Sistemas de recuperación e internet

Metadescripción, procesamiento, webcrawling, técnicas de consulta avanzada, hacking documental y posicionamiento web

Manuel Blázquez Ochando

Monografías electrónicas
mblazquez.es

025.4.036:004 **BLA** **sis**	BLÁZQUEZ OCHANDO, Manuel Sistemas de recuperación e internet: metadescripción, procesamiento, webcrawling, técnicas de consulta avanzada, hacking documental y posicionamiento web / Manuel Blázquez Ochando .– Madrid: mblazquez.es, 2013. 133p. ; 21cm.– (Libros y manuales de la Documentación; 3) ISBN 978-84-695-7019-7 1. Biblioteconomía y Documentación 2. Recuperación de Información I. Título II. Series

UNIVERSIDAD COMPLUTENSE DE MADRID
Facultad de Ciencias de la Documentación

1ªed. enero 2013, Madrid

Publicado por mblazquez.es

ISBN 978-84-695-7019-7

Índice

1. Introducción 4
2. Dublin Core básico: principios y fundamentos 5
3. Dublin Core avanzado: etiquetado completo 11
4. Dublin Core: referencia de codificación 15
5. Dublin Core en RDF 23
6. MADS: metadatos para la descripción de autoridades 33
7. MODS: metadatos para la descripción de objetos bibliográficos 42
8. METS: metadatos para la descripción de metadatos 48
9. Lectura de metadatos: programas parser 53
10. Webmetría y análisis de páginas web 58
11. Técnicas de consulta dinámica GET en Google 66
12. Búsqueda con operadores avanzados y directorios de servidores 71
13. Extensión de consultas avanzadas y recuperación de volcados de datos 77
14. Tácticas de posicionamiento web – SEO search engine optimization 80
15. Ejercicios prácticos 94

Práctica1. Metadatos y descripción Dublin Core 94
Práctica2. Descripción bibliográfica Dublin Core 98
Práctica3. Dublin Core RDF y generadores de metadatos 100
Práctica4. Descripción de autoridades MADS 102
Práctica5. Descripción bibliográfica con MODS 104
Práctica6. Análisis y recuperación parser de metadatos 105
Práctica7. Análisis webcrawler 107
Práctica8. Consultas dinámicas URL en Google 110
Práctica9. Operadores avanzados y directorios de servidores 114
Práctica10. Recuperación de volcados de datos 116
Práctica11. Configuración de archivo robots.txt y sitemap.xml 119

Práctica12. Cálculo de PageRank 122

16. Índice de tablas 125

17. Índice de figuras 127

18. Bibliografía y referencias 128

1. Introducción

La búsqueda y recuperación en Internet consta de métodos y técnicas complementarios a los empleados en el desarrollo de algoritmos en los motores de búsqueda, véase blog de la asignatura Técnicas Avanzadas de Recuperación de Información. En este sentido, se estudiarán con detenimiento todos los metadatos Dublin Core básicos y extendidos, para su aplicación en páginas web en forma de etiquetas embebidas dentro del código fuente de una página web y en formato RDF. Tales sistemas de meta-descripción favorecen los procesos de indexación y recuperación de cualquier página web, siendo considerados como uno de los factores que permite un posicionamiento en los ranking de resultados en los principales buscadores. Unido a este posicionamiento, se encuentran la webmetría y la cibermetría que estudian cuantitativa y cualitativamente las características de los sitios web y sus páginas, así como su nivel de enlazamiento, topografía y grafo correspondientes. Estas técnicas de estudio de la web son de utilidad para elaborar investigaciones que determinan la importancia de cada sitio web, así como para desarrollar una base de conocimiento útil para su explotación, mediante minería de datos, por ejemplo.

Pero también se consideran de importancia, las técnicas de consulta avanzadas por medio del protocolo REST, empleando variables dinámicas en la URL de consulta de los principales buscadores, que en muchos casos actúa con una enorme versatilidad para resolver problemas de búsqueda más especializados. En este sentido, el conocimiento de técnicas básicas de hacking, pueden facilitar la recuperación de información en directorios, la localización de documentos y versiones de páginas web que resultan de difícil acceso. En resumen, puede definirse la búsqueda en Internet, como un verdadero campo de pruebas en continua expansión, cuyas tácticas y métodos se mantienen en continuo cambio y progresión.

2. Dublin Core básico: principios y fundamentos

Qué es un metadato

El origen de los metadatos, se encuentra en el ámbito de la automatización y el desarrollo de bases de datos para la gestión de información. En los años 60 Jack Myers acuño el concepto metadato para referirse al conjunto de campos que permitían describir un producto para su puesta en circulación dentro del mercado. Esta primera aproximación fue de especial relevancia para aumentar y afinar el ámbito de aplicación al entorno web y más especialmente en el ámbito biblioteconómico y documental. se puede definir un metadato como "una descripción de modelos de descripción o catalogación de una serie de elementos, objetos, documentos e incluso etiquetas de descripción". Ésta definición tan amplia, proporciona una idea de la dificultad de acotación, téngase pues como ejemplos, los campos de un ficha catalográfica, los factores de análisis de un estudio comparado, los indicadores de riqueza de un país, los campos de tipificación y parameterización de campos de las bases de datos e incluso la estructura de un artículo científico, pueden ser considerados metadatos. En este sentido, cualquier sistema o convención que permita describir un dato, estructurarlo u organizarlo, es un metadato. Aledaña a esta concepción, se sitúa el concepto meta-información e incluso meta-conocimiento que hacen alusión a estadios de abstracción más elaborados, partiendo del dato original. Por estos motivos los profesionales de la información, así como la propia Documentación, deben estudiar qué métodos de meta descripción aplicar para cada caso. Según (GILLILAND, A.J.; GILL, T.; WHALEN, M.; WOODLEY, M.S. 2008) existen diversos contextos de aplicación de los metadatos, véase *tabla1*.

Contexto	Aplicación	Ejemplos
Administrativo	Gestión y administración de recursos de información	– Adquisición de información – Derechos y reproducción – Requerimientos legales para el acceso – Localización de información – Criterios de selección – Control de versiones
Descriptivo	Representación de recursos de información	– Registros catalográficos – Asistencia a la recuperación – Índices especializados – Relaciones hipertextuales

Preservación	Salvaguardar recursos de información	– Condiciones de uso – Estado de conservación – Medidas de preservación – Copias de seguridad de la información
Técnico	Funcionamiento de sistemas de información	– Documentación y ayuda de programas informáticos – Digitalización de la información – Autentificación y datos de seguridad – Control de tiempo de respuesta
Uso	Nivel y tipo de uso de los recursos informativos	– Información de versiones – Reutilización de la información

Tabla 1. Contextos de aplicación de los metadatos

Un ejemplo de la importancia de los metadatos es la consideración como tal, de las principales normas de descripción y clasificación utilizadas en Biblioteconomía y Documentación, como se demuestra en la siguiente comparativa de metadatos normalizados, en los que se identifica claramente la CDWA, CCO, VRA, MARC, MODS, Dublin Core, DACS, EAD, OBJECT ID, CIMI, FDA.

Meta-etiquetas en la Web

Los metadatos y las meta-etiquetas han sido confundidos en múltiples ocasiones, debido a que muchos metadatos son introducidos en el código fuente de las páginas web, mediante esta solución. Las meta-etiquetas son etiquetas <meta> HTML multipropósito anidadas entre las etiquetas de cabecera <head></head>. Constituyen en definitiva el soporte de los metadatos que se emplearán en cualquier tipo de descripción según contextos técnicos, meta-descriptivos, de uso, etc. La estructura sintáctica de las mismas puede observarse en la siguiente *tabla2*.

Meta-etiqueta	Ejemplo
Sintaxis básica	<meta name="" content="">
Técnica – Tipo de contenido y codificación	<meta http-equiv='content-type' content='text/html; charset=UTF-8'/>
Técnica – Control de cache en página web	<meta http-equiv='pragma' content='no-cache'/>
	<meta http-equiv='cache-control' content='no-cache'/>
Descriptivo – Esquema de ISBN, autoría, derechos, palabras clave, fecha de publicación, descripción	<meta scheme="ISBN" name="identifier" content="84-8181-494-1">
	<meta name="author" content="Nombre Apellidos">
	<meta name="copyright" content="© 2012 Autor">

	<meta name="keywords" content="palabras, clave">
	<meta name="date" content="2012-06T08:49:37+00:00">
	<meta name="description" content="descripción del recurso">

Tabla 2. Principales meta-etiquetas

El origen de las meta-etiquetas, comienza con el lenguaje de marcado HTML en 1999, cuando se concibe un método para la descripción sencilla de páginas web. Las especificaciones oficiales de las meta-etiquetas elaboradas por el W3C Consortium, contemplan diversos usos para la etiqueta meta, más allá del carácter descriptivo, como la instrucción "*cache-control*" que permite que el contenido de la página web no se guarde en la memoria cache del navegador del cliente, la instrucción "*content-type*" que identifica el tipo de página web editada y su correspondiente codificación con un set de caracteres determinado. En tales casos, aparte de describir una serie de características técnicas, la meta-etiqueta indica unos patrones de edición y comportamiento para el funcionamiento de la web. Por otro lado, también se contemplan meta-etiquetas de tipo "*meta-descriptivo*" del sitio web. Las convenidas internacionalmente y más normalizadas son "*author, copyright, keywords, date, description*", aunque es posible encontrar otras carentes de normalización, incluso especiales, diseñadas ad-hoc. Por este motivo, las meta-etiquetas HTML tienen limitaciones importantes frente a lo que se consideran verdaderos metadatos de aplicación profesional y documental, como resulta Dublin Core.

Metadatos Dublin Core

Los metadatos más utilizados y conocidos para la meta-descripción de recursos, sitios y páginas web son los metadatos Dublin Core, desarrollados por el consorcio DCMI, del que forman parte instituciones de gran relevancia para la Biblioteconomía y Documentación a nivel mundial como la JISC o las bibliotecas nacionales de Nueva Zelanda, Finlandia, Singapur o Korea entre otros. Los metadatos Dublin Core se basan en la idea de los niveles de interoperabilidad con el objetivo de conseguir el mayor grado de normalización y correspondencia entre las metadescripciones de los recursos de diversas instituciones.

- *Nivel 1. Definiciones de términos compartidas*. Los metadatos empleados son compartidos y comunes en sus definiciones para cualquier aplicación de

descripción utilizada en las instituciones mediante federación o acuerdo de especificaciones. Este nivel considera el uso de las estructuras de meta-etiquetas para introducir los metadatos correspondientes para la descripción.

– *Nivel 2. Interoperabilidad semántica*. Se basa en el empleo de metadatos mediante archivos RDF, aprovechando la teoría del "Linked Data" o lo que es lo mismo la identificación de y descripción de recursos, páginas web, documentos, etc, usando URIs unívocas de cada uno de ellos. Ello permite llevar los metadatos a un nivel de interoperabilidad con motores de búsqueda semánticos.

– *Nivel 3. Interoperabilidad de la descripción sintáctica*. Implica que existan aplicaciones informáticas compatibles con el modelo "Linked Data" y que por lo tanto utilicen el esquema de metadatos mediante RDF.

– *Nivel 4. Interoperabilidad de los perfiles de descripción*. Supone la suma de todos los recursos descritos mediante RDF, aprovechados para la recuperación mediante inferencia semántica e inteligencia artificial.

Usando metadatos Dublin Core en formato XHTML/HTML

Un requisito imprescindible para introducir metadatos Dublin Core en el código fuente de un documento HTML/XHTML, es la identificación del "*framework*" o marco de trabajo para los prefijos DC que se emplearán en la notación de los metadatos. Por este motivo la estructura del sitio o página web deberá ser similar a la reseñada en la *tabla3 y 4*, véase además especificaciones de implementación sintáctica.

Código fuente XHTML

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
        <head profile="http://dublincore.org/documents/2008/08/04/dc-html/">
                <title>Universidad Complutense de Madrid</title>
                <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/"/>
                <meta name="DC.title" content="Universidad Complutense de Madrid"/>
        </head>
<body>
</body>
</html>
```

Tabla 3. Estructura XHTML para la identificación del marco de trabajo Dublin Core

Código fuente HTML

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html>
        <head profile="http://dublincore.org/documents/2008/08/04/dc-html/">
                <title>Universidad Complutense de Madrid</title>
                <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/"/>
                <meta name="DC.title" content="Universidad Complutense de Madrid"/>
        </head>
<body>
</body>
</html>
```

Tabla 4. Estructura HTML para la identificación del marco de trabajo Dublin Core

Una vez preparada la identificación de los metadatos Dublin Core, se emplea la conocida estructura de meta-etiquetas para introducir el conjunto de metadatos normalizados en lo que se denomina el "*Metadata Element Set*". Éstas especificaciones indican cuáles son las denominaciones de los 15 términos más normalizados que se emplean a modo de metadato dentro de los atributos "*name*" de las meta-etiquetas, véase *tabla5*.

Etiqueta	Modo de empleo	Descripción
Contributor	DC.Contributor	Entidad o persona física responsable de las colaboraciones documentales e informativas en un recurso determinado.
http://purl.org/dc/elements/1.1/contributor		
Coverage	DC.Coverage	Cobertura espacial o temporal de un recurso, incluyendo aspectos de jurisdicción, siempre que sea de relevancia para la descripción del mismo.
http://purl.org/dc/elements/1.1/coverage		
Creator	DC.Creator	Entidad o persona física responsable en primera instancia de la creación de un recurso.
http://purl.org/dc/elements/1.1/creator		
Date	DC.Date	Punto temporal o periodo de tiempo asociado al recurso, así como su ciclo de vida.
http://purl.org/dc/elements/1.1/date		
Description	DC.Description	Contenidos del recurso de información y documentación.
http://purl.org/dc/elements/1.1/description		
Format	DC.Format	Formato del archivo correspondiente al recurso, su medio o soporte físico, dimensiones, etc.
http://purl.org/dc/elements/1.1/format		
Identifier	DC.Identifier	Identificador unívoco para la desambiguación del recurso mediante URI, para un determinado contexto dado.
http://purl.org/dc/elements/1.1/identifier		

Language	DC.Language	Idioma en el que se expresa el contenido del recurso de información y documentación.
http://purl.org/dc/elements/1.1/language		
Publisher	DC.Publisher	Entidad responsable de la publicación del recurso de información y documentación.
http://purl.org/dc/elements/1.1/publisher		
Relation	DC.Relation	Recursos relacionados o vinculados con el presente en la descripción.
http://purl.org/dc/elements/1.1/relation		
Rights	DC.Rights	Derechos de acceso y explotación del recurso.
http://purl.org/dc/elements/1.1/rights		
Source	DC.Source	Fuente de información de la que es deudor el recurso analizado.
http://purl.org/dc/elements/1.1/source		
Subject	DC.Subject	El asunto o temática del recurso analizado.
http://purl.org/dc/elements/1.1/subject		
Title	DC.Title	Título propiamente dicho del recurso.
http://purl.org/dc/elements/1.1/title		
Type	DC.Type	Naturaleza o género del documento que permite describir el formato del documento.
http://purl.org/dc/elements/1.1/type		

Tabla 5. Conjunto de elementos básicos de Dublin Core

3. Dublin Core avanzado: etiquetado completo

En Dublin Core, existen términos para el refinamiento de los metadatos, que amplían sobremanera las posibilidades de los 15 elementos de descripción original, según se comprobó en el artículo anterior. Tales refinamientos pueden ser consultados en el glosario o vocabulario de términos, clases y elementos de Dublin Core disponible en http://dublincore.org/documents/2012/06/14/dcmi-terms/. Los refinamientos y vocabularios extendidos de Dublin Core, tienen un método de articulación o sintáxis de aplicación muy concreta, que responde a una casuística de uso ya establecida contemplada en su normativa de expresión http://dublincore.org/documents/dcq-html/. En este sentido las reglas de construcción de Dublin Core distinguen los siguientes elementos dentro de su vocabulario de cara a su correcta codificación.

- *Recurso.* Es un objeto o elemento que tiene identidad propia, como por ejemplo un documento o página web.

- *Propiedad.* Es una característica, atributo o relación que se utiliza para describir un recurso determinado.

- *Registro.* Son metadatos que estructuran la información de un recurso en una o más propiedades. Se puede considerar un conjunto de propiedades que sirve para describir un recurso.

- *Valor.* Es el contenido que toma un determinado metadato.

- Etiqueta meta. Etiqueta HTML utilizada para embeber la notación del metadato y su contenido.

- *Atributo name.* Atributo de la etiqueta <meta> que contiene la notación del metadato propiamente dicho.

- *Atributo content.* Atributo de la etiqueta <meta> que contiene el valor, dato o contenido del metadato.

- *Atributo scheme.* Atributo de la etiqueta <meta> que determina el esquema o patrón que adquieren los valores o contenidos de un determinado metadato, normalmente éstos serán normalizados y son estándares internacionales de descripción.

- *Atributo lang o hreflang.* Atributos de la etiqueta <meta> o <link> empleados para identificar el idioma o lengua en el que se expresa el contenido o descripción del metadato.

- *Namespace.* Espacio de nombres o marco de trabajo por el que los metadatos Dublin Core expresados en un determinado código fuente, cumplen con las especificaciones oficiales establecidas.

La codificación básica de metadatos, puede verse perfectamente definida en la *tabla6*, distinguiendo perfectamente la etiqueta <meta> y sus atributos "name" y "content". Éstos son los principales componentes por los que se articulan.

```
<meta name="DC.Date" content="Valor"/>
<meta name="DC.Date" content="2012-10-08"/>
```

Tabla 6. Construcción básica de metadatos

No obstante, los refinamientos de Dublin Core, véase *tabla7*, permiten utilizar un vocabulario mucho más extenso que hace superar el medio centenar de metadatos disponibles. Para hacer uso de este vocabulario es necesario utilizar el prefijo "*DCTERMS*" seguido de punto y el término del vocabulario de metadatos previstos en http://dublincore.org/documents/2012/06/14/dcmi-terms.

```
<meta name="DCTERMS.modified" content="2012-10-09"/>
```

Tabla 7. Refinamientos con los términos Dublin Core

Los esquemas de codificación en Dublin Core constituyen otra de las características adicionales que le confieren gran capacidad de normalización. Para hacer uso de ello, se emplea el atributo "*scheme*" y se define el prefijo de refinamiento adecuado. Resulta

significativa la presencia del término "*UDC*" referido a la Clasificación Decimal Universal, véase *tabla8*.

```
<meta name="DC.date" scheme="DCTERMS.W3CDTF" content="2012-09-23"/>
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text"/>
<meta name="DC.subject" scheme="DCTERMS.UDC" content="02"/>
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Software"/>
```

Tabla 8. Esquemas de codificación de los valores o contenidos

Una de las características que permiten la interrelación de recursos mediante metadatos es el atributo "*rel*" en etiquetas "*link*", tal como puede observarse en la *tabla9*.

```
<link rel="DC.relation" href="http://www.ucm.es/"/>
<link rel="DCTERMS.references" href="http://www.ucm.es/"/>
```

Tabla 9. Enlazar otros recursos

Las posibilidades de enlazar recursos relacionados, permite a la postre identificar la tipología de enlace que se define a continuación de la notación del metadato y viene especificado por los tipos de enlaces de las especificaciones oficiales de HTML4 del W3C Consortium, véase *tabla10*.

```
<link rel="DC.rights Copyright" href="http://www.ucm.es/copyright.html"/>
<link rel="DCTERMS.tableOfContents Contents" href="http://www.ucm.es/contents.html"/>
```

Tabla 10. Enlazar recursos especificando el tipo de enlaces
http://www.w3.org/TR/html4/types.html#type-links

Otro atributo de importancia para su empleo en etiquetas <*meta*> y <*link*> es el correspondiente al idioma cuyo valor estará conforme a la norma ISO639-1. En caso contrario deberá ser especificado con el atributo "*scheme*" que identifique la norma que estandariza la notación de idiomas correspondiente, véase *tabla11*.

```
<meta name="DC.subject" lang="es" content="temática"/>
<link rel="DCTERMS.references" hreflang="es" href="http://www.ucm.es/contents.html"/>
```

Tabla 11. Especificar idioma o lengua en el que está escrito un recurso o la descripción del metadato. Empleando la norma ISO639-1

Los metadatos pueden ser repetidos, según las necesidades de la descripción, tal como se muestra a continuación en la *tabla12*. Esto posibilita la introducción de varios títulos, formas alternativas de título, títulos atribuidos, subtítulos o subtítulos paralelos.

```
<meta name=”DC.title” lang=”es” content=”El Documentalista Audaz”/>
<meta name=”DC.title” lang=”es” content=”Cuan terrible es lo que los documentalistas guardan en sus portafolios”/>
```

Tabla 12. Los metadatos pueden ser repetidos si es necesario aludir a distintas formas o descripciones

En Dublin Core al igual que en otros vocabularios y lenguajes de marcado, se requiere la identificación de los “*namespaces*”, también conocidos como espacio de nombres o marco de trabajo. Su presencia o ausencia, no altera el correcto funcionamiento de los metadatos que fueren codificados. No obstante, marca la diferencia a la hora de determinar la correcta validación de los mismos, de acuerdo a las normas y reglas de construcción sintáctica, por lo que resulta obligada su implantación y uso, véase *tabla13*.

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
```

Tabla 13. Enlaces para utilizar los nombres de espacio (Namespaces) en Dublin Core

Finalmente, la *tabla14*, muestra cómo es posible emplear otros esquemas de metadatos distintos a Dublin Core y combinarlos sin interferencia o incompatibilidad alguna. En este caso el ejemplo muestra la introducción del estándar de metadatos *AGLS* del gobierno Australiano, para la descripción e identificación de documentación producida por su administración pública.

```
1º se define el Namespace de los metadatos a utilizar
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<link rel="schema.AGLS" href="http://www.naa.gov.au/recordkeeping/gov_online/agls/1.2" />

2º se emplean normalmente
<meta name="DC.title" content="Services to Government" />
<link rel=”DCTERMS.references” hreflang=”es” href=”http://www.ucm.es/contents.html”/>
<meta name="AGLS.Function" scheme="AGIFT" content="recordkeeping standards" />
```

Tabla 14. Utilizar otros metadatos con otros esquemas

4. Dublin Core: referencia de codificación

Para un mayor conocimiento de los elementos y términos de los metadatos Dublin Core, se presenta la siguiente tabla que reúne todas las notaciones utilizadas en la codificación y que resume todos los aspectos de la descripción de recursos, previstos para Dublin Core. Estos recursos pueden ser páginas web, monografías, artículos científicos, publicaciones periódicas, materiales didácticos, catálogos, documentos bibliográficos, etc.

Tipo de elemento	Prefijo	Elemento / Término	Notación
Dublin Core básico (elementos del Namespace)Ver referencia	DC.	**contributor** – Entidad o persona física responsable de las colaboraciones documentales e informativas en un recurso determinado.	DC.contributor
		coverage – Cobertura espacial o temporal de un recurso, incluyendo aspectos de jurisdicción, siempre que sea de relevancia para la descripción del mismo.	DC.coverage
		creator – Entidad o persona física responsable en primera instancia de la creación de un recurso.	DC.creator
		date – Punto temporal o periodo de tiempo asociado al recurso, así como su ciclo de vida.	DC.date
		description – Contenidos del recurso de información y documentación.	DC.description
		format – Formato del archivo correspondiente al recurso, su medio o soporte físico, dimensiones, etc.	DC.format
		identifier – Identificador unívoco para la desambiguación del recurso mediante URI, para un determinado contexto dado.	DC.identifier
		language – Idioma en el que se expresa el contenido del recurso de información y documentación.	DC.language
		publisher – Entidad responsable de la publicación del recurso de información y documentación.	DC.publisher
		relation – Recursos relacionados o vinculados con el presente en la descripción.	DC.relation
		rights – Derechos de acceso y explotación	DC.rights

		del recurso.	
		source – Fuente de información de la que es deudor el recurso analizado.	DC.source
		subject – El asunto o temática del recurso analizado.	DC.subject
		title – Título propiamente dicho del recurso.	DC.title
		type – Naturaleza o género del documento que permite describir el formato del documento.	DC.type
Dublin Core avanzado (términos del Namespace) Ver referencia	DCTERMS.	**abstract** – Resumen informativo, analítico, descriptivo, sintético o documental del recurso o documento.	DCTERMS.abstract
		accessRights – Derechos de acceso, indicaciones o información sobre las restricciones, privacidad o políticas de seguridad para acceder a la información del recurso.	DCTERMS.accessRights
		accrualMethod – Método de adquisición del recurso por el que es anexionado a la colección. (Donación, Intercambio…)	DCTERMS.accrualMethod
		accrualPeriodicity – Periodo temporal por el que el recurso o documento ha sido adquirido.	DCTERMS.accrualPeriodicity
		accrualPolicy – Reglas, derechos y normas a las que está sujeta la adquisición del recurso o documento.	DCTERMS.accrualPolicy
		alternative – Título o nombre alternativo del recurso o documento. Por ejemplo, el subtítulo u otras formas del título.	DCTERMS.alternative
		audience – Público objetivo al que está destinado el recurso o documento.	DCTERMS.audience
		available – Fecha de disponibilidad del recurso o documento	DCTERMS.available
		bibliographicCitation – Referencias bibliográficas del documento. Se recomienda el empleo del estilo y elementos necesarios para la desambiguación de las obras referenciadas.	DCTERMS.bibliographicCitation
		conformsTo – Implica la conformidad del recurso o documento de acuerdo a normas, reglas o estándares.	DCTERMS.conformsTo
		contributor – Entidad o persona física	DCTERMS.contributor

		responsable de las colaboraciones documentales e informativas en un recurso determinado.	
		coverage – Cobertura espacial o temporal de un recurso, incluyendo aspectos de jurisdicción, siempre que sea de relevancia para la descripción del mismo.	DCTERMS.coverage
		created – Fecha de creación del recurso	DCTERMS.created
		creator – Entidad o persona física responsable en primera instancia de la creación de un recurso.	DCTERMS.creator
		date – Punto temporal o periodo de tiempo asociado al recurso, así como su ciclo de vida.	DCTERMS.date
		dateAccepted – Fecha de aceptación del recurso o documento. De aplicación directa en los artículos de revistas científicas.	DCTERMS.dateAccepted
		dateCopyrighted – Fecha de copyright	DCTERMS.dateCopyrighted
		dateSubmitted – Fecha de envío del documento o recurso. De aplicación directa en los artículos remitidos a las revistas científicas.	DCTERMS.dateSubmitted
		description – Contenidos del recurso de información y documentación.	DCTERMS.description
		educationLevel – definición del nivel de estudios o educativo al que está orientado el contenido del recurso.	DCTERMS.educationLevel
		extent – Tamaño, extensión o duración del recurso o documento.	DCTERMS.extent
		format – Formato del archivo, medio físico, dimensiones del recurso.	DCTERMS.format
		hasFormat – Indicación de que el recurso vinculado con este metadato tiene el mismo contenido pero con un formato distinto.	DCTERMS.hasFormat
		hasPart – Indicación de que el recurso vinculado con este metadato está incluido o forma parte del que se está describiendo. Dicho de otra forma, el recurso tiene una o varias partes en el recurso vinculado.	DCTERMS.hasPart
		hasVersion – El recurso o documento vinculado es una versión, edición o adaptación del documento que está siendo descrito.	DCTERMS.hasVersion

		identifier – Identificador unívoco para la desambiguación del recurso mediante URI, para un determinado contexto dado.	DCTERMS.identifier
		instructionalMethod – Método de enseñanza, instrucción, presentación del documento. Por ejemplo, prácticas, trabajo, teoría, presentación, esquemas, etc.	DCTERMS.instructionalMethod
		isFormatOf – Recurso relacionado o vinculado cuyo contenido es idéntico al del documento descrito, menos en su formato.	DCTERMS.isFormatOf
		isPartOf – El recurso relacionado es parte del documento que está siendo descrito.	DCTERMS.isPartOf
		isReferencedBy – Documento o recurso que referencia, cita o describe el recurso que está siendo descrito.	DCTERMS.isReferencedBy
		isReplacedBy – El recurso vinculado reemplaza al que está siendo descrito. De aplicación en versiones de estándares y normas.	DCTERMS.isReplacedBy
		isRequiredBy – El recurso relacionado requiere descripción o soporte que proporciona el documento que se está describiendo. De aplicación en documentos técnicos, estándares y normas.	DCTERMS.isRequiredBy
		issued – Fecha de emisión o publicación del documento.	DCTERMS.issued
		isVersionOf – El recurso vinculado que describe el documento es una versión, edición o adaptación.	DCTERMS.isVersionOf
		language – Idioma en el que se expresa el contenido del recurso de información y documentación.	DCTERMS.language
		license – Recurso o documento vinculado con las normas legales para el empleo y uso de la información y contenidos.	DCTERMS.license
		mediator – Institución, persona física o privada que media en el acceso al recurso o documento.	DCTERMS.mediator
		medium – Material o soporte del recurso o documento.	DCTERMS.medium
		modified – Fecha de modificación del recurso o documento.	DCTERMS.modified

		provenance – Sujeto productor y cambios en la propiedad y custodia del documento. Es decir, procedencia, origen y seguimiento del documento a partir de sus propietarios y creadores.	DCTERMS.provenance
		publisher – Entidad responsable de la publicación del recurso de información y documentación.	DCTERMS.publisher
		references – Recurso o documento relacionado que es referenciado o citado por el documento que se está describiendo.	DCTERMS.references
		relation – Recursos relacionados o vinculados con el presente en la descripción.	DCTERMS.relation
		replaces – Recurso relacionado que es reemplazado por el documento que se está describiendo.	DCTERMS.replaces
		requires – Recurso relacionado, necesario para el documento que está siendo descrito, para completar sus contenidos, funciones o coherencia.	DCTERMS.requiresT
		rights – Derechos de acceso y explotación del recurso.	DCTERMS.rights
		rightsHolder – Persona o entidad encargada de la gestión de los derechos sobre el recurso o documento.	DCTERMS.rightsHolder
		source – Recurso relacionado del cual se deriva el documento que está siendo descrito.	DCTERMS.source
		spatial – Cobertura espacial, localizaciones topográficas.	DCTERMS.spatial
		subject – El asunto o temática del recurso analizado.	DCTERMS.subject
		tableOfContents – Lista de contenidos, epígrafes, capítulos del recurso o documento.	DCTERMS.tableOfContents
		temporal – Características temporales del recurso, cobertura temporal del documento.	DCTERMS.temporal
		title – Título propiamente dicho del recurso.	DCTERMS.title
		type – Naturaleza o género del documento que permite describir el formato del documento.	DCTERMS.type

		valid – Fecha de validación de un recurso, documento o artículo.	DCTERMS.valid
Esquemas de datos Dublin Core	DCTERMS.	**DCMIType** – Naturalezas o géneros del recurso definidos por el Dublin Core Metadata Initiative. Valores (Collection | Dataset | Event | Image | InteractiveResource | MovingImage | PhysicalObject | Service | Software | Sound | StillImage | Text).	DCTERMS.DCMIType
		DDC – Clasificación Decimal Dewey	DCTERMS.DDC
		IMT – Tipos de documento multimedia MIME. Valores (application | audio | example | image | message | model | multipart | text | video).	DCTERMS.IMT
		LCC – Clasificación de la Library of Congress.	DCTERMS.LCC
		LCSH – Clasificación de encabezamientos de materia de la Library of Congress.	DCTERMS.LCSH
		MESH – Clasificación médica de encabezamientos de material.	DCTERMS.MESH
		NLM – Vocabulario de clasificación de la National Library of Medicine.	DCTERMS.NLM
		TGN – Tesauro de Nombres Geográficos del instituto de investigación Getty.	DCTERMS.TGN
		UDC – Clasificación Decimal Universal.	DCTERMS.UDC

Tabla 15. Referencia de términos y elementos Dublin Core

Un conocimiento exhaustivo de todos los términos y elementos, posibilita que el profesional de la información sea capaz de codificar y describir mediante metadatos Dublin Core cualquier tipo de documento. Véase ejemplo de codificación en la *tabla16.*

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html>
<head profile="http://dublincore.org/documents/2008/08/04/dc-html/">
<title>BUCM Catálogo: Tipología documental en las estadísticas del ISBN</title>
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/"/>
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.title" lang="spa" content="Tipología documental en las estadísticas del ISBN"/>
<meta name="DC.title" lang="eng" content="Document type in the ISBN Statistics"/>
<meta name="DC.creator" content="Giordanino, Eduardo Pablo (egiordan@filo.ub.ar)"/>
<meta name="DCTERMS.provenance" content="Universidad de Buenos Aires"/>
```

```
<meta name="DCTERMS.alternate" content="Una perspectiva de la edición española"/>
<meta name="DCTERMS.abstract" lang="spa" content="Las tipologías documentales usadas en las agencias del ISBN poseen cierto grado de indefinición que afecta a las estadísticas de la producción del sector editorial español y argentino. Se estudian las definiciones de libro presentadas por la legislación de ambos países y la función del control bibliográfico nacional. Analiza las estadísticas de la Agencia Española y la Agencia Argentina del ISBN. Propone una nueva tipología que abarque las categorías de formatos y soportes en uso en el sector editorial digital."/>
<meta name="DC.subject" lang="spa" content="libro electrónico, ISBN, industria editorial, tipología documental"/>
<meta name="DCTERMS.abstract" lang="eng" content="This paper studies the documental types used in the ISBN system of Spain and Argentina. It first describes how new technologies affect the registry of the book production of the publishing sector. Thereafter it considers the book's definition of the legislation of both countries and the function of the national bibliographic control. Analyzes the statistic data of the Spanish and Argentine ISBN agencies about ebook production. Final remarks are related to the proposal of a new documental tipology for the registry of new formats and sources used in the digital publishing industry."/>
<meta name="DC.subject" lang="eng" content="ebook, ISBN, publishing industry, documental tipology"/>
<meta name="DC.description" content="El objetivo fundamental de este trabajo es analizar la tipología documental usada en las agencias de registro del sistema ISBN. A partir de las nuevas formas de edición surgen nuevos tipos de documentos, como los libros electrónicos, presentados en distintos formatos, fenómeno denominado como el "libro hipocondríaco" (Giordanino, 2010). Entre las nuevas formas de edición existe la modalidad de Print on demand (POD, impresión a pedido), no contemplada adecuadamente por las agencias de registro del ISBN, tanto en España (Sánchez Paso, 2004) como en Argentina. Se partirá de un análisis de las definiciones de los materiales editoriales presentes en la legislación y de un estudio de las estadísticas oficiales."/>
<meta name="DC.subject" scheme="DCTERMS.UDC" content="027.7(72)"/>
<meta name="DC.subject" scheme="DCTERMS.UDC" content="025.17"/>
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text"/>
<meta name="DC.publisher" content="Documentación de las Ciencias de la Información"/>
<link rel="DC.publisher" content="http://revistas.ucm.es/index.php/DCIN"/>
<meta name="DCTERMS.dateSubmitted" scheme="DCTERMS.W3CDTF" content="2011-02-03"/>
<meta name="DCTERMS.dateAccepted" scheme="DCTERMS.W3CDTF" content="2011-03-05"/>
<meta name="DC.language" scheme="DCTERMS.ISO639-2" content="spa"/>
<meta name="DC.rights" content="Copyright Universidad Complutense de Madrid 2011"/>
<link rel="DCTERMS.accrualPolicy" content="http://revistas.ucm.es/index.php/DCIN/about/editorialPolicies"/>
<link rel="DCTERMS.hasformat" content="http://revistas.ucm.es/index.php/DCIN/article/download/36456/35304"/>
<meta name="DC.identifier" content="http://dx.doi.org/10.5209/rev_DCIN.2011.v34.36456"/>
<meta name="DCTERMS.references" content="Agencia Internacional del ISBN. Manual del usuario del ISBN. Edición ajustada para Iberoamérica. 5ª ed. Bogotá: CERLALC, 2007."/>
<meta name="DCTERMS.references" content="BÉHAR, Patrick; Laurent Colombani y Sophie Krishnan. Les écrits à l'heure du numérique. Une étude Bain & Company pour le Forum d'Avignon - Culture, Economie, Média. Paris, 2010. http://www.forum-avignon.org/fr/edition-2010/publications. Consultado: 17/12/2010."/>
<meta name="DCTERMS.references" content="BECERRA, Martín; Pablo Hernández y Glenn Postolski. "La concentración de las industrias culturales". En: Héctor Schargorodsky, dir. Industrias culturales: mercado y políticas públicas en Argentina. Buenos Aires: Secretaría de Cultura de la Nación, Ediciones CICCUS, 2003."/>
<meta name="DCTERMS.extent" content="241-254p."/>
<meta name="DCTERMS.format" content="21cm"/>
<meta name="DCTERMS.format" content="text/html"/>
</head>
<body>

<!-- [Artículo o Documento completo] -->

</body>
</html>
```

Tabla 16. Ejemplo de aplicación de metadatos Dublin Core para la descripción de un artículo científico

Según se muestra en el ejemplo anterior, los metadatos Dublin Core permiten un alto nivel de descripción para los materiales bibliográficos, incluyendo los artículos

publicados por las revistas científicas, lo cual supone una adaptación de los campos de fecha para el control del estado de los escritos; por ejemplo la fecha de envío, recepción y aceptación del artículo. Por otro lado destaca la capacidad para introducir clasificaciones especializadas como la CDU, las políticas de la revista para la aceptación de artículos, la institución de procedencia y filiación del autor, las referencias bibliográficas, la extensión y distintos datos de los parámetros del documento.

5. Dublin Core en RDF

Los metadatos Dublin Core también pueden ser expresados mediante RDF (Resource Description Framework), también conocido como el principal modelo normalizado para el intercambio de datos en la web, con el que es posible elaborar la web semántica. RDF es un lenguaje de marcado derivado de XML, diseñado para relacionar recursos, objetos y elementos en la web. Un ejemplo del concepto de relaciones en RDF, podría ser la propia descripción catalográfica de un documento, véase *figura1*.

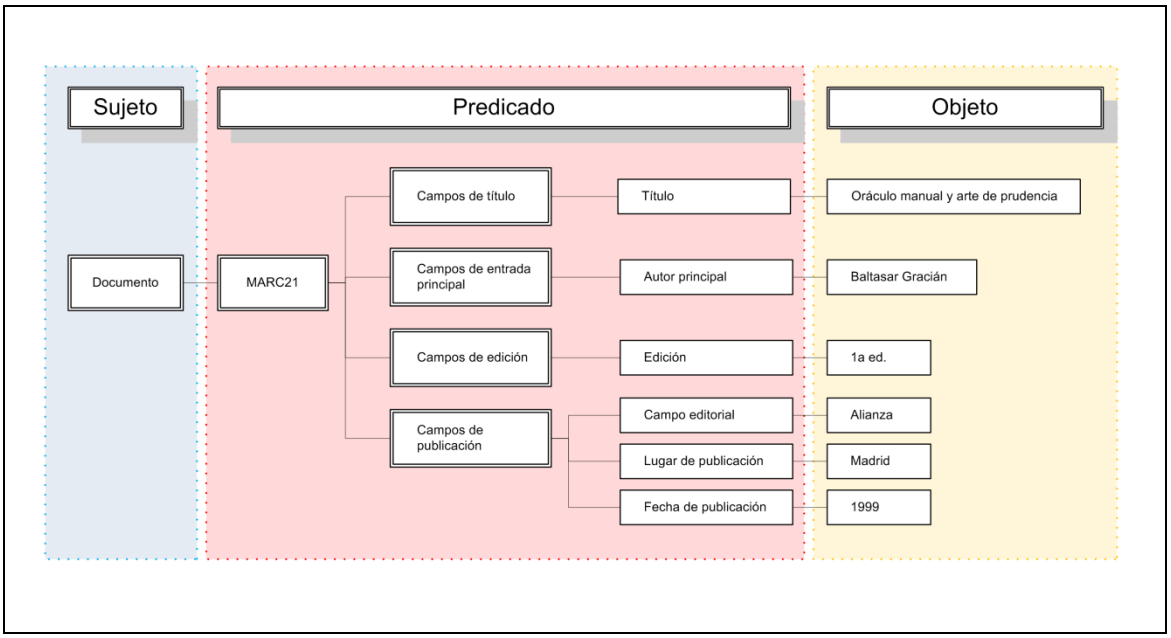


Figura 1. Relaciones entre un documento, sus metadatos de descripción y sus valores descriptivos. Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/esquemas/esquema001-rdf.png

En tal caso un documento puede ser descrito por una determinada norma, que consigna una serie de áreas de descripción, que a la vez contienen una serie de campos de descripción. Todos los elementos que intervienen en este proceso de descripción pueden ser expresados mediante relaciones que utilicen como base su propia URI identificativa, véase *figura2*. Esta teoría de enlazamiento a través de la identificación unívoca de los elementos, se la conoce con el nombre de teoría del Linked Data o teoría de los Datos Enlazados, elaborada por Tim Berners-Lee y presentada por primera vez en el Congreso TED en el año 2009.

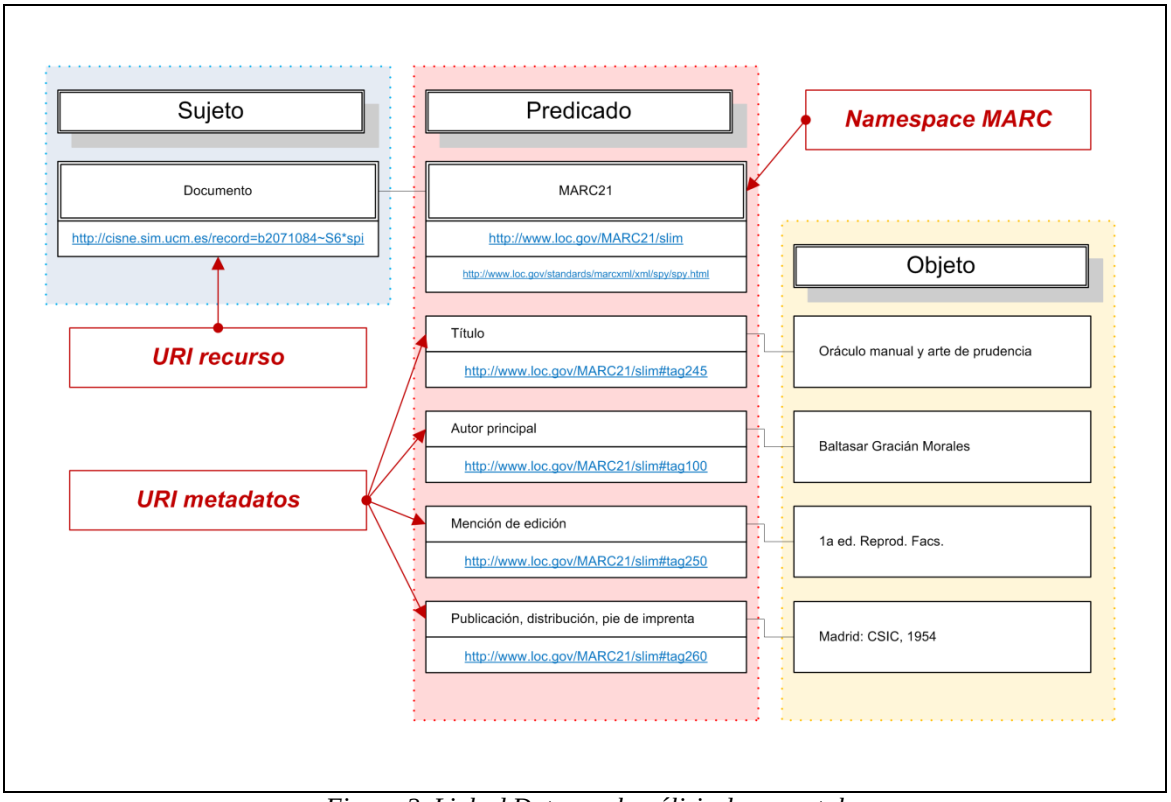


Figura 2. Linked Data en el análisis documental

Como puede observarse, el ejemplo de la *figura2*, muestra la URI (Uniform Resource Identifier) del recurso documental "*Oráculo manual y arte de prudencia*" que corresponde a la URL permanente del catálogo bibliográfico en línea de la Biblioteca Complutense para dicho documento. Por otro lado, los metadatos de descripción empleados proceden de las especificaciones oficiales para MARC21, identificadas por su namespace.

Se produce la circunstancia de que cada metadato o campo de descripción catalográfico consta de su correspondiente identificación URI, derivada de la URL del namespace de MARC21, lo que permite identificar todos los predicados que conforman aquello que se va a decir del sujeto identificado anteriormente. Los valores que describen al sujeto se denominan objetos, proporcionando una información o datos que permiten comprender las características del mismo. Este sencillo modelo de relación entre sujetos, predicados y objetos es la base sobre la que se asienta la web semántica, denominándolo en tal caso como estructura de triples o tripletes, por sus tres componentes principales.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:marc="http://www.loc.gov/MARC21/slim"
         xmlns:dcterms="http://www.purl.org/dc/terms/">

<rdf:Description rdf:about="http://cisne.sim.ucm.es/record=b2071084~S6*spi">
         <marc:tag245a dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd245.html"
rdf:object="Oráculo manual y arte de prudencia"/>
         <marc:tag100a dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd100.html"
rdf:object="Gracián y Morales, Baltasar"/>
         <marc:tag250a dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd250.html"
rdf:object="1a ed. Reprod. Facs."/>
         <marc:tag260a dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd260.html"
rdf:object="Madrid"/>
         <marc:tag260b dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd260.html"
rdf:object="CSIC"/>
         <marc:tag260c dcterms:instructionalMethod="http://www.loc.gov/marc/bibliographic/bd260.html"
rdf:object="1954"/>
</rdf:Description>

</rdf:RDF>
```

Nota: *El código presentado en la tabla ha sido validado en la aplicación oficial del W3C sin que se obtuviera ningún tipo de fallo. No obstante, se advierte al alumno y cualquier lector que las etiquetas marc empleadas corresponden a una adaptación personal de la normativa con efectos didácticos, sin que por ello suponga la norma a seguir. También se debe prestar especial cautela al espacio de nombres o namespace de la Library of Congress en torno a MARC-XML, ya que temporalmente no se encuentra disponible para su consulta.*

Tabla 17. Ejemplo de descripción RDF / MARC-XML de un documento

En la *tabla17*, se muestra la evolución del ejemplo de las *figuras 1 y 2*, comprobándose la importancia de reseñar en todo caso los namespaces (destacados en color verde) de los formatos de marcado (metadatos) que se emplearán para la descripción del documento. El sujeto que será objeto de la descripción (destacado en color azul), corresponde a la URI de la ficha del documento en el catálogo de la Biblioteca Complutense.

También se destaca cómo cada etiqueta del archivo XML, se compone del prefijo correspondiente del tipo de metadato al que corresponde, por ejemplo *<marc:tag245a>*, *<marc:datafield>*, *<rdf:Description>*. Este ejemplo indica que es posible emplear tantos tipos de metadatos como se deseen, incluso combinarlos para perfeccionar la capacidad analítica de la descripción. El resultado del proceso de validación del código es el que sigue a continuación en la *figura3*, comprobándose un esquema de datos similar al planteado en las *figuras 1 y 2*.

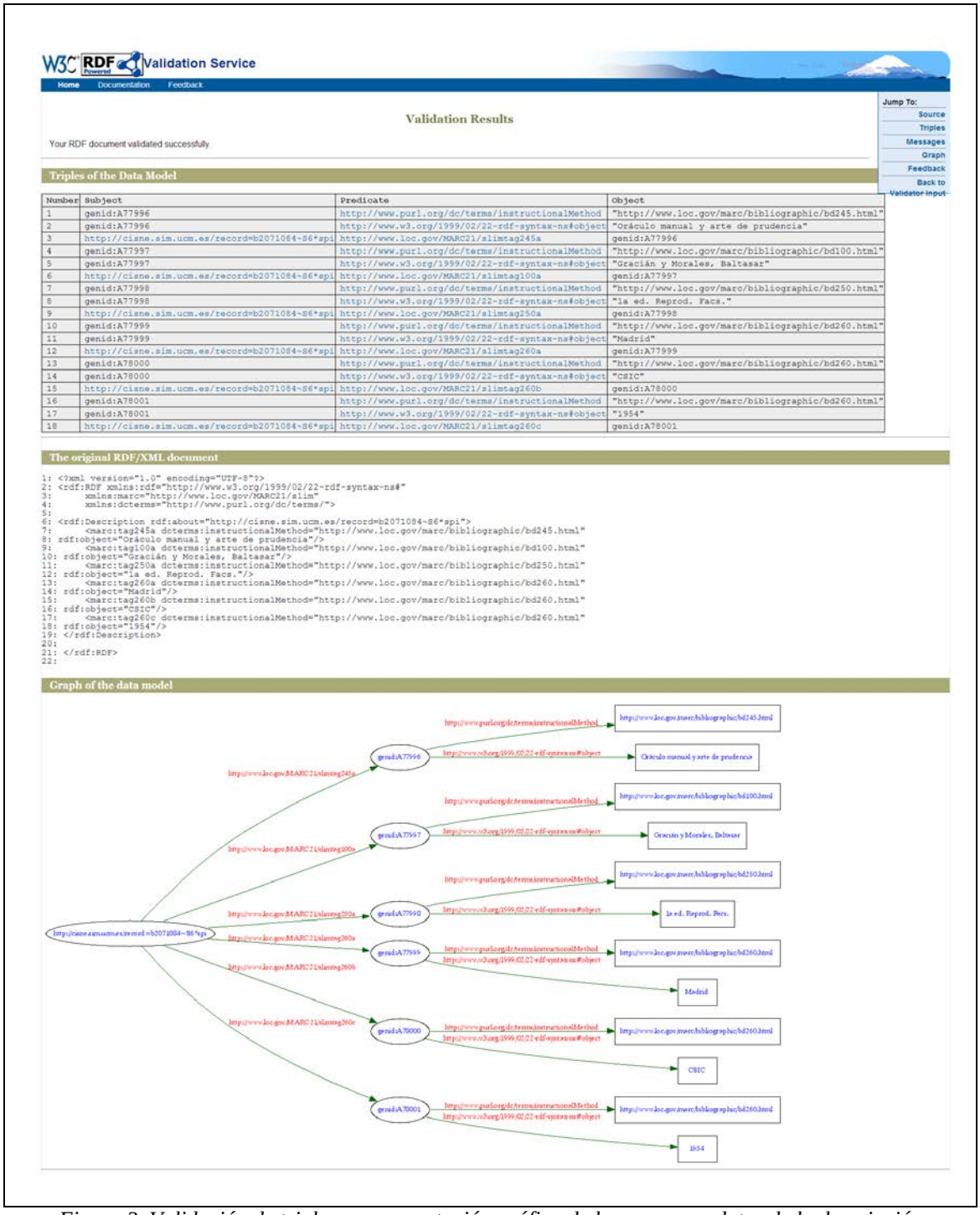


Figura 3. Validación de triples y representación gráfica de los campos y datos de la descripción

En el caso de los metadatos Dublin Core, al igual que otros formatos y normas que han sido diseñadas para ser empleadas en entornos *RDF/XML*, utilizan su propio sistema de prefijos "*DC*" ó "*DCTERMS*" para configurar su etiquetado en forma de lenguaje de marcado. Ello viene especificado en el namespace o espacio de nombres, véase ejemplo de la *tabla18*.

Prefijo del espacio de nombres		URI del espacio de nombres (namespace)
HTML	XML	
DC	dc	http://purl.org/dc/elements/1.1/
DCTERMS	dcterms	http://purl.org/dc/terms/
	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
	xsd	http://www.w3.org/2001/XMLSchema#
	owl	http://www.w3.org/2002/07/owl#
	rss	http://purl.org/rss/1.0/

Tabla 18. Ejemplos de espacios de nombre con sus prefijos de aplicación

Teniendo en cuenta las nociones básicas relativas al empleo de metadatos en RDF, el papel que juegan los espacios de nombres o namespaces, la estructura básica de la construcción de triples, el sujeto, predicado y objeto, ahora se puede definir el modelo sintáctico para utilizar metadatos Dublin Core en RDF, de acuerdo a las normas definidas por DCMI, disponibles en: http://www.dublincore.org/documents/dc-rdf/

En primer lugar es necesario comprender que Dublin Core se aplica en RDF gracias a lo que se denomina el Dublin Core Abstract Model, también conocido como DCAM, por el que se determinan las reglas de construcción básicas en la web semántica mediante metadatos Dublin Core. El modelo dispone que el recurso web (*sujeto*) deba ser descrito mediante propiedades (*predicado*) para los que se consignan unos valores (*objeto*) que pueden ser de tipo literal o no literal, véase *figura4*.

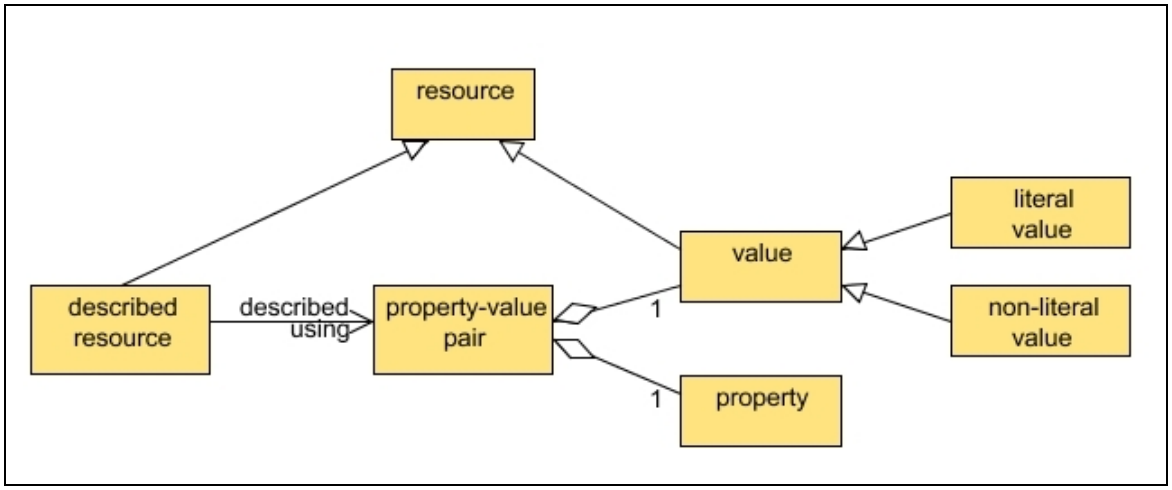


Figura 4. Modelo DCAM básico en Dublin Core. Fuente: http://dublincore.org/documents/2007/04/02/abstract-model/resource-model.jpg

Conceptos básicos del modelo DCAM

- ***Description set = Conjunto de descripción***. Es el conjunto de descripciones que describe un recurso. En el caso de un documento monográfico de una biblioteca, el conjunto de descripciones configuraría la ficha catalográfica completa, considerándose el conjunto de descripción. La cardinalidad es de una o más propiedades para un único recurso.

- ***Las declaraciones de cada par propiedad/valor*** se compone de un URI que identifica dicha propiedad de forma unívoca y un valor literal compuesto por una cadena de caracteres o no literal (*0,1*) de acuerdo al esquema de codificación del vocabulario.

- ***La cadena de caracteres del literal*** puede ser de lenguaje asociado o normalizado como por ejemplo "*spa*" de acuerdo al "*schema ISO639-2*" o contener una cadena de caracteres comprensible, por ejemplo "*Oráculo manual y arte de prudencia*".

Codificación de Dublin Core en RDF

- Primera línea del archivo RDF = ***<?xml version="1.0" encoding="UTF-8"?>***. Indica el tipo de documento, ya que RDF es un formato derivado de XML, por ello se indica como cabecera su versión y codificación de caracteres. Habitualmente se suele emplear la codificación universal *UTF-8* más normalizada para cualquier variedad idiomática y caracteres especiales que puedan ser necesarios referir como datos en la metadescripción.

- ***Identificación del formato empleado y líneas de namespace*** o espacio de nombres utilizados. Lo constituyen las líneas 2 - 4 de la *tabla19*. Se designa el tipo de formato que se emplea como base para la descripción mediante Dublin Core. En este caso, tal como se viene explicando, RDF. La declaración de tipo de documento RDF, se expresa mediante la etiqueta de apertura *<rdf:RDF>* y cierre *</rdf:RDF>*. La etiqueta de apertura constará habitualmente de los atributos de los distintos

namespaces o espacio de nombres de los formatos de metadatos que se emplearán en la descripción. Un ejemplo de atributo con su valor de namespace es *xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"* donde *"xmlns"* significa (XML namespace) y el valor que toma es la URL de la normativa y reglas sintácticas que rigen la construcción y terminología del formato que se especifica a través del prefijo especificado tras los dos puntos rdf. En el caso de Dublin Core, el namespace es *xmlns:dc="http://purl.org/dc/elements/1.1/"*, cuyo prefijo es *"dc"* y sus especificaciones normativas son las especificadas en http://purl.org/dc/elements/1.1/

- Declaración del sujeto, recurso o documento que será descrito, véase línea 6 de la *tabla19*. Dicha operación se efectúa utilizando la etiqueta de apertura *<rdf:Description>* y *</rdf:Description>* de cierre. Entre ambas etiquetas se circunscribirán las etiquetas de descripción Dublin Core propiamente dichas. Obsérvese que la etiqueta de apertura consta de un atributo *rdf:about=""* utilizado para especificar el identificador o sujeto de la descripción, habitualmente una URI permalink del recurso que se describe.

- La descripción Dublin Core puede comprobarse en las líneas 8 - 13:

  - ***<dcterms:subject>Literatura</dcterms:subject>*** Corresponde a la codificación mediante anidamiento del valor literal "literatura". Se pueden crear etiquetas Dublin Core de apertura y cierre con el texto de la descripción embebido.

  - ***<dcterms:subject rdf:type="dcterms:UDC" rdf:object="860"/>*** Otra forma de expresar la materia de literatura del documento es mediante una etiqueta unimembre. Se considera una etiqueta unimembre, aquella que no requiere de etiqueta de cierre y que cumple la condición de que contenga una barra oblicua (*/ backslash*) antes del cierre del etiquetado, como por ejemplo (/>). En este ejemplo se muestra cómo mediante el uso del atributo *rdf:type="dcterms:UDC"* se determina el esquema de codificación de la información del metadato *dcterms:subject*. El literal del metadato se expresa mediante el atributo *rdf:object="860"* que porta el valor numérico de la

Clasificación Decimal Universal para la clasificación de literatura. No obstante, se debe advertir que existen múltiples sintaxis para expresar el mismo contenido, tal como se muestra a continuación:

- *<dcterms:subject xml:scheme="dcterms:UDC">860</dcterms:subject>*
- *<dcterms:subject xml:scheme="dcterms:UDC" rdf:object="860"/>*

– ***<dcterms:title xml:lang="spa"> Oráculo manual y arte de prudencia </dcterms:title>*** Al igual que en casos anteriores se emplean etiquetas de apertura y cierre para contener el dato correspondiente al título del documento. Para especificar atributos propios de las etiquetas <*meta*> en HTML, se emplean atributos con prefijo (*xml:[atributo]*), como por ejemplo *xml:lang="spa"*.

– ***<dcterms:type xml:scheme="dcterms:IMT">application/pdf</dcterms:type>*** En este caso se emplea el atributo scheme propio de la codificación <meta> de HTML, mediante la especificación del prefijo *xml:scheme* necesario para su correcta validación y codificación. Obsérvese que *xml:scheme* actúa como atributo equivalente a *rdf:type*, tal como se describió anteriormente. De la misma manera el atributo *xml:content* sería equivalente a *rdf:object*, para expresar el valor, texto o contenidos del metadato.

– ***<dcterms:creator>Baltasar Gracián y Morales</dcterms:creator>*** La sintaxis utilizada para expresar el autor, de acuerdo a los modelos descritos, también podría ser definido de las siguientes formas:

- <dcterms:creator rdf:object="Baltasar Gracián y Morales"/>
- <dcterms:creator xml:content="Baltasar Gracián y Morales"/>
- <dcterms:creator rdf:resource="http://id.loc.gov/authorities/names/n50047296">Baltasar Gracián y Morales</dcterms:creator>

– ***<dc:identifier rdf:resource="978-84-00-08180-5"/>*** En el caso de establecer algún tipo de identificador URI o referencia URL de control, se emplea el atributo *rdf:resource=""* que contendrá el enlace a otro recurso o página web que identifica el contenido que se pretende describir. En el ejemplo se muestra el número ISBN13 y en la línea anterior la dirección URI de la autoridad de Baltasar Gracián en el índice de autoridades de la Library of Congress.

– ***<dc:relation rdf:resource="http://viaf.org/viaf/177839895/"/>*** Para establecer relaciones con URLs o URIs de terceros recursos, se pueden emplear los atributo *rdf:resource="", xml:href="", xml:content="", rdf:object="".*

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:dcterms="http://www.purl.org/dc/terms/">

<rdf:Description rdf:about="http://cisne.sim.ucm.es/record=b2071084~S6*spi">

<dcterms:subject>Literatura</dcterms:subject>
<dcterms:subject rdf:type="dcterms:UDC" rdf:object="860"/>
<dcterms:title xml:lang="es">Oráculo manual y arte de prudencia</dcterms:title>
<dcterms:type xml:scheme="dcterms:IMT">application/pdf</dcterms:type>
<dcterms:creator>Baltasar Gracián y Morales</dcterms:creator>
<dc:identifier rdf:resource="978-84-00-08180-5"/>
<dc:relation rdf:resource="http://viaf.org/viaf/177839895/"/>

</rdf:Description>

</rdf:RDF>
```

Tabla 19. Ejemplo de Dublin Core en RDF

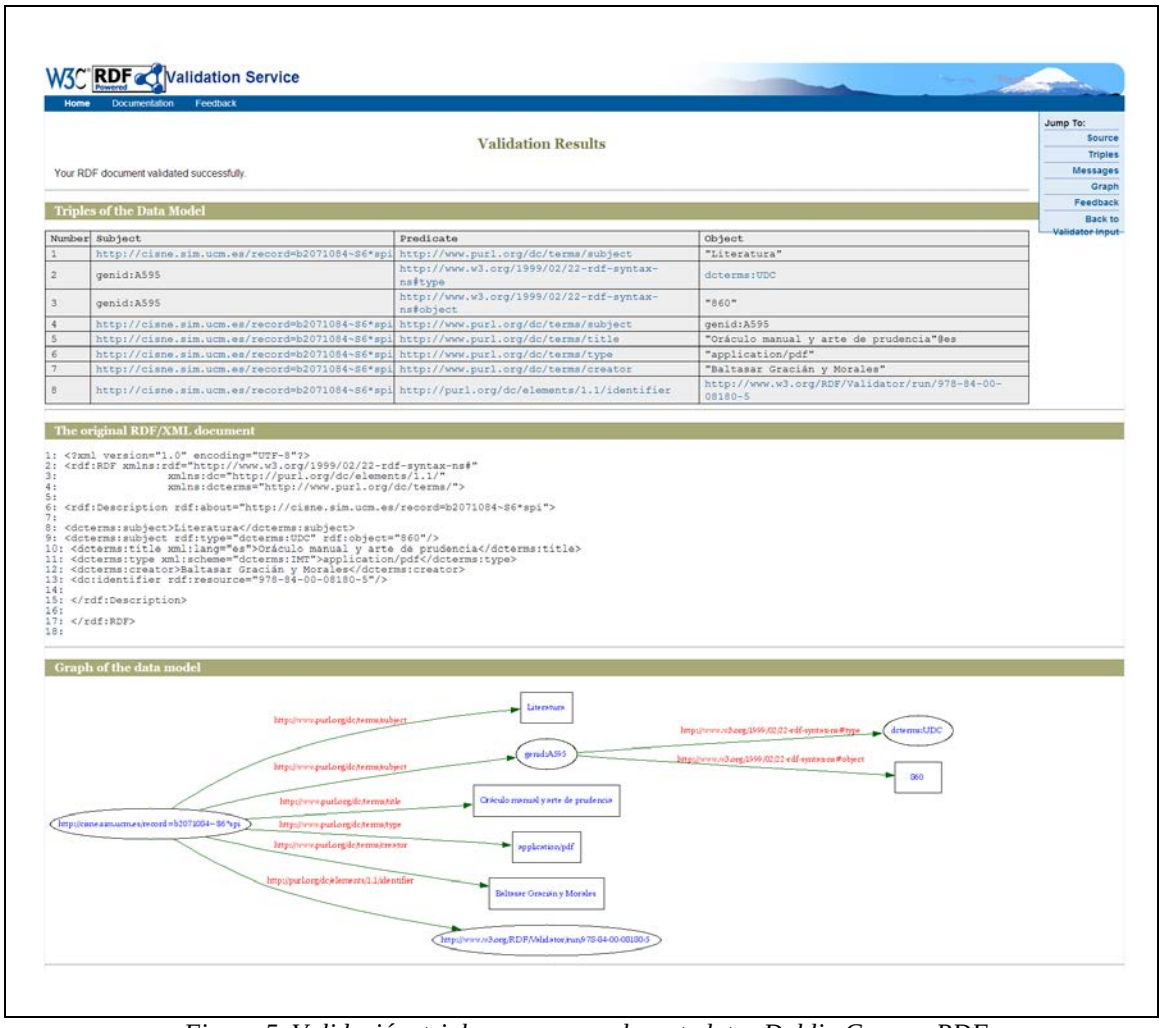


Figura 5. Validación, triples y esquema de metadatos Dublin Core en RDF

Embeber y vincular Dublin Core RDF en HTML

Para lograr la indexación y recopilación de la información en webcrawlers y buscadores resulta necesario vincular de algún modo las metadescripciones elaboradas en RDF. Existen varias técnicas para conseguir este objetivo, por un lado embebiendo el código fuente de RDF y Dublin Core en el documento HTML del recurso web que se está describiendo o bien vinculándolo al archivo RDF de la descripción.

– *Técnica de embebido de Dublin Core RDF en HTML*

```
<html>
<head>
<title>Página web del recurso</title>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:dcterms="http://www.purl.org/dc/terms/">

<rdf:Description rdf:about="http://cisne.sim.ucm.es/record=b2071084~S6*spi">

<dcterms:subject>Literatura</dcterms:subject>
<dcterms:subject rdf:type="dcterms:UDC" rdf:object="860"/>
<dcterms:title xml:lang="es">Oráculo manual y arte de prudencia</dcterms:title>
<dcterms:type xml:scheme="dcterms:IMT">application/pdf</dcterms:type>
<dcterms:creator>Baltasar Gracián y Morales</dcterms:creator>
<dc:identifier rdf:resource="978-84-00-08180-5"/>
<dc:relation rdf:resource="http://viaf.org/viaf/177839895/"/>

</rdf:Description>

</rdf:RDF>

</head>
<body>
<!-- [Artículo o Documento completo] -->
</body>
</html>
```

Tabla 20. Ejemplo de técnica de embebido de Dublin Core RDF en XML

– *Vinculación de archivo Dublin Core RDF en HTML*

```
<html>
<head>
<title>Página web del recurso</title>

<link rel="meta" type="application/rdf+xml" href="descripcion.rdf"/>

</head>
<body>
<!-- [Artículo o Documento completo] -->
</body>
</html>
```

Tabla 21. Ejemplo de vinculación de archivo Dublin Core RDF en HTML

6. MADS: metadatos para la descripción de autoridades

La utilización de metadatos en el campo de la biblioteconomía y documentación resulta más extensa que lo que en principio pudiera pensarse; por ejemplo la conocida descripción bibliográfica o análisis documental. En tal caso existen múltiples componentes que cohabitan en tales descripciones. Éstas son las autoridades del análisis, que constituyen en si mismas los puntos de acceso a los documentos. Ello indica que las descripciones bibliográficas y de las autoridades están vinculadas de forma inseparable, requiriendo para cada caso el tipo de metadato más adecuado.

Por ejemplo la descripción bibliográfica y catalogación de los documentos puede llevarse a cabo mediante los formatos MARC-XML y Dublin Core, tal como se viene especificando en artículos anteriores; (Consiguiendo distintos niveles de exhaustividad: MARC permite la descripción más completa y Dublin Core más abreviada)

En el caso de las autoridades, también existen una serie de metadatos para la descripción de las autoridades. Los más importantes son los conocidos como MADS (Metadata Authority Description Schema). Son utilizados para la descripción de personas físicas, entidades o personas jurídicas, congresos, conferencias y encuentros, títulos colectivos o uniformes, categorías temáticas, géneros y forma, nombres geográficos y extensiones. Las descripciones elaboradas conforme al modelo de metadatos MADS son compatibles con los principales modelos de metadatos y formatos de descripción bibliográfica, como MARC-XML y Dublin Core. Además y no menos importante es su capacidad para ser expresados de igual modo que Dublin Core, como una extensión más en RDF.

Fundamentos de MADS

MADS es un esquema diseñado principalmente en XML para poder describir mediante metadatos las autoridades que pueden encontrarse habitualmente en la descripción o análisis documental. A parte de ser un complemento para formatos como MARC-XML y Dublin Core, fue diseñado principalmente como complemento de los metadatos

MODS (Metadata Object Description Schema) destinados a la descripción bibliográfica, tal como se estudiará en próximos artículos.

Para poder emplear todos los metadatos MADS en un archivo XML o RDF es necesario utilizar los siguientes espacios de nombres o namespaces, véase *tabla22*. (*xmlns="http://www.loc.gov/mads/"*)

Versión	Prefijo XML/RDF	URI del espacio de nombres (namespace)
1.0	mads	http://www.loc.gov/mads/
2.0	mads	http://www.loc.gov/mads/v2
	xlink	http://www.w3.org/1999/xlink
	xsi	http://www.w3.org/2001/XMLSchema-instance

Tabla 22. Espacios de nombre con sus prefijos aplicados en la codificación de MADS

Además existen una serie de estructuras básicas para su construcción en formato XML, que se resumen en la siguiente *tabla23*.

Notación Básica	XML	RDF	Descripción
mads	<mads>	<mads:mads>	Contiene la descripción completa de una autoridad mediante etiquetas <authority>, <variant>
http://www.loc.gov/standards/mads/userguide/generalapp.html#mads			
madsCollection	<madsCollection>	<mads:madsCollection>	Contiene un conjunto de autoridades descritas en etiquetas <mads>, constituyendo el elemento raíz
http://www.loc.gov/standards/mads/userguide/generalapp.html#madscollection			
authority	<authority>	<mads:authority>	Contiene los elementos de descripción de la autoridad <name>, <titleinfo>, <topic>, <temporal>, <genre>, <geographic>, <hierarchicalGeographic>, <occupation>
http://www.loc.gov/standards/mads/userguide/authority.html			
variant	<variant>	<mads:variant>	Contiene las distintas variantes del nombre o título de la autoridad. Serán utilizadas para permitir diversos puntos de acceso a la autoridad y facilitar su recuperación
http://www.loc.gov/standards/mads/userguide/variant.html			
related	<related>	<mads:related>	Contiene las referencias a cualquier otra autoridad o registro relacionado.
http://www.loc.gov/standards/mads/userguide/related.html			

affiliation	<affiliation>	<mads:affiliation>	Contiene los elementos de descripción de la afiliación de la autoridad <position>, <organization>, <address>, <email>, <phone>, <fax>, <hours>, <dateValid>
http://www.loc.gov/standards/mads/userguide/affiliation.html			
classification	<classification>	<mads:classification>	Define el número de clasificación decimal para la materia de la autoridad
http://www.loc.gov/standards/mads/userguide/classification.html			
fieldOfActivity	<fieldOfActivity>	<mads:fieldOfActivity>	Utilizado cuando la autoridad es de tipo personal para indicar el área de conocimiento o especialidad, competencia, responsabilidad, jurisdicción.
http://www.loc.gov/standards/mads/userguide/fieldOfActivity.html			
identifier	<identifier>	<mads:identifier>	Etiqueta repetible para establecer los distintos identificadores o URIs de la autoridad, números de control, etc.
http://www.loc.gov/standards/mads/userguide/identifier.html			
language	<language>	<mads:language>	Etiqueta para la descripción del idioma que se utiliza en la descripción de la autoridad. Contiene los elementos <languageTerm>, <scriptTerm>
http://www.loc.gov/standards/mads/userguide/language.html			
note	<note>	<mads:note>	Etiqueta para la descripción de notas sobre las fuentes de información utilizadas, la historia de la autoridad o sobre la información no localizada y reseñada
http://www.loc.gov/standards/mads/userguide/note.html			
url	<url>	<mads:url>	Etiqueta que permite reseñar enlaces URL relativos o pertenecientes a la autoridad que se está describiendo. Algunos de ellos pueden ser también expresados mediante <identifier>
http://www.loc.gov/standards/mads/userguide/url.html			
extension	<extension>	<mads:extension>	Etiqueta utilizada para contener metadatos de descripción no contemplados en MADS, de cara a la descripción de conceptos y propiedades específicas, mediante otros metadatos o formatos.
http://www.loc.gov/standards/mads/userguide/extension.html			

recordInfo	<recordInfo>	<mads:recordInfo>	Etiqueta contenedora de elementos para la descripción de los datos de gestión de la autoridad. Por ejemplo, la fuente de información de los contenidos, la fecha de creación y modificación de la autoridad, idioma de la catalogación y normas de descripción empleadas. Estos elementos contenidos son <recordContentSource>, <recordCreationDate>, <recordChangeDate>, <recordIdentifier>, <recordOrigin>, <languageOfCataloguing>, <descriptionStandard>
http://www.loc.gov/standards/mads/userguide/recordInfo.html			

Tabla 23. Referencia básica del modelo de metadatos MADS

Reglas básicas de construcción

- El elemento raíz en MADS es (*<madsCollection></madsCollection>* || *<mads:madsCollection></mads:madsCollection>*) que contendrá todas las autoridades definidas a partir de las etiquetas de registro (*<mads></mads>* || *<mads:mads></mads:mads>*)

- La descripción de un registro de autoridad se efectúa mediante las etiquetas (*<mads></mads>* || *<mads:mads></mads:mads>*) que contienen los distintos metadatos de la descripción propiamente dicha. Los elementos contenidos pueden ser *<authority>, <variant>, <related>, <affiliation>, <classification>, <fieldOfActivity>, <identifier>, <language>, <note>, <url>, <extension> y <recordInfo>*

- Los metadatos MADS pueden ser empleados junto con otros formatos y metadatos de descripción bibliográfica como MODS, MARC-XML, Dublin Core.

- Se pueden emplear los signos de puntuación convenidos internacionalmente como por ejemplo en ISBD, para reseñar correctamente la información y datos de las autoridades. En caso contrario MADS dispone de hojas de estilo de tipo XSLT para la visualización y correcta representación de la información.

- El identificador principal MADS para cada autoridad se establece utilizando las etiquetas *<recordInfo><recordIdentifier>*. Los identificadores complementarios o secundarios se establecen con la etiqueta *<identifier>*

- Son elementos obligatorios de primer nivel *<mads>*, *<authority>*, *<recordInfo>*. También existen elementos opcionales que contienen subelementos obligatorios.

- El orden de la notación de los metadatos no afecta al resultado final de la descripción, ya que existen hojas de estilo tipo XSLT diseñadas para interpretar la información, campos y orden correcto de la información consignada.

- Todos los elementos o etiquetas de MADS son repetibles, excepto *<madsCollection>*

Tal como se ha explicado MADS puede construirse con un etiquetado XML básico o ser combinado con otros formatos a través de su empleo en RDF. Seguidamente se muestran algunos ejemplos de codificación en el que se podrán observar sus distintos modos de empleo.

- *Autoridad personal descrita con MADS y expresada en XML*

```
<?xml version="1.0" encoding="UTF-8"?>

<mads>

<authority>
  <name type="personal" authority="abne">
    <namePart>Cervantes Saavedra, Miguel de</namePart>
    <namePart type="date">(1547-1616)</namePart>
    <description>Príncipe de los Ingenios</description>
  </name>
</authority>

<variant type="abbreviation">
  <name type="personal">
    <namePart>Cervantes</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Cervantes di Saavedra, Michele</namePart>
  </name>
</variant>
```

```
<variant type="translation">
  <name type="personal">
    <namePart>Сервантес Сааведра, Мигель де</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Servantes Saavedra, Migel</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Therbantes, Minkel nte</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Zerbantes eta Saabedra, Mikel</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Sirfantis Saafedrā, Mīgīl dī</namePart>
  </name>
</variant>

<variant type="translation">
  <name type="personal">
    <namePart>Sewantisi Saweidela, Migai'er de</namePart>
  </name>
</variant>

<fieldOfActivity>Novelista</fieldOfActivity>
<fieldOfActivity>Poeta</fieldOfActivity>
<fieldOfActivity>Dramaturgo</fieldOfActivity>
<identifier type="viaf">17220427</identifier>

<language>
  <languageTerm type="code" authority="iso639-2b">spa</languageTerm>
  <languageTerm type ="text">Español</scriptTerm>
</language>

<note type="source">Dic. de literatura española e hispanoamericana, 1993 (Cervantes Saavedra, Miguel de (Alcalá de Henares, Madrid, 1547-Madrid, 1616)</note>

<url displayLabel="Ficha VIAF de Miguel de Cervantes Saavedra">http://viaf.org/viaf/17220427</url>

<recordInfo>
  <recordIdentifier source="BNM">XX1718747</recordIdentifier>
</recordInfo>

</mads>
```

Tabla 24. Autoridad personal descrita con MADS y expresada en XML básico

– *Autoridad personal descrita con MADS y expresada en XML con prefijo de espacio de nombres*

```
<?xml version="1.0" encoding="UTF-8"?>

<mads:mads xmlns:mads="http://www.loc.gov/mads/v2"
                xmlns:xlink="http://www.w3.org/1999/xlink"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://www.loc.gov/mads/v2 http://www.loc.gov/standards/mads/mads-2-0.xsd"

                version="2.0">

<mads:authority>
  <mads:name type="personal" authority="abne">
    <mads:namePart>Cervantes Saavedra, Miguel de<mads:namePart>
    <mads:namePart type="date">(1547-1616)<mads:namePart>
    <mads:description>Príncipe de los Ingenios<mads:description>
  <mads:name>
<mads:authority>

<mads:variant type="abbreviation">
  <mads:name type="personal">
    <mads:namePart>Cervantes<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Cervantes di Saavedra, Michele<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Сервантес Сааведра, Мигель де<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Servantes Saavedra, Migel<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Therbantes, Minkel nte<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Zerbantes eta Saabedra, Mikel<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Sirfantis Saafedrā, Mīgīl dī<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Sewantisi Saweidela, Migai'er de<mads:namePart>
  <mads:name>
<mads:variant>

<mads:fieldOfActivity>Novelista<mads:fieldOfActivity>
<mads:fieldOfActivity>Poeta<mads:fieldOfActivity>
<mads:fieldOfActivity>Dramaturgo<mads:fieldOfActivity>
<mads:identifier type="viaf">17220427<mads:identifier>
```

```
<mads:language>
  <mads:languageTerm type="code" authority="iso639-2b">spa<mads:languageTerm>
  <mads:languageTerm type ="text">Español<mads:scriptTerm>
<mads:language>

<mads:note type="source">Dic. de literatura española e hispanoamericana, 1993 (Cervantes Saavedra, Miguel de (Alcalá de Henares, Madrid, 1547-Madrid, 1616)<mads:note>

<mads:url displayLabel="Ficha VIAF de Miguel de Cervantes Saavedra">http://viaf.org/viaf/17220427<mads:url>


<mads:recordInfo>
  <mads:recordIdentifier source="BNM">XX1718747</recordIdentifier>
</mads:recordInfo>

<mads:mads>
```

Tabla 25. Autoridad personal descrita con MADS y expresada en XML con prefijo de espacio de nombres

– *Autoridad personal descrita con MADS y expresada en RDF con prefijo de espacio de nombres*

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:dcterms="http://www.purl.org/dc/terms/"
         xmlns:mads="http://www.loc.gov/mads/v2"
         xmlns:xlink="http://www.w3.org/1999/xlink"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.loc.gov/mads/v2 http://www.loc.gov/standards/mads/mads-2-0.xsd">

<rdf:Description rdf:about="http://viaf.org/viaf/17220427">

<mads:mads version="2.0">

<mads:authority>
  <mads:name type="personal" authority="abne">
    <mads:namePart>Cervantes Saavedra, Miguel de<mads:namePart>
    <mads:namePart type="date">(1547-1616)<mads:namePart>
    <mads:description>Príncipe de los Ingenios<mads:description>
  <mads:name>
<mads:authority>

<mads:variant type="abbreviation">
  <mads:name type="personal">
    <mads:namePart>Cervantes<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Cervantes di Saavedra, Michele<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Сервантес Сааведра, Мигель де<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
  <mads:name type="personal">
    <mads:namePart>Servantes Saavedra, Migel<mads:namePart>
  <mads:name>
<mads:variant>

<mads:variant type="translation">
```

```
   <mads:name type="personal">
      <mads:namePart>Therbantes, Minkel nte<mads:namePart>
   <mads:name>
<mads:variant>

<mads:variant type="translation">
   <mads:name type="personal">
      <mads:namePart>Zerbantes eta Saabedra, Mikel<mads:namePart>
   <mads:name>
<mads:variant>

<mads:variant type="translation">
   <mads:name type="personal">
      <mads:namePart>Sirfantis Saafedrā, Mīgīl dī<mads:namePart>
   <mads:name>
<mads:variant>

<mads:variant type="translation">
   <mads:name type="personal">
      <mads:namePart>Sewantisi Saweidela, Migai'er de<mads:namePart>
   <mads:name>
<mads:variant>

<mads:fieldOfActivity>Novelista<mads:fieldOfActivity>
<mads:fieldOfActivity>Poeta<mads:fieldOfActivity>
<mads:fieldOfActivity>Dramaturgo<mads:fieldOfActivity>
<mads:identifier type="viaf">17220427<mads:identifier>

<mads:language>
   <mads:languageTerm type="code" authority="iso639-2b">spa<mads:languageTerm>
   <mads:languageTerm type ="text">Español<mads:scriptTerm>
<mads:language>

<mads:note type="source">Dic. de literatura española e hispanoamericana, 1993 (Cervantes Saavedra, Miguel de (Alcalá de Henares, Madrid, 1547-Madrid, 1616)<mads:note>

<mads:url displayLabel="Ficha VIAF de Miguel de Cervantes Saavedra">http://viaf.org/viaf/17220427<mads:url>


<mads:recordInfo>
   <mads:recordIdentifier source="BNM">XX1718747</recordIdentifier>
</mads:recordInfo>

<mads:mads>

</rdf:Description>

</rdf:RDF>
```

Tabla 26. Autoridad personal descrita con MADS y expresada en RDF con prefijo de espacio de nombres

7. MODS: metadatos para la descripción de objetos bibliográficos

Los metadatos MODS son un esquema de descripción bibliográfica, derivado de MARC21, incluyendo los conjuntos de etiquetas basadas en MARC-XML. Los metadatos MODS pueden emplearse junto con MADS para la descripción de las autoridades, siendo completamente compatibles por las propiedades de extensibilidad de XML, lenguaje que utiliza como base de su sintaxis, igual que el formato MADS. Esto supone una ventaja que le permite una fácil interoperabilidad con cualquier formato de descripción bibliográfica e incluso archivística (BOUNTOURI, L.; GERGATSOULIS, M. 2009).

Por ejemplo es factible la transformación de los registros archivísticos en EAD a formato MODS, incluso la transformación de descripciones bibliográficas en Dublin Core a MODS y a la inversa, lo cual implica una sencillez en el mapeado y enrutamiento de la información de unos campos de descripción a otros. Por otra parte, al igual que los formatos anteriormente expuestos, MODS puede emplear lenguajes de consulta como XQuery, implementar fácilmente los protocolos SRU y OAI-PMH, e incluso exportar el catálogo bibliográfico como si de un canal de sindicación RSS se tratara, véase (BLÁZQUEZ OCHANDO, M. 2010) y (MCCALLUN, S.; GUENTHER, R. 2010). Otras ventajas es la evidente capacidad para enlazar las descripciones de las distintas autoridades con las descripciones bibliográficas, mediante la técnica de linked data.

Fundamentos de MODS

Al igual que MADS, los metadatos MODS constan de una estructura de elementos y subelementos con los que se especifican los distintos campos de descripción bibliográfica. Al igual que Dublin Core, RDF y MADS, los metadatos MODS constan de su propio espacio de nombres, véase la *tabla27*.

Versión	Prefijo XML/RDF	URI del espacio de nombres (namespace)
1.0	mods	http://www.loc.gov/mods/
3.0	mods	http://www.loc.gov/mods/v3

	xlink	http://www.w3.org/1999/xlink
	xsi	http://www.w3.org/2001/XMLSchema-instance
xsi:schemaLocation		http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd

Tabla 27. Espacios de nombre con sus prefijos aplicados en la codificación de MADS

Cuando MODS se combina con MADS es importante declarar cada autoridad y registro bibliográfico con su correspondiente espacio de nombres, véase *tabla28*.

```
<!-- Ejemplo de uso del namespace de mads y mods-->
<mods xmlns:mods="http://www.loc.gov/mods/v3"></mods>
<mads xmlns:mads="http://www.loc.gov/mads/v2"></mads>

<!-- Ejemplo de namespace de mads y mods con prefijos -->
<mods:mods xmlns:mods="http://www.loc.gov/mods/v3"></mods>
<mads:mads xmlns:mads="http://www.loc.gov/mads/v2"></mads>
```

Tabla 28. Ejemplo de namespace aplicado a registros MADS y MODS

La mayor parte de los elementos de MODS constan de su equivalente en MARC21, lo cual facilita la exportación e interoperabilidad de la información entre los registros de un formato a otro según las necesidades. Por ejemplo la implementación de MODS, resulta más sencilla de aplicar en un catálogo bibliográfico en línea que el propio MARC-XML debido a su estructura compleja.

Muchos de los elementos definidos en MODS son comunes a los metadatos MADS. No obstante, existen diferencias, ya que los metadatos para la descripción de objetos bibliográficos son mucho más extensos en su esquema principal, véase *tabla29*.

Notación Básica	XML	RDF	Descripción
mods	<mods>	<mods:mods>	Etiqueta raíz de la descripción completa de un documento. Contiene a su vez los subelementos <titleInfo>, <name>, <typeOfResource>, <genre>, <originInfo>, <language>, <physicalDescription>, <abstract>, <tableOfContents>, <targetAudience>, <note>, <subject>, <classification>, <relatedItem>, <identifier>, <location>, <accessCondition>, <part>, <extension>, <recordInfo>
http://www.loc.gov/standards/mods/userguide/			
titleInfo	<titleInfo>	<mods:titleInfo>	Etiqueta contenedora de la información del título. Engloba los siguientes

			subelementos <title>, <subTitle>, <partNumber>, <partName>, <nonsort>
http://www.loc.gov/standards/mods/userguide/titleinfo.html			
name	<name>	<mods:name>	Etiqueta contenedora de la información de la persona, entidad, organización o evento que encabeza el registro bibliográfico y que por ende es el autor principal del documento. Contiene los siguientes subelementos <namePart>, <displayForm>, <affiliation>, <role>, <description>
http://www.loc.gov/standards/mods/userguide/name.html			
typeOResource	<typeOfResource>	<mods:typeOfResource>	Definición del tipo de recurso o documento. Puede contener los valores: text, cartographic, notated music, sound recording, sound recording-musical, sound recording-nonmusical, still image, moving image, three dimensional object, software, multimedia, mixed material
http://www.loc.gov/standards/mods/userguide/typeofresource.html			
genre	<genre>	<mods:genre>	Contiene los términos que especifican el tipo de recurso o documento.
http://www.loc.gov/standards/mods/userguide/genre.html			
originInfo	<originInfo>	<mods:originInfo>	Información sobre el origen de los recursos, sus fuentes de información relacionadas, datos de publicación, etc. Contiene los siguientes subelementos <place>, <publisher>, <dateIssued>, <dateCreated>, <dateCaptured>, <dateValid>, <dateModified>, <copyrightDate>, <dateOther>, <edition>, <issuance>, <frequency>
http://www.loc.gov/standards/mods/userguide/origininfo.html			
language	<language>	<mods:language>	Contiene los subelementos <languageTerm>, <scriptTerm> para identificar el idioma o lengua de la descripción del recurso.
http://www.loc.gov/standards/mods/userguide/language.html			
physicalDescription	<physicalDescription>	<mods:physicalDescription>	Etiqueta contenedora de la descripción física del documento. Engloba los subelementos <form>, <reformattingQuality>, <internetMediaType>, <extent>, <digitalOrigin>, <note>
http://www.loc.gov/standards/mods/userguide/physicaldescription.html			
abstract	<abstract>	<mods:abstract>	Resumen del documento o recurso que se está describiendo.
http://www.loc.gov/standards/mods/userguide/abstract.html			
tableOfContents	<tableOfContents>	<mods:tableOfContents>	Relación de temas ó índice del documento que se está describiendo

			.
http://www.loc.gov/standards/mods/userguide/tableofcontents.html			
targetAudience	<targetAudience>	<mods:targetAudience>	Público objetivo al que está destinado el documento o recurso que se describe.
http://www.loc.gov/standards/mods/userguide/targetaudience.html			
note	<note>	<mods:note>	Información general relative al recurso o documento, para completer alguno de sus aspectos. Por ejemplo nota histórica, de la mención de responsabilidad, de autoridades secundarias ó notas generales.
http://www.loc.gov/standards/mods/userguide/note.html			
subject	<subject>	<mods:subject>	Designación temática, término ó frase que representa el tema principal ó enfoque de la obra, documento o recurso. Contiene los siguientes subelementos <topic>, <geographic>, <temporal>, <titleInfo>, <name>, <genre>, <hierarchicalGeographic>, <cartographics>, <geographicCode>, <occupation>
http://www.loc.gov/standards/mods/userguide/subject.html			
classification	<classification>	<mods:classification>	Designación de una materia para clasificar el recurso ó documento, de acuerdo a un esquema, instrumento, lenguaje documental por ejemplo clasificaciones decimales, encabezamientos de materia, etc.
http://www.loc.gov/standards/mods/userguide/classification.html			
relatedItem	<relatedItem>	<mods:relatedItem>	Información que identifica otros recursos relacionados con el que está siendo descrito. La información consignada puede ser una URL, URI o identificador del recurso relacionado.
http://www.loc.gov/standards/mods/userguide/relateditem.html			
identifier	<identifier>	<mods:identifier>	Contiene la URI que identifica el recurso.
http://www.loc.gov/standards/mods/userguide/identifier.html			
location	<location>	<mods:location>	Localización topográfica, signatura, localización electrónica, del documento objeto de la descripción. Contiene los siguientes subelementos <physicalLocation>, <shelfLocator>, <url>, <holdingSimple>, <holdingExternal>
http://www.loc.gov/standards/mods/userguide/location.html			
accessCondition	<accessCondition>	<mods:accessCondition>	Restricciones y condiciones de acceso impuestas para la lectura, descarga, análisis o préstamo del documento.

http://www.loc.gov/standards/mods/userguide/accesscondition.html			
part	<part>	<mods:part>	Designación de las partes físicas del recurso que está siendo descrito. Por ejemplo, volumen, páginas e incluso párrafos. Contiene los siguientes subelementos <detail>, <extent>, <date>, <text>
http://www.loc.gov/standards/mods/userguide/part.html			
extension	<extension>	<mods:extension>	Etiqueta utilizada para contener metadatos de descripción no contemplados en MODS, de cara a la descripción de conceptos y propiedades específicas, mediante otros metadatos o formatos.
http://www.loc.gov/standards/mods/userguide/extension.html			
recordInfo	<recordInfo>	<mods:recordInfo>	Etiqueta contenedora de elementos para la descripción de los datos de gestión de la autoridad. Por ejemplo, la fuente de información de los contenidos, la fecha de creación y modificación de la autoridad, idioma de la catalogación y normas de descripción empleadas. Estos elementos contenidos son <recordContentSource>, <recordCreationDate>, <recordChangeDate>, <recordIdentifier>, <recordOrigin>, <languageOfCataloguing>, <descriptionStandard>
http://www.loc.gov/standards/mods/userguide/recordinfo.html			

Tabla 29. Referencia básica del modelo de metadatos MODS

– Ejemplo de codificación MODS de las Novelas ejemplares de Miguel de Cervantes Saavedra. Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/documentos/mods-book-bne.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<modsCollection>
<mods xmlns:xlink="http://www.w3.org/1999/xlink"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.loc.gov/mods/v3"
      xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd"
      version="3.4">

<titleInfo>
   <title>Novelas ejemplares</title>
</titleInfo>
<name type="personal" authorityURI="http://viaf.org/viaf/17220427">
   <namePart>Cervantes Saavedra, Miguel de [1547-1616]</namePart>
      <role>
         <roleTerm type="code" authority="marcrelator">aut</roleTerm>
         <roleTerm type="text" authority="marcrelator">Autor</roleTerm>
      </role>
</name>

<name type="personal" authorityURI="http://viaf.org/viaf/44309117">
   <namePart>García Lorenzo, Luciano</namePart>
```

```
      <role>
         <roleTerm type="code" authority="marcrelator">aui</roleTerm>
         <roleTerm type="text" authority="marcrelator">Prologuista</roleTerm>
      </role>
</name>

<name type="personal" authorityURI="http://viaf.org/viaf/27078638">
   <namePart>Menéndez Onrubia, Carmen</namePart>
      <role>
         <roleTerm type="code" authority="marcrelator">ann</roleTerm>
         <roleTerm type="text" authority="marcrelator">Anotador</roleTerm>
      </role>
</name>

<typeOfResource>text</typeOfResource>

<originInfo>
   <edition>24a. ed</edition>
   <place>
      <placeTerm type="text">Madrid</placeTerm>
   </place>
   <publisher authorityURI="http://viaf.org/viaf/125385703">Espasa Calpe</publisher>
   <dateIssued keyDate="yes" encoding="w3cdtf">1986</dateIssued>
   <issuance>monografía</issuance>
</originInfo>

<language>
   <languageTerm authority="iso639-2b">spa</languageTerm>
   <languageTerm type="text">Español</languageTerm>
</language>

<subject>
   <titleInfo type="collection" authorityURI="http://viaf.org/viaf/174191679">
      <title>Colección Austral. Clásicos ; 29</title>
   </titleInfo>
</subject>

<physicalDescription>
   <extent>331 p. ; 18 cm</extent>
</physicalDescription>

<classification authority="udc">860</classification>

<identifier type="isbn">84-239-0029-0</identifier>

<tableOfContents>La gitanilla ; El amante liberal ; Rinconete y Cortadillo ; La española inglesa ; El licenciado
Vidriera ; La fuerza de la sangre</tableOfContents>

<location>
<url usage="primary display" access="preview">http://cisne.sim.ucm.es/record=b1181747~S6*spi</url>
</location>

<recordInfo>
   <recordSource>Universidad Complutense de Madrid</recordSource>
   <recordOrigin>Catálogo Cisne UCM - AECID http://cisne.sim.ucm.es/</recordOrigin>
   <languageOfCataloging>
      <languageTerm type="code" authority="iso639-2b">spa</languageTerm>
   </languageOfCataloging>
</recordInfo>

</mods>
</modsCollection>
```

Tabla 30. Registro bibliográfico descrito con metadato MODS

8. METS: metadatos para la descripción de metadatos

METS es un esquema de metadatos para la descripción de objetos de una biblioteca digital. Para ello se emplea XML como lenguaje de marcado base. A diferencia de los metadatos MADS especializados en autoridades y MODS especializados en objetos de tipo bibliográficos y recursos de la web, los metadatos METS están enfocados a la descripción de objetos digitales para la 1) transmisión de información bibliográfica, 2) archivo de la información, 3) difusión de la información. Esto significa que actúan como método de meta-descripción de los objetos y autoridades que se describen mediante MADS y MODS, pero también para otros formatos como MARC, EAD, VRA, Dublin Core, NISOIMG, TEIHDR, DDI, FGDC.

Fundamentos de METS

Los metadatos METS constan de su propio espacio de nombres o namespace, véase *tabla31*, lo que permite su uso complementado en archivos codificados con MADS o MODS, véase *tabla32*.

Versión	Prefijo XML/RDF	URI del espacio de nombres (namespace)
1.0	mets	http://www.loc.gov/METS/
xsi:schemaLocation		http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd

Tabla 31. Espacio de nombres con sus prefijo aplicado en la codificación de METS

```
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:rights="http://www.loc.gov/rights/" xmlns:xlink="http://
www.w3.org/1999/xlink"xmlns:lc="http://www.loc.gov/mets/profiles" xmlns:bib="http://www.loc.gov/mets/profiles/bibRe
cord"xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" OBJID="loc.afc.afc9999005.1153"xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-2.xsd" PROFILE="lc:bibRecord">
<mets:dmdSec ID="dmd1">
<mets:mdWrap MDTYPE="MODS">
<mets:xmlData>
<mods:mods ID="mods1">
<mods:titleInfo>
<mods:title>título del documento</mods:title>
</mods:titleInfo>
```

Tabla 32. Inicio de la codificación del registro bibliográfico con METS y MODS

Por otro lado, METS consta de diversas secciones y estructuras para su construcción, véase *tabla33*.

Etiqueta raíz <mets></mets>	
Sección 1. Cabecera METS <metsHdr></metsHdr>	
Definición: Metadatos descriptivos mínimos para identificar la fecha de creación de los metadatos METS, el estado de la descripción del recurso, objeto o documento, el agente o especialista encargado de la verificación y transferencia (ciclo vital), así como su función exacta.	
Contiene	<agent>, <altRecord>

```
<metsHdr CREATEDATE="2010-09-14T19:00:00" RECORDSTATUS="Complete">
        <agent ROLE="CREATOR" TYPE="INDIVIDUAL">
          <name>Juan Diego Álvarez</name>
        </agent>
        <agent ROLE="DISSEMINATOR" TYPE="INDIVIDUAL">
          <name>Javier Alvarado Martínez</name>
        </agent>
 </metsHdr>
```

Sección 2. Metadatos descriptivos <dmdSec></dmdSec>

Definición: Permite introducir metadatos externos mediante referencia de puntero <mdRef> e internos mediante embebido de los mismos <mdWrap>. Tales metadatos permiten la descripción del documento, objeto o recurso. Todos los metadatos <dmdSec> deben tener un atributo (ID) que permita identificar esa sección en la estructura del archivo.

Contiene	<mdRef>, <mdWrap>

```
<!-- Ejemplo de mdRef y puntero de metadatos -->
    <dmdSec ID="dmd001">
        <mdRef LOCTYPE="URI" MIMETYPE="text/html" MDTYPE="MARC"
        LABEL="La lógica de la investigación científica">http://cisne.sim.ucm.es/record=b2774781~S6*spi</mdRef>
     </dmdSec>
```

```
<!-- Ejemplo de mdWrap que permite embeber metadatos Dublin Core -->

    <dmdSec ID="dmd002">
          <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core Metadata">
            <xmlData>
              <dc:title>La lógica de la investigación científica</dc:title>
              <dc:creator>Karl R. Popper (1902-1994)</dc:creator>
              <dc:date>2011</dc:date>
              <dc:publisher>Tecnos</dc:publisher>
              <dc:subject>Filosofía de la ciencia</dc:subject>
            </xmlData>
          </mdWrap>
     </dmdSec>
```

```
<!-- Ejemplo de mdWrap con metadatos MARC expresados o codificados en Bas64 -->

   <dmdSec ID="dmd003">
         <mdWrap MIMETYPE="application/marc" MDTYPE="MARC" LABEL="OPAC Record">

<binData> TEVBREVSIDAwMDAwbmFtICAyMjAwMDAwIGEgNDUwMCANCjAwOCAgICAxMjA0MTJyMjAxMTIwMDhz
cCAgICAgICAgICAgIDAwMCAgIHNwYSBjIA0KMDE3ICAgIEQuTC5NMjI0ODAtMjAxMSANCjAyMCAg
ICA5NzgtODQtMzA5LTQ2MDctNSANCjA0MCAgICBTcE1hVUNGU3xic3BhIA0KMDgwICAgIDAwMS44
OTF8MjIwMDQgDQowODAgICAgMDAxOjEgDQoxMDAgMSAgUG9wcGVyLCBLYXJsIFJhaW11bmQsfGQx
OTAyLTE5OTQgDQoyNDUgMTMgTGEgbPNnaWNhIGRlIGxhIGludmVzdGlnYWNp824gY2llbnTtZmlj
YSAvfGNLYXJsIFIuIFBvcHBlcg0KMjUwICAgIDJhLiBlZC4sIHJlaW1wLiANCjI2MCAgICBNYWRy
aWQgOnxiVGVjbm9zLHxjMjAxMSANCjMwMCAgICA1NzAgC4gO3xjMjMgY20gDQo0OTAgMCAgRXN0
cnVjdHVyYSB5IGZ1bmNp824gDQo2NTAgMDQgSW52ZXN0aWdhY2nzbiBjaWVudO1maWNhIA0KNjUw
IDA0IEZpbG9zb2btYSB5IGNpZW5jaWEgDQo2NTAgMDggUGhpbG9zb3BoeSBhbmQgc2NpZW5jZSAN
CjY1MCAwOCBTY2llbnRpZiByZXNlYXJjaCANCjkwNyAwMCB6bHxiMCA=
            </binData>
         </mdWrap>
   </dmdSec>
```

Sección 3. Metadatos administrativos <amdSec></amdSec>

Definición: Permiten la organización de metadatos para la administración y gestión de los archivos digitales. En este sentido se pueden identificar 4 tipos (metadatos técnicos <techMD>, metadatos de derechos y propiedad intelectual <rightsMD>, metadatos sobre el origen del documento <sourceMD> y metadatos sobre la procedencia digital <digiprovMD>) Para cada uno de los subelementos se aplica se aplica el etiquetado <mdWrap> y <mdRef> para reseñar los metadatos administrativos. Al igual que en los metadatos descriptivos es necesario identificar cada apartado con el atributo ID, con objeto de crear una estructura del documento METS.

Contiene	<techMD>, <rightsMD>, <sourceMD>, <digiprovMD>

```
   <techMD ID="amd001">
         <mdWrap MIMETYPE="text/xml" MDTYPE="OTHER" LABEL="PRISM Tamaño de la información">
           <xmlData>
             <prism:byteCount>36.864 bytes</prism:byteCount>
             <prism:wordCount>6822</prism:wordCount>
           </xmlData>
         </mdWrap>
         <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core soporte y formato">
           <xmlData>
             <dcterms:format>21cm</dcterms:format>
             <dcterms:medium>papel</dcterms:medium>
           </xmlData>
         </mdWrap>
   </techMD>

   <rightsMD ID="amd002">
         <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core derechos">
           <xmlData>
             <dcterms:rights>Uso comercial, reproducción y difusión sometida a tasas de explotación</dcterms:rights>
             <dcterms:license>http://www.tecnos.es/legal_tecnos/condiciones.html</dcterms:license>
           </xmlData>
         </mdWrap>
   </rightsMD>

   <sourceMD ID="amd003">
         <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core fuente original">
           <xmlData>
             <dcterms:source>Catálogo CISNE Biblioteca Universidad Complutense</dcterms:source>
           </xmlData>
         </mdWrap>
   </sourceMD>
```

Sección 4. Archivo <fileSec></fileSec>

Definición: Sección que permite agrupar los distintos archivos relacionados con el documento u objeto digital. Por ejemplo su ficha descriptiva, su documento digitalizado, imagen, audio, video.

Contiene	<fileGrp>

```
<fileSec>
        <fileGrp ID="v1">
          <file ID="file001" MIMETYPE="application/xml" SIZE="257537" CREATED="2011-06-10">
            <FLocat LOCTYPE="URL">http://dominio.es/biblioteca/bib1_9788430946075.xml</FLocat>
          </file>
        </fileGrp>
</fileSec>
```

Sección 5. Mapa estructural <structMap></structMap>

Definición: La sección mapa estructural define la estructura jerárquica del documento de acuerdo a los IDS establecidos a lo largo de la descripción METS. La jerarquía de las partes del documento METS se define mediante elementos que a su vez anidan diversos tipos de subelementos <mptr> (puntero de METS) y <fptr> (puntero de archivo) que permite identificar los contenidos correspondientes a cada archivo definido en la sección 4.

Contiene	

```
<structMap TYPE="logical">
        <div ID="div1" LABEL="Ficha catalográfica codificada en MARC-XML" ORDER="1">
                <fptr FILEID="file001">
                        <area FILEID="file001"/>
                </fptr>
         </div>
</structMap>
```

Sección 6. Enlace estructural <structLink></structLink>

Definición: Sección para el registro de enlaces a lo largo de la descripción del recurso o documento objetivo (Archivo de sitios web)

Contiene	<smLink>

```
<smLink from="img001" to="file001"/>
<smLink from="img002" to="file002"/>
<smLink from="img003" to="file004"/>
```

Sección 7. Comportamiento <behaviorSec></behaviorSec>

Definición: Permite definir el comportamiento del programa de lectura PARSER para la interpretación del contenido codificado. Ésta sección puede contener uno o varios elementos <behavior> que a su vez constan de subelementos <mechanism> que permiten ejecutar un módulo de lectura en el PARSER de lectura para que trate un determinado tipo de información. En el siguiente ejemplo se observa cómo se alude a una función de carga y lectura de imágenes.

Contiene	<behaviorSec>, <behavior>

```
<METS:behavior ID="DISS1.1" STRUCTID="S1.1" BTYPE="uva-bdef:stdImage"
        CREATED="2002-05-25T08:32:00" LABEL="UVA Std Image Disseminator"
        GROUPID="DISS1" ADMID="AUDREC1">
```

```
    <METS:interfaceDef LABEL="UVA Standard Image Behavior Definition"
        LOCTYPE="URN" xlink:href="uva-bdef:stdImage"/>
    <METS:mechanism LABEL="A NEW AND IMPROVED Image Mechanism"
        LOCTYPE="URN" xlink:href="uva-bmech:BETTER-imageMech"/>
  </METS:behavior>
```

Tabla 33. Referencia de codificación de METS

9. Lectura de metadatos: programas parser

Los metadatos estudiados, Dublin Core, MADS, MODS y METS, así como cualesquiera que sean basados en XML, pueden ser explotados gracias a la existencia de programas de lectura cuyos patrones de funcionamiento permiten un análisis correcto de acuerdo a la norma de construcción de tales metadatos. Esto es lograr recuperar la información que contienen embebida entre etiquetas de apertura y cierre, así como en sus correspondientes atributos. Los programas parser no son herramientas fácilmente visibles para el documentalista, sólo se observan los resultados de los mismos. Su presencia en casi todas las herramientas de la web, los convierte en indispensables y su conocimiento habilita al documentalista para un mejor aprovechamiento de la información publicada en los catálogos bibliográficos, sistemas de información y documentación de las distintas UID. Pero para comprender su función es necesario definirlos, así como establecer cuál es su patrón de funcionamiento básico.

Qué es un parser

Un parser es un analizador sintáctico de patrones o estructuras predefinidas, que actúa sobre un archivo, cadena de caracteres, códigos, formatos o texto, de forma tal que es capaz de generar una pila ordenada de los elementos coincidentes con dicho patrón según su jerarquía y posición original, para su posterior acceso, selección y recuperación. Con independencia del patrón y de la fuente de datos que el parser analiza, también existen otras características que definen su funcionamiento, como el recorrido ascendente bottom-up-parsing o descendente top-down-parsing, por derivación LL left to right, leftmost derivation o por ampliación LR left to right, rightmost derivation.

Un parser es de recorrido ascendente, cuando parte de los elementos básicos de una estructura jerárquica de tal forma que desconoce por completo sus posibles relaciones ascendentes, con terceros elementos padres o ancestros, por lo que su orden de inferencia se basa en la ampliación de tales estructuras con las de los niveles superiores, hasta alcanzar el primer elemento de la jerarquía. Por este motivo un parser de tipo ascendente bottom-up-parsing, también será de tipo LR left to right, rightmost derivation (CHAPMAN, N.P, 1987). Un parser es de recorrido descendente, cuando parte del primer elemento de la estructura jerárquica de tal forma que establece

relaciones con los elementos hijos y nietos mediante la derivación del análisis en cada uno de ellos de forma recursiva, hasta alcanzar los niveles inferiores de la jerarquía. Por este motivo un parser de tipo descendente top-down-parsing, también será de tipo LL left to right, leftmost derivation (GRUNE, D. and Jacobs, C., 1998). En el caso de los metadatos y del análisis de páginas web, los analizadores sintácticos parten de las estructuras propias de XML como patrones conocidos de comparación y análisis. Dado que XML es un lenguaje estructurado y anidado, resulta eminentemente jerárquico y de contexto gramatical conocido. Por estos motivos, los parser aplicados a XML son del tipo LL left to right, leftmost derivation.

Funcionamiento general de un parser XML

Un parser XML es un analizador sintáctico de estructuras anidadas de etiquetas. Ello significa que cualquier formato de sindicación es susceptible de ser analizado por este tipo de programas por el mero hecho de estar basados en XML. Por lo tanto el primer requisito para el funcionamiento de un parser XML es la disposición de un canal de sindicación que actúa como fuente de datos para el análisis. A continuación el sistema carga el archivo XML en búfer de memoria para empezar el análisis descendente de la estructura jerárquica. El primer paso es la detección de la cabecera XML, indicativa de que el archivo posee dicho dominio gramatical. Este paso resulta fundamental, puesto que determina la validación del lenguaje XML. A continuación se procede con un análisis descendente de la estructura jerárquica propia del formato de sindicación. Ello significa que tomará como punto de partida la primera etiqueta de apertura y cierre del formato. Esta primera confrontación también resulta clave, dado que las etiquetas de apertura del canal de sindicación contienen atributos xmlns para definir su propio namespace y el de los módulos que utilicen en todo caso, lo que puede facilitar la identificación de los juegos e etiquetas necesarios para interpretar el contenido del formato. No obstante, el parser por sí solo no entiende estas disquisiciones y únicamente anotará en su pila de elementos la existencia de atributos adscritos a la primera etiqueta del canal de sindicación. Recuérdese que al tratarse de un parser de análisis por derivación, comprobará las etiquetas de primer nivel jerárquico propias de la descripción del canal de sindicación y de las entradas de contenidos que lo conforman de forma secuencial y ordenada hasta agotar todas las alternativas posibles de derivación con el primer elemento del primer nivel jerárquico y sus sucesivos.

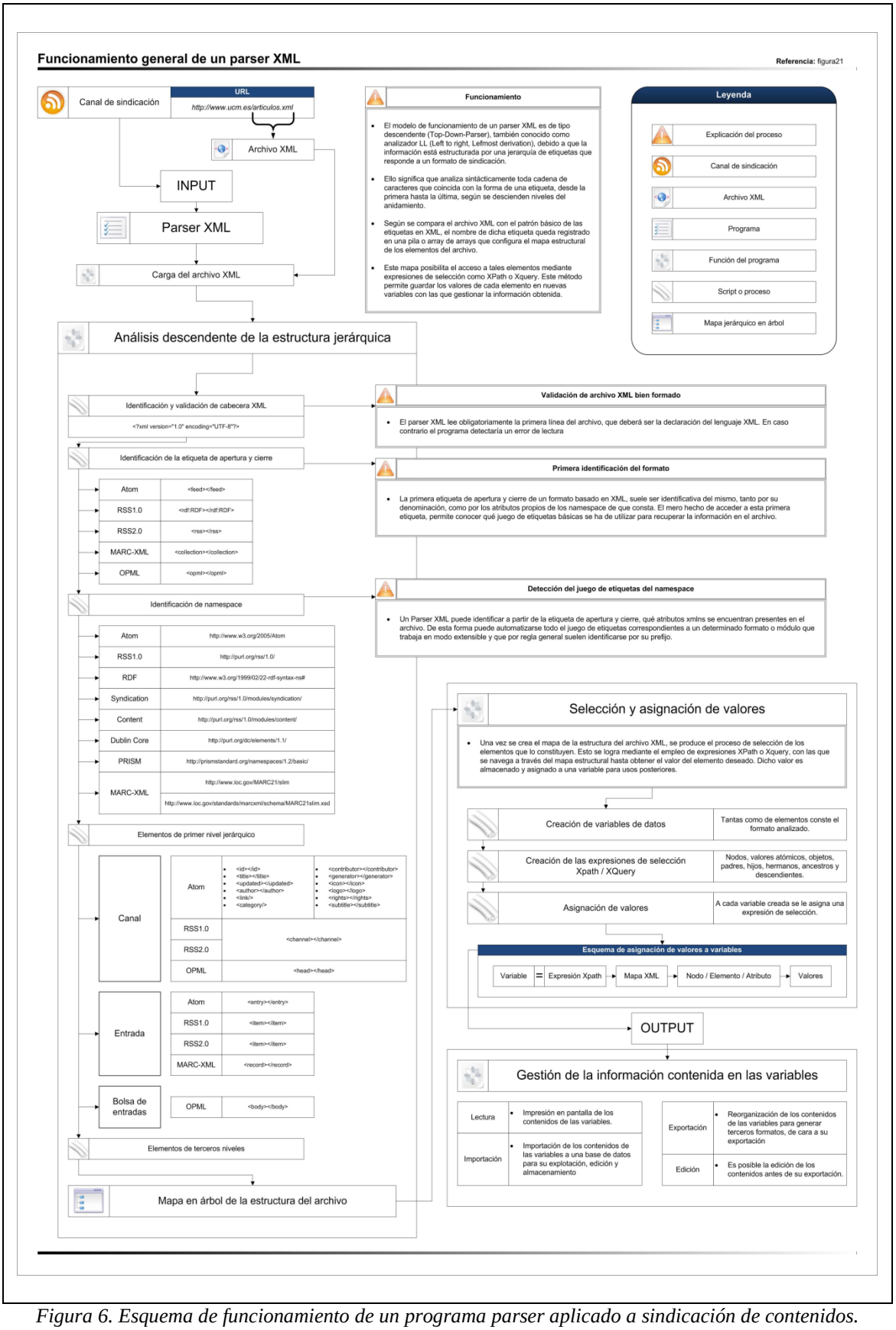


Figura 6. Esquema de funcionamiento de un programa parser aplicado a sindicación de contenidos. Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/esquemas/esquema005-parser.png

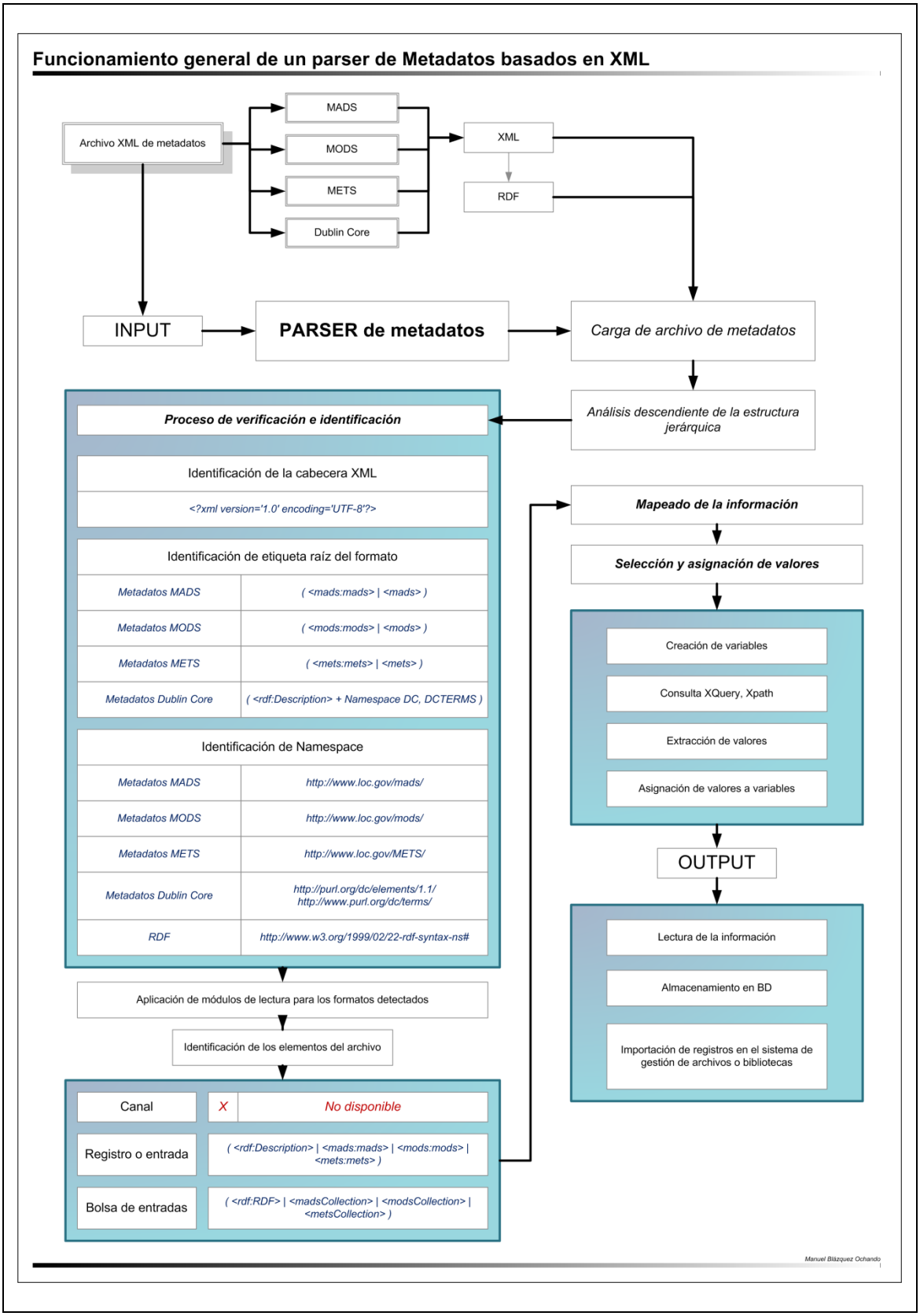


Figura 7. Esquema de funcionamiento de un programa parser especializado en metadatos. Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/esquemas/esquema006-parser.png

De esta forma, el parser configura una pila de elementos que es esencialmente un array de arrays o una matriz de matrices que a modo de mapa de la estructura del archivo de metadatos, permite acceder a sus contenidos, mediante expresiones de selección preferiblemente compuestas en lenguaje XPath o XQuery. Para almacenar sus contenidos, también se necesita un proceso denominado asignación de valores a variables, ello implica crear las variables propias de cada etiqueta para almacenar la información del formato. El método más ordenado y efectivo para lograrlo es la identificación de la variable a una función de selección expresada en XPath que actúa sobre el mapa estructural creado por el parser.

El resultado de su ejecución es la selección del contenido de dicha etiqueta y su asignación a la variable. Finalmente una vez asignados los valores a las variables, pueden emplearse para constituir un servicio de lectura del canal de sindicación, pueden importarse a una base de datos para su posterior edición o exportarse los contenidos a terceros formatos.

10. Webmetría y análisis de páginas web

Dada la importancia y extensión que ha adquirido la cibermetría en los últimos tiempos, se hace necesario conocer algunas definiciones de conceptos relacionados con la Bibliometría aplicada a la Web o Internet, es decir, la cibermetría. Este área de conocimiento, está en pleno estudio y desarrollo, por lo que existe una variación semántica bastante notable en muy corto espacio de tiempo, según se avanza en las investigaciones y pruebas.

Analizando el término "cibermetría", al descomponerlo en (ciber-) y (-metria), se indica la medición cuantitativa de la red virtual o la web. Es por ello, que se puede deducir también que la cibermetría es la aplicación de las técnicas bibliométrico estadísticas a la información recopilada en la web. Partiendo de la metodología como base para distinguir la ciencia que mide la web, cabe distinguir otro concepto emergente en muchos estudios especializados, se trata del término "webmetría". Si bien cibermetría corresponde al estudio cuantitativo de la web, ¿a qué corresponde el término webmetría? Según Björneborn, distingue el área de aplicación.

- **Cibermetría** (BJÖRNEBORN. 2004): Es el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías en Internet, desde perspectivas bibliométricas e informétricas.

- **Webmetría** (BJÖRNEBORN. 2004): Es el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías de una parte concreta de Internet, por regla general a una web o portal, desde perspectivas bibliométricas e informétricas.

Esto significa que la cibermetría acoge todo el espectro de análisis de la web y la webmetría selecciona una parte de ella, una sección o localización muy concreta. Por ejemplo el análisis de la web de española, corresponde a un estudio de tipo cibermétrico. Pero el estudio de la web de la universidad española es mucho más reducido y localizado lo que corresponde según Björneborn a un enfoque webmétrico. (ARROYO, N. 2005)

Factores que pueden influir en los estudios cibermétricos y webmétricos

- Frecuencia de actualización de los sitios web citantes y citados
- La modificación y actualización de los contenidos de una página o sitio web
- La difusión y el nivel de enlazamiento de una web con el resto
- La tipología documental de los recursos electrónicos en constante cambio

Qué es un webcrawler

El término webcrawler, también conocido con las denominaciones rastreador, araña, robot de búsqueda, crawler, spider, bot es un programa que cumple múltiples propósitos de análisis y extracción de información de la web. Constituye el instrumento de investigación principal con el que se realizan los estudios cibermétricos y webmétricos, lo que implica una estrecha relación entre la información que es capaz de recuperar y las técnicas de análisis, tabulación y medición de la metría. Pero ¿Cómo funciona un webcrawler? ¿Qué información puede recuperar? ¿Qué utilidad tiene para el documentalista, de cara a la elaboración de estudios webmétricos?

Cómo funciona un webcrawler

A continuación se presenta un diagrama que explica el funcionamiento del webcrawler Mbot, véase *figura8*. Se trata de un programa especializado en el desarrollo de análisis webmétricos para un determinado área del conocimiento en la web, o grupo demarcado de sitios y páginas web. El mecanismo de funcionamiento se basa en diversos pasos. En primer lugar es necesario elaborar un archivo denominado "*semilla.txt*" que contiene la muestra inicial de direcciones URL que se pretenden analizar. Ello implica un proceso manual de selección de las páginas y sitios web que serán objeto de estudio. Definido el marco de estudio y con ello el listado de direcciones, se realiza un proceso de configuración del webcrawler en el que se determina la profundidad del análisis según los niveles de enlazamiento de los sitios y páginas definidas en la semilla. Esto es analizar los vínculos de los sitios y páginas de la semilla de forma sucesiva, hasta finalizar el proceso.

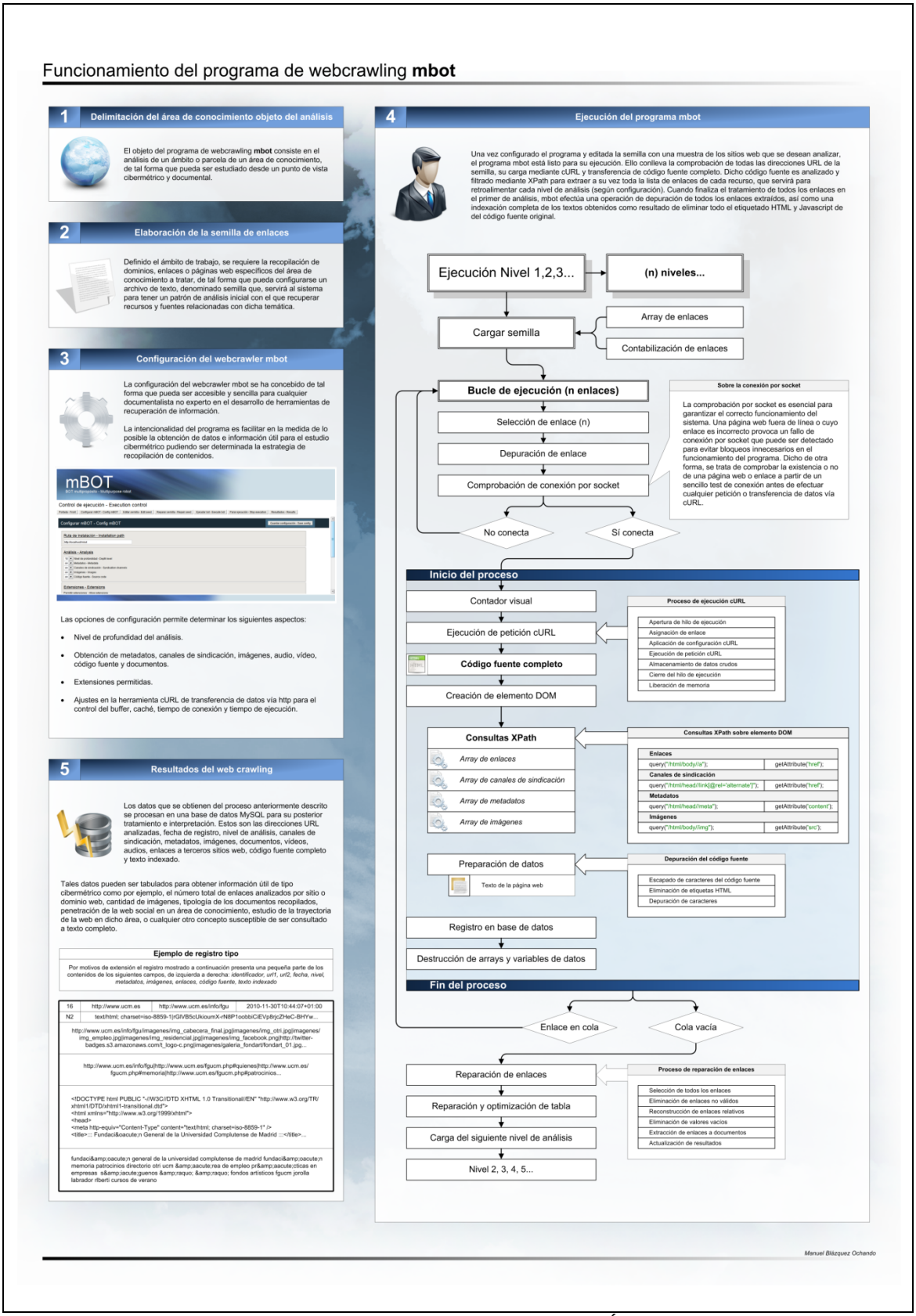


Figura 8. Funcionamiento del programa de webcrawling Mbot. (BLÁZQUEZ OCHANDO, M. 2011). Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/esquemas/esquema007-mbot.png

Cada salto de una página a otra, se denomina nivel de profundidad, de tal manera que es posible navegar de una página a través de sus vínculos, determinando un recorrido que puede ser trazado y reflejado en un sistema de información, como un webcrawler. Pero también pueden y deben configurarse otros ajustes de importancia, como por ejemplo el buffer (que permite retener la información del proceso de extracción de datos), el tiempo en cache (para determinar el número de segundos que el sistema mantiene las entradas DNS en memoria), el tiempo de conexión (define el número de milisegundos que el sistema espera cuando está intentando conectar con la dirección URL especificada) y el tiempo de ejecución máximo por URL. Finalmente también es relevante determinar filtros y extensiones para que el análisis del webcrawler sea más especializado y rápido. Por ejemplo aplicar restricciones por sitio web, por extensión o por patrones o *regexp*, permite diferenciar más fácilmente el tipo de enlace que se pretende recuperar en el análisis webmétrico y obviar o no aceptar aquel que fue especificado.

El proceso de webcrawling, consiste en la extracción de toda la información de la dirección URL de una página o sitio web objetivo que de forma secuencial presenta el programa Webcrawler. Mbot por ejemplo, descarga todo el código fuente de una página web objetivo y lo filtra para obtener todos sus elementos de forma ordenada de cara a un correcto almacenamiento. Dicho almacenamiento puede realizarse de múltiples formas, a través de archivos de texto plano, separados por comas CSV e incluso bases de datos como MySQL. En este sentido se crea un registro tipo que permite almacenar la información ordenada y filtrada por el webcrawler para su posterior tratamiento, tabulación y análisis. De hecho pueden existir registros completos de todo el contenido de una página web, o existir diversos archivos, bases de datos especializados en el almacenamiento de enlaces, imágenes, documentos que fueron recopilados durante la extracción de datos. Para realizar la extracción de la información de una página web y dividir sus correspondientes elementos, se emplean programas de tipo parser dentro del webcrawler, capaces de reconocer las etiquetas e instrucciones de HTML y con ello desarrollar el proceso de extracción de los datos contenidos en ellas. De la misma forma que existen parsers capaces de analizar archivos de metadatos en XML, existen parsers capaces de recopilar los enlaces de una página, los párrafos, metadatos en HTML, canales de sindicación, imágenes, documentos, archivos multimedia, entre otros. No

obstante, la información extraída en este estadio no está definitivamente preparada, ya que en muchos casos los enlaces (unidad fundamental del análisis webmétrico) resultan ser de tipo relativo, lo que dificulta el acceso a subsiguientes niveles de análisis durante el proceso de webcrawling. Ello implica que se desarrolle un proceso de depuración y preparación de las direcciones URL relativas, convirtiéndolas en absolutas, de forma tal que aseguren el acceso a la información que en tal caso vinculan.

Qué información se puede recuperar

La información que en un webcrawler se pueda recuperar, marca en muchos casos los posibles estudios webmétricos que se puedan realizar. A priori es posible recuperar cualquier elemento o contenido de un sitio o página web. Suelen ser objetivo de extracción los títulos de los sitios y páginas, sus metadatos y meta-etiquetas, sus canales de sindicación, las imágenes, documentos, archivos multimedia, código fuente y texto completo párrafo a párrafo.

Qué utilidad tiene para el documentalista, de cara a la elaboración de estudios webmétricos

La utilidad de un webcrawler para la elaboración de estudios webmétricos es capital, dado que la información recopilada por este tipo de programas posibilita la elaboración de una muestra de datos lo suficientemente cualificada y completa como para obtener datos directos sobre los siguientes aspectos:

- Banco de datos de imágenes, documentos, metadatos, canales de sindicación.
- Colección de textos para la recuperación de información.
- Nº total de enlaces analizados (incluyendo duplicaciones).
- Nº total de enlaces únicos analizados (sin duplicaciones).
- Nº total de enlaces analizados según niveles de profundidad.
- Nº de dominios, sitios y páginas web analizadas en cada nivel de profundidad.
- Distribución de dominios de tipo genérico y geográfico, según sitios y páginas web.
- Distribución de tipos de documentos según su extensión o formato. Por ejemplo documentos ofimáticos, audiovisuales, imágenes, web dinámica y estática.

- Análisis de macroestructura de la web. Determinación de los componentes de la web Main, Out, In, Island, Tunnel, Tentacle In, Tentacle Out, según el enlazamiento de los vínculos entre sitios y páginas web del análisis llevado a cabo.
- Ranking de sitios y páginas con más metadatos.
- Distribución de la tipología de metadatos más utilizada.
- Ranking de sitios web con más enlaces únicos y páginas.
- Ranking de sitios web con más documentos, imágenes, archivos audiovisuales, etc.
- Ranking de sitios web con más canales de sindicación.
- Análisis de coenlaces. Sitios y páginas más coenlazados.
- Sitios web más enlazados.
- Páginas web más enlazadas.
- Trazado de hipervínculos entre sitios y páginas web que permite la elaboración de gráficas topográficas de la web analizada.

Análisis de enlaces

Como se puede comprobar, cualquier análisis cibermétrico y webmétrico requiere ineludiblemente de un análisis de los enlaces. Ello significa que la citación entendida en el ámbito de la bibliometría, puede encontrarse igualmente en los documentos de naturaleza electrónica, publicados en la web, añadiendo la variable del enlazamiento. Dicho de otra forma, se pueden aplicar las técnicas de análisis bibliométricas, pero requerirán de un aumento de los vínculos enlazados, concretamente de los "links" que el documento tenga. de esta forma la citación bibliográfica no es el único objeto de análisis y el enlace hipertextual juega un papel determinante para definir la correlación entre varias páginas web, incluso si se trata de una referenciación bibliográfica. De esta forma, pueden existir diversos tipos de análisis de enlaces:

- Análisis de "sitas" o "links que vinculan sitios web" comprobando cuáles son los sitios web de mayor relevancia por el número de enlaces externos e internos que reciben.

- Análisis de "co-citas" que mide el número de veces que aparecen dos documentos referenciados recíprocamente, lo que indica su aproximación temática.
- Análisis de "co-enlaces" que identifica si dos sitios web están referenciados recíprocamente en sus páginas web, midiendo el número de enlaces que sí co-enlazan y el número de enlaces que no co-enlazan.
- Análisis de "Co-ocurrencia por palabras" que determina cuantos documentos tienen en común una serie de descriptores, frases o palabras clave, contabilizando su frecuencia en el número de coincidencias ocurridas para cada término.
- Análisis de macroestructura de la web, véase *figura9*.

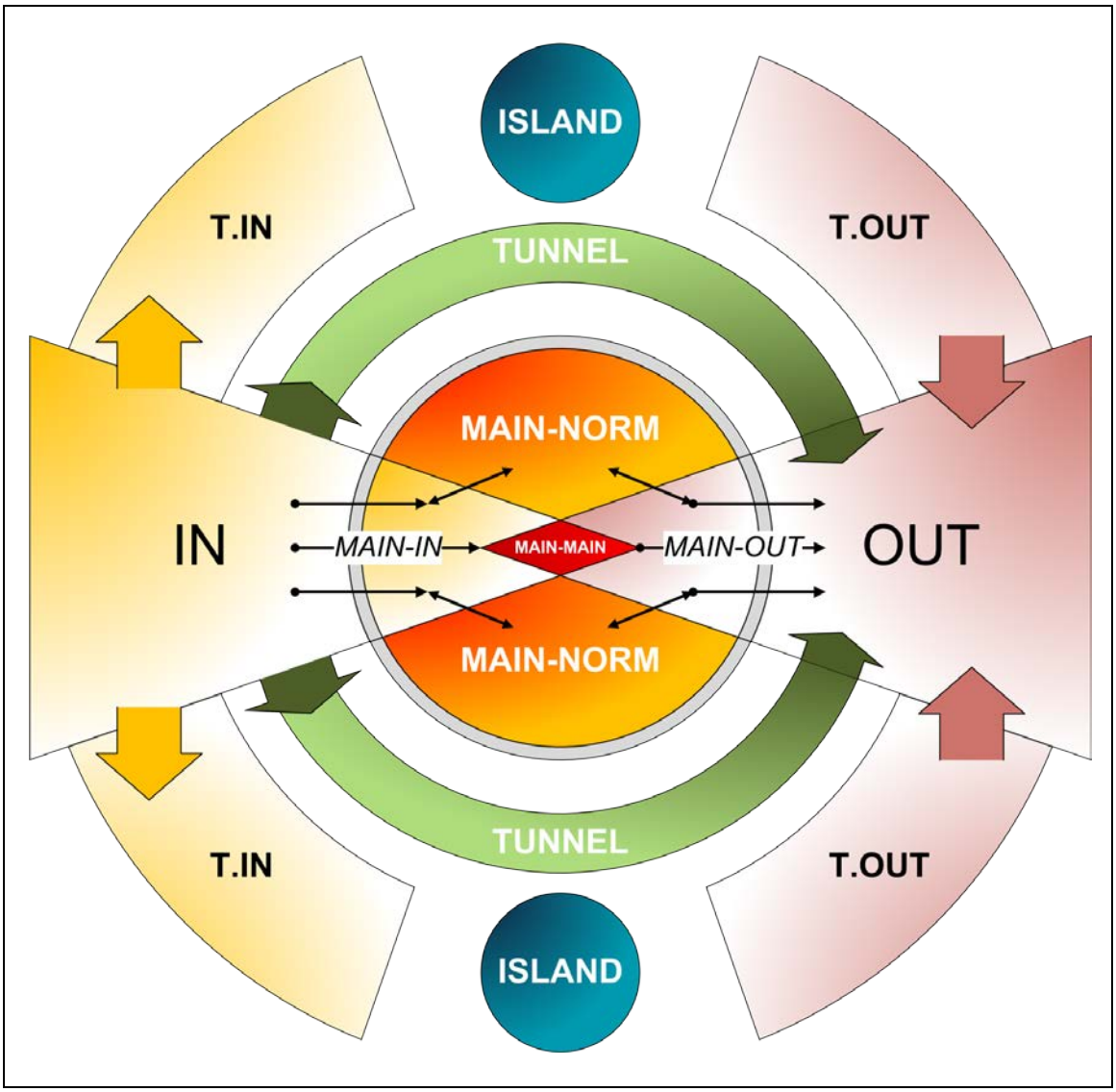


Figura 9. Representación de la macroestructura web, o análisis de grafo. Disponible en: http://www.mblazquez.es/blog_ccdoc-busqueda-internet/esquemas/esquema008-grafo.png

Componente	Descripción	Esquema
MAIN	Sitios web fuertemente conectados en todas las páginas web.	url1←→ url main ←→ url2 linkmap url ←→ url main ←→ linkmap url
IN	Sitios web que enlazan el componente MAIN, pero no lo son de manera recíproca.	url1 → url main url2 → url main linkmap url → url main
OUT	Sitios web que son enlazados desde MAIN pero no de forma recíproca.	url main → url1 url main → url2 url main → linkmap url
ISLAND	Sitios web desconectados de los demás o con un pobre nivel de enlazamiento. Pueden ser alcanzados por el resto de componentes, pero ellos no enlazan a ninguno de ellos.	→ url1 ← → url2 ← → linkmap url ←
TENTACLE IN	Sitios web que sólo conectan el componente IN	(url1 → url main) → linkmap url (url2 → url main) → linkmap url
TENTACLE OUT	Sitios web que sólo conectan el componente OUT	(url main → url1) → linkmap url (url main → url2) → linkmap url
TUNNEL	Sitios web que vinculan el componente IN y OUT sin necesidad de enlace a través de MAIN	url1 → url2 url2 → url1 linkmap url → linkmap url

Tabla 34. Descripción de componentes de la macroestructura web

11. Técnicas de consulta dinámica GET en Google

Las consultas dinámicas en los buscadores son consideradas una de las herramientas más versátiles para la recuperación de información. Se considera que una consulta es dinámica cuando el sistema que hace posible la transmisión de las cadenas de consulta por medio del método GET o POST, permite su correcto procesamiento para su resolución en una base de conocimiento o colección. En este sentido, se pueden realizar consultas empleando el método POST propio de los formularios de consulta, o bien por medio de la construcción de una dirección URL que posea la información de la consulta. Ambos métodos son extensivos en la mayoría de buscadores de la web, especialmente en Google, que posibilita una amplia lista de opciones de recuperación utilizando para ello sus variables.

Qué es el método GET y POST

Los métodos GET y POST forman parte del sistema de peticiones y comunicaciones del protocolo HTTP utilizado extensivamente en la web. El método de petición POST, permite el envío de datos a un servidor web mediante una cabecera y cuerpo de mensaje para su posterior almacenamiento y tratamiento. Esto significa un bloque de datos enviado con la solicitud o petición en el cuerpo del mensaje, especificando el tipo de contenido y datos. Su uso más conocido se encuentra en cualquier formulario de la web dentro de los cuáles se indica el método de transmisión de datos, la página web de destino encargada del procesamiento de la información y los campos de texto, campos seleccionables y áreas de texto que contendrán los datos. De esta forma es posible enviar datos recuperables a través de los nombres dados a dichos campos de texto. Véase en la *tabla35* un ejemplo de formulario con campos definidos.

```
<form action='pagina-web-destino.php' method='post'>
<input type='text' name='data1' value=''/>
<input type='text' name='data2' value=''/>
<input type='text' name='data3' value=''/>
<input type='text' name='data4' value=''/>
<input type='text' name='data5' value=''/>
<input type='submit' name='send' value='Enviar datos'/>
</form>
```

Tabla 35. Ejemplo de formulario web que emplea el método POST

En cambio el método HTTP GET se emplea para efectuar peticiones de información o recuperación de información en un servidor web objetivo. Esto es una petición que se efectúa por medio de la URL de una página de consulta, que está habilitada para recopilar los términos de una búsqueda o bien sus parámetros (contexto en el que se explica el presente artículo). Además, las peticiones GET están diseñadas para no cubrir otros cometidos como podría ser el envío de datos seguros para su almacenamiento, tal como funcionaría con el método POST.

De hecho el método GET es público, carece de cifrado y no se considera un método seguro para el envío de datos. Por este motivo su correcto diseño e interacción en los servicios de búsqueda y recuperación de información resulta esencial para evitar problemas de alteración o modificación de los datos del servidor. De hecho, cuando un usuario consulta con su navegador web cualquier página web como por ejemplo http://www.google.es/ está efectuando una consulta o petición a un servidor remoto, véase *tabla36*.

Dicha petición permite al servidor HTTP de Google interpretar la consulta que se está realizando por método GET, concretamente la aplicación de consulta del navegador (*source=search_app*), que no existen más peticiones (*Connection:close*), que el idioma, codificación y tipo de datos del navegador del cliente son los especificados, no permitiendo el control de versiones del cache.

Petición de cabecera HTTP
GET /webhp?source=search_app HTTP/1.1 Host: www.google.es Connection: close Accept-Encoding: gzip Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8 Accept-Language: es-ES,es;q=0.8 Accept-Charset: ISO-8859-1,UTF-8;q=0.7,*;q=0.7 Cache-Control: no-cache

Tabla 36. Ejemplo de petición de cabecera HTTP

Ante la petición del cliente, existe una respuesta de cabecera HTTP, véase *tabla37*. En ella se muestra el estado de recepción de la petición, normalmente el mensaje suele ser 200 OK que significa una correcta recepción de los datos de la petición. Por otro lado la fecha y hora de la respuesta, identificación de que la respuesta no expira (-1), datos de sesión (Cookies), plataforma de privacidad de las consultas (P3P), identificación del

servidor (GWS Google Web Server), frame de carga y cierre de conexión. Junto con la cabecera se remite el cuerpo del mensaje de respuesta enviado en formato HTML como página web.

Respuesta de cabecera HTTP	
Status	HTTP/1.1 200 OK
Date	Mon, 03 Dec 2012 07:19:55 GMT
Expires	-1
Cache-Control	private, max-age=0
Content-Type	text/html; charset=ISO-8859-1
Set-Cookie	PREF=ID=a2379f825779c3c2:FF=0:TM=1354519195:LM=1354519195:S=3xPq3l_pqAmIbJWo; expires=Wed, 03-Dec-2014 07:19:55 GMT; path=/; domain=.google.es
Set-Cookie	NID=66=ZC7AMosQZKkONqSiAvUIdkVNisns4FgofwpJJaalT3KvKcJoHTkWr1wnStTSLw80-X9cjHahtjNz7lqj5kH2bET1JaOjjkEjMV_Dyts_AyX8smZnmbijRhul1o8K-sKB; expires=Tue, 04-Jun-2013 07:19:55 GMT; path=/; domain=.google.es; HttpOnly
P3P	CP="This is not a P3P policy! See http://www.google.com/support/accounts/bin/answer.py?hl=en&answer=151657 for more info."
Server	gws
X-XSS-Protection	1; mode=block
X-Frame-Options	SAMEORIGIN
Connection	close

Tabla 37. Ejemplo de respuesta de cabecera HTTP

Las variables

Si bien, el envío de datos por medio del método GET y POST de HTTP, posibilita la recuperación, consulta y almacenamiento de la información en la web, también es cierto, que para poder manejar dicha información, ésta debe ser correctamente identificada. Esto es determinar cuál es el nombre de las variables que contendrán los datos o informaciones que se transmiten. De hecho una variable es un método de almacenamiento de cadenas de caracteres, que se utiliza para recuperar la información enviada por los distintos métodos de HTTP, véase *tabla38*.

Lenguaje	Variable GET	Variable POST
PHP	$_GET[nombre-variable]	$_POST[nombre-variable]
ASP	Request.QueryString("nombre-variable")	
RUBY	CGIMethods.parse_query_paremeters (@request.query_string)	CGIMethods.parse_query_parameters (@request.raw_post)

Tabla 38. Métodos de recuperación de variables

Por ejemplo, de acuerdo al formulario de la tabla1, las variables POST que la pagina-de-destino.php recuperaría serían: *$_POST[data1]*, *$_POST[data2]*, *$_POST[data3]*, *$_POST[data4]*, *$_POST[data5]*. La notación de las variables corresponde al lenguaje de programación PHP y se observa cómo el nombre de los campos del formulario, identifica el nombre de la variable transmitida por método POST. Con el método GET sucede lo mismo, sólo que el envío de datos se produce a través de la URL.

Consultas básicas en Google mediante URL y método GET

Consulta básica dinámica en Google
Variable de consulta básica de Google (q)
https://www.google.com/search?q=consulta
Variable de idioma (hl)
https://www.google.com/search?hl=idioma&q=consulta
Conversiones ASCII a hexadecimal. Véase: http://www.mikezilla.com/exp0012.html
(Espacio en blanco) = (%20)

Tabla 39. Consulta dinámica básica en Google

Consultas avanzadas en Google mediante URL y método GET

Consulta avanzada dinámica en Google		
Variable	**Valor/es**	**Descripción**
q	Consulta	Cadena de consulta completa del usuario
start	Número superior a 0	Muestra un determinado número de resultados desde 0
num	1-100	Muestra determinado número de resultados por página
filter	0-1	Filtrar resultados duplicados. Si se usa 0 mostrará resultados únicos.

restrict	Código de restricción	Código de restricción de resultados por país. Por ejemplo:*countryES, countryFR, countryDE, countryUK, countryUS*
hl	Código de idioma	Idioma en el que Google muestra la información de los resultados. Por ejemplo: *es, en, de, fr, it*
lr	Código de idioma	Recupera páginas web escritas en el código de idioma especificado. Por ejemplo: *lang_es, lang_en, lang_fr, lang_de, lang_de, lang_it*
ie	UTF-8	Codificación del input de la consulta en Google
oe	UTF-8	Codificación del output de resultados de Google
as_epq	Consulta	Frase de consulta exacta, evita tener que emplear las comillas dobles en la variable (q)
as_ft	i - e	El valor (i) permite incluir en la búsqueda el tipo de archivo determinado. El valor (e) permite excluir de la búsqueda el tipo de archive determinado
as_filetype	Extensión de archive	Define el tipo de archivo por su extensión que se aplicará en la búsqueda
as_qdr	m - y	Determina el periodo de tiempo dentro del cual se recuperarán los contenidos. Por ejemplo: *m1, m2, m3* (corresponde al último mes, últimos 2 meses y últimos tres meses), *y2, y3* (corresponde a los últimos dos y 3 años)
as_nlo	Número inferior	Busca resultados con números comprendidos entre *as_nlo y as_nhi*
as_nhi	Número superior	
as_oq	consulta	Recupera páginas que contengan alguna de las palabras indicadas
as_occt	any, title, body, url, links	Recupera páginas que contengan alguna de las palabras indicadas en un punto específico. Por ejemplo el valor (any) indica en cualquier punto de la página web, (title) especifica en el título, (body) corresponde al cuerpo de la página web, (url) en la dirección de la propia página y (links) en alguno de los enlaces propios de la página web.
as_dt	i - e	El valor (i) permite incluir búsquedas del dominio especificado. El valor (e) permite excluirlas
as_sitesearch	Consulta	Permite incluir o excluir de la consulta un dominio o sitio web que se especifique.
safe	active off	Permite realizar búsquedas seguras
as_rq	URL	Permite realizar búsquedas de páginas similares a la URL especificada
as_lq	URL	Recupera páginas vinculadas a la URL especificada

Tabla 40. Consulta dinámica avanzada en Google

12. Búsqueda con operadores avanzados y directorios de servidores

Las variables observadas en el artículo anterior permiten elaborar consultas mediante la construcción de la URL y el envío de las cadenas de consulta mediante el método GET. A continuación se exponen en profundidad algunos de los operadores avanzados de Google que mayor interés tienen para el documentalista, así como su aplicación.

Operadores avanzados

- ***Buscar dentro del título de una página web. Operador "intitle".*** El título de las páginas web está contenido entre las etiquetas *<title></title>*. Mediante el operador intitle pueden recuperarse todas las páginas web que contengan en el título algunas de las palabras especificadas en la consulta. Por ejemplo intitle:"diario" que recuperaría todos los diarios de prensa, diarios personales, blogs o páginas que contengan la palabra diario. Esta consulta es de especial utilidad en la restricción de búsquedas y para la concreción de recursos especializados. Además pueden emplearse más de un término de consulta, siempre y cuando esté entrecomillado y separado por espacios. Por ejemplo intitle:"diario" "prensa", permite mejorar la precisión de las páginas web de diarios como medios de comunicación, discriminando mejor los resultados.

- ***Buscar dentro del texto de una página web. Operador "allintext".*** Permite buscar en el contenido textual de las páginas web entre las etiquetas *<body></body>*, lo que significa que la recuperación se realice exclusivamente en este apartado y no en el título, enlaces o url de la propia página web. Por ejemplo allintext:"noticias" "ministerio de economía" "españa" permitiría recuperar todas las noticias publicadas en páginas web sobre el ministerio de economía.

- ***Buscar en URL. Operador "inurl" y "allinurl".*** Existen diferencias en el uso de los operadores de búsqueda URL como inurl y allinurl. En principio la recuperación en ambos casos está focalizada en la URL del recurso o página

web objetivo y no en sus enlaces. Por otra parte el operador "*inurl*" indica que la cadena de consulta debe estar presente en la dirección URL o en cualquier parte del nombre de archivo de dicha URL. Por ejemplo una búsqueda *inurl:diario oficial* permitirá obtener casi 15 millones de resultados. Ello se debe a que los resultados pueden contener las palabras diario oficial en cualquier punto de su URL incluyendo el nombre del archivo o página web. Sin embargo si se emplea el operador *allinurl:"diario oficial"*, el número de resultados se reduce por debajo de los 2 millones. Ello ocurre porque el operador allinurl restringe los resultados, mostrando únicamente aquellos que contengan las palabras diario oficial en la URL principal del sitio web, dominio o subdominio, pero no en los nombres de los archivos, documentos, páginas o recursos que enlazan.

- ***Buscar dentro de un sitio web concreto. Operador "site".*** Al igual que con la variable as_sitesearch, Google dispone del operador site cuyo propósito es exactamente el mismo. La recuperación mediante el operador site es compatible con el uso de otros operadores como por ejemplo inurl o allinurl. Esto hace posible recuperar páginas web muy concretas en un espacio de un dominio o sitio web previamente acotado. Por ejemplo *site:boe.es inurl:2012/12/04* permite recuperar todas las páginas del Boletín Oficial del Estado publicadas en dicha fecha.

- ***Buscar archivos o documentos según su extensión. Operador "filetype".*** El operador filetype, realiza las mismas funciones que la variable as_filetype. Su aplicación permite una recuperación precisa de determinados tipos de documentos con cualquier tipo de extensión, más o menos frecuentes. En este sentido una página web de referencia para encontrar el tipo de extensión adecuada para la recuperación de una determinada información es el sitio web http://filext.com/ que contiene una de las mayores bases de datos de extensiones del mundo, con su correspondiente descripción, programas que utilizan o generan ese tipo de archivos, aplicación, datos que contienen, etc. Actualmente contiene más de 51.000 tipos de archivos registrados. En cuanto a la búsqueda y recuperación de información dentro de un tipo de documento, ésta puede ser realizada utilizando los términos de consulta correspondientes después de especificar la extensión del documento objetivo. Por ejemplo *filetype:pdf* tesis

doctoral, permite recuperar todos los documentos con extensión PDF en cuyo texto aparezcan las palabras tesis doctoral.

- ***Buscar enlaces a una página. Operador "link".*** El operador link no es equivalente a la variable as_qdr=links pero sí causa el mismo efecto que la variable as_lq. Se utiliza para obtener aquellos recursos o páginas que enlazan la URL especificada. Por ejemplo *link:http://www.ucm.es* y *https://www.google.com/search?as_lq=http://www.ucm.es* proporcionan los mismos resultados, permitiendo observar por cuántas páginas web es vinculada el sitio de la Universidad Complutense.

- ***Buscar texto dentro de los enlaces de un sitio web. Operador "inanchor".*** Permite recuperar aquellas páginas web que contengan el texto de la consulta en la representación textual de sus enlaces. Concretando este aspecto, dentro de las etiquetas anchor *representación textual*. Por ejemplo *inanchor: "recursos biblioteconomía"*, permitiría recuperar todos los portales y directorios de recursos que estuvieran enlazados mediante el texto especificado.

- ***Obtener información de una página web. Operador "info".*** Para obtener todas las posibilidades de acción con respecto a una página web, así como acceder a sus correspondientes resúmenes, se requiere el empleo del operador info. Por ejemplo *info:http://ccdoc-sistemasrecuperacioninternet.blogspot.com.es/* muestra todas las opciones de interacción, como la vista en cache del sitio web, páginas similares, páginas enlazadas, páginas del sitio, y consulta por frase exacta de la URL.

- ***Páginas similares o relacionadas. Operador "related".*** Muestra sitios web relacionados o similares al especificado. Realiza las mismas funciones que la variable *as_rq*. El operador *relataded* es incompatible con el uso de otros operadores.

Directorios de servidores

La localización de listados de directorios es una técnica de gran interés para el Documentalista experimentado, debido a que le permite recuperar documentos que de otro modo sería imposible. Todos los documentos que se divulgan en la web a través de artículos y páginas web están almacenados en alojamientos de servidores, que en muchos casos pueden ser accesibles de forma sencilla utilizando las consultas adecuadas. Por ejemplo si se desean recuperar trabajos, investigaciones, artículos científicos relativos a "*Information retrieval*" existen dos alternativas.

En primer lugar utilizar las conocidas bases de datos bibliográficas como la Web Of Science o por el contrario utilizar métodos menos ortodoxos como la técnica de consulta de listados de directorios de servidores. En tal caso, la consulta *intitle:index.of "parent directory"* acompañada de los términos o palabras clave adecuados permite listar cualquier directorio abierto. En muchos casos, esta técnica ha permitido encontrar brechas de seguridad en muchos sistemas y redes institucionales, por lo que se ruega un correcto uso de la misma.

Por ejemplo, para resolver la consulta "*Information retrieval*", podría emplearse la sentencia *intitle:index.of "parent directory" information retrieval* que devolvería documentación producida por algunos de los principales grupos de investigación especializada en el área de conocimiento e incluso encontrar directorios de servidor abiertos de profesores e investigadores concretos como por ejemplo el del profesor Mike Thelwall *http://www.scit.wlv.ac.uk/~cm1993/papers/*. Esta demostración permite poner de relieve la importancia de este tipo de búsquedas, incluso cuando se trata de la localización de archivos concretos, en cuyo caso la consulta se vería modificada de la siguiente forma *intitle:index.of "parent directory" ws_ftp.log information retrieval*. El objetivo de la sentencia es introducir el nombre del archivo log tipo que permite visualizar todos los movimientos de carga de archivos en el alojamiento.

No obstante se debe tener en cuenta que no existe una única manera de buscar listados de directorios en servidores. Esto viene determinado por la variedad en la tipología y versiones de los distintos servidores, véase *tabla41*.

Listado de directorios de servidores web	
"AnWeb/1.42h" intitle:index.of "Apache Tomcat/" intitle:index.of "Apache-AdvancedExtranetServer/" intitle:index.of "Apache/df-exts" intitle:index.of "Apache/" intitle:index.of "Apache/AmEuro" intitle:index.of "Apache/Blast" intitle:index.of "Apache/WWW" intitle:index.of "Apache/df-exts" intitle:index.of "CERN httpd 3.0B (VAX VMS)" intitle:index.of "CompySings/2.0.40" intitle:index.of "Davepache/2.02.003 (Unix)" intitle:index.of "DinaHTTPd Server/1.15" intitle:index.of "HP Apache-based Web "Server/1.3.26" intitle:index.of "HP Apache-based Web "Server/1.3.27 (Unix) mod_ssl/2.8.11 OpenSSL/0.9.6g" intitle:index.of "HP-UX_Apache-based_Web_Server/2.0.43" intitle:index.of "httpd+ssl/kttd" * server at intitle:index.of "IBM_HTTP_Server" intitle:index.of "IBM_HTTP_Server/2.0.42" intitle:index.of "JRun Web Server" intitle:index.of "LiteSpeed Web" intitle:index.of "MCWeb" intitle:index.of "MaXX/3.1" intitle:index.of "Microsoft-IIS/* server at" intitle:index.of "Microsoft-IIS/4.0" intitle:index.of "Microsoft-IIS/5.0 server at" intitle:index.of "Microsoft-IIS/6.0" intitle:index.of "OmniHTTPd/2.10" intitle:index.of "OpenSA/1.0.4" intitle:index.of "OpenSSL/0.9.7d" intitle:index.of "Oracle HTTP Server/1.3.22" intitle:index.of "Oracle-HTTP-Server/1.3.28" intitle:index.of "Oracle-HTTP-Server" intitle:index.of "Oracle HTTP Server Powered by Apache" intitle:index.of "Patchy/1.3.31" intitle:index.of "Red Hat Secure/2.0" intitle:index.of "Red Hat Secure/3.0 server at" intitle:index.of "Savant/3.1" intitle:index.of "SEDWebserver *" "server at" intitle:index.of "SEDWebserver/1.3.26" intitle:index.of "TcNet httpsrv 1.0.10" intitle:index.of "WebServer/1.3.26" intitle:index.of "WebTopia/2.1.1a " intitle:index.of "Yaws 1.65" intitle:index.of "Zeus/4.3" intitle:index.of "Apache/1.0" intitle:index.of "Apache/1.1" intitle:index.of "Apache/1.2" intitle:index.of "Apache/1.2.0 server at" intitle:index.of	"Apache/1.2.4 server at" intitle:index.of "Apache/1.2.6 server at" intitle:index.of "Apache/1.3.0 server at" intitle:index.of "Apache/1.3.2 server at" intitle:index.of "Apache/1.3.1 server at" intitle:index.of "Apache/1.3.1.1 server at" intitle:index.of "Apache/1.3.3 server at" intitle:index.of "Apache/1.3.4 server at" intitle:index.of "Apache/1.3.6 server at" intitle:index.of "Apache/1.3.9 server at" intitle:index.of "Apache/1.3.11 server at" intitle:index.of "Apache/1.3.12 server at" intitle:index.of "Apache/1.3.14 server at" intitle:index.of "Apache/1.3.17 server at" intitle:index.of "Apache/1.3.19 server at" intitle:index.of "Apache/1.3.20 server at" intitle:index.of "Apache/1.3.22 server at" intitle:index.of "Apache/1.3.23 server at" intitle:index.of "Apache/1.3.24 server at" intitle:index.of "Apache/1.3.26 server at" intitle:index.of "Apache/1.3.27 server at" intitle:index.of "Apache/1.3.27-fil" intitle:index.of "Apache/1.3.28 server at" intitle:index.of "Apache/1.3.29 server at" intitle:index.of "Apache/1.3.31 server at" intitle:index.of "Apache/1.3.33 server at" intitle:index.of "Apache/1.3.34 server at" intitle:index.of "Apache/1.3.35 server at" intitle:index.of "Apache/2.0 server at" intitle:index.of "Apache/2.0.32 server at" intitle:index.of "Apache/2.0.35 server at" intitle:index.of "Apache/2.0.36 server at" intitle:index.of "Apache/2.0.39 server at" intitle:index.of "Apache/2.0.40 server at" intitle:index.of "Apache/2.0.42 server at" intitle:index.of "Apache/2.0.43 server at" intitle:index.of "Apache/2.0.44 server at" intitle:index.of "Apache/2.0.45 server at" intitle:index.of "Apache/2.0.46 server at" intitle:index.of "Apache/2.0.47 server at" intitle:index.of "Apache/2.0.48 server at" intitle:index.of "Apache/2.0.49 server at" intitle:index.of "Apache/2.0.49a server at" intitle:index.of "Apache/2.0.50 server at" intitle:index.of "Apache/2.0.51 server at" intitle:index.of "Apache/2.0.52 server at" intitle:index.of "Apache/2.0.55 server at" intitle:index.of "Apache/2.0.59 server at" intitle:index.of

Tabla 41. Consultas para mostrar directorios de distintos tipos y versiones de servidores

Por otra parte, también pueden aplicarse otras técnicas de consulta de directorios, denominadas "*de recorrido de directorios*". Consisten en el empleo del operador inurl combinado con las consultas anteriores. Ello permite especificar otros directorios que estén contenidos dentro del principal, efectuando un recorrido completo en todos sus apartados, por ejemplo *intitle:index.of "parent directory" inurl:"/paper/" information retrieval* permite recuperar todos los directorios que contengan artículos o papers

especializados en recuperación de información. Para ello los resultados deberán cumplir la condición de contener una carpeta denominada "*paper*".

13. Extensión de consultas avanzadas y recuperación de volcados de datos

Extensión de consultas avanzadas

- ***Filtrar extensiones y archivos en un sitio web.*** En muchos casos, la exigencia en las búsquedas de datos y documentos plantean el empleo de diversos operadores que identifiquen cuáles son los formatos válidos y desde qué sitio deben ser recuperados. Por ejemplo la consulta *-ext:html -ext:htm -ext:shtml -ext:asp -ext:php site:csic.es* permite obtener todos los subdominios de la página web del Consejo Superior de Investigaciones Científicas CSIC. Ello es debido a que se indica claramente qué archivos no son deseados entre los resultados. En tal caso se emplea el signo menos (-) precedido del operador de extensiones (*ext:*) y la extensión correspondiente. De esta forma entre los resultados no estarán presentes ninguna página html, asp o php que configuran todas las páginas web del sitio del CSIC. Por el contrario, el resultado obtenido serán todos los subdominios que contenga "*csic.es*". Para obtener un determinado tipo de documento dentro del dominio y subdominios del CSIC, tan sólo sería necesario modificar la consulta dada por la *siguiente ext:pdf -ext:html -ext:htm -ext:shtml -ext:asp -ext:php site:csic.es* en la que se indica la presencia de los archivos de extensión (*pdf*), que a su vez es equivalente a la expresión *+ext:pdf -ext:html -ext:htm -ext:shtml -ext:asp -ext:php site:csic.es* ya que el signo más (+) se emplea para indicar el cumplimiento obligatorio de la condición, filtro u operador que se está utilizando. Como se podrá observar existen múltiples formas de aludir a un mismo objetivo, como por ejemplo el operador (*filetype:*) y (*ext:*) cuya finalidad es la misma.

- ***Recuperar copias de seguridad y archivos temporales.*** En muchas ocasiones, puede ser necesario realizar consultas sobre archivos, documentos, bases de datos o páginas web publicadas en el pasado o cuya copia de seguridad alberga información de interés. En esos casos es posible realizar consultas para recuperar tales copias de seguridad y archivos temporales de forma sencilla mediante el operador (*inurl:*) utilizando las palabras claves y extensiones adecuadas,

utilizadas por los principales archivos de seguridad y almacenamiento. Por ejemplo *inurl:temp, inurl:tmp, inurl:backup, inurl:bak.* Estos casos pueden reproducirse en combinación con las consultas de directorios de servidores, como por ejemplo *intitle:index.of "parent directory" inurl:backup site:mit.edu* que permitiría observar los directorios de backup de los dominios, subdominios y páginas del MIT. De esta forma y mejorando la combinación de los operadores se pueden obtener los archivos backup en formato sql de un sitio web completo. Por ejemplo al realizar la consulta *ext:sql inurl:backup* se obtiene el enlace http://www.dpm-cultura.org/static/files.bk/backup.sql que contiene la copia de seguridad del sitio web de la Delegación de Cultura de la Diputación de Málaga, que pudiera contener información de interés para el trabajo documental.

- ***Combinación de extensiones y operadores.*** En muchas ocasiones, las consultas requieren diversas alternativas entre múltiples extensiones o cadenas de texto. En estos casos, la combinación de extensiones y operadores se realiza a modo de expresión regular REGEXP, tal como se muestra en el siguiente ejemplo *ext:(doc | pdf | xls | txt | ps | rtf | odt | sxw | psw | ppt | pps | xml) (intext:information retrieval | intext:"retrieval models") inurl:book.* Es posible determinar distintas alternativas por medio de valores separados por barras verticales (|) contenidos entre paréntesis. De esta forma se recuperan todos los libros que versan sobre recuperación de información en todos los formatos posibles y con diversos textos entre sus contenidos.

Recuperación de volcados de datos

Los volcados de datos constituyen una fuente de información muy importante para obtener catálogos, registros, tablas, bancos de datos completos sobre un tema o área de conocimiento determinada. Las consultas más eficientes en este sentido son las de tipo SQL (Structured Query Language) y CSV (Comma Separated Values). La información en gran medida se exporta en tales formatos y conviene conocer algunas cadenas de texto claves para su recuperación automática.

- ***Volcados de datos SQL.*** En el caso de los volcados de datos en formato SQL, existe un método de migración de datos denominado "*dumping data*". En tales casos, los programas gestores de bases de datos MySQL, generan archivos

automáticos con los contenidos de las tablas, estructuras y registros de la base de datos objetivo. Estos archivos pueden utilizarse para generar backups, copias de seguridad o servir de plataforma para la importación de los registros en terceros sistemas de información. En tales casos, ese proceso de automatización permite en un alto porcentaje, poder recuperar el texto predeterminado "MySQL *dump"*, *"Dumping data"* y *"phpMyAdmin MySQL-Dump"* con alto poder discriminatorio con respecto al resto de archivos. De esta forma, las consultas quedarían como *ext:sql "MySQL dump", ext:sql "Dumping data" y ext:sql "phpMyAdmin MySQL-Dump"*. No obstante pueden obtenerse resultados muy similares, utilizando la sentencia *ext:sql "INSERT INTO"*, ya que la instrucción *"Insert Into"* se refiere al proceso de inserción de registros que habitualmente es utilizado en los volcados de datos, delatando la presencia de registros y datos que se pretenden recuperar. Conociendo las distintas instrucciones del lenguaje de consulta SQL, es posible modificar las búsquedas dependiendo de la finalidad de uso de los resultados. Por ejemplo, el desarrollo de nuevos diseños de bases de datos, estructuras de campos y sus características en MySQL, dependen de archivos de instalación en formato PHP o SQL que contienen instrucciones como *"Update Set"*, *"Create Table"* o *"Alter Table"*. Esto hace que las consultas puedan ser del tipo *inurl:install ext:sql intext:"update set"*.

- ***Volcados de datos CSV.*** En el caso de los volcados de datos en formato CSV, resulta interesante comprobar cómo la búsqueda genérica *ext:csv* o *filetype:csv*, produce millones de resultados entre los que se puede obtener todo tipo de información, incluyendo catálogos bibliográficos y registros de bases de datos. Por ejemplo, la búsqueda de revistas científicas puede automatizarse con búsquedas similares a la siguiente *filetype:csv -github intext:"journal"*. Entre los resultados obtenidos, se encuentran listados completos de revistas científicas como por ejemplo la proporcionada por el Instituto de Investigación Scripps, http://www.scripps.edu/library/open/vivo_data/vivo_journal_holdings.csv, especializado en la investigación médica.

14. Tácticas de posicionamiento web – SEO search engine optimization

Qué es posicionamiento web o search engine optimization

El posicionamiento web, también conocido como SEO (search engine optimization) tiene su origen en el año 2001 en el contexto del marketing como aquellos métodos que permitían una mejor visibilidad de los productos anunciados a través de los buscadores en la web (SULLIVAN, D. 2004), aunque su aplicación en el contexto de la recuperación de la información y la Documentación, no llegará hasta el año 2005, momento aproximado en el que se comienza a explotar y estudiar de forma científica los métodos para obtener un mejor puesto entre las páginas de resultados del principal buscador en la web, Google. De hecho Lluis Codina lo define de la siguiente forma:

- *Posicionar es colocar alguna cosa en su lugar óptimo. En el ámbito de la world wide web, posicionar un sitio significa optimizarlo para que aparezca en las primeras posiciones de las páginas de resultados de los motores de búsqueda. Así mismo, podemos definir posicionamiento web como el conjunto de procedimientos y técnicas que tienen como finalidad dotar a un sitio o a una página web de la máxima visibilidad en Internet.* (CODINA, L.; MARCOS, M.C. 2005)

- *Academic Search Engine Optimization (ASEO) es la creación, publicación y modificación de documentos académicos de una manera que hace que sea más fácil para los motores de búsqueda académicos a ambos lo rastreo y el índice.* (BEEL, J.; GIPP, B.; WILDE, E. 2010)

- Def1: *Conjunto de procedimientos y técnicas que estudian las características que proporcionan a un sitio o una página web la máxima visibilidad en Internet.* Def2: *Conjunto de procedimiento que permiten colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda. Por extensión: Optimizar una página web de cara a los resultados proporcionados por los motores de búsqueda.* (CODINA, L. 2004)

Posicionamiento Web puede definirse como *el conjunto de métodos de programación, etiquetado, descripción, promoción y enlazamiento (legítimos o no), que permiten inferir en el cálculo de un algoritmo de ordenación, con el objetivo de ordenar un recurso o página web entre las primeras posiciones de las páginas de resultados de un buscador y ante un conjunto de consultas (conocidas o no) dadas por el usuario.* BLÁZQUEZ OCHANDO, M. 2013.

Qué aspectos favorecen la promoción y mejor posicionamiento de un sitio web

- Uso de etiquetas *<title></title>*. La presencia o ausencia de texto en las etiquetas *<title>* marca la diferencia entre recuperar o no una página web, ya que cualquier webcrawler los indexa, pondera y alfabetiza. De hecho, las páginas web en cuyo título se encuentra algún término de la consulta del usuario son posicionadas preferentemente sobre las que sólo los contienen en alguna parte del texto.

- Uso de meta-etiquetas *<meta name='' content=''/>*. Ayudan a describir de forma básica el contenido de la página web atendiendo a sus diversas opciones (*name='title', name='author', name='description', name='keywords'*). El texto del atributo (*content=''*) es indexado convenientemente y ponderado para su recuperación efectiva.

- Uso de metadatos Dublin Core, RDF. Los webcrawler son capaces de reconocer metadatos en formato Dublin Core e incluso detectar los archivos RDF vinculados, con la descripción de las páginas del sitio web. Esto permite no sólo procesar la página con más información que con la meta-etiqueta habitual, sino procesar un sitio web como parte de la red semántica de un buscador. Ello hace que la presencia y la visibilidad sea en un doble plano.

- Realizar descripciones únicas para cada página. En muchos casos, los sitios web contienen una única descripción que coincide en todas sus páginas web. Esto es un error, ya que dificulta la distinción de los contenidos particulares, de la generalidad del sitio web. Por ello cada página deberá tener sus propias descripciones con metadatos y meta-etiquetas.

- Estructura de direcciones URL más comprensiva, que incluya parte del título del contenido enlazado. Esto es el empleo de direcciones URL canónicas, optimizadas para su indexación con webcrawlers, que permitan la identifican del contenido y el título por sus palabras clave. Además las direcciones URL deberán estar escritas a ser posible en minúsculas, dado que son de más fácil lectura, evitando espacios o sustituyéndolos por guiones medios (-) o bajos (_)

- Crear estructuras de directorios sencillas. No deberá anidarse más de 2 niveles de carpetas. Cuando se amplía el número de carpetas y contenidos anidados, la navegación se dificulta sobremanera y con ello el seguimiento y acceso de los webcrawlers a los contenidos.

- Facilitar la navegabilidad de un sitio web a través de breadcrumbs (migas de pan) que permitan observar la ruta de navegación del usuario. Esto es que en cada momento se pueda retroceder sobre los enlaces cargados por el usuario, conocer la sección o categoría temática que se está consultando y la dirección permalink del contenido que se muestra.

- Crear mapas de sitio para el usuario y para los motores de búsqueda. El diseño de un mapa comprensivo de los contenidos y su organización en secciones a modo de directorio facilita la navegación del usuario a través de la red de páginas del sitio, mejorando el número de enlaces internos de la página y con ello cierta influencia en los cálculos de PageRank de terceras páginas enlazadas. Por otra parte de cara a los webcrawler el diseño de un mapa de todos los enlaces del sitio web, denominado sitemap.xml, que comprende el valor o peso de los contenidos de cada una de las páginas, según las estimaciones y especificaciones dadas por el administrador del sitio web.

- Usar texto en vez de objetos hechos en flash o javascript. El abuso de objetos de vídeo hechos en flash, shockwave o javascript no facilitan la labor de indexación de los webcrawler. Siempre que sea posible se debe minimizar el uso de tales medios y emplear visores adaptados para su reproducción. Por ejemplo mediante código HTML5 o JQuery, que constituyen métodos de codificación mejor adaptados a la problemática que entraña el posicionamiento web. En otros casos el empleo de códigos en javascript es inevitable, desde el punto de vista de la redirección de páginas web, el envío de variables de datos, en formularios, etc. Tales casos están reconocidos y en principio no suponen penalización en los principales buscadores.

- Crear página de error 404 con redireccionamiento. Resulta importante que todos los enlaces de un sitio web funcionen correctamente y no estén rotos. En caso contrario conviene crear una página de error 404, personalizada capaz de reconocer el contenido que pretendía cargar el usuario, reconocer sus palabras claves de consulta y redirigirle a un contenido aproximado. Esto evita que un webcrawler acceda a una página sin contenido, incrementando las posibilidades de indexación de contenidos y con ello un mejor posicionamiento.

- Contenidos actualizados y servicios de calidad. Textos de fácil lectura. Temática bien tratada y centrada en el contexto. Contenidos únicos, actualizados y originales. Diseñar textos comprensibles para el usuario y no para el robot de búsqueda. Enriquecer textos mediante anclas o enlaces, consiguiendo un valor hipertextual. Publicar nuevos contenidos de forma reiterada, por ejemplo diariamente, favorecen el posicionamiento de las páginas involucradas y el sitio web en general.

- Usar formatos CSS para mejorar la visibilidad y legibilidad de los textos y de los enlaces. La diferenciación de los enlaces de los distintos menús, de los enlaces propios del texto, marca la diferencia para el usuario y en la accesibilidad de los contenidos.

- Crear índices en los documentos extensos para facilitar su navegación. Uso de anclas para los enlaces internos. (*enlace a punto1 ... Punto1*)

- Uso del atributo "*alt*" en imágenes y del atributo "*title*" en los enlaces. La información complementaria que se puede aportar en cada contenido favorece su accesibilidad y su indexación por medio de webcrawlers. De esta forma, se pueden introducir variaciones y sinónimos en la descripción de los contenidos y enlaces para conseguir más puntos de acceso concordantes con la hipotética consulta del usuario. Todo ello favorece la visibilidad del contenido del sitio web y aumenta las posibilidades de posicionarse mejor que otra página web similar.

- Agrupar contenidos en carpetas especializadas. Esto es almacenar todas las imágenes, documentos, archivos de subida, descarga, esquemas, formularios, etc. en carpeta "*images*" y "*documents*" para favorecer que el acceso a todos los contenidos sea unificado.

- Uso de las etiquetas de cabecera *<h1>, <h2>, <h3>, <h4>, <h5>, <h6>* para intitular distintos párrafos y contenidos del texto según su importancia. Ello es reconocido por los webcrawler y establece una ponderación determinada para las palabras y textos comprendidos en las mismas.

- Uso de etiquetas especializadas. * (énfasis), (negrita), <address> (dirección), <abbr> (abreviatura), <article> (artículo), <aside> (contenido adjunto al documento principal), <base> (dirección URL base a partir de la que se generan todas las direcciones URL relativas), <blockquote> (sección en la que se cita una fuente de información o texto), <cite> (cita del título de un trabajo o documento), <code> (sección definida como código fuente), <details> (define detalles adicionales), <dfn> (definición de un término), <footer> (define el pie de página), <figure> (especifica la sección que ocupan las imágenes, figuras o ilustraciones del contenido), <figcaption>*

(determina el título de la figura), <nav> (define la sección de enlaces que permiten la navegación en la página web)

- Crear archivo robots.txt para especificar las restricciones de acceso para los webcrawler en el directorio de carpetas y contenidos del servidor correspondiente a un sitio web. Su correcta configuración debería facilitar el acceso a las carpetas de páginas, documentos e imágenes y prohibir la indexación de páginas de configuración, acceso de usuarios, login, backup, instalación, etc. En este sentido también se emplea el atributo (*rel='nofollow'*) en los enlaces de las páginas del sitio web que no deban ser rastreadas por el webcrawler. Por otra parte, se emplea la meta-etiqueta *<meta name="robots" content="noindex" />* para especificar que la página portadora de la instrucción no deberá ser indexada.

- Promocionar el sitio web creando canales de sindicación, servicios de distribución de correos y alertas, redes sociales. Ello permite transmitir y publicar enlaces al contenido que se pretende promocionar, favoreciendo su lectura y la creación de nuevos backlinks.

- Aumentar el número de enlaces entrantes para aumentar el PageRank del sitio web. Intercambiar enlaces, crear comunidades de blogs, webs, son métodos que ayudan a mejorar la visibilidad y aumentan la popularidad y tráfico (visitas) de los contenidos de un sitio web.

- Uso de herramientas estadísticas de la web. Por ejemplo Google Analytics, ayudan a orientar la estrategia de publicación de contenidos, según las páginas más visitadas, la indagación de las palabras clave de consulta utilizadas por el usuario para localizar el sitio web, e incluso el análisis de sitios web, permiten mejorar la visibilidad y la ordenación planificada en las SERPs.

- PageRank. Lograr un alto valor de PageRank favorece que la web se posicione mejor entre las páginas de resultados, ante una determinada consulta. Por lo tanto, planificar las hipotéticas consultas del usuario, cuidar el contenido y sobre

todo incidir en una correcta proporción de enlaces entrantes (inbounds o backlinks) y salientes es la clave para obtener una posición dominante en cualquier buscador.

Aspectos que penalizan el ranking de un sitio web

- Relleno y repetición de palabras clave. Consiste en repetir frases, textos o palabras clave, para mejorar la relevancia de una página web y aumentar su visibilidad, incidiendo en la frecuencia de aparición de las palabras.

- Texto oculto. Cuando el texto de relleno o la repetición de palabras clave, se oculta utilizando un tamaño de letra muy pequeño o se utiliza un color de fuente idéntico al color de fondo de la página. La ocultación se puede conseguir cuando se introduce dicho texto en los comentarios del código fuente, en atributos de etiquetas poco frecuentes, entre otros. Véase: http://support.google.com/webmasters/bin/answer.py?answer=66353

- Contenidos generados automáticamente. Por ejemplo textos traducidos automáticamente sin revisión, textos generados a partir de búsquedas o tomando como base contenidos de canales de sindicación, ocultación de textos, combinación de contenidos de páginas sin valor o peso. Véase: http://support.google.com/webmasters/bin/answer.py?answer=2721306

- Esquemas de enlaces. Esto es usar enlaces entrantes para mejorar la valoración de una página objetivo, manipulando dichos enlaces, o haciéndolos provenir del mismo sitio web, mediante técnicas de auto-enlace. Véase: http://support.google.com/webmasters/bin/answer.py?hl=es&answer=66356

- Encubrimiento. Consiste en la disposición de diferentes versiones de una página web para el webcrawler y para los usuarios. Ello se consigue al distinguir el tipo de visitante (agente o usuario), para los que se automatizan distintas variables de contenidos. De esta forma es posible construir una página con textos de relleno para los buscadores y conseguir así un mejor posicionamiento y construir una

página normal y correcta para el usuario. Véase: http://support.google.com/webmasters/bin/answer.py?answer=66355

- Redireccionamiento engañoso. Consiste en utilizar métodos de redireccionamiento javascript o meta, que en vez de cargar la página solicitada por el usuario, cargan la página objetivo, obteniendo un mayor número de visitas, tráfico y dinero en el caso de anuncios por visitas o clics. Véase: http://support.google.com/webmasters/bin/answer.py?answer=2721217

- Páginas puerta. Constituyen un compendio de páginas consideradas "*spam*" que tienen como objetivo promocionar una palabra, frase o texto con enlace a la página objetivo de posicionamiento. Éstas se activan en forma de popups, o páginas instantáneas con redireccionamiento automático. Véase: http://support.google.com/webmasters/bin/answer.py?answer=2721311

- Duplicación de sitios. Técnica consistente en duplicar los contenidos de una página web con diversas copias y redireccionamientos entre sí, para mejorar su posicionamiento. Esto es emplear el mismo código fuente y efectuar leves variaciones en los textos, manteniendo los enlaces entrantes de las páginas clonadas sobre la página original objetivo.

- Actualización automática permanente. Consiste en una actualización de los contenidos de una página web de forma rotativa, permanente y periódica. Ello permite introducir texto oculto, de relleno así como repetición de palabras claves, antes de que el webcrawler reconozca el engaño. Mientras, se mantiene la indexación del sitio web.

- Consultas automáticas reiteradas. Consiste en repetir "n" veces una serie de consultas predefinidas con una serie de palabras clave y operadores de forma tal, que un buscador las registre y ayude a posicionar las páginas web que coincidan con tales cadenas de consulta. Esta operación puede ser automatizada por diversos métodos en javascript y php, para realizar una repetición masiva del orden de varias decenas de consultas por minuto. En muchos casos la petición de

búsquedas abusivas en un buscador, puede llegar a confundirse con los ataques de denegación de servicio DoS (Denial of Service), por saturación en el número de conexiones simultáneas.

Archivo robots.txt

El archivo robots.txt es un archivo de texto que permite determinar la configuración de acceso de los webcrawler para la indexación de un sitio web y con ello sus contenidos y documentos. En teoría un robot de búsqueda es capaz de leer en primera instancia el archivo robots.txt para comprobar las rutas de acceso a los directorios y archivos de un determinado sitio web.

Su uso no significa que se respeten todas las restricciones que se establezcan, especialmente cuando se traten de webcrawlers maliciosos, por otra parte su acceso es público, lo que supone que las restricciones establecidas en su configuración son públicas. A pesar de todo, la mayoría de los motores de búsqueda sí respetan las instrucciones que se establecen, protegiendo de la indexación aquellos contenidos y páginas que podrían desvirtuar la recuperación de información en el sitio web, véase *tabla42* y siguiente referencia https://developers.google.com/webmasters/control-crawl-index/docs/robots_txt.

Restringir todo el sitio web	
User-agent: * # comentario Disallow: /	Se utiliza la instrucción *User-agent:* para identificar los motores de búsqueda afectados. En caso de utilizar el asterisco, se refiere que es de aplicación para todos. Si se desea especificar uno en concreto, éste se define con su nombre normalizado. Por ejemplo *User-agent: Google*
Permitir el acceso completo a todo el sitio web	
User-agent: * Disallow:	Para permitir el acceso al sitio web o restringirlo, se emplea la instrucción *Disallow:* o *Disallow: /*. El slash actúa como signo de restricción.
Restringir carpetas y sus contenidos en el sitio web	
User-agent: * Disallow: /interface/ Disallow: /java/ Disallow: /install.php Disallow: /login.php	Cuando se desea restringir el acceso de directorios y subdirectorios se emplea la instrucción *Disallow: /nombre-directorio/*. Obsérvese que se utiliza un slash al final del nombre del directorio para indicar que la prohibición implica también a los subdirectorios y archivos anidados. Por otra parte también pueden especificarse archivos concretos para su restricción específica, siguiendo la instrucción *Disallow: /ruta-completa/nombre-archivo.extension*

Permitir el acceso de carpetas y páginas del sitio web	
User-agent: * Allow: /documents/ Allow: /images/ Allow: /output/ Allow: /index.php Allow: /opac.php	Para permitir el acceso a determinados directorios o archivos, basta con no mencionarlos en la configuración Disallow, o bien precisarlos mediante la instrucción *Allow:*

Tabla 42. Aspectos básicos de la configuración de un archivo robots.txt

Archivo sitemap.xml

Es un protocolo y convención internacional para motores de búsqueda que permite al administrador de un sitio web determinar cuál es el mapa de enlaces vigente y la importancia de los mismos, asignando un valor o peso que incide en el posicionamiento de los contenidos en dichos buscadores. Este mapa de enlaces del sitio web, se elabora en formato XML, y se edita conforme a las reglas establecidas en su página web oficial sitemap0.90.

Link bait - linkbaiting (cebo de enlaces)

Cualquier página de contenidos que dada su calidad, singularidad y método de difusión consigue ser una de las más citadas y enlazadas, con el doble objetivo de mejorar el posicionamiento del sitio web que la aloja y aumentar el número de visitas. No se debe confundir con el concepto "*Granja de enlaces*". Se pueden diferenciar distintas técnicas para conseguir el link bait;

- Permitir la descarga completa o parcial de los contenidos, de tal manera que la página web sea enlazada por proporcionar un contenido libre y compartido, atrayendo en enlazamiento del recurso y por ende el aumento en el número de backlinks.

- Posicionar el contenido en listas de recomendaciones, propias de la web 2.0 y la web social. Al lograr visibilidad en este tipo de recursos, el número de usuarios potenciales aumenta y con ello la posibilidad de que se logre enlazar el contenido objetivo.

- Envío de notas de prensa, ponencias, congresos. En tales casos se puede referir la página web de contenidos, consiguiendo una lectura casi obligada de los usuarios.

- Publicación de reseñas, noticias y novedades en listas de distribución, de correos, foros, comunidades y canales de sindicación de contenidos. Hace posible que las personas que suscriben tales grupos y medios reciban el enlace a la página de contenidos, promocionando así su difusión.

Link farm - link farming (granja de enlaces)

Se denomina granja de enlaces al conjunto de páginas web construidas y gestionadas por un único administrador entre cuyos enlaces se vincula a una página web objetivo para mejorar su PageRank. Las granjas de enlaces pueden ser lícitas cuando el contenido de sus páginas es original y responde al mismo contexto de contenidos, siendo necesario su backlink. Las granjas de enlaces también pueden ser fraudulentas, cuando se generan miles de páginas web de forma automática con contenidos aleatorios que enlazan o redirigen a la página web objetivo. Otro tipo de granjas de enlaces son aquellas basadas en el spam o spammers que lejos de aumentar el PageRank de la página web objetivo a la que enlazan, logran disminuir su coeficiente utilizando el parámetro de Google que penaliza a las páginas con backlinks fraudulentos.

Pagerank

El PageRank de una página web es un número ponderado que representa la importancia relativa de dicha página en función del número de enlaces entrantes. La formulación del algoritmo de PageRank fue descrita por primera vez por (BRIN, S.; PAGE, L. 2000), planteando un método ordenación de páginas web, basado en sus contenidos y en los enlaces que dichas páginas recibían, sus backlinks. No obstante, la formulación original, (PAGE, L. 2001) no representa el único método de ordenación de los SERPs definitivo. Esto es debido a que con el tiempo, los webmasters, comenzaron a desarrollar técnicas de posicionamiento para influir directamente en el algoritmo, desvirtuando el propósito original de PageRank. Este hecho hizo que a la fórmula original se le añadieran decenas de factores que aún hoy siguen ampliándose, pudiéndose comprobar en la *tabla43*. La fórmula del PageRank, véase *figura10*, tiene en cuenta la relevancia de una página web,

partiendo del principio de que es más útil tanto en cuento es más visitada por los usuarios y recibe más enlaces entrantes o backlinks.

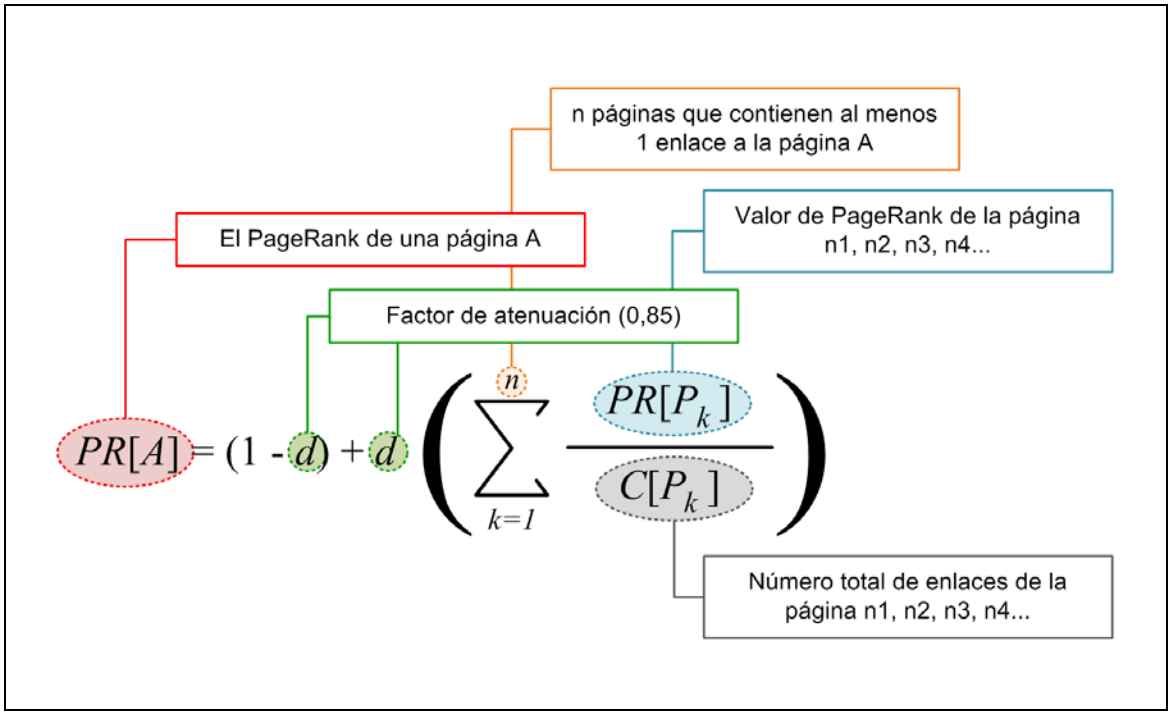


Figura 10. Fórmula estándar para el cálculo del PageRank

De esta forma, el PageRank de una página web objetivo denominada "A" es igual a 1 menos el factor de atenuación (que originalmente tuvo el valor 0,85 por defecto, pero que en realidad es desconocido y secreto, aunque se especula con la posibilidad de que tenga que ver con el tráfico o visitas de la página), más dicho factor de atenuación, por la suma de todas las páginas que contengan al menos 1 enlace a la página objetivo "A" entre su número total de enlaces. Se puede deducir que los factores que operan en la fórmula del PageRank son externos, no pudiendo ser manipulados por el webmaster de la página objetivo "A", aunque sí de forma fraudulenta, a través de técnicas como el link farm.

Factores que determinan el posicionamiento web en Google
Dominio
– Edad del dominio – Información del nombre de dominio – Tipo de dominio (nivel superior, subdominio, geográfico, genérico) – Número de veces que cambia la dirección IP – Palabras claves del dominio
Servidor
– Localización geográfica del servidor que mantiene la página web

– Número de caídas del servicio del servidor y de la página web
Programación
– URL canónica, optimizada para SEO, comprensible en su lectura, con palabras claves identificativas – HTML validado y bien construido. Uso de etiquetado específico – Uso de atributos title=’’ y alt=’’ – Empleo de web semántica – Uso de estilos CSS que faciliten la accesibilidad y usabilidad
Contenido
– Idioma empleado para la difusión del contenido – Uso de metadatos – Especialización y originalidad de los contenidos – Extensión del texto – Densidad del texto enlazado. Por ejemplo texto enlazado – Densidad de texto puro, sin enlaces, imágenes y código fuente. – Nivel de actualización, renovación o publicación de nuevos contenidos. Tiempo medio de renovación. – Número de meta-etiquetas empleadas, su descripción, extensión y representatividad – Número de metadatos empleados, su descripción, extensión y representatividad – Titulación de enlaces, capas, secciones, párrafos, figuras e imágenes – Contenido inapropiado o plagiado – Contenido generado automáticamente, basado en consultas y otras formas de promoción fraudulentas – Corrección en la puntuación y gramática del contenido – Nuevas frases o cadenas de texto no registradas
Enlaces
– Número de enlaces internos (enlazan otras páginas del mismo sitio web) – Número de enlaces internos que contengan el mismo texto de enlace – Número de enlaces que contienen el atributo “nofollow” – Número de enlaces Salientes por dominio y por página – Calidad de las páginas web vinculadas o enlazadas – Redireccionamiento en páginas de error 404 – Número de enlaces a imágenes – Número de enlaces entrantes o “backlinks” (enlaces que recibe la página web objetivo desde otras páginas web) – Importancia y calidad de los backlinks que enlazan la página web – Conlinks o coenlaces entre la página web enlazada y sus backlinks – Backlinks de marcadores sociales – Backlinks de obras de referencia en línea, directorios y buscadores especializados.
Sitio web
– Presencia de archivo robots.txt – Número de contenidos a los que tiene acceso el buscador desde robots.txt – Frecuencia de actualización media del sitio web – Número total de páginas del sitio web – Edad del sitio web. Desde que comenzó a ser indexado – Presencia de archivo sitemap.xml – Presencia de contenidos convenidos (sección “about us”, “location”, “contact”, etc.) – Tipología de sitio web (blog, portal, website, directorio, buscador, etc.)
Páginas web
– Meta-etiquetas para control de indexación y robots – Edad de la página – Número de cambios realizados en la página – Duplicación de contenidos con otras páginas dentro del sitio web – Audiencia o público objetivo al que va destinado el contenido – Tiempo de carga de la página web – Número de errores de la página – Extensión del texto – Número de enlaces internos – Número de enlaces externos
Palabras clave
– Cadena de texto del título de la página

– Cadena de texto de los atributos title=’’ y alt=’’ – Palabras de los textos enlazados (en enlaces internos y externos) – Palabras contenidas en etiquetas de tipo header como <h1>, <h2>… negrita, cursiva, énfasis, subrayado – Palabras del título de la página y de su dirección URL – Palabras empleadas en los comentarios del código HTML
Tráfico y visitas
– Número de visitas – País de origen de las visitas. Demografía de los visitantes – Porcentaje de rebote. – Sitios o páginas web similares – Tendencias de las visitas (al alza o a la baja) – Posición de la página o sitio web en las páginas de resultados. SERPs (Search Engine Results Pages)
Sanciones
– Relleno fraudulento de palabras clave – Compra de links entrantes – Spamming (posicionamiento basado en técnicas de spam) – Cloacking (modalidades de ocultación de texto) – Duplicación y plagio de contenidos – Historial de sanciones – Redirección y enlaces a páginas inexistentes – Redirección a páginas generadas automáticamente con la consulta del usuario – Links ficticios – Suplantación de páginas web mediante Phising – Enlaces a páginas con contenido malware (virus, troyanos, código malicioso) – Enlaces a páginas comprendidas en la lista negra de sitios y páginas de Google – Granjas de enlaces para mejorar el PageRank o posicionamiento
Otros factores
– Cumplimiento de las directrices de Google para Webmasters – Presencia de contenidos en herramientas de Google. Por ejemplo Google News, Google Books, Google Scholar – Presencia entre las SERPs de blogs de Google – Uso de servicios de anuncios como Google AdWords – Uso de servicios de monitorización y estadísticas Google Analytics – Citación y referenciación de las fuentes de información

Tabla 43. Factores para el posicionamiento web en Google

15. Ejercicios prácticos

- Práctica1. Metadatos y descripción Dublin Core
- Práctica2. Descripción bibliográfica Dublin Core
- Práctica3. Dublin Core RDF y generadores de metadatos
- Práctica4. Descripción de autoridades con MADS
- Práctica5. Descripción bibliográfica con MODS
- Práctica6. Análisis y recuperación parser de metadatos
- Práctica7. Análisis webcrawler
- Práctica8. Consultas dinámicas URL en Google
- Práctica9. Operadores avanzados y directorios de servidores
- Práctica10. Recuperación de volcados de datos
- Práctica11. Configurar archivos robots.txt y sitemap.xml
- Práctica12. Cálculo de PageRank

Práctica1. Metadatos y descripción Dublin Core

Según se ha explicado en el capítulo de introducción a los metadatos, éstos son empleados en todo tipo de normas de descripción, incluso en las propias reglas de catalogación, puesto que son todos aquellos campos que definen un documento. En la presente práctica, el alumno deberá indagar sobre el propósito y finalidad de una serie de metadatos, detectando su ámbito de aplicación, URL de especificaciones y breve descripción. A continuación, se plantea el diseño de metadatos para la descripción de "mobiliarios", "alimentos", "herramientas", "edificios" y "documentos", trabajo que ejercitará la capacidad para resolver problemas dentro y fuera del entorno documental. Finalmente, se deberá generar una página HTML con la descripción de los 15 metadatos oficiales de Dublin Core sobre una página web de una universidad que seleccione previamente.

1. Define los siguientes metadatos, respondiendo a cada uno de los títulos de cada columna.

Tipo Metadato	Denominación completa	Aplicación	Muestra de 5 metadatos	Sitio web de las especificaciones
Dublin Core				
CDWA				
CCO				
VRA				
MARC				
AACR				
MADS				
MODS				
DACS				
EAD				
OAI-PMH				
Object ID				
CIMI				
FDA				

2. Diseña una serie de 10 metadatos para la descripción de “mobiliario”, “alimentos”, “herramientas”, “edificios”, “documentos”.

Mobiliario		
Metadato	**Tipo de información recopilada**	**Descripción**
Alimentos		
Metadato	**Tipo de información recopilada**	**Descripción**

Herramientas		
Metadato	**Tipo de información recopilada**	**Descripción**
Edificios		
Metadato	**Tipo de información recopilada**	**Descripción**
Documentos		
Metadato	**Tipo de información recopilada**	**Descripción**

3. Elige uno de los siguientes recursos de la lista y descríbelo creando una página web básica en HTML con las 15 etiquetas de descripción básicas de Dublin Core. Pegar en el siguiente recuadro el código fuente resultante.

Recursos disponibles	
http://www.harvard.edu/ http://www.mit.edu/ http://www.stanford.edu/ http://www.berkeley.edu/ http://www.cornell.edu/ http://www.umich.edu/ http://www.umn.edu/ http://www.washington.edu/ http://www.wisc.edu/ http://www.utexas.edu/ http://www.upenn.edu/ http://www.psu.edu/ http://www.columbia.edu/ http://www.cmu.edu/ http://www.uiuc.edu/ http://www.ucla.edu/ http://www.tamu.edu/ http://www.umd.edu/ http://www.purdue.edu/ http://www.unc.edu/ http://www.msu.edu/ http://www.cam.ac.uk/ http://www.ufl.edu/ http://www.rutgers.edu/ http://www.indiana.edu/ http://www.nyu.edu/ http://www.ncsu.edu/ http://www.virginia.edu/ http://www.yale.edu/ http://www.arizona.edu/ http://www.utoronto.ca/ http://www.ucsd.edu/ http://www.pitt.edu/ http://www.princeton.edu/ http://www.duke.edu/ http://www.caltech.edu/ http://www.vt.edu/	http://www.ubc.ca/ http://www.gatech.edu/ http://www.ethz.ch/ http://www.ox.ac.uk/ http://www.osu.edu/ http://www.uchicago.edu/ http://www.usc.edu/ http://www.uga.edu/ http://www.ucdavis.edu/ http://www.uci.edu/ http://www.iastate.edu/ http://www.colorado.edu/ http://www.uconn.edu/ http://www.u-tokyo.ac.jp/ http://www.ed.ac.uk/ http://www.uio.no/ http://oregonstate.edu/ http://www.jhu.edu/ http://www.asu.edu/ http://www.ucsb.edu/ http://www.utah.edu/ http://www.helsinki.fi/ http://www.umontreal.ca/ http://www.ntnu.no/ http://www.uiowa.edu/ http://www.ucl.ac.uk/ http://www.wustl.edu/ http://www.unl.edu/ http://www.sfu.ca/ http://www.gmu.edu/ http://www.ualberta.ca/ http://www.univie.ac.at/ http://www.unam.mx/ http://www.bu.edu/ http://www.epfl.ch/ http://www.anu.edu.au/ http://umass.edu/

Recurso elegido

Código fuente resultante

Práctica2. Descripción bibliográfica Dublin Core

La siguiente práctica tiene como objetivo, emplear todos los conocimientos de codificación Dublin Core estudiados hasta el momento, para catalogar y describir de forma exhaustiva los principales tipos documentales más habituales. En este caso se presenta un artículo, una revista científica y finalmente una monografía. La finalidad y objetivo del ejercicio es utilizar el mayor número de metadatos posibles empleando los refinamientos aprendidos anteriormente. El resultado de la codificación deberá reseñarse en los cuadros dispuestos debajo de cada pregunta. Se valorará positivamente la introducción de elementos de descripción puramente bibliográficos, como por ejemplo, citas bibliográficas, referencias bibliográficas, fechas, relación con otros recursos, entre otros. Seguidamente se propone la resolución de varias preguntas en torno al uso y aplicación de algunos de los metadatos más relevantes.

1. Describe con elementos y términos Dublin Core el siguiente artículo de una revista científica: *CLAUSÓ GARCÍA, A.; CARPALLO BAUTISTA, A. 2011. Estudio de la producción científica de las publicaciones en México referentes a la catalogación de documentos: 1990-2009*

2. Describe con elementos y términos Dublin Core la ficha principal de la siguiente revista científica: *Revista General de Información y Documentación*

3. Describe con elementos y términos Dublin Core la siguiente obra de referencia: *WILFRID LANCASTER, F. 1979. Information retrieval systems: characteristics, testing, and evaluation.*

4. Demuestra de qué dos formas puede representarse el título y el subtítulo de un documento mediante Dublin Core.

5. Qué diferencias existen entre los metadatos “isReferencedBy” y “references”. Plantea un ejemplo en el que puedan observarse tales diferencias.

6. Qué diferencias existen entre los metadatos “hasPart” y “isPartOf”. Plantea un ejemplo en el que puedan observarse tales diferencias.

7. Qué metadato Dublin Core se utiliza para la identificación de la procedencia y el sujeto productor. ¿Hasta qué punto los metadatos permitirían efectuar una descripción archivística completa? Razona tus respuestas.

Práctica3. Dublin Core RDF y generadores de metadatos

Los generadores de metadatos se han convertido en una herramienta habitual del documentalista para generar de forma rápida la codificación Dublin Core. En esta práctica se trata de comprobar su funcionamiento y resultados, comparando sus códigos, problemas, ventajas y aciertos. Por otra parte, se propone una práctica de codificación manual en RDF.

1. Utilizando el generador de metadatos http://dublincoregenerator.com/generator_nq.html describir la siguiente obra:

    a. *WILFRID LANCASTER, F. 1979. Information retrieval systems: characteristics, testing, and evaluation. Disponible en: http://books.google.es/books/about/Information_retrieval_systems.html?id=JOhTAAAAMAAJ&redir_esc=y*

    b. *Pegar el código fuente resultante en su modalidad XML, con Namespaces*

2. Utilizando el generador de metadatos http://www.ukoln.ac.uk/metadata/dcdot/ describir la obra anteriormente referida.

    a. Pegar el código fuente resultante en su modalidad XML

    b. Pegar el código fuente resultante en su modalidad RDF

3. ¿Qué generador de metadatos permite una descripción más exhaustiva? ¿Cuál de ellos genera una mejor codificación (más correcta de acuerdo a la norma)? ¿Qué diferencias existen entre la versión XML de ambos generadores? ¿Hasta qué punto requiere revisión humana la generación automática de metadatos?

4. Codifica manualmente con elementos y términos Dublin Core en formato RDF la siguiente referencia: *BONNET, J. 2010. Bibliotecas llenas de fantasmas. Disponible en: http://cisne.sim.ucm.es/record=b2652015~S6*spi*

Práctica4. Descripción de autoridades MADS

A partir de las normas de construcción e implementación de los metadatos MADS, llevar a cabo la descripción de las siguientes autoridades utilizando como base el formato RDF. En segundo lugar, la práctica propone razonar con diversas preguntas la aplicación del modelo de metadatos MADS en proyectos como VIAF de la OCLC.

1. Elaborar una descripción con metadatos MADS sobre el profesor Frederick Wilfrid Lancaster, para continuar con el ejemplo de prácticas anteriores (*WILFRID LANCASTER, F. 1979. Information retrieval systems: characteristics, testing, and evaluation*). Utilizar la entrada de la autoridad personal disponible en la wikipedia.

2. Introducir en la metadescripción MADS de Lancaster una extensión de su descripción con Dublin Core para describir o vincular su obra:

3. Buscar la ficha de autoridad de Lancaster en Viaf y responder a las siguientes preguntas.

    a. Cuál es su URI permalink en Viaf

    b. Si se descarga la versión de la autoridad en formato RDF, ¿qué metadatos se emplean en dicha descripción?

    c. ¿Puede MADS describir con el mismo nivel de detalle que Viaf en formato RDF las distintas autoridades? ¿Todos los campos de de descripción en VIAF fueron considerados en MADS?

d. ¿Por qué la descripción de autoridades con MADS es más sencilla y práctica? Por qué la OCLC en su proyecto VIAF no emplea MADS, ¿qué razones se pueden aducir? Téngase en cuenta que MADS es un formato de metadatos especializado en la descripción de autoridades y VIAF utiliza múltiples formatos para conseguir el mismo propósito.

Práctica5. Descripción bibliográfica con MODS

A partir de las normas de construcción e implementación de los metadatos MODS, llevar a cabo la descripción principal de una revista científica, un artículo y una monografía.

1. Elaborar una descripción bibliográfica con metadatos MODS de la siguiente monografía: (*WILFRID LANCASTER, F. 1979. Information retrieval systems: characteristics, testing, and evaluation*). Obtener la información bibliográfica del catálogo CISNE.

2. Elaborar una descripción bibliográfica con metadatos MODS de la ficha principal de la revista: (*Revista General de Información y Documentación*)

3. Elaborar una descripción bibliográfica con metadatos MODS del siguiente artículo científico: (*CLAUSÓ GARCÍA, A.; CARPALLO BAUTISTA, A. 2011. Estudio de la producción científica de las publicaciones en México referentes a la catalogación de documentos: 1990-2009)*

4. ¿Para qué se podrían emplear los metadatos Dublin Core en el caso de la descripción del artículo científico? ¿Qué lagunas pueden solucionar? ¿Qué modificaciones supondría en el código de la pregunta 3?

Práctica6. Análisis y recuperación parser de metadatos

Los metadatos Dublin Core, MADS, MODS, METS, MARC-XML, EAD, entre otros, pueden ser analizados y aprovechados para su recuperación, edición y gestión, por medio de programas de análisis "parser", ya citados en el apartado anterior. Sus empleos en los sistemas de gestión de contenidos, buscadores, catálogos bibliográficos, bases de datos especializadas y directorios, están muy extendidos.

De hecho, sin este tipo de herramientas no se podrían leer las noticias publicadas por los medios de comunicación en tiempo real, compartir información mediante archivos basados en formato XML, no existiría la web semántica y todos sus contenidos no tendrían sentido, pues no podrían ser en tal caso recuperados. En la presente práctica, se propone comprender el funcionamiento de los programas parser especializados en metadatos, a través de una herramienta de lectura básica, capaz de interpretar las consultas mediante XPath realizadas por el usuario sobre un determinado código XML. Según los códigos varíen, también deberán variar las estrategias de consulta y filtrado del usuario.

- ***Herramienta de análisis parser de metadatos***
  http://www.mblazquez.es/blog_ccdoc-busqueda-internet/programas/parser-metadata.php

1. Recuperar el título, subtítulo, autor y rol del siguiente registro en MODS:
   http://www.loc.gov/standards/mods/v3/mods99042030_linkedDataAdded.xml

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

2. Recuperar la clasificación LLC del siguiente registro en MODS:
   http://lccn.loc.gov/97129132/mods

Campo1:
[Pegar impresión de pantalla]

3. Recuperar el título de la revista, género, número y fecha a la que pertenece el siguiente artículo en el registro MODS:
   http://www.loc.gov/standards/mods/v3/modssupplement.xml

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

4. Obtener los cuatro identificadores del registro en MODS: http://lccn.loc.gov/94759273/mods

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

5. Recuperar la entrada principal de la autoridad y sus tres variantes del siguiente registro en MADS: http://id.loc.gov/authorities/names/n81018141.madsxml.xml

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

6. Recuperar la fecha de la segunda variante de la autoridad y sus notas en el siguiente archivo MADS: http://id.loc.gov/authorities/names/n81018141.madsxml.xml

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

7. Recuperar el título, autor, fecha de publicación e identificador del código Dublin Core RDF de la práctica3, pregunta 2b.

Campo1:	***Campo3:***
Campo2:	***Campo4:***
[Pegar impresión de pantalla]	

Práctica7. Análisis webcrawler

Se propone el empleo de una herramienta webcrawler básica, desarrollada originalmente para demostrar las posibilidades de recuperación de información y análisis webmétrico con métodos plenamente automatizados. El programa Mbot que se utilizará, corresponde a su fase beta1.0, que aún no estando perfeccionada, permite llevar a cabo pequeños análisis de páginas web, con los que obtener conclusiones y razonamientos valiosos para el aprendizaje.

En la práctica se empleará para obtener los distintos componentes de páginas web, analizando varios niveles de enlazamiento desde un sitio web de una universidad escogida en el ranking web de universidades http://www.webometrics.info/ Una vez obtenidos los datos, el alumno deberá razonar y estudiar los resultados, para tratar de responder a las preguntas planteadas sobre el número de enlaces, metadescripción obtenida y variabilidad en el número de contenidos y componentes obtenidos.

- ***Webcrawler Mbot beta 1.0***
  http://www.mblazquez.es/documents/articulo-pruebas1-mbot.html

1. Selecciona la URL de una universidad entre todas las disponibles en el ranking web de universidades http://www.webometrics.info/en/world y analiza sus componentes mediante el webcrawler Mbot beta1.0. Si la universidad elegida no produce resultados, se insta seleccionar otra distinta que permita analizar los componentes mínimos de enlaces, código fuente, palabras y título. Una vez elegida la universidad reseña en el siguiente cuadro la URL de la universidad seleccionada.

2. Obtener los componentes del sitio web de portada de la universidad en el nivel 1. Seleccionar 10 enlaces de la portada de la universidad y analizar sus componentes en el nivel 2. Seleccionar 10 enlaces obtenidos en el análisis del nivel 2 y obtener los componentes correspondientes al nivel 3.

Recurso web	Nivel	Metadatos	Meta-etiquetas	Enlaces	Documentos	Imágenes	Multimedia	Canales de sindicación	Título	Titulares	Párrafos	Código fuente (caracteres)	Palabras	Tiempo
	1													
	2													
	2													
	2													
	2													
	2													
	2													
	2													
	2													
	2													
	2													
	3													
	3													
	3													
	3													
	3													
	3													
	3													
	3													
	3													

3. Qué páginas están mejor descritas mediante metadatos y meta-etiquetas. ¿A qué es debido? ¿Por qué esas y no otras?

4. Qué nivel de análisis tiene más enlaces. Razona tu respuesta.

5. Qué nivel de análisis tiene más documentos. Razona tu respuesta.

6. Qué nivel tiene más canales de sindicación. Razona tu respuesta.

7. Cuantas palabras se han extraído durante el análisis en total.

8. Cuanto tiempo tardó el programa en analizar las 21 páginas.

9. Cuantos enlaces fueron extraídos en total en cada nivel.

10. Existe algún enlace mejor conectado que el resto. ¿Cómo averiguarlo? Si existe, ¿cuál sería?

Práctica8. Consultas dinámicas URL en Google

Las consultas dinámicas mediante URL y método GET permiten efectuar búsquedas precisas y rápidas sin necesidad de utilizar el interfaz gráfico de consulta avanzada de Google. Comprender su funcionamiento, permite conocer con qué variables y entrada de datos trabaja el buscador y qué ventajas pueden ser obtenidas con su uso. Por ejemplo, descargar todos los archivos y documentos sobre una temática, publicados por una institución, en un determinado sitio web, obtener información de volcados de datos de bases de datos, entre otros.

1. Buscar “Blogs de biblioteconomía” de México

[URL de la consulta]
[Imprimir pantalla]

2. Recuperar los primeros 10 resultados de la consulta “bibliotecas” en el dominio “http://www.csic.es/”

[URL de la consulta]
[Imprimir pantalla]

3. Recuperar archivos en formato SQL, en inglés con la consulta “Information Retrieval”

[URL de la consulta]
[Imprimir pantalla]

4. Recuperar por la frase exacta “application/rss+xml” y algunas de las siguientes palabras “bibliotecas archivos museos documentalista bibliotecario archivero libro documentación ciencias”. Además comprobar que el formato sea “xml”

[URL de la consulta]
[Imprimir pantalla]

5. Recuperar páginas brasileñas con la consulta “mídia brasileira” en portugués en cuyo título figure la palabra “diário”

[URL de la consulta]
[Imprimir pantalla]

6. Recuperar los 100 primeros documentos en formato “mdb” en inglés cuya frase exacta sea “isbn” y al menos contenga alguna de las siguientes palabras “books library Information Science university”

[URL de la consulta]
[Imprimir pantalla]

7. Recuperar todos los documentos en formato “xls” con algunas de las siguientes palabras “thesaurus, tesauro, clasificación, classification, category”

[URL de la consulta]
[Imprimir pantalla]

8. Recuperar todos los enlaces que contengan algunas de las palabras indicadas “ontology” en sus enlaces y sólo mostrar aquellos resultados en formato “rdf”

[URL de la consulta]
[Imprimir pantalla]

9. Buscar las últimas 100 noticias del periódico “El mundo” durante el último mes.

[URL de la consulta]
[Imprimir pantalla]

10. Buscar todas las páginas similares a la especificada http://www.webometrics.info/

[URL de la consulta]
[Imprimir pantalla]

11. Buscar todos los resultados de “Max Planck” duplicados del dominio http://www.mpg.de/

[URL de la consulta]
[Imprimir pantalla]

12. Buscar todos los resultados de “Max Planck” únicos del dominio http://www.mpg.de/

[URL de la consulta]
[Imprimir pantalla]

13. Recuperar los 100 primeros documentos en formato “pdf” cuya temática sea “Library and Information Science”, siempre y cuando consten de “isbn”

[URL de la consulta]
[Imprimir pantalla]

14. Recuperar los 100 primeros documentos únicos publicados en Alemania y escritos en inglés sobre la consulta exacta “information retrieval” siempre y cuando contengan además alguno de los siguientes términos “statistics biblography library research users”

[URL de la consulta]
[Imprimir pantalla]

15. Mostrar los últimos 100 documentos en formato “.sql” publicados en España

[URL de la consulta]
[Imprimir pantalla]

16. Buscar los 50 primeros resultados a partir de la página 10 con la consulta “Information processing” publicados en EEUU, especialmente en el dominio http://www.mit.edu/

[URL de la consulta]
[Imprimir pantalla]

17. Recuperar todos los canales de sindicación del dominio http://www.medworm.com/

[URL de la consulta]
[Imprimir pantalla]

18. Recuperar los últimos 100 documentos en formato “pdf” publicados en el dominio http://www.doaj.org/

[URL de la consulta]
[Imprimir pantalla]

19. Recuperar todos los documentos en formato “pdf” del dominio http://eprints.rclis.org/

[URL de la consulta]
[Imprimir pantalla]

20. Recuperar todos los enlaces de universidades disponibles en http://www.webometrics.info/

[URL de la consulta]
[Imprimir pantalla]

Práctica9. Operadores avanzados y directorios de servidores

La detección de directorios de servidores abiertos con documentación científica, constituyen una fuente de recursos de información poco conocida y explotada en el mundo de la Documentación.

El objetivo de la práctica será poner a prueba los conocimientos adquiridos sobre consulta y detección de directorios de servidores en algunas de las instituciones más prestigiosas a nivel académico.

1. Buscar directorios abiertos especializados en recuperación de información

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

2. Buscar directorios abiertos científicos en el dominio del CSIC

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

3. Buscar directorios abiertos científicos en el dominio del MIT

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

4. Buscar directorios abiertos científicos en la Universidad de Stanford

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

5. Buscar directorios abiertos de artículos especializados en biblioteconomía y documentación

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

6. Buscar directorios abiertos de la Administración Pública Española

[Texto de la consulta]	
[Impresión de pantalla]	
[Copiar 5 resultados más pertinentes]	
¿Contienen fichero log?	
Elige un directorio y aplica la técnica de recorrido, reseñando la consulta correspondiente	

Práctica10. Recuperación de volcados de datos

Aplicando todas las técnicas de consulta estudiadas, se propone la recuperación de los volcados de datos de diversas temáticas en formatos CSV y SQL. Se reseñará la cadena de consulta utilizada en cada caso, la dirección URL del archivo de volcado de datos y una muestra textual del archivo en la que aparezca reflejada la temática adecuada al tema propuesto.

1. Recuperar los siguientes volcados de datos en formato CSV y SQL:

1	**Medicina**	**CSV**	***Consulta: [Cadena de consulta]*** ***URL: [Reseñar la URL del archivo CSV]*** ***Muestra de datos: [Copiar y pegar una muestra de los datos del volcado que demuestren que la temática del contenido es la adecuada al tema propuesto]***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
2	**Catálogo bibliográfico**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
3	**Artículos de un blog**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
4	**Datos estadísticos**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
5	**Ontología**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***

6	**Recuperación de información**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
7	**Revistas científicas**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
8	**Química**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
9	**Matemáticas y física**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
10	**Economía**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
11	**Biología**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***
12	**Informática**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***

13	**Educación**	**CSV**	***Consulta:*** ***URL:*** ***Muestra de datos:***
		SQL	***Consulta:*** ***URL:*** ***Muestra de datos:***

Práctica11. Configuración de archivo robots.txt y sitemap.xml

La configuración de los archivos robots.txt y sitemap.xml constituyen uno de los principios por los que se determinan las rutas de indexación y acceso de los webcrawler de los motores de búsqueda. A partir de las instrucciones y documentación oficial analizada, resolver la práctica editando tales archivos.

1. Edita un archivo sitemap.xml, de acuerdo a las especificaciones oficiales http://www.sitemaps.org/es/index.html y siguiendo los siguientes pasos:

    a. Seleccionar un sitio web de universidad o centro de investigación del ranking web de universidades y centros de investigación disponible en: http://www.webometrics.info/

Centro	
URL	

    b. Probar el siguiente generador automático de sitemap.xml (http://www.xml-sitemaps.com/) con la dirección URL seleccionada, aplicando las siguientes especificaciones. 1) Cálculo de prioridad automática, 2) Frecuencia de actualización semanal. Pegar el código fuente resultante.

    c. Probar el siguiente generador automático de sitemap.xml (http://www.xmlsitemapgenerator.org/) con la dirección URL seleccionada, aplicando las siguientes especificaciones. 1) Utilizar opciones avanzadas del generador, 2) Frecuencia de actualización semanal, 3) mantener marcadas las opciones por defecto.

[Explica para qué sirven las distintas opciones avanzadas del generador]

[Imprimir pantalla de configuración que se utiliza para generar el archivo sitemap.xml]

[Pegar código fuente resultante]

d. Cuál de los dos generadores automáticos de sitemap.xml genera un código XML más fidedigno respecto a las normas oficiales de construcción. Justifica tu respuesta.

e. Qué papel juega la etiqueta de prioridad de los enlaces, cómo se deben puntuar las distintas páginas de un sitio web.

2. Edita un archivo de permisos de acceso para motores de búsqueda "robots.txt" a partir del siguiente directorio raíz y de acuerdo a las siguientes especificaciones

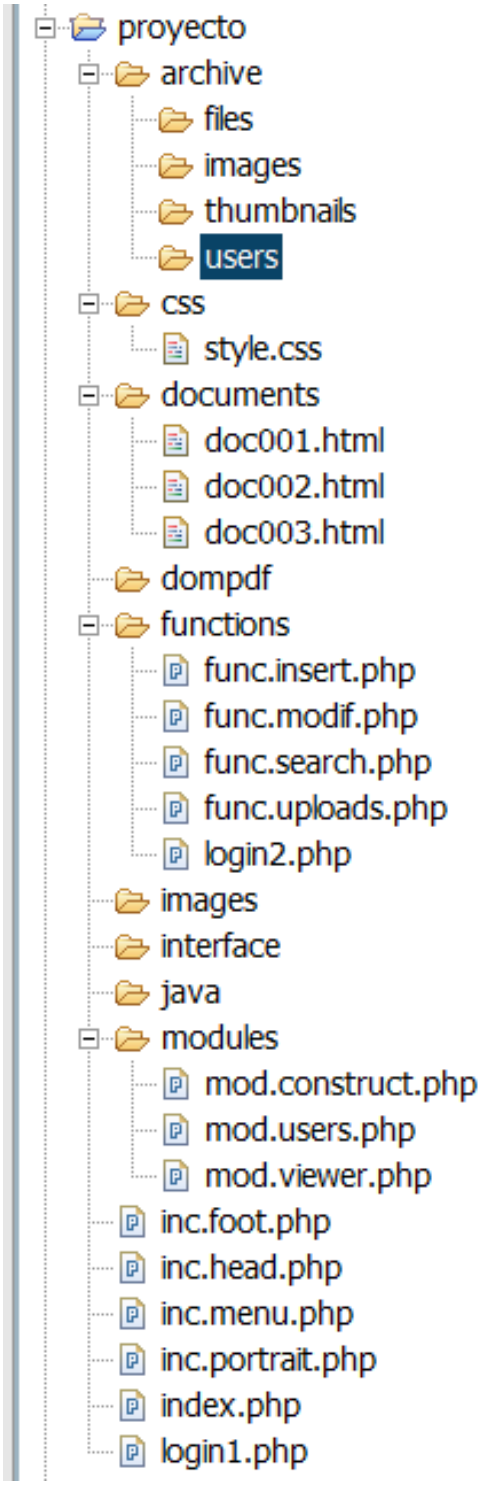
   a. Permitir la indexación de los motores de búsqueda Google y Yahoo, pero denegar el acceso al resto.

   b. Los directorios, subdirectorios y archivos de instalación, interfaz, accesos de usuario, applets de java y cualquier otro elemento sensible deben estar restringidos para todos los buscadores.

   c. Permitir acceso a los directorios, subdirectorios y archivos de contenidos, que deberían ser indexados.

Configuración del archivo robots.txt

Directorio raíz de muestra para la elaboración del ejercicio de configuración



Práctica12. Cálculo de PageRank

El cálculo de PageRank es uno de los factores que determinan el posicionamiento de una página web en Google. Atendiendo a este aspecto de gran importancia para el ordenamiento de los SERPs, se propone la resolución de una práctica de cálculo asistido por un simulador de PageRank, que permite realizar las 10 primeras iteraciones del cálculo, o bien utilizar la fórmula básica por defecto para efectuar un cálculo puntual (a modo de prueba) de un sitio web determinado

- ***Simulador de PageRank***
  http://www.mblazquez.es/blog_ccdoc-busqueda-internet/documentos/simulador-pagerank.xlsx

1. Descargar el simulador de PageRank y analizar los datos de la prueba de PageRank por iteración que por defecto aparecen en la hoja de cálculo. Responde a las siguientes preguntas:

¿Cuántos enlaces entrantes tienen la página A y la página J?	
Qué página web tiene mayor PageRank y porqué	
Cómo afecta el muestreo por iteración. Analiza estadísticamente qué iteración provoca mayores cambios en el valor de PageRank y en qué páginas. Introduce una gráfica que ayude a explicarlo.	
Si el factor de atenuación se reduce a 0,5 qué efecto se produce sobre el cálculo de PageRank por iteraciones. Crea una gráfica que permita establecer una comparativa con respecto al caso anterior.	

2. Probar el simulador de PageRank por iteración, modificando los valores iniciales por defecto de la iteración a “2”, factor de atenuación “0,85”, no modificar distribución de enlaces de la red de páginas.

Qué cambios se producen en las iteraciones del PageRank, ¿se produce un incremento o amplificación de los valores con respecto a los casos anteriores? Crea una gráfica que permita establecer una comparativa con respecto al caso anterior.	

3. Probar el simulador de PageRank por iteración, modificando los valores iniciales por defecto de la iteración a “1”, factor de atenuación “0,85” y aplicar el siguiente número de enlaces en el cuadro de distribución.

Columna A – Fila J	**1**

¿Qué efectos se producen? Explica qué sucede en el cálculo de PageRank y porqué.	
Crea una gráfica de las iteraciones de PageRank para el documento A y J y compáralas con los casos anteriores	

4. Modifica la distribución de enlaces de la red de páginas web para lograr que la página A y B sean las que mayor valor de PageRank adquieran sobre el resto.

¿Qué cambios y modificaciones hay que realizar? Aportar valores introducidos y explicar la alteración de los resultados	
[Impresión de pantalla con los cambios y resultados]	

5. Calcula el PageRank de la Universidad Complutense http://www.ucm.es/ , utilizando la herramienta del simulador “PageRank web”. Sigue las siguientes instrucciones.
    a. Buscar en Google las primeras 10 páginas que enlazan a la web de la UCM
    b. Comprobar cuál es el número de enlaces de cada una de esas 10 páginas y su PageRank (Utiliza http://www.calcularpagerank.com/index.php)
    c. Introduce los datos en el simulador y determina el valor de PageRank

[Imprimir pantalla del simulador con los datos introducidos]
[Pegar valor de PageRank]

6. Seguir los mismos pasos que en la pregunta4 para calcular el PageRank de http://www.doaj.org/

[Imprimir pantalla del simulador con los datos introducidos]
[Pegar valor de PageRank]

7. Seguir los mismos pasos que en la pregunta4 para calcular el PageRank de http://www.loc.gov/index.html

[Imprimir pantalla del simulador con los datos introducidos]
[Pegar valor de PageRank]

16. Índice de tablas

Tabla 1. Contextos de aplicación de los metadatos 6
Tabla 2. Principales meta-etiquetas 7
Tabla 3. Estructura XHTML para la identificación del marco de trabajo Dublin Core ... 8
Tabla 4. Estructura HTML para la identificación del marco de trabajo Dublin Core 9
Tabla 5. Conjunto de elementos básicos de Dublin Core 10
Tabla 6. Construcción básica de metadatos 12
Tabla 7. Refinamientos con los términos Dublin Core 12
Tabla 8. Esquemas de codificación de los valores o contenidos 13
Tabla 9. Enlazar otros recursos 13
Tabla 10. Enlazar recursos especificando el tipo de enlaces 13
Tabla 11. Especificar idioma o lengua empleando la norma ISO639-1 13
Tabla 12. Los metadatos pueden ser repetidos 14
Tabla 13. Namespaces en Dublin Core 14
Tabla 14. Utilizar otros metadatos con otros esquemas 14
Tabla 15. Referencia de términos y elementos Dublin Core 20
Tabla 16. Aplicación de Dublin Core para la descripción de un artículo científico 21
Tabla 17. Ejemplo de descripción RDF / MARC-XML de un documento 25
Tabla 18. Ejemplos de espacios de nombre con sus prefijos de aplicación 27
Tabla 19. Ejemplo de Dublin Core en RDF 31
Tabla 20. Ejemplo de técnica de embebido de Dublin Core RDF en XML 32
Tabla 21. Ejemplo de vinculación de archivo Dublin Core RDF en HTML 32
Tabla 22. Espacios de nombres en la codificación de MADS 34
Tabla 23. Referencia básica del modelo de metadatos MADS 36
Tabla 24. Ejemplo de MADS y XML básico 38
Tabla 25. Ejemplo de MADS y XML con prefijo de espacio de nombres 40
Tabla 26. Ejemplo de MADS y expresada en RDF con prefijo de espacio de nombres 41
Tabla 27. Espacios de nombre con sus prefijos aplicados en la codificación de MADS 43
Tabla 28. Ejemplo de namespace aplicado a registros MADS y MODS 43
Tabla 29. Referencia básica del modelo de metadatos MODS 46
Tabla 30. Registro bibliográfico descrito con metadato MODS 47
Tabla 31. Espacio de nombres con sus prefijo aplicado en la codificación de METS ... 48
Tabla 32. Inicio de la codificación del registro bibliográfico con METS y MODS 48

Tabla 33. Referencia de codificación de METS 52
Tabla 34. Descripción de componentes de la macroestructura web 65
Tabla 35. Ejemplo de formulario web que emplea el método POST 66
Tabla 36. Ejemplo de petición de cabecera HTTP 67
Tabla 37. Ejemplo de respuesta de cabecera HTTP 68
Tabla 38. Métodos de recuperación de variables 69
Tabla 39. Consulta dinámica básica en Google 69
Tabla 40. Consulta dinámica avanzada en Google 70
Tabla 41. Consultas de directorios de distintos tipos y versiones de servidores 75
Tabla 42. Aspectos básicos de la configuración de un archivo robots.txt 89
Tabla 43. Factores para el posicionamiento web en Google 93

17. Índice de figuras

Figura 1. Relaciones entre documento, metadatos y valores descriptivos. 23
Figura 2. Linked Data en el análisis documental 24
Figura 3. Validación de triples y representación gráfica de los campos y datos..... 26
Figura 4. Modelo DCAM básico en Dublin Core. 27
Figura 5. Validación, triples y esquema de metadatos Dublin Core en RDF..... 31
Figura 6. Esquema de un programa parser aplicado a sindicación de contenidos..... 55
Figura 7. Esquema de un programa parser especializado en metadatos..... 56
Figura 8. Funcionamiento del programa de webcrawling Mbot. 60
Figura 9. Representación de la macroestructura web, o análisis de grafo..... 64
Figura 10. Fórmula estándar para el cálculo del PageRank 91

18. Bibliografía y referencias

- ABRAHAM, R.H. 1996. Webometry: measuring the complexity of the World Wide Web. World Futures, 50, 785-791. Disponible en: http://www.ralph-abraham.org/articles/MS%2385.Web1/

- ABRAHAM, R.H. 1998. Webometry: measuring the synergy of the World Wide Web. Biosystems. 46(1-2), 209-212.

- ALONSO BERROCAL, J.L.; GARCÍA FIGUEROLA, L.C.; ZAZO RODRÍGUEZ, F. 2004. Cibermetría: Nuevas Técnicas de Estudio Aplicables al Web. Madrid: Trea.

- AMARAL, M. 2010. METS for Transferable Metadata. Disponible en: http://easydigitalpreservation.wordpress.com/2010/06/30/mets-for-transferable-metadata/

- Apache Software Foundation. 2008. Class GetMethod [Especificaciones oficiales del método GET]. Disponible en: http://hc.apache.org/httpclient-3.x/apidocs/org/apache/commons/httpclient/methods/GetMethod.html

- ARROYO, N.; ORTEGA, J.L. PAREJA, V.; PRIETO, J.A.; AGUILLO, I. 2005. Cibermetría: Estado de la cuestión. En: 9as Jornadas Españolas de Documentación, FESABID (Madrid 14 y 15 de abril). Disponible en: http://digital.csic.es/bitstream/10261/4296/1/R-17.pdf

- BEEL, J.; GIPP, B.; WILDE, E. 2010. Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co. Disponible en: http://www.sciplore.org/publications/2010-ASEO--preprint.pdf

- BJÖRNEBORN, L. 2004. Small-world link structures across an academic web space: a library and information science approach. Copenhage: Department of Information Studies, Royal School of Library and Information Science.

- BLÁZQUEZ OCHANDO, M. 2010. Aplicaciones de la sindicación para la gestión de catálogos bibliográficos. Disponible en: http://eprints.ucm.es/11233/1/T32065.pdf

- BOUNTOURI, L.; GERGATSOULIS, M. 2009. Interoperability between archival and bibliographic metadata: an EAD to MODS crosswalk. Disponible en: http://eprints.rclis.org/14598/1/bountouri_interoperability.pdf

- BRIN, S.; PAGE, L. 2000. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Disponible en: http://infolab.stanford.edu/~backrub/google.html

- CARROLL, N. 2010. Search engine optimization and user behaviour. Disponible en: http://www.hastingsresearch.com/net/09-SEO-ELIS-encyclopedia-article.html

- CHAPMAN, J. (et.al.). 2005. MODS Implementation Guidelines for Cultural Heritage Materials. Disponible en: http://old.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf

- CODINA, L. 2004. Posicionamiento Web: Conceptos y Ciclo de Vida. Disponible en: http://www.geocities.ws/.../Posicionamiento_Web_Conceptos_y_Ciclo_de_Vida.pdf

- CODINA, L.; MARCOS, M.C. 2005. Posicionamiento web: conceptos y herramientas. Disponible en: http://www.elprofesionaldelainformacion.com/contenidos/2005/marzo/1.pdf

- CODINA, L.; MARCOS, M.C. 2005. Posicionamiento web: conceptos y herramientas. Disponible en: http://www.elprofesionaldelainformacion.com/.../2005/.../1.pdf

- CYGANIAK, R. 2012. prefix.cc: namespace lookup for RDF developers. Disponible en: http://prefix.cc/

- DCMI. 2008. [Especificaciones Oficiales]. Expressing Dublin Core metadata using HTML/XHTML Meta and link elements. Disponible en: http://dublincore.org/documents/dc-html/

- DCMI. 2012. [Especificaciones Oficiales]. Dublin Core Metadata Element Set, Version 1.1. Disponible en: http://dublincore.org/documents/dces/

- FABA PÉREZ, C.; GUERRERO BOTE, V.P.; F. MOYA ANEGÓN. 2004. Fundamentos y técnicas cibermétricas. Badajoz: Consejería de Educación, Ciencia y Tecnología. Junta de Extremadura.

- FLEISS, W. 2007. SEO in the Web 2.0 Era: The Evolution of Search Engine Optimization. BKV. pp7. Disponible en: http://www.bkv.com/redpapers-media/SEO-in-the-Web-2.0-Era.pdf

- GILLILAND, A.J.; GILL, T.; WHALEN, M.; WOODLEY, M.S. 2008. Introduction to metadata. Getty.

- GONZALO, C. 2006. Tipología y análisis de enlaces web: aplicación al estudio de los enlaces fraudulentos y de las granjas de enlaces. Disponible en: http://www2.ub.edu/bid/consulta_articulos.php?fichero=16gonza2.htm

- GOOGLE. 2012. Directrices para webmasters. Disponible en: http://support.google.com/webmasters/bin/answer.py?hl=es&answer=35769

- GOOGLE. 2012. Guía para principiantes sobre optimización para motores de búsqueda. Disponible en: http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.es/es/es/webmasters/docs/guia_optimizacion_motores_busqueda.pdf

- GRAELLS, E.; R. BAEZA YATES. 2007. Características de la Web Chilena 2007. Santiago de Chile. Disponible en: http://alumnos.dcc.uchile.cl/~egraells/wp-content/uploads/2008/10/estudio-ecc.pdf

- HEATH, T. 2012. Linked Data - Connect Distributed Data across the Web. Disponible en: http://linkeddata.org/

- LOC. 2010. MODS Full Record Examples. Disponible en: http://www.loc.gov/standards/mods/v3/mods-userguide-examples.html

- LOC. 2012. MADS 2.0 User Guidelines. Disponible en: http://www.loc.gov/standards/mads/userguide/index.html

- LOC. 2012. Metadata Authority Description Schema: schema and documentation. Disponible en: http://www.loc.gov/standards/mads/

- LOC. 2012. MODS 3.4 User Guidelines. Disponible en: http://www.loc.gov/standards/mods/userguide/

- LOC. 2012. Outline of Elements and Attributes in MADS 2.0. Disponible en: http://www.loc.gov/standards/mads/mads-outline.html

- LOC. 2012. Outline of Elements and Attributes in MODS Version 3.4. Disponible en: http://www.loc.gov/standards/mods/mods-outline.html

- LOC. 2012. PREMIS: Preservation Metadata Maintenance Activity. Disponible en: http://www.loc.gov/standards/premis/

- LOC; DLF. 2010. METS: Metadata encoding and transmission standard: primer and reference manual. Disponible en: http://www.loc.gov/standards/mets/METSPrimerRevised.pdf

- LOC; Eito Brun, R. (trad.). 2012. METS: Introducción y tutorial. Disponible en: http://www.loc.gov/standards/mets/METSOverview_spa.html

- LONG, J. 2005. Hacking con Google. Anaya Multimedia.

- LONG, J. 2008. Google Hacking 2. Mitp

- LONG, J. 2012. Google Hacking for Penetration Testers. OReilly

- MCCALLUN, S.; GUENTHER, R. 2010. Using MODS for discovery of LC's rich collections. Disponible en: http://presentations.ala.org/images/e/e3/Mccallum_guenther.pdf

- MÉNDEZ, E.; SENSO, J.A. 2004. Introducción a los metadatos: estándares y aplicación. Disponible en: http://www.sedic.es/autoformacion/metadatos/programa.htm

- NILSSON, M.; POWELL, A.; JOHNSTON, P. NAEVE, A. 2008. [Especificaciones Oficiales]. Expressing Dublin Core metadata using the Resource Description Framework (RDF). Disponible en: http://www.dublincore.org/documents/dc-rdf/

- NILSSON, M.; POWELL, A.; JOHNSTON, P.; NAEVE. A. 2008. Expressing Dublin Core metadata using the Resource Description Framework (RDF). Disponible en: http://www.dublincore.org/documents/dc-rdf/

- NISO. 2004. Understanding Metadata. Disponible en: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

- OCLC. 2012. VIAF: Fichero de Autoridades Virtual Internacional. Disponible en: http://viaf.org/

- OGDEN, LR. 2008. Black-Hat SEO Practices. Disponible en: http://eprints.rclis.org/3986/

- PAGE, L. 2001. Method for node ranking in a linked database. Disponible en: https://docs.google.com/a/google.com/viewer?url=www.google.com/patents/US6285999.pdf

- PALMER, S.B. 2002. RDF in HTML: Approaches. Disponible en: http://infomesh.net/2002/rdfinhtml/

- PEARCE, J.; PEARSON, D.; WILLIAMS, M.; YEADON, S. 2008. The Australian METS Profile – A Journey about Metadata. En: D-Lib Magazine. Vol.14 (n3/4). Disponible en: http://www.dlib.org/dlib/march08/pearce/03pearce.html

- Rovira, Cristòfol and Fernández-Cavia, José and Pedraza-Jimenez, Rafael and Huertas, Assumpció Posicionamiento en buscadores de las webs oficiales de capitales de provincia españolas. El profesional de la información, 2010, vol. 19, n. 3, pp. 277-283. [Journal Article (Print/Paginated)]. Disponible en: http://eprints.rclis.org/14658/1/Rovira-Fdez-Cavia-Pedraza-Huertas.pdf

- SHREEVES. S.L. (et.al.). 2009. Digital Library Federation - Aquifer Implementation Guidelines for Shareable MODS Records. Disponible en: https://wiki.dlib.indiana.edu/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf

- STEVEN, S.J. 2011. Dublin Core and MODS Element Comparison Examples. Disponible en: http://www.neal-schuman.com/metadata-digital-collections/MDC_DC-MODS_Element_Comparison_Examples.pdf

- W3C. MANOLA, F.; MILLER, E. 2004. [Especificaciones Oficiales]. RDF Primer. Disponible en: http://www.w3.org/TR/rdf-primer/

- WALSH, B. 2010. Building readership - Create Good Linkbait. En: Clear Bloggin. Springer. pp287. Disponible en: http://link.springer.com/content/pdf/10.1007%2F978-1-4302-0321-6_13

- West. A.W. 2012. Search engine optimization. En: HTML5. Disponible en: http://link.springer.com/content/pdf/10.1007%2F978-1-4302-4276-5_16

- WU, B.; DAVISON, B.D. 2005. Identifying Link Farm Spam Pages. Disponible en: http://www.ra.ethz.ch/CDstore/www2005/docs/p820.pdf