ISSN 0121–1129

Modelo para búsqueda y recuperación semántica en bibliotecas digitales

A Semantic Model for Digital Libraries to Search and to Retrieve

Fecha de recepción: 14 de septiembre de 2010
Fecha de aprobación: 10 de noviembre de 2010

Jorge Eliécer Giraldo Plaza*
Maryem A. Ruiz*
Sandra Patricia Mateus*

Resumen

Las tecnologías de la Web Semántica, aplicadas a las bibliotecas digitales, permiten representar, buscar y recuperar la información. En este artículo se propone un modelo de búsqueda y recuperación de información para objetos digitales en bibliotecas digitales, haciendo uso de una ontología de dominio.

Palabras clave: Bibliotecas digitales, Recuperación de información, Web semántica.

Abstract

The Semantic Web Technology applied to digital libraries allows to represent, to search and to retrieve information. Here a search and information retrieval model for Digital Objects in Digital Libraries is proposed, using an ontology domain.

Key words: Digital Libraries, Information Retrieval, Semantic Web.

* Politécnico Colombiano Jaime Isaza Cadavid, Medellín - Colombia.
GRINSOFT: Grupo de Investigación en Desarrollo de Software. Línea de Investigación en Inteligencia Computacional. grinsoft@elpoli.edu.co

I. Introducción

Actualmente, la Web es un espacio preparado para intercambiar información, diseñado para el consumo humano; las páginas web son creadas por personas para ser entendidas por personas. No existe un formato común para mostrar la información, por lo cual, los desarrolladores de páginas web las crean dependiendo de los potenciales usuarios que van a visitarlas. En los últimos años, algunas empresas han realizado anotaciones de datos introducidas dentro de este código HTML, siguiendo algún esquema de anotación común, normalmente basado en XML [1]. Es así como se quiere aprovechar esta tecnología y ampliar su descripción por medio de lenguajes de marcado semántico, para así lograr que se puedan describir los objetos desde un punto de vista conceptual, que luego permita su búsqueda y recuperación efectiva.

Por medio de las tecnologías de la Web Semántica [2] se puede hacer uso de metadatos descritos semánticamente, que permiten un procesamiento basado en conceptos de los objetos descritos [3]. Se propone entonces un modelo de representación y búsqueda semántica de objetos digitales en bibliotecas digitales que permita, por medio de una ontología de dominio, realizar y describir semánticamente los objetos digitales, y, del mismo modo, buscarlos y recuperarlos; para ello se realizó un estudio comparativo de los lenguajes de recuperación semántica, propios de RDF, y posteriormente se desarrolló un prototipo funcional que permita realizar las funciones de representación y búsqueda. Tanto la representación como la búsqueda emplean una librería apropiada para su manejo en la Web.

Algunos proyectos similares se han desarrollado en los últimos años, sin embargo, se especializan en dominios específicos, propios de un contexto; tal es el caso de *OntoGuate* [4], donde se busca la administración de una ontología de turismo con sus respectivas búsquedas, así pues, clasifica las respuestas con base en las instancias de las clases. Otro proyecto es *SIMILE* [5], que se enfoca en realzar los aspectos de integración de los metadatos, subrayando la importancia de la representación semántica como medio de estandarización hacia la accesibilidad. *Bicks* [6], por su parte, se refiere a la infraestructura de red detrás de las bibliotecas digitales como punto clave en la distribución de información entre usuarios, empleado una ontología de dominio relacionada con el tema de la cultura. *JeromeDL* [7] es una biblioteca digital semántica social, su potencial radica en permitir el intercambio de información de manera colaborativa entre usuarios de librerías; se considera social, ya que permite la inclusión de herramientas de la Web 2.0.

El documento se estructura como sigue: en la sección 2 se presenta la aplicación de la Web Semántica por medio de la construcción de la ontología de objetos digitales en bibliotecas digitales y la selección del lenguaje apropiado para su búsqueda; en la sección 3 se presenta el diseño del modelo de representación y su vista arquitectónica; en la sección 4 se presenta el diseño del modelo de búsqueda y recuperación, y en la sección 5 se tratan aspectos de la implementación técnica del prototipo de validación; por último se presentan las conclusiones y el trabajo futuro.

II. Aplicación de la Web Semántica

En esta sección se presenta la construcción de la ontología de objetos digitales en bibliotecas digitales y su posterior validación con el lenguaje de consulta seleccionado a partir de un estudio comparativo.

A. Construcción de ontología de dominio

Un objeto digital se caracteriza por los siguientes atributos, que por cuestiones de espacio no se entrarán a detallar: Comunicación, Localización, Formato, Legalidad, Estabilidad y Granularidad, y sobre él se pueden realizar los siguientes procesos: Digitalización, Preservación Digital y Definir Metadatos. Por su parte, una biblioteca digital se caracteriza por los siguientes atributos: Disponibilidad, Recuperación, Autenticidad, Utilización, Asequibilidad y Tecnología.

En la Fig. 1 se presenta el diagrama conceptual de los objetos digitales modelados desde el punto de

vista de las bibliotecas digitales, es decir, no todos los atributos de un objeto digital se acomodan a lo que se describe acerca de una biblioteca, así pues, lo que se busca es determinar cuáles serían los atributos apropiados que puedan ser modelados semánticamente.

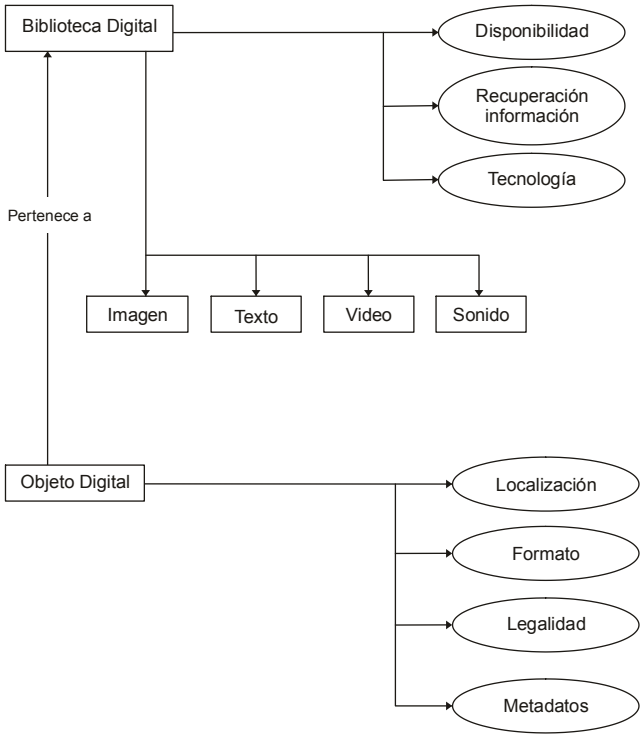

Fig. 1. Diagrama conceptual objeto digital en biblioteca digital

La metodología empleada es la *Ontology-101* [8], que es una de las más acogidas y facilita encontrar bastante documentación que sirva como guía para realizar un modelado propio. La metodología 101 propone los siguientes pasos:

Dominio y alcance de la ontología: El dominio de la ontología apunta a información referente a objetos digitales en bibliotecas digitales desde el punto de vista de sus propiedades, atributos, clasificación, representación y comunicación. La ontología se usará para realizar consultas semánticas en un posterior desarrollo de un módulo de consulta; la estructura de una consulta se fundamenta en las propiedades de cada uno de los objetos por consultar, en este caso específico, objetos de bibliotecas digitales. La ontología responderá a preguntas como: ¿los textos componen un tipo de objeto digital?, ¿los objetos digitales son considerados documentos?, ¿una imagen es un documento?, ¿qué lenguaje de consulta se aplica para los videos?, ¿qué tipo de consultas puedo realizar sobre las imágenes?

Considerar reutilización: Se realizó una búsqueda por palabras claves sobre la biblioteca de ontologías de DAML (DARPA Agent Markup Language) –disponible en http://www.daml.org/ontologies/–, que arrojó los siguientes resultados: Cuatro ontologías de documentos, de las cuales se eligió la ubicada en la URL http://www.daml.org/ontologies/132, dado que ofrece una descripción de un documento desde el punto de vista de tipo de escrito, sea publicación, tesis, informe o periódico. Se encontró la Ontología Dublín Core, especializada en el manejo de documentos; su principal aporte es el manejo de los derechos del documento, donde propone manejarlo

únicamente como un comentario, ya que si se restringe a un proceso automático puede que dependiendo del país se pueda o no trabajar con un documento.

Enumeración de términos: Los términos aquí enunciados refieren a un objeto digital desde el punto de vista de una biblioteca digital, estos son: objeto digital, origen, digital, físico, procesos, metadatos, digitalización, preservación digital, tipos, atributos, comunicación, localización, restricciones, protocolos, naturaleza, compuesta, heterogéneos, complejos, dinámicos, textos, bases de datos, imágenes, video, disponibilidad, utilización, protección, tecnología, arquitectura, editores, recursos, métodos búsqueda, sonido, imagen, combinados, derechos, tesis, periódico, tesis, informe.

Establecer clases y jerarquía de clases: Para este paso y los posteriores se hará uso de la herramienta Protégé, elegida debido a su relación con la metodología, ya que desde su creación por parte de la Universidad de Stanford es un complemento entre sí. A lo anterior se presentará un pantallazo de la herramienta y su aplicación para establecer la jerarquía de clases según la enumeración de términos (ver Fig. 2).

Fig. 2. Jerarquía de clases en Protégé

Debido a que hay que utilizar una herramienta que determina un lenguaje de representación, se ha elegido RDF y RDF-SCHEMA; el primero para representar las instancias, y el segundo para las clases que las definen.

B. Selección del lenguaje de consulta

Con el fin de determinar el lenguaje de recuperación de información apropiado en ambientes de bibliotecas digitales con características semánticas, se propone en este documento un estudio comparativo desde el punto de expresividad de los siguientes lenguajes de consulta y recuperación: XPath, XQuery, XQL, RQL, SPARQL, SeRQL y OWL-QL. Los criterios de evaluación se orientan al manejo y manipulación de los datos, al análisis de los grafos de representación del conocimiento y a las inferencias que se realicen sobre estos; tales criterios son: expresiones condicionales, operadores matemáticos, manejo de clases y objetos, análisis de atributos mediante sus rangos y dominios, determinación de rutas entre nodos que representan los conceptos, y criterios orientados al manejo de las expresiones relacionadas con dichos conceptos.

XQuery [9], también conocido como XML Query, es un leguaje de consultas estándar que utiliza la notación XML para definir consultas y manejar los resultados; se centra en encontrar y extraer elementos y atributos de documentos XML. XQuery está definido en términos de un modelo formal abstracto, no en términos de texto XML. Cada entrada a una consulta es una instancia de un modelo de datos, y la salida de una consulta también. El núcleo estructural que ofrece XQuery para realizar las expresiones se conoce como FLWOR, que es al XQuery lo que las distintas cláusulas dentro de una sentencia Select (select, from, where, etc.) son al SQL; el nombre viene de **F**or, **L**et, **W**here, **O**rder by y **R**eturn. A continuación se explica cada uno de estos bloques de consulta.

XQL [10] (XML Query Language) es una notación para obtener información de un documento; la información puede ser un conjunto de nodos o información sobre las relaciones entre nodos o valores derivados. XQL es una extensión natural del sistema de patrones XSL; es un lenguaje que ofrece la posibilidad de realizar consultas flexibles para extraer datos de documentos XML en la Web; se basa en operadores de búsqueda de un modelo de datos para documentos XML, que puede realizar consultas en infinidad de tipos de documentos, como son documentos estructurados, colecciones de documentos, bases de datos, estructuras DOM [11], catálogos, etc.

SPARQL [12] (pronunciado "sparkle") es un lenguaje de recuperación basado en RDF; su nombre es un acrónimo recursivo del inglés SPARQL Protocol and RDF Query Language. Se trata de una recomendación para crear un lenguaje de consulta dentro de la Web Semántica que está ya implementada en muchos lenguajes y bases de datos. Similar a otros lenguajes de consultas sobre conjuntos de datos, SPARQL permite a los usuarios declarar específicamente las condiciones requeridas para los datos a ser recuperados, más que describir explícitamente los pasos orientados a la descripción de una ruta para recuperarlos.

RQL [12] es un lenguaje de consulta declarativo para RDF; explícitamente captura esta semántica en su diseño; fue desarrollado en el instituto ICS-FORTH, y su potencia semántica está basada en la evaluación de caminos de expresiones sobre grafos RDF. RQL permite el uso de variables tanto para denotar clases como propiedades, y consultar esquemas RDF y RDF-S [13], y descripciones RDF en una misma consulta. RQL está definido por medio de un conjunto de consultas básicas e iteradores que se permiten construir otras consultas a través de una composición funcional con base en la teoría planteada por OQL [14] (Bases de datos orientadas a objetos).

SeRQL [15] (Sesame RDF Query Language, pronunciado como "circle") es un lenguaje de recuperación para RDF/RDFS desarrollado por Aduna como parte del software Sesame; combina características de otros lenguajes (principalmente RQL, RDQL [16], N-Triples [17] y N3 [18]), y añade otras propias.

Con base en [19], se presenta en el Cuadro 1 un resumen de los lenguajes (primera columna) y los criterios de evaluación (primera fila), haciendo uso de las siguientes convenciones:

EC = Expresiones condicionales
CE = Cuantificadores existenciales
OM = Operadores matemáticos
CO = Clases y objetos
RD = Rango y dominio
RA = Recursos adyacentes
PR = Predicados sobre recursos
DR = Distancia entre recursos
RF = Reificación

Cuadro 1
Comparación de lenguajes

Lenguaje/Criterios	EC	CE	OM	CO	RD	RA	PR	DR	RF
XQUERY	+	+	+	-	-	-	-	-	+-
XQL	+	+	+	-	-	-	-	-	+-
SPARQL	+	+-	+	+-	+-	+	+	+	+
RQL	+	+-	+	+-	+-	+	+	+	+
SeRQL	+	+-	+	+-	+-	+	+	+	+

La ventaja de los lenguajes que realizan consultas sobre documentos RDF consiste en que se complementan efectivamente con los modelos RDF, sin embargo, no permiten hacer consultas de tipo semántico, en el sentido de que no tienen por qué basarse necesariamente en elementos (conceptos, atributos y relaciones) de una ontología, sino exclusivamente en el modelo RDF. No obstante, el que presenta un apropiado comportamiento es el RQL.

III. Modelo propuesto

A. Representación

Para representar los objetos digitales se empleó una representación basada en tripletas semánticas propuestas por el modelo RDF, las cuales se componen de un Sujeto, un Predicado y un Objeto o Literal, como se aprecia en la Fig. 3.

En la Fig. 4 se presenta un ejemplo básico de dos tripletas RDF. La primera se compone del sujeto, que es el recurso http://sitioweb.com/jegiraldo, con propiedad *Creador* y con sujeto el recurso http://www.ingenieriainformatica.tk. La segunda tripleta es la que existe entre el recurso y el literal "colombiano", que están unidos por la propiedad *nacionalidad*. De esta misma manera se representaron los objetos digitales; basados en la ontología de dominio se propone la siguiente representación. El objeto digital, mediante su código de representación única, sería el *Objeto* de la tripleta, mientras cada uno de sus atributos, representados por recurso en la ontología, es un valor literal. Lo anterior se ilustra en la Fig. 4.

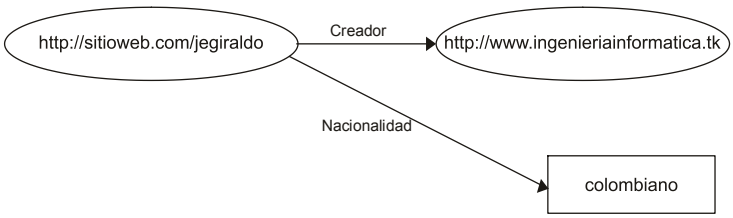

Fig. 3. Tripleta semántica

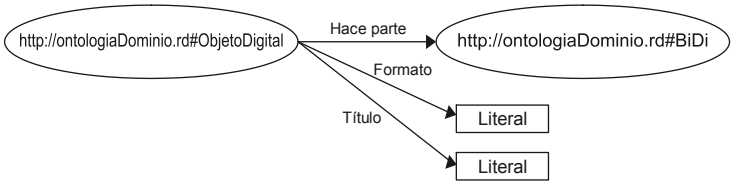

Fig. 4. Tripletas semánticas de objeto digital

B. Búsqueda

Para efectos de la búsqueda, el sistema permitirá consultar clases a partir de literales ingresados; así, el sistema pide al usuario el valor literal de determinada propiedad y se busca la coincidencia; después de encontrarla se continúa con la exposición de los demás atributos, entre ellos la ubicación física.

Los tipos de búsqueda posible son:

Búsqueda por atributos. El usuario selecciona el atributo del objeto digital, por ejemplo: título o autor; el sistema retorna un atributo específico de la clase, cuya propiedad tiene el valor ingresado como literal en la propiedad seleccionada. En la sección de implementación se darán detalles de los atributos retornados.

Búsqueda por palabra clave. En este caso el usuario digita en el campo de búsqueda una palabra, pero sin relacionarla con algún atributo; el sistema retorna un atributo específico. En la sección de implementación se darán detalles de los atributos retornados.

Búsqueda por palabra clave y atributos. Esta búsqueda se especializa por buscar en los valores de los atributos, en los nombres de los atributos y en los nombres de las clases, retornando en varios casos la clase (sujeto) con algunos atributos.

C. Arquitectura propuesta

En la Fig. 5 se expone la arquitectura general de la plataforma computacional que da soporte a la validación del modelo propuesto.

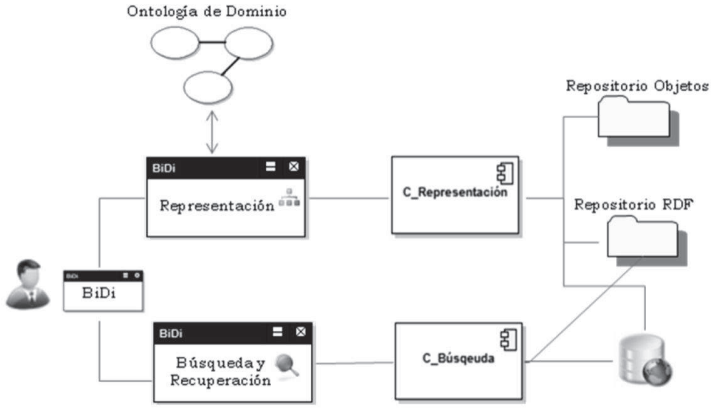

Figura 5. Arquitectura del sistema

La arquitectura está compuesta por las interfaces gráficas de usuario para la representación (carga de documentos) y la búsqueda y recuperación; este formulario presentará la respuesta con base en los atributos Título, Autor, Ubicación y Descripción, así mismo, tendrá la oportunidad de visualizar los metadatos relacionados con el objeto recuperado.

Cada interfaz se conecta con su respectiva capa de proceso. El componente de representación recibe el objeto cargado y construye automáticamente una descripción semántica del objeto y la escribe en RDF. Esta actividad se complementa con la carga del objeto digital y su almacenamiento en un repositorio.

Aunque en la descripción semántica es posible almacenar los pares documento RDF y objeto digital, se recomienda que para efectos de la organización, el registro del almacenamiento de los objetos y, especialmente, el soporte a la persistencia de los datos se emplee un sistema de base de datos en donde los registros de las bases de datos se cargan con las sentencias generadas, y así las consultas se pueden optimizar. Sin embargo, se realiza una validación

mediante consultas semánticas directamente en los archivos RDF, así pues, se abordan los metadatos directamente en el archivo que contiene su descripción por medio de RDF.

IV. Aspectos de implementación

La aplicación se desarrolló con el lenguaje para la Web PHP, empleando la biblioteca especializada para RDF – RAP (RDF API PHP). La interfaz de usuario admite el ingreso de los atributos básicos de los objetos digitales, permitiendo la construcción de las tripletas semánticas a partir de ello, y también por información que se captura a partir de la información propia del archivo que se carga.

La aplicación soporta todo tipo de objetos digitales, dado que simplemente les hace una descripción semántica en términos de tripletas, permitiendo la variedad de objetos; así mismo, los objetos se categorizan dependiendo de la extensión del archivo. Por cada objeto que se carga por medio del sistema, es almacenado el archivo primitivo y se construye automáticamente una representación en RDF, que también es almacenada. En la Fig. 6 se expone la interfaz principal, que aún se encuentra en desarrollo.

Fig. 6. Biblioteca Digital – BiDI_API

V. Conclusiones y trabajo futuro

El uso de la semántica para representar objetos permitirá una apropiada búsqueda, ya que se emplearán los conceptos, más que la sintaxis. Desarrollar la aplicación con soporte en la Web facilita el acceso de las personas y no se depende de tecnologías propietarias o complicadas para su instalación o soporte.

Como trabajo futuro se tiene pensado mejorar el sistema de almacenamiento de los archivos, pensando en una solución mixta entre un gestor de bases de datos y los archivo en RDF que representan los objetos; del mismo modo, se espera refinar las consultas necesarias en términos de tripletas semánticas, específicamente por especialidad de atributo.

Referencias

[1] T. Bray, *Extensible Markup Language (XML) 1.0, World Wide Web Consortium* October 2000. Disponible en: http://www.w3.org/TR/REC-xml.

[2] T. Berners-Lee, J. Hendler y O. Lassila, *The Semantic Web*. Scientific American.com. May 17, 2001.

[3] S. Alotaibi, "Semantic Web Technologies for Digital Libraries: From Libraries to Social Semantic Digital Libraries (SSDL), Over Semantic Digital Libraries (SDL)". In: *The 4th Saudi International Conference*, Friday 30 and Saturday 31 July 2010, The University of Manchester, UK.

[4] L. Espino, *Sistema de gestión del conocimiento basado en una ontología, aplicado a rutas turísticas*. Ontoguate. Universidad San Carlos de Guatemala. 2008.

[5] S. Kruk, R. Haslhofer, B. Kneževic, *Tutorial 7-Semantic Digital Libraries*, JCDL 2007.

[6] BRICKS Project. *Building Resources for Integrated Cultural Knowledge Services*. European Commission • Sixth Framework Programme. 2004.

[7] S. R. Kruk, S. Decker, L. Zieborak, *JeromeDL-Managing Digital Library Database with the Semantic Web Technologies, Proceedings DEXA2005*. Copenhagen, Denmark, August 2005.

[8] M. D. Sánchez, J. M. Cavero, E. Marcos, *Ontologías y MDA: una revisión de la literatura. DSDM '05*. II Taller sobre Desarrollo Dirigido por Modelos. MDA y aplicaciones. 2005.

[9] M. Marchiori, *XML Query (XQuery), World Wide Web Consortium,* 23 September 2003. Disponible en: http://www.w3.org/XML/Query.

[10] J. Robie, J. Lapp, y D. Schach. *XML Query Language (XQL). World Wide Web Consortium.* Disponible en: http://www.w3.org/TandS/QL/QL98/pp/xql.html.

[11] World Wide Web Consortium (W3C). *Document Object Model*. 2003. Disponible en: http://www.w3.org/dom.

[12] E. Prud'hommeaux y A. Seaborne, *SPARQL Query Language for RDF. World Wide Web Consortium (W3C)*. 2006. Disponible en: http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406.

[12] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis y M. Schol, "RQL: A Declarative Query Language for RDF". In *Proceedings of the Eleventh International World Wide Web Conference (WWW'02)*, USA, May 7-11 2002.

[13] D. Brickley, y R. Guha, *RDF vocabulary description language 1.0: RDF Schema. World Wide Web Consortium (W3C)*, 2003. Disponible en: http://www.w3.org/TR/rdf-schema.

[14] G. Cattell, D. Barry, M. Berler, J. Eastman, D. Jordan, D. Russell, O. Schadow, T. Stanienda y F. Vélez, *The Object Database Standard ODMG 3.0*. Morgan Kaufmann, January 2000.

[15] J. Broekstra y A. Kampman, *SeRQL: An RDF Query and Transformation Language*. Submitted to the International Semantic Web Conference, ISWC 2004.

[16] A. Andy Seaborne, *Rdql - A Query Language for RDF, W3C Member Submission*, January 2004. Disponible en: http://www.w3.org/Submission/2004/SUBM-RDQL-2004.

[17] M. Sintek and S. Decker, "TRIPLE - an RDF Query, Inference and Transformation Language". In *Deductive Databases and Knowledge Management (DDLP)*, 2001.

[18] T. Berners-Lee, *Notation 3. 2001*. Disponible en: http://www.w3.org/DesignIssues/Notation3.

[19] J. Giraldo, S. Mateus y M. Ruiz, "Lenguajes de recuperación de información sobre la Web Semántica". *Revista Politécnica*, enero-junio de 2009, Año 5, n.° 8, pp. 39-46.