción de información en Internet, las equivalencias semánticas multilingües, etc., además sería un proyecto inabarcable en términos económicos y de tiempo, e incluso injustificado, dada la potencia de los algoritmos de recuperación de las herramientas que no usan lenguajes controlados (v. gr. Altavista). Además, si no se ha llegado a un consenso de vocabularios controlados en la época previa a la web, ¿cómo se podría llegar a un estándar de organización del conocimiento que abarcase todas las disciplinas, todos los conceptos, incluso los que están por descubrir, y todos los países y lenguas? Sin embargo, de igual forma que cada vez más en la Red están surgiendo proyectos de control terminológico como el tesauro de la *California Environmental Resources Evaluation System* [CERES-THES, 2000] o *Zthes* [TAYLOR, 1999] para la búsqueda de información basada en metadatos en la recuperación por materias, las intranets corporativas deben asimilar estos esfuerzos para dotar a al modelo de metadatos adoptado, coherencia semántica, teniendo en cuenta además, que:

- La producción de contenidos en una Intranet es controlable, las materias próximas y el idioma uniforme.
- Existen aplicaciones sencillas, que pueden utilizar todo el personal de la Intranet para asignar la misma estructura de metadatos (p. ej. Verity ha desarrollado lo que denominan *Knowledge Organizer* <<http://www.verity.com/products/ko/index.html>> para clasificar y organizar el conocimiento corporativo; o la compañía Computer Associates, que ha desarrollado una aplicación similar, *Platinum Repository* <<http://www.cai.com/products/decisionbase/repository.htm>> para dar mayor coherencia global a la información de una empresa, estructurándola en metadatos.
- El esfuerzo de crear un sistema de organización del conocimiento será rentable y garantizará la calidad del servicio de información de la Intranet.

Con todo, los tesauros en el contexto de información distribuida de la Intranet, estructurada por metadatos proporcionarán un soporte a la recuperación de información basado en el conocimiento y facilitará la combinación de múltiples bases de datos o la unificación del acceso a diversos contenidos. Los tesauros en estos sistemas de información serán pues, algo más que una mera herramienta para la indización, son el soporte semántico de la metainformación.

3.2. RDF/XML para la estructuración de tesauros

Es obvio que en los sistemas de información electrónica (de acceso web o no) no tienen ningún sentido los tesauros tradicionales impresos, ni siquiera muchos de los que pueden existir en formato electrónico; es preciso contar con una herramienta de control terminológico, con un tesoro que pueda adaptarse a las necesidades de escalabilidad e interoperabilidad de la Intranet. A esto hay que añadir que, cada vez más los entornos web están apostando por el estándar XML (*eXtensible Markup Language*), también en el desarrollo de intranets corporativas [Orzech, 1999]. XML es un metalenguaje que, desarrollado por el Consorcio web (W3C) permite concebir lenguajes de marca específicos para estructurar el conocimiento en distintos tipos de documentos; además de otras DTDs (*Document Type Definition*) que se pueden desarrollar para los distintos tipos de información de la Intranet, XML puede ser el soporte de las herramientas de control terminológico. Si bien existen distintos tesauros hipertextuales en Internet, y distintas herramientas para crear tesauros en HTML (v.gr. <http://www.multites.com>) y ponerlos en entornos Internet/Intranet, HTML es insuficiente para compartir toda la potencia semántica de un tesoro. A esta realidad del estado de los estándares para el desarrollo de tecnología web, hay que añadir la tentativa de revisar la norma americana de construcción de tesauros ANSI/NISO Z39.19 (1993 –R1998) para desarrollar un nuevo estándar que recoja los criterios y metodología para el desarrollo automatizado de tesauros, así como el conjunto de herramientas que muestren las relaciones semánticas entre los términos, un estándar que soporte, tanto una gran variedad de presentaciones de tesauros electrónicos, como los protocolos interoperables y estructura/semántica aplicable a éstos [MILSTEAD, 1999]

En los últimos dos o tres años han estado apareciendo tentativas para desarrollar productos y servicios basados en la organización del conocimiento para la web, tanto desde el mundo de la investigación (como el proyecto CERES o el Zthes, que ya hemos

mencionado, o el grupo de trabajo sobre ontologías, que han desarrollado un lenguaje propio de marcado —*Ontology Markup Language*, OIL <<http://www.ontoknowledge.org/oil/>>, o los trabajos del grupo para implementar la gestión del contenido semántico a través de la norma ISO de registros de metadatos ISO/IEC 11176 <<http://hmrha.hirs.osd.mil/mrc/>>, etc.), como de iniciativas privadas como la de Interconnect <http://www.interconnect.com/>, Metacode <<http://www.metacode.com/>> o VHG Consulting <<http://www.vhg.org.uk>>. Sin embargo, y a pesar de esta plétora de modelos y proyectos, existe un convenio más o menos generalizado (auspiciado tanto por las recomendaciones del W3C, como por las indicaciones del grupo de trabajo en la revisión de la norma de construcción de tesauros Z39.19 y del grupo sobre Sistemas y Servicios de conocimiento en red (NKOS) de la utilización del *Resource Description Framework* (RDF) para la estructuración y mantenimiento de tesauros.

RDF, además de ser un modelo para la estructuración de metadatos, permite describir cualquier recurso que pueda asignársele un URI (*Uniform Resource Identifier*) y podemos asignarle URI a los términos de un tesoro, por ello puede considerarse un esquema para la representar lenguajes jerárquicos y mapas de conocimiento. Al igual que XML, se trata de un estándar desarrollado por el W3C (actualmente el consorcio está trabajando en la unificación de los esquemas de RDF y de XML para simplificar su sintaxis). La semántica funcional de RDF está formada por: un modelo de datos, una sintaxis y un esquema. La especificación del modelo y la sintaxis (Recomendación del Consorcio Web desde febrero de 1999 [W3C-RDF-MS]), es un estándar estructural de metainformación diseñado para servir como fundamento para la interoperabilidad en el procesamiento de metadatos. La especificación del esquema (desde el 27 de marzo del 2000, Candidato para la Recomendación) *proporciona los recursos suficientes para crear modelos RDF que representen la estructura lógica de un tesoro* [W3C-RDF-S]. Sin embargo, lo más significativo de RDF es la utilización de los *namespaces* que permiten la utilización de vocabularios distribuidos.

Utilizar RDF/XML para el desarrollo de tesauros o vocabularios controlados en el entorno de información distribuida de una Intranet, implica "utilizar metadatos para describir metadatos", aprovechar la flexibilidad de XML para gestionar el conocimiento corporativo tanto a nivel estructural como semántico, así como por la versatilidad de los enlaces y la posibilidad de compartir distintos tesauros.

4. Conclusiones

Con las reflexiones que hemos planteado en esta breve comunicación, destacando el papel de los vocabularios controlados y listas de autoridad para determinar el contenido de los metadatos, se desprende la necesidad de redefinir el concepto de herramientas terminológicas en un contexto de información distribuido. Podemos resumir, además, las siguientes conclusiones:

- El modelo de organización de la información en Intranets que defendemos aquí: esquema de metadatos + vocabulario controlado, es factible y operativo en Intranets corporativas en tanto que su planteamiento, como sistema de recuperación de información, es asimilable a las *subject-gateways* de Internet.
- La asignación de metadatos normalizados es la solución para la falta de semántica entendible por máquina. Para una recuperación temática consistente en Intranets, es necesario contar no sólo con modelos de metadatos que interoperen a nivel técnico, sino que el contenido de esas estructuras de información sea lo más uniforme posible. Para ello, no sólo el desarrollo tecnológico será suficiente para mejorar la recuperación de información, sino también una representación ortodoxa del conocimiento distribuido, lo que, hasta hoy, sólo es posible mediante el lenguaje.
- Las herramientas terminológicas, como los tesauros, son un fundamento importante en la creación, mantenimiento y reutilización de fuentes de información de valor añadido en entornos online.
- El esfuerzo de interoperabilidad terminológica en la construcción y mantenimiento de tesauros, debe apoyarse en estándares de marcado de documentos como RDF/XML que demuestran la adecuación para representar estructuras jerárquicas, así como la posibilidad de compartir esquemas de datos en las descripciones.

Bibliografía

[BACA, 1998] Introduction to metadata: pathways to Digital Information. Murta Baca, ed. Los Angeles: Getty Information Institute, 1998

[CERES-THES, 2000] The CERES/NBII Thesaurus partnership Project [sitio web]. California Environmental Resources Evaluation System, 24 de julio de 2000. Disponible en: <http://ceres.ca.gov/thesaurus/> (consultado el 24 de julio de 2000)

[CHEN, et al, 1999] Michael Chen, Marti Hearst, Jason Hong, James Lin. Cha-cha: a System for Organize Intranet Search Results [documento www]. En: *Proceedings of the 2nd USENIX Symposium on Internet Technologies and SYSTEMS*. Boulder, October 1999. Disponible en: <http://cha-cha.berkeley.edu/papers/usits99/index.html> (consultado el 24 de julio de 2000)

[DOERR AND FUNDULAKI, 1998] Martin Doerr and Irini Fundulaki. *SIS-TMS: A thesaurus management system for distributed digital collection* [documento ps]. Crete, Greece: Institute of Computer Science. Information Systems and Software Technology Division, 1998. Disponible en: http://www.ics.forth.gr/proj/isst/Publications/paperlink/edl_98_springer.ps.gz (consultado el 20 de julio de 2000)

[DORAN, 1999]. Kelly Doran. Metadata for a corporate Intranet. *Online, january/february, 1999*, p.43-50

[FICHTER, 1999] Darlene Fichter. Administrative and factual metadata for intranets: issue and options. *Online*, vol. 3, n.6, 1999

[GOGLIN, 1998] Jean François Goglin. *La construction du datawarehouse: du datamart au dataweb*. Paris: Hermes, 1998

[KOCH and VIZINE-GOETZ, 1998]. Traugott Koch, Diane Vizine-Goetz. Automatic Classification and content navigation support for web services. *Annual Review of OCLC Research*, 1998. Disponible en: http://www.oclc.org/oclc/research/publications/review98/koch_vizine-goetz/automatic.htm (consultado el 20 de julio de 2000)

[MILLER, 2000] Paul Miller. I say what I mean, but do I what I say? [documento www] *Ariadne*, issue 23, 22 March 2000. Disponible en: <http://www.ariadne.ac.uk/issue23/metadata> (consultado el 20 de julio de 2000)

[MILSTEAD, 1999] Jessica Milstead. *NISO/APA/ASI/ALCTS Workshop on Electronic Thesauri: Planning for a Standard* [documento www]. National Information Standard Organization, 1999. Disponible en: <http://www.niso.org/thes99rprt.html> (consultado el 20 de julio de 2000)

[ORZECZ, 1999] Dan Orzech. XML is here to stay [documento www]. *Intranet Journal*, 26 July 1999. Disponible en: http://intranetjournal.earthweb.com/development/xml_intranet_072699.html (consultado el 20 de julio de 2000)

[ROSA, 1999] Antonio de la Rosa. Instrumentos terminológicos en el www: xml. *El profesional de la Información*, vol. 8, n.10, octubre 1999, p. 14-36

[SEINER, 2000] Robert S. Seiner. *Questions metadata can answer* [documento pdf]. Computer Associates Products, CAI, 14 de julio de 2000. Disponible en: http://www.cai.com/products/decisionbase/questions_metadata_can_answer.pdf (consultado el 24 de julio de 2000)

[TAYLOR, 1999] Mike Taylor. *Zthes: a Z39.50 Profile for Thesaurus Navigation* [documento www]. Washington: Library of Congress, 28 de febrero de 1999. Disponible en: <http://lcweb.loc.gov/z3950/agency/profiles/zthes-02.html> (consultado el 24 de julio de 2000)

[W3C-NS] World Wide Web Consortium. *Namespaces in XML* [documento www]. Tim Bray, Dave Hollander, Andrew Layman, eds. W3C, 14 de enero de 1999. Disponible en: <http://www.w3.org/TR/REC-xml-names> (consultado el 20 de julio de 2000)

[W3C-RDF-MS] World Wide Web Consortium. *Resource Description Framework (RDF): Model and Syntax Specification. W3C Recommendation, 22 February 1999* [documento www].

Ora Lassila and Ralph R. Swich, eds. 22 feb. 1999. Disponible en: <<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>> (consultado el 20 de julio de 2000)

[W3C-RDF-S] World Wide Web Consortium.. *Resource Description Framework (RDF) Schema Specification 1.0 W3C Candidate Recommendation 27 March 2000* [documento www]. Dan Brickley, R.V. Guha, eds. 27 mar. 2000. Disponible en Internet <<http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>> (consultado el 20 de julio de 2000)

[WHEATLEY AND AMSTRONG, 1997] A. WHEATLEY and C. J. Armstrong. Metadata, recall, and abstracts: can abstracts ever be reliable indicators of document value?. *Aslib Proceedings*, September 1997, vol. 49, nº 8, p. 206-213