VNiVERSiDAD D SALAMANCA

MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

TRABAJO FIN DE MÁSTER

EL USO DE METADATOS EN WEBS INSTITUCIONALES

Autor: Pedro Cortés García

Vº Bº

Tutor: Carlos García Figuerola

Salamanca, 2019

MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

UNIVERSIDAD DE SALAMANCA
FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN
MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

Trabajo Fin de Máster

EL USO DE METADATOS EN WEBS INSTITUCIONALES

Autor: Pedro Cortés García

Tutor: Carlos García Figuerola

Salamanca, 2019

ASIENTO CATALOGRÁFICO ADAPTADO AL FORMATO DEL REPOSITORIO INSTITUCIONAL GREDOS

- **Título**:

El uso de metadatos en webs institucionales

- **Autor:**

Cortés García, Pedro

- **Director:**

G. Figuerola, Carlos

- **Palabras clave:**

[ES] Metadato, Metaetiqueta, Esquema de metadatos, Dublin Core, Web institucional, Recuperación de información, Descripción de información, Crawler

[EN] Metadata, Metatag, Metadata standard, Dublin Core, Institutional website, Information retrieval, Information description, Crawler

- **Clasificación UNESCO:**

57 Lingüística : 5701 Lingüística aplicada : 570102 Documentación automatizada

- **Fecha:**

2019-07-05

- **Resumen:**

[ES] El presente Trabajo Fin de Máster pretende servir de acercamiento teórico al concepto, características, tipología y funciones del metadato (y a algunos de sus esquemas) y su inclusión en recursos digitales, en este caso en la World Wide Web, concretamente en los sitios web institucionales de cada Comunidad Autónoma española.

[EN] This Master's End Paper aims to serve as a theoretical approach to the concept, characteristics, typology and functions of metadata (and some of its schemes) and its inclusion in digital resources, in this case in the World Wide Web, specifically in the institutional websites of each Spanish Autonomous Community.

- **Descripción**:

Trabajo de Fin de Máster en Sistemas de Información Digital, 2019.

SUMARIO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

GLOSARIO DE SIGLAS

DTD Document Type Definition

Especificación formal de los elementos estructurales y de las definiciones necesarias para la codificación de ciertos tipos de documentos en SGML (véase la definición más abajo). Entre los DTD se incluyen MARC y TEI (véanse las definiciones más abajo).

EAD Encoded Archival Description

Norma que utiliza un vocabulario XML (véase la definición más abajo) para la codificación de instrumentos de descripción de archivos con el fin de publicar, intercambiar y usar la información archivística a través de internet.

FTP File Transfer Protocol

Protocolo de red para la transferencia de archivos entre sistemas conectados a una red TCP (véase la definición más abajo), basado en la arquitectura cliente-servidor.

GNU GNU's Not Unix

Colección de programas informáticos y un sistema operativo de tipo Unix, desarrollado por y para el Proyecto GNU, y auspiciado por la Free Software Foundation.

HTML HyperText Markup Language

Lenguaje "markup" (o de marcado) derivado del SGML (véase la definición más abajo) que se usa para crear documentos para las aplicaciones de la World Wide Web. El HTML enfatiza el diseño más que la representación de la estructura del documento o de los elementos de los datos.

HTTP HyperText Transfer Protocol

Protocolo estándar que permite a los usuarios con navegadores acceder a documentos HTML y a medios externos.

ISP Internet Service Provider

Organización o proveedor comercial que da acceso a Internet.

MARC MAchine-Readable Cataloging

Conjunto de estructuras de datos estandarizadas que se usan para describir los materiales bibliográficos que facilitan la catalogación cooperativa y el intercambio de datos en sistemas de información bibliográfica.

RDF Resource Description Framework

Familia de especificaciones de la W3C (véase la definición más abajo) originalmente diseñado como un modelo de datos para metadatos.

SGML Standard Generalized Markup Language

Estándar del ISO (International Organization for Standarization) que sirve para definir, especificar y crear documentos digitales que puedan ser enviados, visualizados, enlazados y manipulados de forma independiente del sistema.

TCP/IP Transmission Control Protocol/InternetProtocol

Conjunto de protocolos de interconexión estandarizados del ISO que facilitan el enlace de sistemas de información institucionales a otros sistemas de información en el Internet, independientemente de su plataforma de ordenadores. El TCP y el IP son dos estándares de comunicación para software que permiten la comunicación entre múltiples ordenadores en un contexto sin errores.

TEI Text Encoding Initiative

Iniciativa colaborativa internacional encaminada a crear directrices genéricas para un esquema de codificación estándar de textos académicos.

URL Uniform Resource Locator (también llamado Universal Resource Locator)

Dirección de Internet que indica al usuario cómo y dónde encontrar una ficha determinada en la World Wide Web. Un URL no sólo incluye el nombre del fichero sino también el nombre del ordenador host, el directorio y el protocolo necesario.

URN Uniform Resource Name (también llamado Universal Resource Name/Number)

El identificador único de una ficha de Internet que es independiente de la localización de la misma. La ficha continúa siendo accesible gracias a su URN, sin importar los cambios que puedan ocurrir en su host o en el recorrido del directorio.

WWW World Wide Web

Una arquitectura de cliente-servidor para acceder a documentos en hipermedia a través de Internet.

W3C World Wide Web Consortium

Consorcio internacional que genera recomendaciones y estándares que aseguran el crecimiento de la World Wide Web a largo plazo.

XML eXtensible Markup Language

Subconjunto simplificado de SGML, designado específicamente para ser usado con la World Wide Web, que proporciona estructuras de datos más sofisticadas que el HTML.

Z39.50

Un protocolo estándar para la localización de información del ANSI (American National Standards Institute) que permite someter una busca a una base de datos independientemente del tipo de hardware o software que ésta use. Fue originalmente implementado en el mundo bibliotecario.

1. INTRODUCCIÓN

“La rápida proliferación de documentos digitales inició una nueva etapa en la organización de la información. Muchas propuestas para la codificación de la información digital surgieron para dar solución a las diferentes necesidades de los usuarios. En este marco, es donde se hace necesaria la aplicación de los metadatos ofreciendo amplias posibilidades para representar de manera estandarizada la descripción e información digital. Dicha incursión de los metadatos en las tareas de descripción de documentos web, supuso un gran avance en materia de recuperación de la información, pues vino a facilitar, pero sobre todo a normalizar la representación de la información digital. La cuestión entonces es ¿cómo hacer para que la recuperación de la información sea efectiva bajo los criterios de búsqueda del usuario? La respuesta puede hallarse en hacer más efectivos los motores de búsqueda que emplean los sistemas de recuperación de información; pero más que eso, la verdadera solución es describir los objetos o documentos digitales de la mejor forma posible, y bajo estándares establecidos, para que, de esta manera, se puedan recuperar de una forma fácil, eficiente y pertinaz. En este sentido donde la aparición de los metadatos aplicados a la representación de documentos digitales para su posterior búsqueda y recuperación fue clave”. (Castro-Ponce, 2013)

Por todo ello y debido a su importancia, el presente Trabajo Fin de Máster pretende servir de acercamiento teórico al concepto, características, tipología y funciones del metadato (y algunos de sus esquemas) y su inclusión en recursos digitales, en este caso en páginas de la World Wide Web. Por tanto, este Trabajo se trata de responder a preguntas tales como: ¿Qué son realmente los metadatos? ¿Y sus esquemas o sistemas? ¿Es un concepto realmente nuevo? ¿Tienen presente? ¿Y futuro? ¿Para qué sirven? ¿Qué normativas han surgido? ¿Qué diferencias y similitudes hay entre ellas? ¿Cómo se aplican en un caso concreto?

Para responder a esta última cuestión y tras las consideraciones teóricas previas, descripción de herramientas especializadas y metodología aplicada se realiza un análisis y evaluación de las webs institucionales de cada Comunidad Autónoma para conocer el estado real de la implementación de los metadatos en cada sitio web. Los resultados obtenidos se estructuran mediante tablas y se comentan con un pequeño texto de cada uno de los casos estudiados. Posteriormente se alcanzarán unas conclusiones o consideraciones generales y particulares. Finalmente, el presente Trabajo Fin de Máster se completa con la bibliografía utilizada y un listado de enlaces de interés.

1.1.- Justificación

La principal justificación del presente Trabajo Fin de Máster es poner en práctica, contrastar y ampliar los conocimientos adquiridos durante el Máster, en el especial en la asignatura de Descripción, representación y organización de contenidos digitales. Para lograrlo se exploran las funciones y procesos vinculados al tratamiento técnico, la representación y la recuperación de información digital, abordando los conceptos y modelos necesarios para la representación y organización de los contenidos digitales y se analizando las técnicas y los modelos que se aplican en el proceso de recuperación de la información digital y la evaluación de, en concreto, los denominados metadatos.

Se pretende, pues, potenciar la capacidad de uso y aplicación de las técnicas, normativas y otros instrumentos estudiados en el Máster utilizados en la organización,

representación, preservación, recuperación, acceso, difusión e intercambio de la información digital y la aplicación de habilidades en la obtención, tratamiento e interpretación de metadatos sobre el entorno de las unidades y servicios de información digital y el estudio, la gestión y la evaluación de los procesos de producción, transferencia y uso de la información digital. Para ello, el presente Trabajo va a versar sobre aquellos metadatos en entorno web que facilitan la: identificación de documentos en un entorno web, la descripción de su contenido, la localización y accesibilidad y la gestión de derechos: copyright, reproducción, restricciones de acceso, etc., ya que son los que más se acercan a la posible solución del problema planteado a lo largo de la introducción de este texto: el exceso de información en Internet y la dificultad de su localización y posterior recuperación. A su vez, otra razón de ser del presente Trabajo es buscar cierta originalidad debido a la inexistencia de estudios o investigaciones sobre el uso y aplicación de metadatos en webs institucionales de cada Comunidad Autónoma española.

1.2.- Objetivos

El objetivo general del presente Trabajo Fin de Máster es aplicar los conocimientos y las competencias adquiridas en las asignaturas del posgrado, en el contexto de la descripción y recuperación de información digital, concretamente mediante el estudio del campo de los metadatos.

Entre los objetivos específicos encontramos los siguientes:

- Analizar los conceptos, tipologías, características y funciones de los metadatos y conocer su forma de implementación.
- Estudiar los esquemas de metadatos, describiendo y realizando una comparación entre algunos de ellos.
- Detectar, extraer, recolectar y gestionar metadatos implementados en las páginas web aplicando las herramientas avanzadas adecuadas a un caso práctico.
- Evaluar y valorar la implementación de dichos metadatos en los sitios web institucionales de las Comunidades Autónomas españolas representando lo analizado mediante tablas y/o gráficos.

1.3.- Estructura del Trabajo

El presente Trabajo Fin de Máster comienza con un acercamiento teórico al concepto de metadato (y a sus esquemas). Posteriormente se realiza una breve descripción general de las herramientas existentes y, más concretamente, de las usadas en el caso práctico. Dicha revisión bibliográfica del tema seleccionado se elaboró con la información obtenida a través de las bases de datos, buscadores académicos, recolectores de repositorios institucionales y con la ayuda del personal de la biblioteca de la Facultad de Traducción y Documentación en la Universidad de Salamanca. Tras la presentación de dichos instrumentos se desgrana la metodología aplicada en ellos para nuestro estudio. Se completa el Trabajo con la presentación y análisis de resultados extrayendo unas conclusiones derivadas de nuestros objetivos. Para finalizar se incluye la bibliografía utilizada y unos enlaces de interés.

2. CONTEXTO TEÓRICO

2.1 Definición de metadato

“Los metadatos, en sí, no suponen algo completamente nuevo dentro del mundo bibliotecario. El término metadatos (del griego μετα, «meta», 'después de, más allá de' y latín datum, 'lo que se da', «dato») fue acuñado por Jack Myers en la década de los 60 para describir conjuntos de datos. La primera acepción que se le dio (y actualmente la más extendida) fue la de “dato sobre el dato”, ya que proporcionaban la información mínima necesaria para identificar un recurso”. (Howe, 1993)

Más adelante, en este mismo Trabajo Fin de Máster se descubrirá que además puede incluir información descriptiva sobre el contexto, calidad y condición o características del dato.

“Teniendo en cuenta que la mayoría de sistemas de metadatos ha sido creada no sólo por profesionales de la información sino también por informáticos, diseñadores de programas, técnicos de sistemas, etc., la utilización de este término puede conllevar una carga excesiva (por ejemplo, reglas de catalogación, clasificaciones de materias...). El concepto de metadato se utiliza como un término neutral, que permite alejarnos de posibles prejuicios por parte de todas aquellas personas menos cercanas al mundo bibliotecario, y que coloca a todos los grupos profesionales implicados en su desarrollo en una posición de igualdad”. (Caplan, 1995)

En el informe de Biblink[1] (Heery, 1996) el metadato se define como “información sobre una publicación en oposición a su contenido. No sólo incluye descripción bibliográfica, sino que también contiene información relevante como materias, precio, condiciones de uso, etc”. Otros autores extienden el concepto de *dato sobre el dato* al afirmar que “incluyen información sobre su contexto, contenido y control, así como todo lo que tenga que ver con el dato" (Pasquinelli, 1997).

La evolución del término desde esta fecha hasta 1997 ha sido descrita por Lange y Winkler (1997) revelando que no existen demasiadas novedades.

Ercegovac (1999) manifiesta que “un metadato describe los atributos de un recurso, teniendo en cuenta que el recurso puede consistir en un objeto bibliográfico, registros e inventarios archivísticos, objetos geoespaciales, recursos visuales y de museos o implementaciones de software. Aunque puedan presentar diferentes niveles de especificidad o estructura, el objetivo principal es el mismo: describir, identificar y definir un recurso para recuperar, filtrar, informar sobre condiciones de uso, autentificación y evaluación, preservación e interoperabilidad”.

En el mundo archivístico, señala Méndez Rodríguez (2003a), que el momento de inflexión se ubica en el Archiving Metadata Forum realizado en el año 2000. En este encuentro se definieron los metadatos como: “Información estructurada o semiestructurada que facilita la creación, gestión y uso de registros a través del tiempo, dentro del dominio en que fue creado o a lo largo de él. Los metadatos para la gestión de documentos digitales pueden usarse para identificar, autentificar y contextualizar registros; y las personas procesos y sistemas que los crean y gestionan y los mantienen y utilizan”.

La Norma UNE-ISO 15489-1:2016 entiende la gestión de metadatos como una parte inseparable de la gestión de documentos, los metadatos se definen “como datos que

[1] Nombre del proyecto puesto en marcha por iniciativa de un grupo de bibliotecas nacionales europeas que tenía como principal objetivo el estudio del rol de las bibliografías nacionales en relación con las publicaciones electrónicas.

describen el contexto, contenido y estructura de los documentos, así como su gestión a lo largo del tiempo".

El artículo 42 del Real Decreto 1671/2009, por el que se desarrolla parcialmente la Ley 11/2007, de acceso electrónico de los ciudadanos a los servicios públicos, define los metadatos como "cualquier tipo de información en forma electrónica asociada a los documentos electrónicos, de carácter instrumental e independiente de su contenido, destinada al conocimiento inmediato y automatizable de alguna de sus características, con la finalidad de garantizar la disponibilidad, el acceso, la conservación y la interoperabilidad del propio documento".

Por su parte, atendiendo a la Norma UNE-ISO 23081-1, los metadatos son "información estructurada o semiestructurada que posibilita la creación, registro, clasificación, acceso, conservación y disposición de los documentos a lo largo del tiempo y dentro de un mismo dominio o entre dominios diferentes. Cada uno de estos dominios, representa un área del discurso intelectual y de la actividad social o de la organización desarrollado por un grupo propio o limitado de individuos que comparten ciertos valores y conocimiento. Los metadatos para la gestión de documentos pueden usarse para identificar, autenticar y contextualizar tanto los documentos como los agentes, procesos y sistemas que los crean, gestionan, mantienen y utilizan, así como las políticas que los rigen".

De todas las definiciones expuestas hasta ahora se pueden extraer algunos puntos básicos -- concepto de objeto, recuperación de información, dato sobre el dato -- que nos son útiles para la realización de una nueva definición que abarque a todas las publicadas anteriormente, de tal forma que resulte posible concluir que metadato es "toda aquella información descriptiva sobre el contexto, calidad, condición o características de un recurso, dato u objeto que tiene la finalidad de facilitar su recuperación, autentificación, evaluación, preservación o interoperabilidad" (Méndez y Senso, 2004)

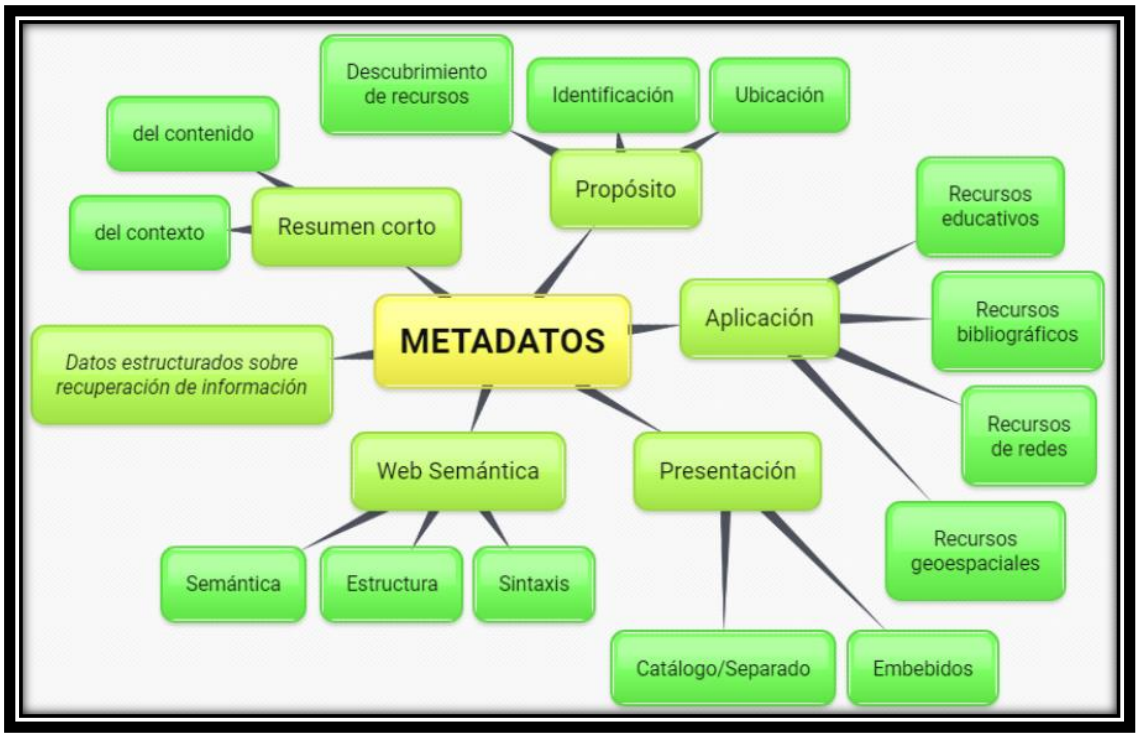

Figura 1.- Mapa conceptual del término metadatos*. Fuente: elaboración propia*

2.2 Tipología de los metadatos

"Existe una gran variedad de formatos en la actualidad. Además, nos encontramos con que la mayoría de las bibliotecas digitales que utilizan metadatos para identificar sus objetos (bien mediante repositorios, bien dentro del mismo objeto) tienden a generar sus propios modelos Esto crea serios problemas a la hora de integrar estos sistemas dentro de un criterio común". (Hípola et al. 2000).

Por eicho motivo, el presente Trabajo Fin de Máster se centrará en aquellos formatos que son de dominio público y que más están siendo descritos por la comunidad científica. Utilizando como sustento el criterio utilizado por Gilliland-Swetland (1998) y traducido por Senso y Piñero (2003), se puede clasificar a los metadatos en los siguientes tipos (*véase Tabla I*):

Tipo	Uso	Ejemplos
Administrativo	Usados en la identificación, gestión y administración de recursos de información	Adquisición de información Derechos y reproducción Requerimientos legales para el acceso Localización de información Criterios de selección para la digitalización Control de la versión
Descriptivo	Utilizados para representar recursos de información	Registros catalográficos Proporcionar ayuda en la búsqueda Índices especializados Hiperenlazar relaciones entre recursos Anotaciones de los usuarios
Preservación	Para salvaguardar los recursos de información	Informar sobre las condiciones de uso de los recursos físicos Informar sobre las acciones llevadas a cabo para preservar versiones físicas y digitales de recursos
Técnico	Relativos a cómo funcionan los sistemas o el comportamiento de los metadatos	Documentación de hardware y software Digitalización de la información (formato, ratio de compresión...) Autentificación y datos de seguridad (encriptación, passwords, etc.) Control de tiempo de respuesta de sistemas
Uso	Relativos al nivel y tipo de uso que se hace con los recursos informativos	Información sobre versiones Reutilización del contenido del recurso

Tabla I.- Tipos de metadatos. Fuente: Senso, J. A., de la Rosa Piñero A. (2003), El concepto de metadato: algo más que descripción de recursos electrónicos, Ciência da Informação, 32, 2, (95-106). Disponible en: http://www.scielo.br/pdf/ci/v32n2/17038.pdf

Siguiendo dicho criterio, se vuelve a consultar a Méndez Rodríguez (2003b) que enumera y completa la definición de estos cinco tipos esenciales de metadatos aceptados por todas las comunidades o dominios de metadatos:

1. Metadatos administrativos o metadatos para la gestión de recursos, los utilizados para la propia gestión y administración de los registros, en el momento de su creación. Las fechas de creación o adquisición, los permisos de acceso, derechos o procedencia, o las pautas para disposiciones como retenciones o retiros, son ejemplos de los derechos que un archivero digital, curador, podría emplear. Metadatos similares serían relevantes para un administrador de base de datos o para los administradores responsables de la captura de flujos de tráfico de redes de datos o telecomunicaciones, o de datos de registros y eventos de seguridad. Todos ellos están relacionados con el contexto administrativo de un determinado recurso de información, facilitando el registro y manejo de diferentes aspectos, tales como los derechos de autor y permisos de acceso, así como las acciones necesarias para su preservación.

2. Metadatos descriptivos. La mayor parte de los esfuerzos dedicados a los metadatos por parte del sector de la archivística se han concentrado en los metadatos descriptivos, como una rama natural de la catalogación tradicional. Sin embargo, parece claro que dedicar demasiada atención a esta área (desarrollando, por ejemplo, etiquetas descriptivas y vocabularios controlados hasta un nivel de refinamiento muy específico) a expensas de otras consideraciones puede causar limitaciones o defectos del sistema en su conjunto.

Los metadatos descriptivos son aquellos que dependen del propio documento y sirven para representar o identificar los objetos de información digital en su fase de organización. La mayor parte de los esfuerzos dedicados a los metadatos por parte del sector de la archivística se han concentrado en los metadatos descriptivos, como una rama natural de la catalogación tradicional.

La siguiente figura (*véase Figura 2*) nos muestra las variadas interdependencias que deben establecerse, donde los metadatos descriptivos son solo una de las diversas subcategorías entre todos los elementos en juego. Se utilizan para describir e identificar los principales atributos o características de los recursos de información, siendo algunos de los tienen mayor relevancia los relacionados con su contenido intelectual o temático.

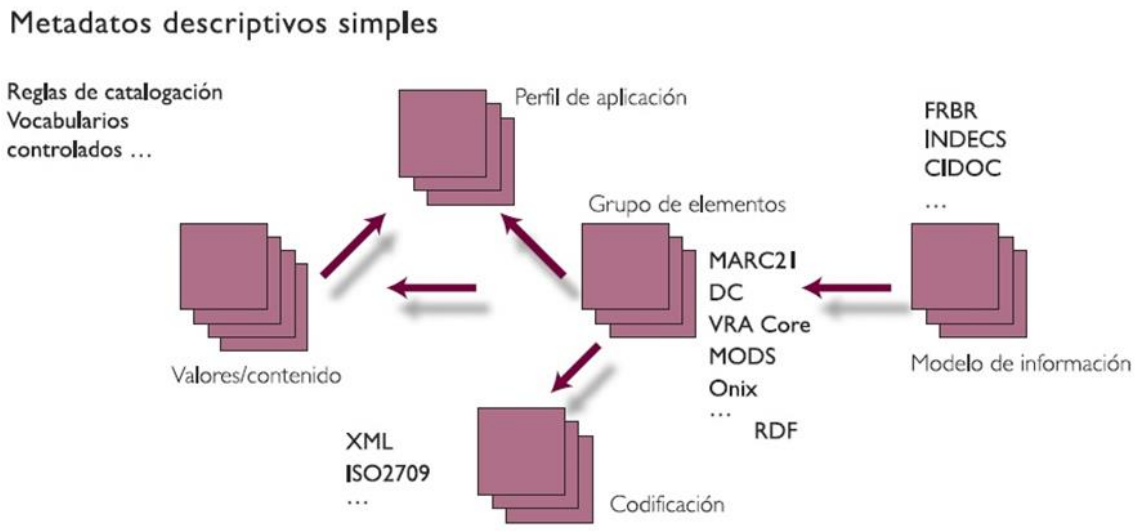

Figura 2.- Esquema de los metadatos descriptivos simples. Fuente: Dempsey, L. (2005). *Metadata: practice and practice*. CLIR/DLF. Managing Digital Assets: A Primer for Library and Information Technology Administrators. Disponible en: https://www.clir.org/wp-content/uploads/sites/6/dempsey.ppt

3. Metadatos para la conservación, aquellos metadatos destinados a gestionar la preservación de las fuentes de información. "Son todos los diversos tipos de datos que nos permitirán la recreación e interpretación de la estructura y el contenido de los datos digitales que se hayan conservado. Definidos de este modo, está claro que los metadatos de conservación tienen que soportar cierto número de funciones relacionadas, pero distintas". (Delgado Gómez, 2005).

Lynch (1999), por ejemplo, escribe que, "dentro de un depósito digital, los metadatos acompañan y hacen referencia a cada objeto digital y proporcionan información asociada descriptiva, estructural, administrativa, de gestión de derechos y de otros tipos. La amplia gama de funciones que los metadatos de conservación pretenden satisfacer significa que la definición de normas de metadatos no es una tarea simple, y que muchos de los esquemas publicados actualmente son relativamente complejos. La situación se complica adicionalmente por la percepción de que diferentes estrategias de

conservación digital o de tipos de información digital requerirán el soporte de diferentes tipos de metadatos".

4. Metadatos técnicos, aquellos creados por, o generados para, un sistema automatizado.

5. Metadatos de uso, generalmente creados de forma automática, relativos al nivel de utilización y al tipo de usuarios de un determinado servicio de información

2.3 Funciones y beneficios del uso de metadatos

FUNCIONES Han sido analizadas a lo largo del tiempo por varios autores. En la siguiente tabla se presenta por autores en orden cronológico las funciones que desempeñan los metadatos:		*Buscar / Localizar / Descubrir*	*Recuperar / Extraer / Acceder*	*Transferir*	*Evaluar / Explorar*	*Administrar*	*Describir / Interpretar*	*Preservar*	*Usar / Utilizar /Explorar*	*Técnicos*	*Gestionar*	*Identificar*
Beard	1996	✓	✓	✓	✓							
Gilliland-Swetland	2000					✓	✓	✓	✓	✓		
Senso y De La Rosa Piñero	2003					✓	✓	✓	✓	✓		
ECNBII	2003	✓	✓						✓			
Nebert	2004	✓			✓				✓			
Oosterom	2004	✓	✓		✓				✓			
Johnston y Powell	2005	✓	✓		✓		✓				✓	
Moellering et al.	2006	✓	✓		✓				✓			
Gayatri y Ramachandran	2007	✓	✓								✓	✓
Díaz et al.	2008	✓			✓				✓			
UNE-ISO 19115	2014		✓		✓				✓			

Tabla II- Funciones de los metadatos. Fuente: elaboración propia

Por tanto, se puede extrapolar de la anterior tabla que las funciones principales del metadato son: proporcionar suficiente información para permitir discernir el contenido, formato y alcance de un conjunto de datos, recuperarlos y usarlos evaluando su idoneidad para satisfacer los requisitos deseados.

Para Carbajal Sardá y Moreno Coatzozón (2011) "los principales beneficios del uso de metadatos son los siguientes:

- Adhieren contenido, contexto y estructura a los objetos de información, asistiendo de esta forma al proceso de recuperación de conocimiento desde colecciones de objetos

- Permiten generar distintos puntos de vista conceptuales para sus usuarios o sistemas, y liberan a estos últimos de tener conocimientos avanzados sobre la existencia o

características del objeto que describen. Estos puntos de vista conceptuales pueden depender del sistema o usuario que los utiliza.

- Permiten el intercambio de la información sin la necesidad de involucrar el intercambio de los recursos mismos. Esta particularidad facilita entre otras cosas las búsquedas sobre colecciones distribuidas. Además, los metadatos permiten una descripción precisa y discreta de los recursos permitiendo la creación de colecciones virtuales de descripciones donde agrupan los objetos de información para satisfacer requerimientos específicos.

- En cada proceso productivo, o en cada etapa del ciclo de vida de un objeto de información, se van generando metadatos para describirlos, y metadatos para describir dichos metadatos (manual o automáticamente). Éstos generan valor agregado al sistema de metadatos y objetos que describen, generando datos desde los cuales es posible extraer conocimiento relevante al sistema mismo y sus procesos.

- Permiten un acceso a los recursos en forma controlada ya que se conoce con precisión el objeto descrito. Es posible entonces establecer sistemas de filtrado y permiten generar bases para una autentificación y mecanismos para definir grados de confianza sobre las fuentes de información.

- Permiten preservar los objetos de información permitiendo migrar (gracias a la información estructural) sucesivamente éstos, para su posible uso por parte de las futuras generaciones. La información semántica de los objetos se mantiene, disminuyendo así la pérdida del conocimiento.

- Los metadatos son esenciales para sostener un crecimiento de una Web a mayor escala, permitiendo búsquedas e integración del conocimiento desde un mayor número de fuentes heterogéneas".

2.4 Implementación e instituciones implicadas en la introducción de metadatos

En el momento de creación de metadatos relativos a un documento, UNE-ISO 23081-1 se suma al modelo del *continuum* del documento indicando que "inicialmente los metadatos definen el documento en el mismo momento de su incorporación, fijándole en su contexto y estableciendo el control de su gestión. Durante la existencia de los documentos o sus agrupaciones, se irán añadiendo nuevas capas de metadatos debido a la existencia de nuevos usos en otros contextos. Esto significa que a lo largo del tiempo los metadatos continúan acumulando información relacionada con el contexto de gestión de los documentos, los procesos de negocio en los que se utilizan, así como sobre los cambios estructurales que les afectan o su apariencia. Los metadatos aplicados a los documentos durante su vida activa pueden también seguir utilizándose cuando no sean necesarios para la gestión, pero sean conservados para facilitar la investigación o debido a otros valores". UNE-ISO 23081-1 cifra "las razones para la implementación de metadatos o la utilidad de estos en los puntos siguientes:

- Proteger los documentos como prueba y asegurar su accesibilidad y disponibilidad a lo largo del tiempo.
- Facilitar la comprensión de los documentos. Servir de base y garantizar el valor probatorio de los documentos.
- Contribuir a garantizar la autenticidad, fiabilidad e integridad de los documentos.
- Respaldar la gestión del acceso, la privacidad y los derechos de propiedad intelectual.
- Servir de base para una recuperación eficiente. Respaldar las estrategias de interoperabilidad, permitiendo que se incorporen oficialmente al sistema

documentos creados en diversos entornos administrativos y técnicos y que se mantengan durante tanto tiempo como sea necesario.

- Proporcionar vínculos lógicos entre los documentos y su contexto de creación, manteniéndolos de forma estructurada, fiable e inteligible.
- Facilitar la identificación del entorno tecnológico en que los documentos digitales fueron creados o se incorporaron al sistema y la gestión del entorno tecnológico en el que se han mantenido, de modo que puedan ser reproducidos como documentos auténticos cuando se necesiten.
- Facilitar la migración eficiente y exitosa de documentos electrónicos de un entorno o plataforma informáticos a otro, o cualquier otra posible estrategia de conservación".

En cuanto a la implementación, UNE-ISO 15489-2, 11.3 aclara que "la atribución de metadatos debería ser tan automática como sea posible. La atribución manual de metadatos, debería hacerse, en la medida de lo posible, utilizando listas predefinidas de selección (no campos abiertos que puedan ser cumplimentados a voluntad). Normalmente, esto ocurrirá con una parte de los metadatos de incorporación, mientras que todos los metadatos de proceso deberían introducirse automáticamente. En el momento de la incorporación, los metadatos definen el documento en el momento en que se incorpora al sistema, fijándole en su contexto de actividad y estableciendo un control sobre la gestión del mismo. Durante la existencia de los documentos o sus agrupaciones, se añadirán nuevas capas de metadatos. Puede que una parte de estos metadatos sea atribuida por un usuario, pero en la medida de lo posible deberían reunirse automáticamente. Los metadatos atribuidos manualmente requieren mayor validación para apoyar la coherencia y la calidad semánticas y sintácticas".

Según Senso y Piñero (2003) "siempre que se habla de metadatos, tarde o temprano, aparece la pregunta: *¿quién debería ser el responsable de introducirlos en los documentos electrónicos?* Estos mismos autores responden que este interrogante nos lo encontramos especialmente en el entorno Internet. Pocos son los productores que hayan propuesto soluciones suficientemente eficaces para este problema. Entre otras cosas, y muy probablemente, porque no se trata de un aspecto técnico o de un problema tecnológico sino de concienciación sobre la necesidad de las cosas. A estas alturas poca gente cuestiona que sean los bibliotecarios los que realicen las descripciones bibliográficas de los ejemplares que componen la colección en sus centros y, en realidad, lo que hacen es describir el contenido de fondos, es decir, introducir metadatos. Con esto no se pretende señalar al bibliotecario como eje fundamental para que funcione un sistema de recuperación de información distribuida".

Como indica Heery (1996), "las organizaciones involucradas en la creación, mantenimiento y actualización de metadatos se pueden categorizar en:

1. Editores: del campo de la edición clásica y del mundo de la edición electrónica.

2. Servicios de información: agencias bibliográficas nacionales (en el entorno británico el British Library National Bibliographic Service), agencias bibliográficas comerciales (Whitaker), agencias que sirven resúmenes o servicios de indización (INSPEC), bases de datos de publicaciones periódicas (Blackwells, CARL).

3. Proveedores: de monografías (Dawson) o de publicaciones periódicas o seriadas (Swets, EBSCO).

4. Bibliotecas: por medio de agencias generadoras o gestoras de catálogos colectivos (OCLC – precisamente Dublin Core nace como iniciativa de este consorcio bibliotecario) y de cada una de las bibliotecas que preste especial atención a los recursos electrónicos (bibliotecas digitales).

Un aspecto que resulta evidente es que, con el aumento de productos en formato electrónico, la generalización de la edición y la aparición constante de nuevas

herramientas, la descripción de estos recursos por medio de metadatos es un hecho que debe tender a globalizarse".

De nuevo, para Heery (1996), "las organizaciones involucradas en esta propuesta se pueden considerar como clásicas dentro del mundo de la edición – ya sea ésta electrónica o no –. Con el aumento de productos en formato electrónico, la introducción de metadatos es fruto del trabajo de varias organizaciones. Además de estas categorías existe un nuevo grupo de entidades implicadas dentro del proceso de edición:

1. Autores: el ejemplo más claro lo tenemos en el caso que nos ocupa, las páginas Web.

2. Servicios de búsqueda en Internet.

3. Servicios de archivos electrónicos: colecciones de materiales electrónicos como Oxford Text Archive, Essex Data Archive, Electronic Text Centre de la University of Virginia, Cervantes Virtual…

4. Depósitos (repositorios) de colecciones de documentos: algo muy común dentro del mundo académico norteamericano (Los Alamos National Psysics Pre-print Archive).

5. Bibliotecas digitales".

De estos cinco grupos, los que más interesan para los resolver los objetivos de este Trabajo Fin de Máster sean los dos primeros.

2.5 Esquema de metadatos

El Real Decreto 4/2010 que regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica entiende esquema de metadatos como "Instrumento que define la incorporación y gestión de los metadatos de contenido, contexto y estructura de los documentos electrónicos a lo largo de su ciclo de vida".

Por su parte, la norma UNE-ISO 23081-1, 3.3 define esquema de metadatos como un "plan lógico que muestra las relaciones entre los distintos elementos del conjunto de metadatos, normalmente mediante el establecimiento de reglas para su uso y gestión y específicamente respecto a la semántica, la sintaxis y la obligatoriedad de los valores".

El capítulo 10 de la UNE-ISO 23081-2 está dedicado específicamente al desarrollo de un esquema de metadatos para la gestión de documentos y en el mismo se subraya: "Los esquemas de metadatos son potentes instrumentos que facilitan la interoperabilidad y ayudan a asegurar el mantenimiento de los documentos a largo plazo. Añade la norma que los elementos de metadatos pueden tomar su valor de esquemas de codificación, fuentes autorizadas, que incluyen listas predefinidas, clasificaciones, vocabularios controlados o taxonomías".

Ejemplos de esquemas de codificación son las herramientas predefinidas para la gestión de documentos que se definen en UNE-ISO 15489-1, 9.2 y 9.5: cuadros de clasificación, tablas de seguridad y acceso, o calendarios de conservación. Añade UNE-ISO 23081-2, 10.3.3 que "para garantizar la interoperabilidad, los esquemas de codificación necesitan definirse con el mismo rigor que los esquemas de elementos de metadatos. Las relaciones entre los términos de los esquemas de codificación deben admitir su procesamiento automatizado".

Alvite Díaz (2014) recoge "los beneficios de los esquemas de metadatos y de los esquemas de codificación como los siguientes:

- Facilitar una gestión de metadatos consistente e integrada.
- Permitir la interoperabilidad mediante la comparación o mapeo de diferentes conjuntos de metadatos.

- Expresar las interrelaciones de los elementos y su semántica.
- Controlar las relaciones entre elementos de metadatos y su semántica inherente.
- Asegurar y mantener la consistencia en sistemas de información (por ejemplo: sistemas de gestión de documentos).
- Favorecer el desarrollo modular, la ruptura o vinculación de sistemas de información.
- Proporcionar una base para el desarrollo de sistemas de información o bases de datos".

Duval et al. (2002) manifiestan en un artículo "los principios de acuerdo compartidos por dos importantes iniciativas de metadatos: la Dublin Core Metadata Initiative (DCMI) y el Institute for Electrical and Electronics Engineers (IEEE) Learning Object Metadata Working Group (LOM). Este acuerdo surgió de una reunión conjunta del grupo de trabajo (taskforce) de metadatos en Ottawa en agosto de 2001 donde se establecieron los principios comunes que pueden servir de base para el diseño de cualquier esquema de metadatos o aplicación. Estos principios son:

- **Modularidad:** la arquitectura de los metadatos debe ser flexible de manera tal que se puedan combinar de manera interoperable, sintáctica y semánticamente elementos provenientes de diferentes esquemas preestablecidos. Cada componente del esquema debe incluirla funcionalidad y los requisitos específicos para una determinada aplicación.
- **Extensibilidad:** los esquemas de metadatos deben ser lo suficientemente flexibles como para poder ajustarse a las necesidades específicas del repositorio, sin perder de vista la interoperabilidad proporcionada por el esquema base. Así pues, el elemento *creador* va a estar presente en todos los esquemas, como así también el *identificador*, mientras que otros serán específicos para una aplicación particular.
- **Refinamiento:** el esquema de metadatos debe permitir un nivel de detalle más exhaustivo mediante, por ejemplo, la adición de calificadores a los elementos. Esto hará más específico su significado.
- **Plurilingüismo:** los esquemas de metadatos deben respetar la diversidad lingüística y cultural y la internacionalización de la Web, poniendo a disposición del usuario los recursos en su idioma nativo".

2.5.1.- Clasificación de los esquemas de metadatos

Existen varias maneras de clasificar el gran número de esquemas de metadatos existentes. Jenn Riley (2010) "analiza más de 100 estándares de metadatos utilizados en el sector del patrimonio cultural y los clasifica mediante el gráfico *Seeing Standards: A Visualization of the Metadata Universe*, según su propósito, su función, su dominio y la comunidad por la cual puede ser usado.

- Propósito: se refiere al tipo de registro para el cual ha sido diseñado el estándar de metadatos.

- Función: se refiere al papel que tiene el estándar en la creación y almacenamiento de los metadatos.

- Dominio: se refiere a los tipos de materiales para los cuales está destinada la norma o podría potencialmente ser útil. No es una lista exhaustiva, y cada estándar puede utilizarse para varios materiales a la vez.

- Comunidad: se refiere al ámbito en que estos estándares se utilizan o potencialmente podrían ser utilizados".

Hay casi tantos esquemas, modelos o estándares de metadatos, como proyectos de creación de sistemas y servicios de información digital en la World Wide Web. En el cuadro siguiente (*véase Tabla III*) se han seleccionado cinco de los esquemas de

metadatos referenciados por Riley (2010) más conocidos. Se indican el propósito, dominio, función y comunidad a la que cada esquema responde con mayor pertinencia.

ESQUEMA	PROPÓSITO	DOMINIO	FUNCION	COMUNIDAD
DC Dublin Core Metadata Initiative	Descriptivos	Textos Imágenes/Vídeos Sonido Datos geoespaciales Conjunto de datos Objetos culturales	Estándar de estructura	Bibliotecas Industria de la información
METS Metadata Encoding and Transmission Standard	Contenedores	Textos Imágenes/Vídeos Sonido Datos geoespaciales Conjunto de datos Objetos culturales	Estándar de estructura Formato de registro	Bibliotecas
MODS Metadata Object Description Schema	Descriptivos	Textos Imágenes/Vídeos Sonido Datos geoespaciales Conjunto de datos Objetos culturales	Estándar de estructura	Bibliotecas
LOM Learning Object Metadata	Descriptivos	Imágenes	Estándar de estructura Formato de registro	Industria de la información
PREMIS PREservation Metadata Implementation Strategies	De preservación	Textos Imágenes/Vídeos Sonido Datos geoespaciales Conjunto de datos Objetos culturales	Estándar de estructura	Archivos Bibliotecas

Tabla III.- Clasificación de esquemas de metadatos. Fuente: Testa, P. M. y Degiorgi, E. H. (2013) *Esquemas de metadatos para los repositorios institucionales de las universidades argentinas* Tesis. Universidad Nacional de Cuyo. Disponible en: http://bdigital.uncu.edu.ar/objetos_digitales/5881/tesisdegiorgitesta.pdf

Como se puede apreciar "existen muchos esquemas de metadatos y cada institución selecciona uno según su objetivo o los recursos que quiera describir. Las propuestas son múltiples y van desde lo más simple a lo más complejo y/o de lo general a lo específico. Existen esquemas de metadatos por ejemplo para: datos geoespaciales, imágenes, de preservación, recursos educativos y otros". (Testa y Degiorgi, 2013).

En los siguientes apartados se explicará más detalladamente algunos de ellos.

2.5.2.- Dublin Core

Existen diversos esquemas de metadatos, los cuales son cada vez más complejos. A manera de ejemplificar esto, se puede mencionar el conjunto de elementos de metadatos del Dublin Core, el cual está orientado sobre todo a recursos Web; no es muy complejo, constituye uno de los esquemas más difundidos y debido a su importancia y pertinencia es que se analizará más profundamente en el presente Trabajo Fin de Máster.

"Dublin Core nace como producto del trabajo cooperativo de ámbito internacional - promovido en su primera fase por la OCLC y el NCSA (National Center for

Supercomputing Applications)- cuyo objetivo principal fue crear un conjunto de elementos que permitieran la descripción de recursos electrónicos con el fin de facilitar su búsqueda y recuperación". (Ortiz-Repiso Jiménez, 1999)

"Originalmente se concibió como un conjunto de etiquetas que deberían ser generadas por el autor del documento HTML con la finalidad de facilitar su identificación y posterior recuperación en Internet. No obstante, este modelo ha llamado la atención de diversas comunidades de profesionales interesados en la descripción de recursos en museos, bibliotecas, organizaciones comerciales... Desde que en 1995 naciera este proyecto en Dublin (Ohio, Estados Unidos) se ha utilizado el mismo esquema de trabajo. Generalmente, cada pocos meses se reúne una serie de expertos del mundo bibliotecario, de la gestión de redes y comunicaciones, informáticos, etc. para estudiar su impacto, analizar modificaciones e investigar en posibles aplicaciones. En su forma simple de quince elementos, Dublin Core ha sido diseminado como parte del Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) y ha logrado la normalización IETF RFC 5013, la norma ANSI/NISO Z39.85-2007 y la norma ISO 15836:2009. Dublin Core Metadata Initiative (DCMI) ha desarrollado un conjunto mayor de elementos y subelementos (términos de metadatos DCMI) y un marco para el desarrollo de perfiles de aplicación (elementos Dublin Core combinados con vocabularios especializados desarrollados para propósitos particulares). Dublin Core puede codificarse con diversas sintaxis, incluidos texto, HTML, XML y RDF". (Senso y Piñero, 2002)

En resumen, Dublin Core es una norma internacional que especifica un conjunto estándar de metadatos que sirven para identificar con mucha precisión el contenido de recursos digitales.

Los objetivos básicos que dieron pie al primer conjunto de metadatos DCMI fueron, según Weibel (1995):

• "Establecer un sistema normalizado para la descripción de documentos distribuidos en el World Wide Web, cualquiera que sea su formato: HTML, PDF, PS, SGML... Dándole un carácter internacional desde su creación y desarrollo hasta su implementación.

• Facilitar a autores y editores de estos documentos la incorporación de elementos que identifiquen y describan sus aportaciones.

• Facultar la integración haciendo uso de sistemas de clasificación, indización y control de autoridades con los que la comunidad bibliotecaria se encuentra familiarizada.

• Corregir las deficiencias de los sistemas de recuperación basados en texto completo.

• Evitar el uso de formatos complejos (MARC, TEI, etc.).

• Crear una serie de elementos:

- Fáciles de crear y actualizar.
- Comprensibles para cualquier usuario.
- Normalizados.
- Ampliables manteniendo la compatibilidad con definiciones anteriores.
- A los que se pueda aplicar diferentes niveles de especificidad y control mediante el uso de cualificadores y subtipos.
- Referidos a las características intrínsecas de los documentos".

Según Méndez y Senso, 2004, "algunas de las fortalezas de este esquema de metadatos son:

- Su simplicidad
- Repetibles tantas veces como sea necesario.
- La independencia sintáctica (que ha permitido que se integre en la estructuración de datos en XML/RDF).

- Alto nivel de normalización formal: ANSI/NISOZ39.85-2001, ISO 15836-2003.
- Crecimiento y evolución del estándar a través de una institución formal consorciada: la DCMI.
- El conjunto de elementos DC se ha convertido en una infraestructura operacional del desarrollo de la Web Semántica".

La norma UNE-ISO 15836 "define el Conjunto de Elementos Dublin Core, o lo que se conoce habitualmente como *DC simple*. Se compone de 15 elementos básicos para describir cualquier objeto de información, se presentan habitualmente divididos en tres grupos que indican la clase o alcance de la información incluida en ellos, y que responden, en cierta medida, a las expectativas que tiene el usuario cuando se enfrenta a la información de la red:

1.- Contenido: title, subject, description, source, language, relation, coverage

2.- Propiedad intelectual: creator, publisher, contributor, rights

3.- Instanciación: date, type, format, identifier

Su uso está previsto, principalmente, en el seno de instituciones académicas, culturales y de investigación, es decir, en general en contextos donde el esfuerzo de descripción venga justificado por la calidad intrínseca y la condición no volátil de los recursos, por lo cual es fácil prever que también será utilizado por todas aquellas instituciones y empresas que deseen beneficiarse de un estándar de descripción de datos de estas características". La siguiente tabla (*véase Tabla IV*) describe los 15 elementos básicos de Dublin Core:

Elemento DC.	Descripción del elemento
Title	**Título: El nombre dado a un recurso.** Típicamente, un título es el nombre formal por el que es conocido el recurso.
Creator	**Autor: La entidad primariamente responsable de la creación del contenido intelectual del recurso.** Entre los ejemplos de un creador se incluyen una persona, una organización o un servicio. Típicamente, el nombre del creador podría usarse para indicar la entidad.
Subject	**Materias y palabras clave: El tema del contenido del recurso.** Un tema será expresado como palabras clave, frases clave o códigos de clasificación que describan el tema de un recurso. Se recomienda seleccionar un valor de un vocabulario controlado o un esquema de clasificación formal.
Description	**Descripción: La descripción del contenido del recurso.** La descripción puede incluir, pero no se limita a: un resumen, tabla de contenidos, referencia a una representación gráfica de contenido o una descripción de texto libre del contenido.
Publisher	**Editor: La entidad responsable de hacer que el recurso se encuentre disponible.** Ejemplos de editores son una persona, una organización o un servicio. Típicamente, el nombre de un editor podría usarse para indicar la entidad.
Contributor	**Colaborador. La entidad responsable de hacer colaboraciones al contenido del recurso.** Ejemplos de colaboradores son una persona, una organización o un servicio. Típicamente, el nombre del colaborador podría usarse para indicar la entidad.

Date	**Fecha: Una fecha asociada con un evento en el ciclo de vida del recurso.** Típicamente, la fecha será asociada con la creación o disponibilidad del recurso. Se recomienda utilizar un valor de datos codificado definido en el documento "Date and Time Formats", que sigue la norma ISO 8601 que sigue el formato YYYY-MM-DD.
Type	**Tipo: la naturaleza o categoría del contenido del recurso.** El tipo incluye términos que describen las categorías generales, funciones, géneros o niveles de agregación del contenido. Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, el *DCMI Vocabulary* -DCMITYPE-). Para describir la manifestación física o digital del recurso, se usa el elemento Formato.
Format	**Formato: la manifestación física o digital del recurso.** El formato puede incluir el tipo de media o dimensiones del recurso. Podría usarse para determinar el *software*, *hardware* u otro equipamiento necesario para ejecutar u operar con el recurso. Ejemplos de las dimensiones son el tamaño y la duración. Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, la lista de Internet Media Types (MIME) que define los formatos de medios de ordenador).
Identifier	**Identificación: Una referencia no ambigua para el recurso dentro de un contexto dado.** Se recomienda identificar el recurso por medio de una cadena de números de conformidad con un sistema de identificación formal, tal como un URI (que incluye el Uniform Resource Locator -URL, el Digital Object Identifier (DOI) y el International Standard Book Number (ISBN).
Source	**Fuente**: **Una referencia a un recurso del cual se deriva el recurso actual.** El recurso actual puede derivarse, en todo o en parte, de un recurso fuente. Se recomienda referenciar el recurso por medio de una cadena o número de conformidad con un sistema formal de identificación.
Language	**Lengua**: **La lengua del contenido intelectual del recurso.** Se recomienda usar RFC 3066 en conjunción con la ISO 639 [ISO639][5], que define las etiquetas de dos y tres letras primarias para lenguaje, con subetiquetas opcionales. Ejemplo: "en" u "eng" para Inglés, "akk" para Acadio, y "en-GB" para inglés usado en Reino Unido.
Relation	**Relación**: **Una referencia a un recurso relacionado.** Se recomienda referenciar el recurso por medio de una cadena de números de acuerdo con un sistema de identificación formal.
Coverage	**Cobertura: La extensión o ámbito del contenido del recurso.** La cobertura incluiría la localización espacial (un nombre de lugar o coordenadas geográficas), el período temporal (una etiqueta del período, fecha o rango de datos) o jurisdicción (tal como el nombre de una entidad administrativa). Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, del Thesaurus of Geographic Names (TGN) y que, donde sea apropiado, se usen preferentemente los nombres de lugares o períodos de tiempo antes que los identificadores numéricos tales como un conjunto de coordenadas o rangos de datos.
Rights	**Derechos**: **La información sobre los derechos de propiedad y sobre el recurso.** Este elemento podrá contener un estamento de gestión de derechos para el recurso, o referencia a un servicio que

	provea tal información. La información sobre derechos a menudo corresponde a los derechos de propiedad intelectual, copyright y otros derechos de propiedad.

Tabla IV.- Elementos de Dublin Core. Fuente: María Jesús Lamarca Lapuente (2006). *Metadatos Dublin Core. Hipertexto: El nuevo concepto de documento en la cultura de la imagen*. Disponible en: http://www.hipertexto.info/documentos/dublin_core.htm

A continuación, se mostrará un ejemplo de aplicación del esquema:

```
<rdf:Description>
  <dc:creator>Peter Noeller</dc:creator>
  <dc:title>Algebra</dc:title>
  <dc:subject>mathematics</dc:subject>
  <dc:date>2008-04-23</dc:date>
  <dc:language>EN</dc:language>
  <dc:description>
     An Introduction to Algebra
  </dc:description>
</rdf:Description>
```

Figura 3.- Ejemplo de esquema Dublin Core. Fuente: Barrueco, J. M. (2011*) Introducción a los metadatos para las colecciones digitales*. Universidad de Valencia. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf

"Además de los elementos básicos existen otros mecanismos que sirven para adaptar el Dublin Core a las necesidades concretas de información y que hacen que este modelo de metadatos sea aplicable a cualquier proyecto de sistema o servicio de información digital. Estos mecanismos son fundamentalmente:

- A través de perfiles de aplicación, desarrollados para el uso del Dublin Core asociado a distintas disciplinas, como por ejemplo el perfil de aplicación para bibliotecas: DC-Lib.

- Mediante la Dublin Core Metadata Initiative terms: son lo que se denomina *términos de metadatos* (metadata terms), donde se incluyen tanto los nuevos elementos que se van incorporando al vocabulario DC, como las matizaciones de elementos ya existentes, esquemas de codificación (antes conocidos como *cualificadores*) y términos de vocabularios controlados. Todos estos términos de metadatos sirven para adecuar y precisar el valor y la utilidad de la metainformación expresada a través de DC". (Méndez y Senso, 2004)

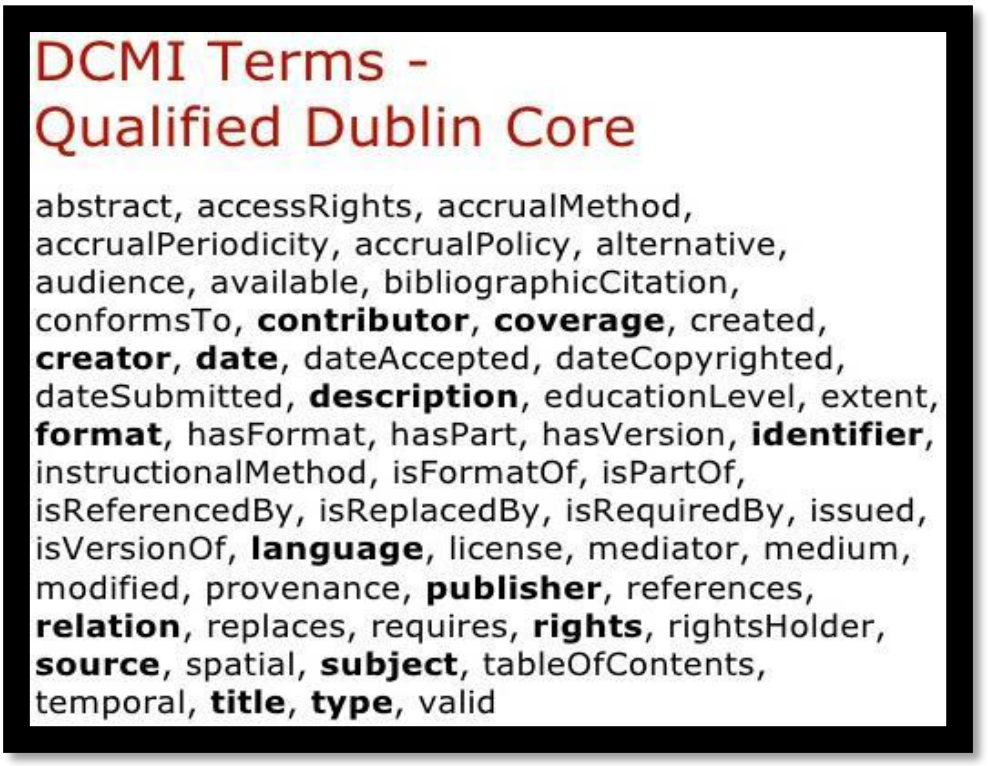

Figura 4.- Términos DCMI *(en negrita los 15 términos originales).* Fuente: Ebner, H. (2009). *Introduction to Dublin Core Metadata.* Knowledge Management Research group of Royal Institute of Technology (KTH), Sweden. Disponible en: https://www.slideshare.net/ebner/introduction-to-dublin-core-metadata

2.5.3.- METS

Según se explica en su propio sitio web oficial[2] la Biblioteca del Congreso de Estados Unidos (Library of Congress) ha desarrollado esquemas, como el Metadata Encoding and Transmision Standard (METS), el cual codifica metadatos administrativos, descriptivos y estructurales para la constitución de objetos digitales de almacenamiento. "Proporciona una estructura jerárquica para codificar metadatos sobre la estructura interna de un ítem. Esto es simplemente una serie de elementos enlazados denominados div (división) cuyo resultado emula la estructura de un objeto digital. Por ejemplo, un libro digitalizado tendrá su estructura de divs organizada para emular su división original en capítulos, secciones, etc. Un documento METS consiste de cinco secciones principales:

- Metadatos descriptivos. Puede simplemente apuntar a una descripción externa al propio documento (por ejemplo, a un registro MARC en un OPAC) o contener embebido los datos o una mezcla de ambos.

- Metadatos administrativos. Proporcionan información sobre cómo se crearon los ficheros, propiedad intelectual, metadatos técnicos, etc. Se puede incluir una o varias entradas de metadatos administrativos por cada documento.

- Grupos de ficheros. Lista todos los ficheros que componen el objeto digital.

- Mapa estructural. Es el corazón de un documento METS. Se encarga de diseñar una estructura jerárquica para el objeto y enlazar los elementos de esta estructura con ficheros y metadatos que pertenecen a dichos ficheros.

[2] Sitio web oficial de METS. Disponible en: https://www.loc.gov/standards/mets/METSOverview_spa.html

- Comportamiento. Se puede utilizar para asociar determinados comportamientos con los contenidos del fichero mets. Por ejemplo, ejecutar alguna aplicación sobre el contenido de un elemento determinado".

A continuación, se muestra un ejemplo de un documento METS:

```
<dmdSec ID="dmd002">
    <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core">
     <dc:title>Tityre tu patulae recubans sub tegmine fagi </dc:title>
     <dc:creator>Virgili Maro´, Publi, 70-19 aC.</dc:creator>
     <dc:title.alternative>Opera</dc:title.alternative>
     <dc:date.created>[ca. 1450] ;</dc:date.created>
    </mdWrap>
</dmdSec>
```

Figura 5.- Metadatos descriptivos METS. Fuente: Barrueco, J. M. (2011) *Introducción a los metadatos para las colecciones digitales*. Universidad de Valencia. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf

```
<amdSec ID="AMD001">
    <mdWrap MIMETYPE="text/xml" MDTYPE="MIX" LABEL="MIX">
            <mix:CameraCaptureSettings>
            <mix:ImageData>
               <mix:fNumber>14</mix:fNumber>
               <mix:exposureTime>1/3.3 seconds</mix:exposureTime>
               <mix:exposureProgram>3</mix:exposureProgram>
               <mix:isoSpeedRatings>100</mix:isoSpeedRatings>
            </mix:ImageData>
            </mix:CameraCaptureSettings>
   </mdWrap>
</amdSec>
```

Figura 6.- Metadatos técnicos METS. Fuente: Barrueco, J. M. (2011) *Introducción a los metadatos para las colecciones digitales*. Universidad de Valencia. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf

```
<fileSec>
  <fileGrp ID="filegrp1" USE="Large images">
     <file ID="file11" MIMETYPE="image/jpg" SEQ="1">
       <FLocat LOCTYPE="URL">
       http://weblioteca.uv.es/europeana/ms/0768/uv_ms_0768_11.jpg
```

```
      </file>
      <file ID="file12" MIMETYPE="image/jpg" SEQ="2">
        <FLocat LOCTYPE="URL">
        http://weblioteca.uv.es/europeana/ms/0768/uv_ms_0768_21.jpg
      </file>
  </fileGrp>
  <fileGrp ID="filegrp2" USE="Small images">
      <file ID="files1" MIMETYPE="image/jpg" SEQ="1">
        <FLocat LOCTYPE="URL">
        http://weblioteca.uv.es/europeana/ms/0768/uv_ms_0768_1s.jpg
      </file>
      <file ID="files2" MIMETYPE="image/jpg" SEQ="2">
        <FLocat LOCTYPE="URL">
        http://weblioteca.uv.es/europeana/ms/0768/uv_ms_0768_2s.jpg
      </file>
  </fileGrp>
</fileSec>
```

Figura 7.- Ficheros METS. Fuente: Barrueco, J. M. (2011) *Introducción a los metadatos para las colecciones digitales. Universidad de Valencia*. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf

```
<structMap TYPE=''physical''>
    <div ORDER=''0'' TYPE=''book'' LABEL=''Tityre tu patulæ ...''
    DMDID=''dmd002''>
      <div ORDER=''1'' TYPE=''page'' LABEL=''Enq.''>
        <fptr FILEID=''file11''/>
        <fptr FILEID=''files1''/>
      </div>
      <div ORDER=''1'' TYPE=''page'' LABEL=''FG 1''>
        <fptr FILEID=''file12''/>
        <fptr FILEID=''files2''/>
      </div>
    </div>
</structMap>
```

Figura 8.- Mapa estructural METS . Fuente: Barrueco, J. M. (2011) *Introducción a los metadatos para las colecciones digitales. Universidad de Valencia*. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf

2.5.4.- LOM

- Learning Object Metadata (LOM), se trata de un estándar abierto de la IEEE (Institute of Electrical and Electronics Engineers) [IEEE 1484.12.1 2002 Standardfor Learning Object Metadata[3]] para la descripción de objetos de aprendizaje. “Se define como tal cualquier entidad, digital o no, que pueda ser usado con propósitos educativos. LOM está compuesto de una jerarquía de elementos. En un primer nivel existen nueve categorías, cada una de las cuales contiene subelementos. Estos pueden ser simples, es decir, almacenan datos directamente, o pueden ser contenedores si almacenan otros elementos. La semántica de un elemento viene determinada por su contexto, por el elemento padre o contenedor en la jerarquía y por otros elementos en el mismo contenedor. Por ejemplo, existen hasta seis elementos denominados Description, el significado de cada uno de ellos vendrá marcado por el elemento padre en el que se incluye cada uno de ellos: General. Description, Rights. Description, Classification, etc. Las nueve categorías principales son: General, Life Cycle, Meta-Metadata, Technical,

[3] IEEE 1484.12.1 2002 Standardfor Learning Object Metadata. Disponible en: https://standards.ieee.org/standard/1484_12_1-2002.html

Classification, Educational, Rights, Relation, Annotation". (Barrueco, 2011) En la siguiente figura (*véase Figura 9*) se muestra la estructura jerárquica del esquema LOM:

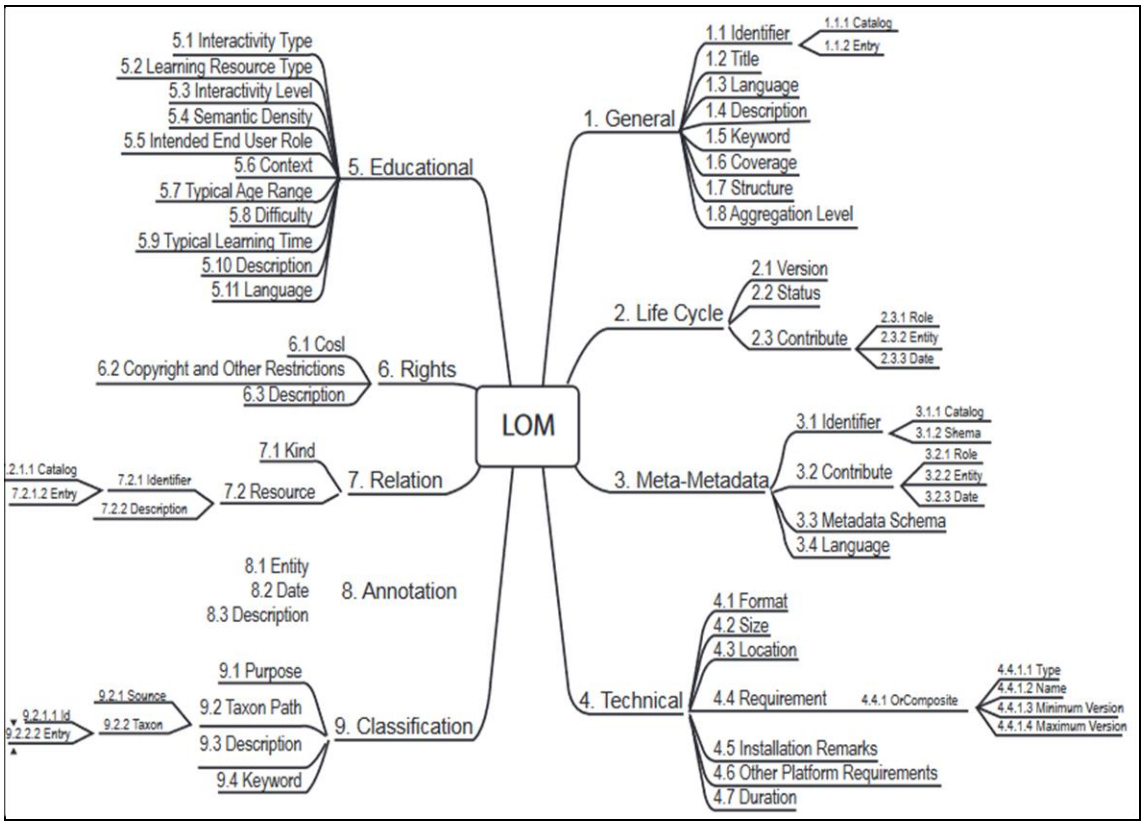

Figura 9.- Estructura jerárquica del esquema IEEE/LOM. Fuente: Barker, P. (2005) What is IEEE Learning Object Metadata / IMS Learning Resource Metadata? Cetis. Standards briefing series. JISC. Disponible en: http://www.dia.uniroma3.it/~sciarro/e-learning/WhatIsLOMscreen.pdf

2.6 Etiquetas meta

La gran mayoría de los metadatos están embebidos dentro de los ficheros de la Web. Esto acota la riqueza estructural de los metadatos aplicables, pero tiene la virtud de la sencillez. La World Wide Web ofrece un sistema que es añadir los metadatos en forma de etiquetas meta (o metatags) HTML y, con los editores actuales, cualquier usuario puede crear una página web y añadir las etiquetas meta fácilmente.

```
<meta name="Generator" content="Drupal 7 (http://drupal.org)" />
    <meta http-equiv="Content-Language" content="es">
  <meta name="Keywords" content="Castilla-La Mancha, administración, adm
  <meta name="Author" content="Gobierno de Castilla-La Mancha">
  <meta name="Owner" content="webmaster@jccm.es">
  <meta name="description" content="Web oficial del gobierno autonómico
```

Figura 10.- Ejemplo de etiquetas meta embebidas. Fuente: código fuente de https://www.castillalamancha.es/

"El origen de las etiquetas meta comienza con el lenguaje de marcado HTML en 1999, cuando se concibe un método para la descripción sencilla de páginas web. El lenguaje HTML era en principio un lenguaje semántico que se convirtió en lenguaje de formato, puesto que elementos como los encabezados indicaban no sólo la importancia de un

texto marcado con caracteres de un primer nivel, sino también la forma y tipografía de los títulos y encabezamientos.

Si queremos crear páginas web con propósito general y facilitar la búsqueda y recuperación de nuestros documentos, no es necesario utilizar un lenguaje muy estructurado ni un sistema muy sofisticado de metadatos, basta con utilizar las etiquetas meta que pueden ser incrustadas dentro del propio documento creado en lenguaje HTML. Caso bien distinto es si pretendemos crear un sistema de información bien estructurado pues, en este caso, sí es necesario utilizar otros sistemas de metadatos y otros lenguajes más elaborados y estructurados. Por este motivo, las etiquetas meta HTML tienen limitaciones importantes frente a lo que se consideran verdaderos esquemas de metadatos de aplicación profesional y documental, como el anteriormente explicado Dublin Core. Incluso los metadatos en sentido estricto y purista y las etiquetas meta han sido confundidos en múltiples ocasiones, debido a que muchos metadatos son introducidos en el código fuente de las páginas web, mediante esta solución". (Lamarca Lapuente, 2006)

Las etiquetas meta fueron muy importantes de cara al posicionamiento web en los inicios de los buscadores. Si creábamos una página web y queríamos tener presencia en Internet y un buen posicionamiento en buscadores o SEO, debíamos incluir información por medio de etiquetas meta para que los robots que sustentan a los buscadores o índices pudieran indizar nuestras páginas.

¿Qué es el SEO? "Se denomina SEO (Search Engine Optimization) u optimización para motores de búsqueda, al conjunto de técnicas encaminadas a subir posiciones en los buscadores de Internet por las palabras clave que nos interesen. Los principales buscadores de Internet ordenan los resultados mediante complejos algoritmos matemáticos que evalúan decenas de factores (entre ellos las etiquetas meta). Muchos de esos factores pueden ser intervenidos u optimizados". (Ramos, 2011)

"El acceso a una web a través de Internet es muy fácil de encontrar cuando se trata de webs de instituciones o empresas que poseen una dirección que coincide con el nombre de la institución o de la empresa. Sin embargo, la mayor parte de las páginas poseen una URL que nada tiene que ver con su nombre o contenido y más aún cuando se trata de páginas que ocupan un tercer o cuarto nivel. Para que las páginas pudiesen ser encontradas por otros usuarios, era preciso no sólo dar de alta la web en los principales buscadores, sino también introducir datos o texto descriptivo en las páginas a través de etiquetas meta que indiquen información sobre el contenido, materia, autor, etc. para que esta información pudiera ser extraída de forma automática por los principales robots, bases de datos e indizadores automáticos de páginas web que existen en Internet". (Méndez Rodríguez, 2002)

Las etiquetas meta 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 tabla (*véase Tabla V*):

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'/>
	<meta http-equiv='pragma' content='no-cache'/>

Técnica – Control de cache en página web	<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="84-8273-674-2">
	<meta name="author" content="Nombre Apellidos">
	<meta name="copyright" content="© 2019 Autor">
	<meta name="keywords" content="palabras, clave">
	<meta name="date" content="2019-06T18:29:57+00:00">
	<meta name="description" content="descripción del recurso">

Tabla V.- Estructura sintáctica de las etiquetas meta. Fuente: elaboración propia

Las especificaciones oficiales de las etiquetas meta 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 o 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 etiqueta meta indica unos patrones de edición y comportamiento para el funcionamiento de la web. Por otro lado, también se contemplan etiqueta meta de tipo descriptivo del sitio web. Éstas son la mínima aportación semántica que HTML, desde su creación, proporciona a los buscadores web. Describen el documento con una serie de datos básico: autor, descripción, palabras clave, codificación de caracteres… (W3C Consortium, 1999)

Dichas etiquetas meta son las analizadas para este Trabajo Fin de Máster.

Las etiquetas meta name convenidas internacionalmente y más normalizadas son "*author*, *copyright*, *description* y *keywords*", aunque es posible encontrar otras carentes de normalización, incluso especiales, diseñadas ad-hoc. En general (*véase la tabla V*) siempre tienen la misma forma:

<meta name=" " content=" " >

donde *name* y *content* son los nombres de los atributos de dicha etiqueta meta.

Así, "la semántica del tag <meta> no define ninguna interpretación específica para los atributos *name* y *content*. Esa información es interpretada por las aplicaciones que leen las páginas, de acuerdo con una semántica que se estableció en base al uso. Un tag <meta> con name='keywords' es interpretado por los robots de los motores de búsqueda indicando que el campo *content* listará una serie de palabras clave que estarán asociadas con el contenido de la página. Para tales convenciones, como la de que *keywords* significa un conjunto de palabras clave, es necesario que sean utilizados vocabularios de referencia, cuyo significado sea comprendido por quien publica la página, de manera tal que la aplicación que la consume pueda conseguir obtener un significado entendido equivalente al significado pretendido. Como se dijo anteriormente, esta forma de incluir metadatos posee una expresividad bastante limitada, pero se hizo sumamente extensiva a partir de la creación de estándares y/o esquemas para la definición de metadatos". (Ceweb.br, 2008)

2.6.1.- Etiqueta meta description

Hace años la etiqueta meta descripción influía enormemente en el posicionamiento en buscadores, y aunque esto ya no es así, sigue teniendo importancia. El contenido que aquí pongamos será la que veamos en el pequeño extracto bajo el título en los resultados de búsqueda, denominado snippet. Esta etiqueta ofrece al robot la información general sobre la página: tema, tipo de información, institución responsable, etc. Esta etiqueta meta es importante ya que es tu tarjeta de presentación en la hoja de resultados de un buscador y no debe contener más de 155 caracteres. Se puede acompañar esta etiqueta meta con el nombre de atributo *lang*, de esta manera es posible ofrecer diferentes versiones en diferentes lenguajes en la descripción de una misma página.

```
<meta name="description" content="Esta es la descripción general de página">
```

2.6.2.- Etiqueta meta keywords

Las keywords son las palabras clave o descriptores del contenido de la página. Se puede utilizar un lenguaje libre, pero si se quiere ser más rigurosos en la descripción, deberemos utilizar una lista de materias o, incluso, un tesauro. Actualmente, esta etiqueta carece de importancia los grandes buscadores. Sin embargo, se sigue implementando para aquellos navegadores minoritarios que sigue haciendo uso de él. Se aconseja no usar más de cinco o seis palabras clave.

```
<meta name="keywords" content="Palabras clave separadas por comas">
```

2.6.3.- Otras etiquetas meta

“Otros dos meta name menos importantes son autor y copyright, de uso opcional desde el punto legal, permiten hacer referencia al diseñador de una página web y al propietario de los derechos del código fuente de una página HTML. En algunos sistemas de gestión de contenido (CMS), las etiquetas de autor se generan automáticamente y en ellas se nombra a la persona que ha trabajado la página en último lugar. De esta forma, los administradores pueden extraer del directorio quién gestiona cada página”. (IONOS, 2019)

El principal inconveniente que encontramos en el uso de esta etiqueta es que la sintaxis para describir los elementos de los metadatos es excesivamente flexible. Así, por ejemplo, es posible usar diferentes términos para expresar el mismo concepto:

```
<meta name="autor" content="Goytisolo, Juan">
<meta name="writer" content="Goytisolo, Juan">
```

Ambas opciones son totalmente válidas, pero a la hora de alimentar una base de datos, el hecho de que existan campos que se denominen igual, aunque tengan el mismo tipo de información, no deja de ser un gran problema. Al mismo tiempo, no existe ninguna norma con respeto al orden que debe asignarse a los elementos dentro de un atributo. Es decir que “Goytisolo, Juan” se considera igual de válido que “Juan Goytisolo”. La dificultad, a la hora de recuperar el documento, resulta evidente.

2.7 Herramientas para metadatos

Méndez y Senso, 2004, aseveran que “cada vez existe más software para la creación, implementación, gestión y extracción de metadatos. Estas herramientas, de índole, objetivos y procedencias diversas, podemos tipificarlas genéricamente como: plantillas que funcionan en modo servidor y aplicaciones cliente, además de distinguir, por supuesto, software libre, de fuente abierta o comercial”.

A continuación, los autores señalan algunos ejemplos de estas herramientas clasificadas en función del uso al que se destinen (creación de etiquetas meta, creación de metadatos propiamente dichos, extracción de metadatos y visor de metadatos):

- Plantillas para la creación de etiquetas meta para incluir en la fuente de un documento electrónico HTML/XHML:

Generator Meta Tags For Your Site: http://www.submitshop.com/metatags/metatags.html
MGAWEB'S Metatag Generator: http://www.mgaweb.com/submit/hhess/metatags.asp

- Plantilla para la creación de metadatos (siguiendo el esquema Dublin Core):

DC-DOT: http://www.ukoln.ac.uk/metadata/dcdot

- Software para la extracción de metadatos:

Foca: https://www.elevenpaths.com/es/labstools/foca-2/index.html

FOCA (Fingerprinting Organizations with Collected Archives) es una herramienta usada para localizar metadatos e información oculta en los documentos que examina. Estos documentos pueden estar en páginas web, y con FOCA se pueden descargar y analizar. Los documentos que es capaz de analizar son muy variados, siendo los más comunes los archivos de Microsoft Office, Open Office o ficheros PDF entre otros.

MediaInfo: https://mediaarea.net/en/MediaInfo

Otro excelente programa para obtener los metadatos de archivos de audio, imagen y video. Una vez obtenidos, podemos guardarlos en archivos de texto, CVS, HTML entre otros. Lo más resaltante de esta alternativa es que con ella podemos analizar toda una carpeta completa de archivos para extraer sus datos, mismos que podemos ver de diferentes formas antes de exportarlos.

Metadata Extraction Tool: https://sourceforge.net/projects/meta-extractor/

Una de las mejores herramientas gratuitas que también toma los datos de las imágenes, audios y videos que tengamos en nuestro ordenador. La información extraída puede ser guardada en un archivo XML o simplemente ser visualizada.

A diferencia de las herramientas anteriores, con este programa se puede extraer datos de archivos del paquete Office, lenguaje de marcado y de Internet como ARC, entre otros.

- Visores online de metadatos en imágenes. Estas herramientas nos permiten la consulta de los datos EXIF (Exchangeable Image File Format), un estándar que almacena metadatos de las fotos hechas con cámaras digitales. Los datos EXIF contienen información relativa a la imagen y a cómo ha sido tomada. Son, por ejemplo, datos relativos a: cámara utilizada, objetivo, enfoque, zoom, fecha, hora, lugar, etc. Estos datos se incrustan en el fichero de imagen, en formato JPG o RAW, por lo que no tenemos un fichero independiente que nos permita acceder a esta información. Aunque siempre podemos verla directamente desde la ventana de propiedades de una foto, existen muchas otras aplicaciones especialmente diseñadas para permitirnos consultar este tipo de información muy fácilmente, aplicaciones como las siguientes:

Jeffrey's Exif Viewer: http://exif.regex.info/exif.cgi
Metapicz: http://metapicz.com/#landing
Online Exif Viewer: http://exif-viewer.com/
Online photo EXIF metadata reader: http://www.findexif.com/#results
EXIF Viewer: http://www.prodraw.net/online-tool/exif-viewer.php
Exif Data: http://exifdata.com/
Camera Summary: https://camerasummary.com/

2.8 Crawler

"Los crawlers (o arañas) son programas software capaces de recorrer la Web automáticamente, recopilando las páginas accedidas para construir un índice que permita búsquedas por palabra clave sobre su contenido. Los crawlers convencionales se inicializan con un conjunto *semilla* de páginas de entrada y obtienen nuevas páginas mediante la navegación automática a las páginas enlazadas desde el conjunto semilla. Las nuevas páginas son añadidas al conjunto semilla y el proceso se repite hasta que o bien no hay más páginas que examinar o bien se sobrepasa algún límite definido por el administrador". (Álvarez Díaz, 2007)

Para Alonso Berrocal et al. (2006) "el procedimiento básico de un crawler consiste en suministrar una URL inicial o un conjunto de ellas, obtener la página web correspondiente y a continuación extraer todos los enlaces existentes en dicha página. A continuación, será necesario comprobar las URL que se habían seguido previamente y, en caso de no haberlos recorrido, introducirlos en una cola de URL a seguir. Después, normalmente, almacenamos la información, bien en bases de datos o en estructuras de ficheros con codificación ASCII. Finalmente, obtenemos la URL del siguiente enlace a seguir y comienza de nuevo el proceso".

En la siguiente figura se aprecia la arquitectura básica de funcionamiento de un crawler, denominado Multithreaded downloader:

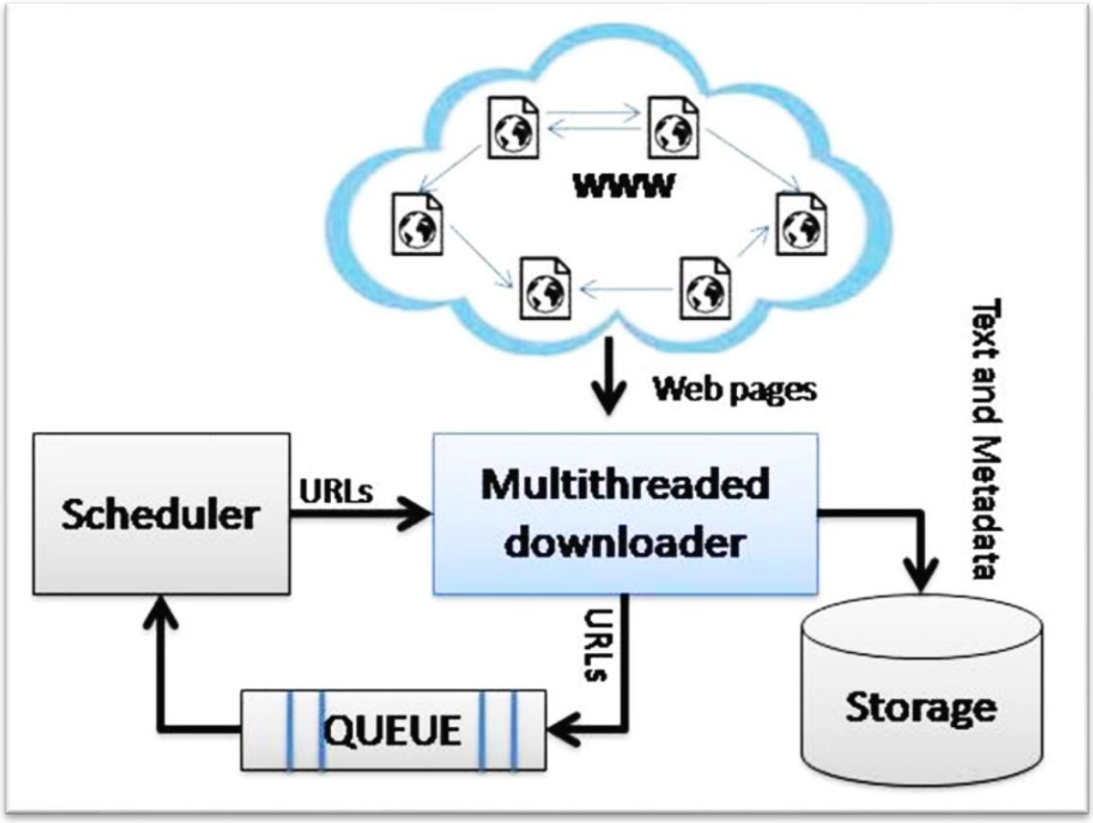

Figura 11.- Arquitectura típica de un web crawler. Fuente: Iqbal et al. (2015). Information Retrieval Process on the Web: A survey on Web Crawler Types & Algorithms. Disponible en: https://www.researchgate.net/publication/282245933_Information_Retrieval_Process_on_the_Web_A_survey_on_Web_Crawler_Types_Algorithms

Se observa (*véase Figura 11*) cómo se realiza dicho proceso: un gestor de descargas denominado *Multithreaded downloader* examina el contenido de una website, creando un documento con los metadatos y almacenando el contenido en un repositorio. A su vez, se busca en la citada website más enlaces, los cuales son enviados a una cola de espera para su posterior procesamiento. Por otra parte, existe un módulo denominado *Scheduler*, que se ocupa de tomar los enlaces de la cola de espera para enviarlos al programador y realizar con él un barrido de segundo nivel.

"Los crawlers más conocidos son UbiCrawler, Viuva Negra, y el módulo de rastreo distribuido de Google, además de otros de naturaleza comercial previos a Google (Altavista, Infoseek, Lycos, Excite y HotBot). En cuanto a crawlers de código abierto, se destacan Heritrix, Nutch, Combine y WIRE". (Camargo y Ordoñez, 2013)

2.8.1.- Evolución histórica del crawler

Se recuerda que "los crawlers o rastreadores son utilizados especialmente por los motores de búsqueda. A lo largo de estos años, lo que siempre han intentado es proporcionar la información más correcta y adecuada a la búsqueda que realiza el usuario por un término concreto. Para conseguir esto, cada motor de búsqueda ha realizado sus propios estudios para la aplicación de diferentes algoritmos de recuperación de información, de indexación de páginas, de lecturas de contenido, etc. Debido al gran número de páginas existentes, y, sobre todo, al crecimiento constante de las mismas, los rastreadores procuran evolucionar y que sus rastreos sean cada vez más rápidos, la información que recoja sea comprimida lo máximo posible y las

estructuras de datos sean elegidas con buen criterio, para que el espacio sea utilizado de forma correcta y eficiente". Benítez Andrades (2010)

Para Camargo y Ordóñez, 2013; "el primer crawler oficialmente reconocido es el *Wanderer* de Matthew Gray; fue un algoritmo de rastreo simple en la Web desarrollado en 1993. Fue presentado como un rastreador para el MIT, con el único propósito de generar estadísticas, y no fue publicado o expuesto a la comunidad científica. Posteriormente aparecieron otros cinco en un periodo de dos años: JumpStation, RBSE, WebCrawler, WWWWorm y MOMspider. Dichos crawlers han dado origen a la mayoría de los actuales. Sin embargo, el algoritmo de referencia obligatoria fue PageRank, propuesto por Lawrence Page y Sergei Brin; en él se expone cuál debería ser la estructura de un motor de búsqueda Web, incluyendo crawler, indexación y búsqueda".

Posteriormente, en otro artículo (Cho et al., 1998) -- resultado de un proyecto de investigación de la Universidad de Stanford -- los autores describen públicamente este algoritmo que sería el fundamento de Google durante sus primeros diez años. En este documento se "definen métricas importantes en el rastreo como similitud de página (medida para descartar páginas iguales en la Web), *conteo regresivo* (para aumentar el peso a una página entre más referenciada esté en otras páginas), PageRank (ponderar mejor las páginas referenciadas por portales importantes), completitud de la información y ubicación (según el lugar del código en el que se encuentre un resultado). Igualmente presentan un algoritmo simple de rastreo basado en estos criterios".

"En el 2002, se presentó un crawler llamado WebRACE, con capacidades de procesamiento distribuido, almacenamiento temporal de objetos y servicio de filtrado. Para ello, se utilizó un motor desarrollado por la Universidad de California, llamado eRACE, capaz de recolectar, anotar y diseminar información de fuentes heterogéneas. Otra propuesta que merece mencionarse es Ubicrawler, un crawler distribuido, programado en Java con todas las funciones descentralizadas. Ya que su algoritmo es distribuido, los autores informan que en cada CPU en la cual se ejecute el crawler se pueden procesar hasta 660 páginas por segundo". (Boldi et al., 2004)

2.8.2.- Wget

GNU Wget[4] (al que para abreviar se llamará Wget), es, según indica el manual oficial, "una utilidad de línea de comandos (carece de interfaz) que no sólo recorre la web, sino que también realiza la descarga no-interactiva de archivos de Internet" (GNU Wget, 1996).

Lara Carrascal, 2008 entiende por descarga no-interactiva al hecho de que "el programa puede funcionar sin la presencia del usuario, trabajando en segundo plano, cancelando la descarga cuando el sistema se apague y pudiendo reanudarla en el próximo inicio".

"Wget es una herramienta libre que permite la descarga de contenidos desde servidores web de una forma simple. Su nombre deriva de World Wide Web (W), y de obtener (en inglés get), lo que podría traducirse como *obtener desde la WWW*. Fue escrito originalmente por Hrvoje Nikšić y por ser un proyecto de software libre tiene una gran cantidad de colaboradores directos e indirectos. Su primera versión se lanzó en 1996, coincidiendo con el boom de popularidad de la web. Es un programa utilizado a través de línea de comandos, principalmente en sistemas tipo UNIX, especialmente en GNU/Linux. Escrito en el lenguaje de programación C, Wget puede ser fácilmente instalado en sistemas derivados de UNIX, y ha sido portado a muchas interfaces gráficas de usuario (GUI) y aplicaciones gráficas de descarga como Gwget2 para GNOME, wGetGUI3 y VisualWget4 para Microsoft Windows, Wget 1.10.2r25 para Mac OS X. Actualmente soporta descargas mediante los protocolos HTTP, HTTPS y FTP. Entre las características más destacadas que ofrece Wget está la posibilidad de fácil descarga de mirrors (repositorios) complejos de forma recursiva, conversión de enlaces para la

[4] **Sitio web de GNU Wget**: https://www.gnu.org/software/wget

visualización de contenidos HTML localmente, soporte para proxies, etc. Wget ha sido diseñado para la robustez en conexiones de red lentas o inestables. Si una descarga no se completa debido a un problema en la red, Wget automáticamente tratará de seguir descargado desde donde acabó, y repetir el proceso hasta que el archivo completo haya sido recuperado. Fue uno de los primeros clientes que hizo uso de la entonces nueva cabecera HTTP Range, para ofrecer esta característica". (Llamaret Heredia, 2019)

"Wget opcionalmente puede trabajar como una araña web extrayendo recursos enlaces de las páginas web HTML y descargarlas en la secuencia, repitiendo el proceso recursivamente hasta que todas las páginas hayan sido descargadas o hasta que haya sido alcanzada una profundidad de repetición máxima especificada por el usuario. Las páginas web descargadas son guardadas en una estructura de directorio que se parece a un servidor remoto. Esta descarga recursiva permite hacer una copia exacta de cualquier sistema de archivos parcial o completamente de un sitio web vía HTTP. Los enlaces de sitios web HTML descargados pueden ser ajustados para indicar el punto o zona de descarga. Cuando se realiza esta clase de copia exacta del sistema de archivos (en inglés mirroring) automática de un sitio web, Wget es compatible con el estándar de exclusión de robots. En resumen, no sólo puede descargar archivos, sino que también puede descargar páginas web en formato HTML y XHTML, recreando la estructura, de forma local, del directorio ubicado en el servidor, respetando el archivo *robots.txt* y convirtiendo los enlaces para poder visualizarlas cuando no se está conectado a internet". (Kerrisk, 2010)

"Es la utilidad preferida de la mayoría de programas de GNU/Linux que requieren de conectarse a internet para la descarga de archivos ya que se adapta a la velocidad de conexión y al consumo de recursos del sistema de una manera asombrosa, y su consumo de CPU y memoria es simplemente ridículo, si lo comparamos con los gestores de descargas en modo gráfico que existen en GNU/Linux, algunos de ellos meras interfaces gráficas de uso de Wget". (Lara Carrascal, 2008)

2.9 Expresiones regulares

"Es bueno conocer un poco de historia sobre la primera aparición de expresiones regulares en el campo de la computación las cuales surgen dentro de un editor de líneas llamado *Ed*, creado por el ingeniero Ken Thompson en los laboratorios de Bell Labs en 1969. Aquel editor de líneas apareció por primera vez como una utilidad para el sistema operativo de UNIX y era el encargado de suplir las labores para editar archivos. En el comando utilizado por *Ed* se colocaba g(global) al inicio y p(print) al final, ambas letras son conocidas como modificadores y lo que queda en la parte central es la combinación de caracteres que conforman la expresión regular. En este punto surge algo muy interesante ya que la instrucción anterior da como resultado la frase *Global Regular Expression Print* cuya abreviatura conforma la palabra grep, a partir de esta característica de Ed nace el comando grep (comando presente en todos los sistemas operativos basados en UNIX) el cual nos permite realizar búsquedas a nivel del sistema de archivos. Actualmente muchos lenguajes de programación modernos cuentan con el soporte necesario para el uso de expresiones regulares, aunque la sintaxis puede cambiar para cada lenguaje, en esencia podremos hacer soluciones similares para cada uno". (Méndez Ortega, 2018)

La definición de expresión regular, también conocida como regex o regexp, es, "en la ciencia de la computación y la teoría de lenguaje formal, una secuencia de caracteres que forma un patrón de búsqueda, principalmente utilizada para la localización de cadenas de caracteres, operaciones de sustituciones u otras modificaciones. Las expresiones regulares son una buena herramienta para cualquier desarrollador o programador. Su potencial es infinito ya que nos pueden resultar útil para la validación

de campos o el parseo de textos a la hora de rastrear web. También resultan muy útiles la hora de editar grandes cantidades de líneas de código, análisis de logs y cientos de utilidades más. Su utilidad principal es describir el conjunto de cadenas que componen un string, especialmente útil en editores de texto e IDEs (Integrated Development Environment o entornos de desarrollo integrado) para búsqueda y manipulación de textos". (Alonso Vega, 2017)

En resumen, "una expresión regular en programación es una forma de representar una secuencia de caracteres que describe un conjunto de cadenas sin enumerar sus elementos y haciendo uso de la sintaxis propia de los diferentes lenguajes de programación. Por ejemplo, el grupo formado por las cadenas *color* y *colour* se puede describir mediante la expresión regular *col(o|ou)r*. El uso de expresiones regulares constituye un mecanismo flexible y eficiente para el procesamiento de textos que requiere el conocimiento y destreza en un lenguaje de programación relativamente sencillo que sirve para hacer búsquedas en un texto determinado o en un corpus lingüístico. Con el uso de algunas herramientas informáticas como motores de búsqueda, las expresiones regulares pueden utilizarse para añadir, aislar o quitar textos o datos de forma rápida y ágil". (Chacón Beltrán, 2008)

"Las expresiones regulares se componen de dos tipos de caracteres, los llamados metacaracteres que son de tipo especial, por ejemplo, * (cuyo significado es: repetición de 0 ó más veces el carácter o subexpresión previos) y los llamados literales o caracteres normales de texto. Para ilustrar esta organización podría considerarse que las expresiones regulares son una lengua con texto literal, que serían las palabras, y metacaracteres que serían la gramática. Las palabras se combinan con la gramática de acuerdo con un conjunto de reglas para crear una expresión que comunica una idea". (Friedl, 2006). Una muestra de las expresiones más sencillas sería (*véase Tabla VI*):

Expresiones simples (patrones de un único símbolo)	
Expresión	**Significado**
x	Caracter x, si es carácter normal
.	Cualquier caracter
[aeiou]	Un caracter del conjunto
[a-z]	Un caracter del rango
[^aeiou-9]	Complementa el conjunto
^	Principio del texto, si va al comienzo
$	Fin del texto, si va al final
Expresiones compuestas (combinaciones de patrones simples)	
xy	Expresión x seguida de y
x+	Una o más repeticiones de x
x*	Cero o más repeticiones de x
x?	Cero o una aparición de x
una|otra	Una u otra expresión
(x)	Expresión x
Ejemplos de expresiones regulares	
[0-9]	Un dígito numérico, decimal
[AEIOUaeiou]	Una vocal
[A-Za-z]	Una letra, mayúscula o minúscula
[A-Za-z][A-Za-z0-9]*	Una palabra que empieza por letra y puede contener números
^[+\-0-9][0-9]*\.[0-9]*	Un número, con punto decimal, al comienzo del texto
(Lu|Ma|Mi|Ju|Vi|Sa|Do)	Abreviatura del nombre de un dia de la semana.

Tabla VI.- Ejemplos de expresiones regulares principales. Fuente: Collado, M. (2013). Lenguaje AWK. Material de estudio de Entornos de programación. Ingeniería informática. UPM. Disponible en: http://lml.ls.fi.upm.es/ep/awk.html

2.9.1.- SED

El programa **SED**[5] (Stream EDitor) según el tutorial de Americati (2007) "es considerado un editor de texto orientado a *flujo* (en contraposición a los clásicos editores interactivos) el cual acepta como entrada un archivo o la entrada estándar; cada línea es procesada y el resultado es enviado a la salida estándar. Su utilidad más obvia es la de describir un conjunto de cadenas para una determinada función, resultando de utilidad en editores de texto y otras aplicaciones informáticas para buscar y manipular textos". SED toma las líneas de una en una, les aplica la transformación o sustitución que le indiquemos y nos las devolverá modificadas. La sintaxis utilizada por SED puede resultar algo oscura al principio, pero un mínimo conocimiento de este comando nos permitirá hacer modificaciones en el texto que de otro modo serían muy complejas. La principal característica de SED es que acepta expresiones regulares como patrones y con ellas puede realizar prácticamente cualquier modificación que podamos imaginar en las diferentes líneas de nuestros ficheros. Por ejemplo, si se quiere cambiar todas las etiquetas <h1> </h1> por <h2> </h2>; SED, apoyándose en las expresiones regulares, es capaz de facilitarnos la labor. A continuación, se expone el esquema del funcionamiento de SED (*véase Figura 12*):

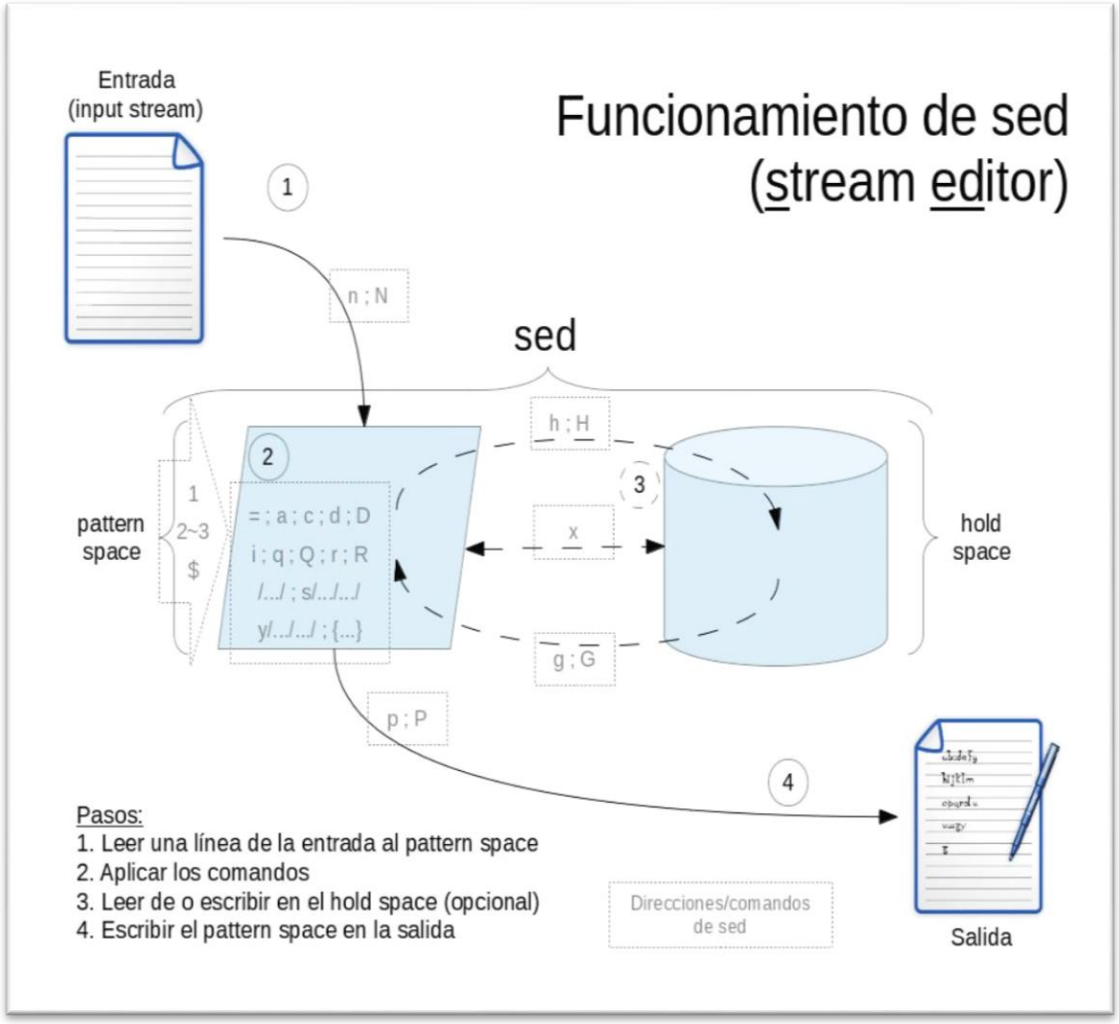

Figura 12.- Esquema del funcionamiento de SED. Fuente: Gmelendez (2013). Wikimedia Commons. Disponible en: https://commons.wikimedia.org/wiki/File:SedDiagram.jpg

[5] **Sitio web de GNU SED**. Disponible en: https://www.gnu.org/software/sed/

3. METODOLOGÍA

3.1 Selección de webs

Para la realización del estudio de la aplicación de metadatos en webs institucionales se seleccionaron, capturaron y analizaron datos de los siguientes 19 sitios. (Los portales oficiales de las 17 Comunidades Autónomas más las 2 Ciudades Autónomas) a fecha de 18 de junio de 2018:

COMUNIDAD AUTÓNOMA	URL
Andalucía	http://www.juntadeandalucia.es
Aragón	http://www.aragon.es
Asturias	http://www.asturias.es
Canarias	http://www.gobiernodecanarias.org
Cantabria	http://www.cantabria.es
Castilla-La Mancha	http://www.castillalamancha.es
Castilla y León	http://www.jcyl.es
Cataluña	http://web.gencat.cat
Comunidad Valenciana	http://www.gva.es
Extremadura	http://www.juntaex.es
Galicia	http://www.xunta.gal
Islas Baleares	http://www.caib.es
La Rioja	http://www.larioja.org
Madrid	http://www.madrid.org
Navarra	http://www.navarra.es
País Vasco	http://www.euskadi.eus
Región de Murcia	http://www.carm.es
Ciudad Autónoma de Ceuta	http://www.ceuta.es
Ciudad Autónoma de Melilla	http://www.melilla.es

Tabla VII.- Sitios web analizados. Fuente: elaboración propia

La estrategia utilizada para la realización del caso estudiado fue el siguiente:

3.2 Descarga de sitios web

Por su facilidad de uso, funcionamiento estable y capacidad de recursividad se decidió realizar la descarga mediante el programa **GNU Wget**. Como se mencionó anteriormente, este programa se considera un **crawler** y permite recuperar o duplicar (crear una copia exacta) de un sitio web completo de forma recursiva.

3.2.1.- Parámetros utilizados

Para realizar nuestra descarga se tuvieron en cuenta varias consideraciones:

- Se escogió realizar una descarga recursiva. Ésto se entiende por la descarga automatizada de todos los archivos vinculados a la página indicada, con el objetivo de conseguir que ésta funcione al usarla offline.

- Debido a la naturaleza del estudio la descarga consistió en la recopilación de todos los documentos HTML de cada sitio web. Esto fue debido a que no se entendió necesaria la descarga de otro tipo de archivos como imágenes, sonidos, vídeos, hojas de estilo, etc.

- Se aplicó hasta cinco niveles de profundidad. Esto es causado al posible enorme tamaño que tiene cada sitio web analizado. Se decretó que cinco niveles sería suficiente volumen de muestra para nuestro estudio.

- Se creó un directorio para cada sitio web con el nombre de la correspondiente Comunidad Autónoma. Dentro de dicho directorio se optó por no crear subdirectorios con cada uno de los niveles para que el manejo de los ficheros fuera más sencillo.

Para satisfacer todas estas demandas se realizó la descarga introduciendo comandos en la consola de CMD o Símbolo del sistema siguiendo este esquema utilizando una serie de parámetros que nos proporciona el programa Wget:

```
wget -r -l 5 -nd -A html,htm URLSitioWeb -P CarpetaDestino
```

-r : descarga recursiva
-l 5 : descarga a 5 niveles
-nd : no crear directorios
- A html,htm : descargar solamente los archivos del tipo HTM o HTM.
URLSitioWeb : la URL de cada sitio web a analizar
-P CarpetaDestino : crea una carpeta de destino con el nombre que deseemos

El resultado de estas descargas fueron la recopilación de 19 carpetas (una por cada sitio web descargado) con sus respectivos ficheros HTML recuperados hasta cinco niveles de profundidad. Obviamente, debido a la diferente organización de cada uno de los sitios web, las carpetas obtenidas tuvieron un tamaño y, por tanto, tiempo de descarga distintos.

3.3 Extracción de metadatos

Para la extracción de los metadatos de todos los ficheros recuperados se optó por el programa SED y el uso de expresiones regulares. En el caso analizado se aplicó la siguiente expresión regular para la obtención de todas las etiquetas cuya cadena de caracteres o patrón fuera tal que <meta name="" content="">:

```
sed -n "/<meta/{s/^.*<meta name=\"\(.*\)\".*content=\"\(.*\)\".*$/\1 \2/;p}" *.html
```

De esta forma se extrajeron todos los valores de cada uno de los atributos mencionados.

3.4 Procesamiento de datos

Debido al volumen de metadatos recolectados se optó por crear una **base de datos** cuya extensión de archivo fue .mdb. Dicha base de datos se gestionó a través del programa **Microsoft Access 2013** debido a su interfaz amigable y a su uso intuitivo siguiendo el siguiente esquema:

Columna 1.- **Comunidad**: Comunidad Autónoma del sitio web al que pertenece.
Columna 2.- **Documento**: Fichero .html o .htm del que se han recuperado los metadatos.
Columna 3.- **Name**: Registro del valor entrecomillado recuperado del identificador *name=*
Columna 4.- **Contenido:** Registro del valor entrecomillado recuperado del identificador *content=*

Metadatos

Comunidad	Documento	Name	Contenido
Andalucia	banhap.htm	author	Consejería de Hacienda y Administración Pública
Andalucia	banhap.htm	description	Sitio web oficial de la Consejería de Hacienda y Admin
Andalucia	banhap.htm	keywords	hacienda, presupuesto, tributos, administracion, andalucia

Figura 13.- Ejemplo de esquema de la base datos. Fuente: elaboración propia

En un breve resumen, Access (parte de la suite Microsoft Office) permitió la creación y gestión de una base de datos. Se manejó datos de tipos texto de manera flexible, realizando las consultas oportunas y se pudo exportar datos a Microsoft Excel para la realización de cálculos, estadísticas, tablas y gráficos.

3.5.- Metadatos analizados

Se realizó una criba desechando las etiquetas meta name erróneas o no reconocibles. Asimismo, se diferenciaron versiones, mayúsculas y/o subatributos. Por ejemplo: *autor / Author" lang="es.* Finalmente, los metadatos a analizar (en total se hallaron 172788 registros) son los siguientes 87 agrupados en distintos tipos. Básicamente fueron halladas etiquetas meta HTML y, en menor medida, metadatos Dublin Core.

Meta-etiquetas web	
abstract	fechaPublicacion
author	Generator
Author" lang="es	keywords
Author" lang="eu	keywords" lang="es
Copyright	Keywords" lang="eu
Copyright" lang="es	keywords" xml:lang="es" lang="es
Copyright" lang="eu	lugar
Coverage.jurisdiction	municipio
Creator	news_keywords
date	nombreOrganismo
DEPARTAMENTO	organismoConvocante
Descripcion	organismoResponsable
description	Owner
description" itemprop="description	provincia
description" lang="es	TEMAS
description" lang="eu	tematica

description" xml:lang="es" lang="es	tematicaACUL
FECHA	Title
FECHA_HORA	TITULO
fechaOrdenacion	TituloGSA

Dublin Core	
DC.contributor	DC.relation
DC.Coverage.t.max	DC.Source
DC.Coverage.t.min	DC.subject
DC.creator	DC.title
dc.date	DC.type
DC.Date.created	DC.type" scheme="DCMITYPE
DC.Date.modified	DCTERMS.created" scheme="iso-8859-1
DC.Date.valid	DCTERMS.creator
DC.description	dcterms.description
DC.format	dcterms.format
DC.identifier	dcterms.identifier
DC.language	DCTERMS.language
DC.language" scheme="iso-8859-1	dcterms.subject
DC.Language" scheme="RFC1766	dcterms.title
DC.publisher	dcterms.type

Etiquetas orientadas a redes sociales		
Open Graph (Facebook)	**Twitter Cards**	
og:description	twitter:card	twitter:image
og:image	twitter:creator	twitter:image:src
og:title	twitter:description	twitter:site
og:type	twitter:dnt	twitter:title
og:url	twitter:domain	

Geo etiquetas
geo.placename
geo.position
geo.region

Tabla VIII.- Metadatos analizados. Fuente: elaboración propia

4. RESULTADOS Y ANÁLISIS

4.1 Comunidades Autónomas

ANDALUCÍA

http://www.juntadeandalucia.es

metadatos por archivo: 0,459

Nombre	#	Nombre	#
abstract	1	DC.type	259
author	39	DC.type" scheme="DCMITYPE	1
Creator	2	DCTERMS.created" scheme="iso-8859-1	1
DC.contributor	1	description	190
DC.creator	10	Generator	329
DC.Date.valid	1	geo.placename	3
DC.description	1	geo.position	3
DC.format	3	keywords	71
DC.identifier	25	ObjectType	2
DC.language	24	Title	448
DC.language" scheme="iso-8859-1	1	twitter:card	2
DC.publisher	24	twitter:description	1
DC.relation	4	twitter:domain	1
DC.subject	1	twitter:image	1
DC.title	24	twitter:title	2
		TOTAL	**1481**
Archivos analizados	*3328*	*Nº Names de metadatos*	*30*

Tabla IX.- Metadatos extraídos en Andalucía. Fuente: elaboración propia

El portal institucional de la Junta de Andalucía posee el número más alto de todo el país en tipos de metadatos (30) pero en un nivel tan bajo que su media de metadato por página es menor a 1. Quiere decir que es un portal web muy atomizado. La mayoría son diversos metatags que aparecen de forma dispersa por todo el sitio web y muchos de ellos apenas una o dos veces. Es de los pocos portales que posee etiquetas Dublin Core, pero en un nivel ínfimo (tan sólo destaca levemente *DC.type)*.

ARAGÓN

http://www.aragon.es

metadatos por archivo: 0,006

Nombre	#	Nombre	#
author	1	Generator	1
Description	2	Keywords	2
		TOTAL	**6**
Archivos analizados	*1070*	*Nº Names de metadatos*	*4*

Tabla X.- Metadatos extraídos en Aragón. Fuente: elaboración propia

La segunda Comunidad Autónoma que menos metadatos aplica, sólo 6.

ASTURIAS

http://www.asturias.es

metadatos por archivo: 0,876

Nombre	#	Nombre	#
description	2564	twitter:description	34
GENERATOR	1	twitter:image:src	34
Keywords	2564	twitter:site	34
twitter:card	34	twitter:title	34
twitter:creator	34		
		TOTAL	**6185**
Archivos analizados	*7064*	*Nº Names de metadatos*	*9*

Tabla XI.- Metadatos extraídos en Asturias. Fuente: elaboración propia

El caso asturiano destaca por el uso de dos metatags básicos de HTML: *description* y *Keywords*. Es reseñable la aparición (nivel ínfimo) de seis metatags de Twitter en 34 archivos analizados.

CANARIAS

http://www.gobiernodecanarias.org

metadatos por archivo: 0,933

Nombre	#	Nombre	#
creator	1	generator	2
date	1	keywords	70
dcterms.format	1	twitter:card	1
dcterms.identifier	1	twitter:image	1
dcterms.title	1	twitter:site	1
dcterms.type	1	twitter:title	1
description	71		
		TOTAL	**153**
Archivos analizados	*164*	*Nº Names de metadatos*	*32*

Tabla XII.- Metadatos extraídos en Canarias. Fuente: elaboración propia

El sitio institucional de Canarias es el más pequeño con amplia diferencia ya que sólo se han descargado 164 archivos. Casi en la mitad de ellos sean extraído los metatags básicos *description* y *keywords.*

CANTABRIA

http://www.cantabria.es

metadatos por archivo: 0,001

Nombre	#	Nombre	#
Generator	1		
		TOTAL	**1**

Nombre	#	Nombre	#
Archivos analizados	*1375*	*Nº Names de metadatos*	*1*

Tabla XIII.- Metadatos extraídos en Cantabria. Fuente: elaboración propia

El portal más pobre. Un meta name en todo el website.

CASTILLA-LA MANCHA

http://www.castillalamancha.es

metadatos por archivo: 5,201

Nombre	#	Nombre	#
Author	1248	twitter:creator	107
description	1248	twitter:description	102
Generator	1340	twitter:image	107
Keywords	1248	twitter:site	107
Owner	1248	twitter:title	107
twitter:card	107		
		TOTAL	**6969**
Archivos analizados	*1340*	*Nº Names de metadatos*	*11*

Tabla XIV.- Metadatos extraídos en Castilla-La Mancha. Fuente: elaboración propia

Interesante caso el del portal institucional de Castilla-La Mancha. Más del 90% del sitio al menos posee cinco metatags de HTML. Author, description, Generator (éste al 100%), Keywords, Owner. Además, le acompañan seis metatags de Twitter.

CASTILLA Y LEÓN

http://www.jcyl.es

metadatos por archivo: 10,390

Nombre	#	Nombre	#
author	4613	keywords	4613
DC.Coverage.t.max	2	lugar	20
DC.Coverage.t.min	4	municipio	2
DC.Date.created	2953	nombreOrganismo	1835
DC.Date.modified	447	organismoConvocante	250
DC.Description	6	organismoResponsable	349
DC.Source	2216	provincia	22
DC.Subject	15	tematica	741
dcterms.description	4586	tematicaACUL	4
dcterms.subject	4590	twitter:card	4605
description	4600	twitter:description	4592
fechaOrdenacion	2960	twitter:domain	4605
fechaPublicacion	254	twitter:image:src	133
		TOTAL	**49017**

Nombre	#	Nombre	#
Archivos analizados	*4718*	*Nº Names de metadatos*	*26*

Tabla XV.- Metadatos extraídos en Castilla y León. Fuente: elaboración propia

Castilla y León posee el portal con el mayor número de metadatos name extraídos con diferencia (casi 50.000) Incorpora en más del 90% del sitio los metatags típicos *autor, keywords y description*. La mayor novedad es que al mismo porcentaje también incorpora dos *dcterms***.** Es la única que lo hace a ese nivel. También destaca el número de metatags Twitter y dos metadatos Dublin Core (éstos sólo en, aproximadamente, el 50% del website).

CATALUÑA

http://web.gencat.cat

metadatos por archivo: 4,333

Nombre	#	Nombre	#
twitter:card	1716	twitter:site	1716
twitter:description	632	twitter:title	1716
twitter:image	1716		
		TOTAL	**7496**
Archivos analizados	*1730*	*Nº Names de metadatos*	*5*

Tabla XVI.- Metadatos extraídos en Cataluña. Fuente: elaboración propia

Destaca por utilizar solamente metadatos de Twitter. Prácticamente al 100%.

COMUNIDAD VALENCIANA

http://www.gva.es

metadatos por archivo: 1,936

Nombre	#	Nombre	#
author	2	keywords	2
description	7981	title	7979
generator	3		
		TOTAL	**15967**
Archivos analizados	*8248*	*Nº Names de metadatos*	*5*

Tabla XVII.- Metadatos extraídos en la Comunidad Valenciana. Fuente: elaboración propia

Más de 8000 archivos hacen de este el portal el mayor analizado. El portal de la Comunidad Valenciana está basado en dos metatags básicos HTML ya conocidos: *description* y *title*. Curioso que en este caso *keywords* no es usado.

EXTREMADURA

http://www.juntaex.es

metadatos por archivo: 1,998

Nombre	#	Nombre	#
description	1661	keywords	1661
		TOTAL	**4983**
Archivos analizados	*1663*	*Nº Names de metadatos*	*2*

Tabla XVIII.- Metadatos extraídos en Extremadura. Fuente: elaboración propia

El portal institucional de la Junta de Extremadura es el caso más nítido. Dos metatags básicos en más del 99% del website.

GALICIA

https://www.xunta.gal

metadatos por archivo: 0,510

Nombre	#	Nombre	#
description	167	keywords	5
		TOTAL	**172**
Archivos analizados	*337*	*Nº Names de metadatos*	*2*

Tabla XIX.- Metadatos extraídos en Galicia. Fuente: elaboración propia

Muy pocos archivos y metadatos extraídos.

ISLAS BALEARES

http://www.caib.es

metadatos por archivo: 0,713

Nombre	#	Nombre	#
author	9	Generator	960
creator	3	keywords	13
date	3	title	4
dcterms.format	2	twitter:card	2
dcterms.identifier	2	twitter:creator	1
dcterms.title	2	twitter:image	2
dcterms.type	2	twitter:site	2
description	252	twitter:title	2
		TOTAL	**1266**
Archivos analizados	*1776*	*Nº Names de metadatos*	*16*

Tabla XX.- Metadatos extraídos en Islas Baleares. Fuente: elaboración propia

Menos de un metadato por página en las que destaca sólo el conocido *generator.*

LA RIOJA

http://www.larioja.org/es

metadatos por archivo: 0,010

Nombre	#	Nombre	#
author	4	generator	7
description	8	keywords	3
description" itemprop="description	1		
		TOTAL	**23**
Archivos analizados	*2344*	*Nº Names de metadatos*	*5*

Tabla XXI.- Metadatos extraídos en La Rioja. Fuente: elaboración propia

Paupérrimo sitio web en metadatos.

MADRID

http://www.madrid.org

metadatos por archivo: 2,676

Nombre	#	Nombre	#
Author	2179	description	2579
Copyright	1	generator	173
DC.CREATOR	4	Keywords	2147
DC.TITLE	2	Owner	2106
Descripcion	36	TITLE	9
		TOTAL	**9236**
Archivos analizados	*3451*	*Nº Names de metadatos*	*10*

Tabla XXII.- Metadatos extraídos en Madrid. Fuente: elaboración propia

El portal insitucional de la Comunidad de Madrid tiene (aproximadamente al 60% del total) una serie de metatags básicos HTML: *Author, description, keywords, owner.*

NAVARRA

http://www.navarra.es

metadatos por archivo: 5,918

Nombre	#	Nombre	#
Author	21	og:image	1
copyright	4	og:title	1
DC.Creator	3	og:type	1
DC.Language" scheme="RFC1766	3	og:url	1
DEPARTAMENTO	82	Owner	3578
Description	3666	TEMAS	80
FECHA	3535	Title	17
FECHA_HORA	3525	TITULO	80
generator	9	twitter:card	3
geo.placename	1	twitter:creator	1
geo.position	1	twitter:description	2
geo.region	1	twitter:image	1

Nombre	#	Nombre	#
Keywords	3572	twitter:site	3
news_keywords	1	twitter:title	3
og:description	1		
			TOTAL **18197**
Archivos analizados	*3700*	*Nº Names de metadatos*	*29*

Tabla XXIII.- Metadatos extraídos en Navarra. Fuente: elaboración propia

Navarra posee el número más alto en variedades de metatags (29) y también un nivel muy alto de promedio de metadatos por página (casi 6). Destacan los clásicos metatags básicos de HTML.

PAÍS VASCO

http://www.euskadi.eus

metadatos por archivo: 7,770

Nombre	#	Nombre	#
author	3675	dc.subject	3117
Author" lang="es	113	dc.title	3430
Author" lang="eu	475	description	2
Copyright" lang="es	108	description" lang="es	131
Copyright" lang="eu	449	description" lang="eu	1630
dc.creator	3412	keywords	1
dc.date	3433	Keywords" lang="es	83
dc.description	3387	Keywords" lang="eu	346
dc.identifier	3433	twitter:dnt	2
dc.language	3433		
			TOTAL **30660**
Archivos analizados	*3946*	*Nº Names de metadatos*	*19*

Tabla XXIV.- Metadatos extraídos en el País Vasco. Fuente: elaboración propia

El sitio web insitucional del País Vasco es el más completo de los analizados ya que utiliza en gran parte del portal algunos de los 15 metadatos el esquema Dublin Core. Usa algunos de los típicos metatags HTML. Al igual que su vecino Navarra, llama la atención el uso de *lenguage* en este tipo de portales bilingües.

Estamos, por tanto, ante el único website con el esquema de metadatos más trabajado Siete etiquetas Dublin Core en tres cuartas partes de las páginas analizadas.

REGIÓN DE MURCIA

http://www.carm.es

metadatos por archivo: 2,890

Nombre	#	Nombre	#
description	10	keywords" xml:lang="es" lang="es	2464
description" xml:lang="es" lang="es	2464	Language	2475

Nombre	#	Nombre	#
keywords" lang="es	11		
		TOTAL	7424
Archivos analizados	*2563*	*Nº Names de metadatos*	*5*

Tabla XXV.- Metadatos extraídos en la Región de Murcia. Fuente: elaboración propia

Caso típico, pocos metatags pero utilizados en casi la totalidad del portal.

CIUDAD AUTÓNOMA DE CEUTA

http://www.ceuta.es

metadatos por archivo: 2,929

Nombre	#	Nombre	#
author	27	keywords	1786
description	1788	keywords" lang="es	1
generator	1824		
		TOTAL	5426
Archivos analizados	*1853*	*Nº Names de metadatos*	*5*

Tabla XXVI.- Metadatos extraídos en Ceuta. Fuente: elaboración propia

Al igual que en el caso de Murcia, se usan los típicos metatags básicos de HTML en casi el 100% del sitio web

CIUDAD AUTÓNOMA DE MELILLA

http://www.melilla.es

metadatos por archivo: 2,960

Nombre	#	Nombre	#
Author	2491	DCTERMS.title	1
DCTERMS.creator	1	Description	2491
DCTERMS.language	1	Keywords	2491
		TOTAL	7476
Archivos analizados	*2526*	*Nº Names de metadatos*	*6*

Tabla XXVII.- Metadatos extraídos en Melilla. Fuente: elaboración propia

Exactamente el mismo caso que la otra ciudad autónoma. Como curiosidad, Melilla cuenta con una página con DCTERMS.

4.2. Total

A continuación, se analizan el número total de archivos y metadatos analizados por CCAA:

Comunidad Autónoma	Nº total de metadatos	Archivos analizados	Metadatos / Archivos analizados	*Nº Names de metadatos*
Andalucía	1481	3328	0,459	*30*
Aragón	6	1070	0,006	*4*
Asturias	5333	6185	0,876	*9*
Canarias	153	164	0,933	*13*
Cantabria	1	1375	0,001	*1*
Castilla-La Mancha	6969	1340	5,201	*11*
Castilla y León	49017	4718	10,390	*26*
Cataluña	7496	1730	4,333	*5*
C. Valenciana	15967	8248	1,936	*5*
Extremadura	3322	1663	1,998	*2*
Galicia	172	337	0,510	*2*
Islas Baleares	1266	1776	0,713	*16*
La Rioja	23	2344	0,010	*5*
Madrid	9236	3451	2,676	*10*
Navarra	18197	3700	5,918	*27*
País Vasco	30660	3946	7,770	*19*
Región de Murcia	7424	2563	2,890	*5*
C. A. de Ceuta	5426	1853	2,929	*5*
C. A. de Melilla	7476	2526	2,960	*6*
TOTAL	***172788***	***53196***	***3,248***	

Tabla XXVIII.- Datos resumidos de metadatos por comunidad autónoma. Fuente: elaboración propia

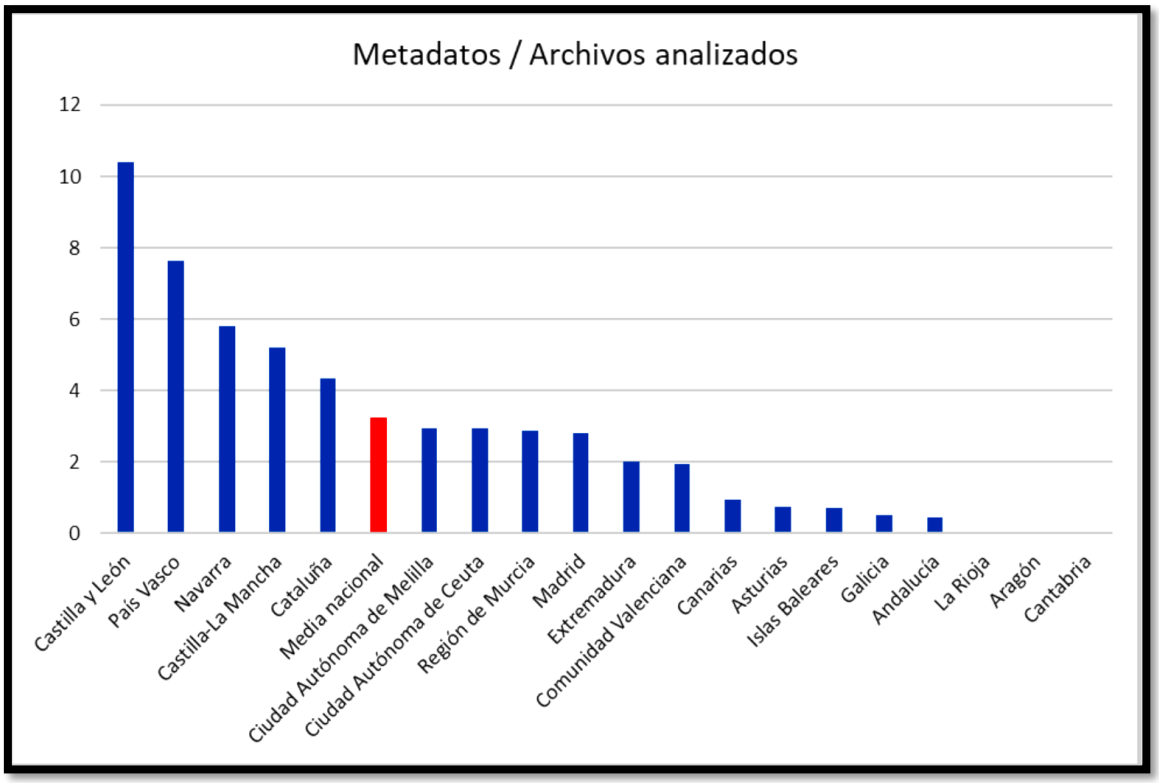

Figura 14.- Número de metadatos / Archivos analizados por CCAA. Fuente: elaboración propia

Nombre	Registros totales *(y porcentaje respecto al total)*	CCAA
description	29280 *(55 %)*	17
keywords	20249 *(38%)*	16
author	14309 *(29 %)*	12
Title	8457 *(16%)*	5
Owner	6932 *(13%)*	3
twitter:card	6470 *(12%)*	8
Twitter:descripion	5363 *(11%)*	6
Generator	4650 *(9%)*	12

Tabla XXVIII.- Nº de registros y nº de CCAA de los metadatos más usados. Fuente: elaboración propia

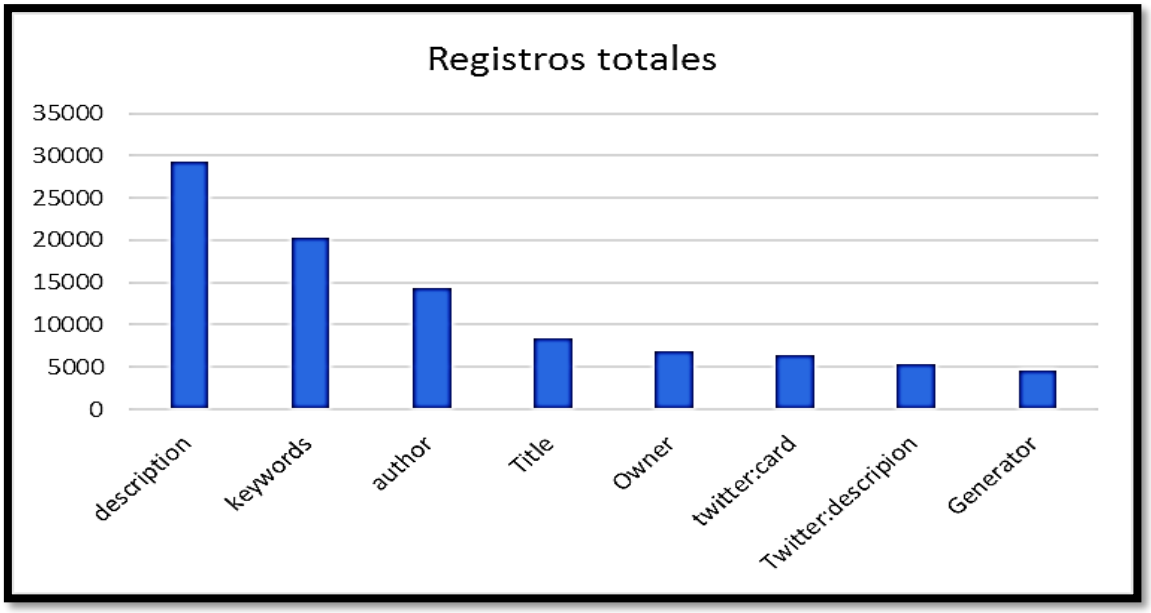

Figura 15.- Frecuencia de metadatos más usados por name. Fuente: elaboración propia

4.3. Análisis de uso

➢ **description**: Debería describir el contenido de la página en particular. El texto debe ser claro y conciso, no es recomendable usar frases genéricas.

29280 registros hallados (en el 55% de las páginas analizadas). 1181 vacíos o nulos (4%), prácticamente todos en Navarra.

En casi ningún caso es utilizado correctamente. Algunos portales utilizan el mismo contenido en todos los casos. Por ejemplo:

- En Canarias: “Gobierno de Canarias”

- En Castilla-La Mancha (incluidos errores en la codificación): “Web oficial del gobierno autonÃ³mico de Castilla-La Mancha con informaciÃ³n sobre actividad administrativa, economÃa, educaciÃ³n, sanidad, servicios sociales, sede electrÃ³nica...”

- En la Región de Murcia: “Sitio web oficial de la Comunidad Autónoma de la Región de Murcia. Bienvenidos.”

- En la Ciudad Autónoma de Ceuta: “Portal Oficial del Ayuntamiento de Ceuta - Boletines, Actas Plenarias, Presupuestos, Planes de la Ciudad, etc..”

- En la Ciudad Autónoma de Melila: “Ciudad Autónoma de Melilla”

- En el caso de la Comunidad Valencia se utiliza un único término que va variando.

➢ **keywords:** son palabras clave separadas por comas que han quedado obsoletas desde 2009 al no ser utilizado en el algoritmo de búsqueda de Google.

20249 registros (en el 38% de las páginas analizadas). 5670 vacíos o nulos (el 28%) pertenecientes a tres portales: Andalucía, Castilla y León y Navarra.

Las palabras clave contenidas en este metadato en algunos portales es siempre el mismo. Por ejemplo:

- En Castilla-La Mancha: “Castilla-La Mancha, administraciÃ³n, administraciÃ³n electrÃ³nica, economÃa, educaciÃ³n, sanidad, eGov, eAdmin”

- En Madrid: “Comunidad de Madrid”

- En la Ciudad Autónoma de Ceuta: “ceuta, ayuntamiento, ciudad autonoma, estrecho gibraltar, bocce, boletin, actas plenarias, real decreto, estatuto, autonomia, presupuestos, plan, sebta, septa, ceuti, caballa”

- En la Ciudad Autónoma de Melilla: “ciudad autónoma de melilla, melilla, 24 horas,”

En otros casos se utilizan unos keywords básicos acompañados en cada caso de otros variados y/o adecuados al documento.

- En la Región de Murcia (bajo el name keywords" lang="es y keywords" xml:lang="es" lang="es): “Comunidad Autónoma Región Murcia, Región Murcia, Murcia”

- En el caso de Asturias se aprecia que existen en la mayoría dos tipos diferenciados: “Trabajastur, empleo, trabajo, formacion, cursos, ofertas” y “Gobierno Principado Asturias, Instituto Asturiano de Administración Pública Adolfo Posada”

- En Navarra, además de en muchos casos estar vacío el content, no se utiliza el metadato adecuadamente, ya que en muchos se contiene una descripción del sitio.

Como en Murcia, la web del País Vasco se diferencia el metatag con las subetiquetas " lang="es coincidiendo en su mayoría estas keywords: “Sede electronica, oficina virtual, ayudas y subvenciones, mis gestiones, tramites administracion, tramites etxebide, tramites familia, tramites becas, asuntos sociales, pasarela pagos, calendario administracion, dias hábiles” y " lang="eu. coincidiendo “Egoitza elektronikoa, bulego birtuala, diru laguntzak, nire gestioak, tramiteak administrazioarekin, izapideak administrazioarekin, izapideak etxebide, izapideak familiak, izapideak bekak, gizarte gaiak, ordainketa online”

➢ **Dublin Core:** Tan sólo dos Comunidades lo utilizan en cierta cantidad: Castilla y León y el País Vasco. Y tan sólo en este último lo utiliza en más de dos tercios de la web. También Andalucía cuenta con algunos DC.Type.

En Castilla y León destaca el uso de los DC.Date.created y DC.Date.modified utilizando correctamente el formato AAAA-MM-DD. También es reseñable el uso de DC.Source en el que se señalan las distintas fuentes.

Por cantidad utilizada, el portal del País Vasco es el más completo de todos en lo referido a etiquetas bajo el esquema Dublin Core, aunque no todos correctamente usados. De los 15 elementos, son usados con asiduidad estos 7:

- dc.creator: mal utilizado. Contiene una cadena de dígitos ilegible. Ejemplo: “r01e00000ff26d46687a470b8a86a04237242b34f,r01e00000ff26d468d9a470b8b31b8fd84a0d5db3,r01epd0122e4ed314423e0db04c97a47b5baa317f,r01e”

- dc.date: no son mayoría, pero algunos campos están vacíos o el formato no es el correcto.

- dc.description: algunos contenidos vacíos. En los rellenados se utiliza principalmente el euskera.

- dc.identifier: secuencias de caracteres utilizados para identificar unívocamente un recurso.

- dc.language: utilizado correctamente bajo la RFC 3066 con las dos letras primarias (*eu* para el euskera y *es* para el español).

- dc.subject: erróneo. Al igual que el dc.creator son líneas de caracteres sin sentido.

- dc.title: bien utilizado. Destaca que sólo se expresan en euskera.

➢ **DCTERMS:** sólo los utiliza en cantidad el sitio web de Castilla y León de forma dispar:

- dcterms.description: los 4586 registros son utilizados correctamente.

- dcterms.subject: los 4590 registros están vacíos.

5. CONCLUSIONES

5.1 Conclusiones generales

Los metadatos son datos estructurados que describen un recurso de información. Son clasificados comúnmente en administrativos, los cuales facilitan la gestión administrativa de los recursos proporcionando información de preservación, técnica, de derecho de autor y de creación del recurso…; estructurales, que facilitan la navegación describiendo la estructura del contenido con el objetivo de relacionar un archivo con otro; y descriptivos, que contribuyen a la identificación del recurso.

En el contexto actual, el desarrollo de los metadatos, y su uso paulatino en los sistemas de recuperación de información, ha logrado incrementar el nivel de interpretación de contenidos en los recursos electrónicos en tiempos donde los dominios científicos son cada vez más complejos, debido a las relaciones interdisciplinares en constante evolución. Esto se ha hecho cada vez más evidente en el perfeccionamiento de la web semántica a través de las ontologías, donde los metadatos (y sus diferentes esquemas) han incidido en el significado correcto de términos o el conjunto de ellos. Dichos metadatos y esquemas son elementales para ratificar que los recursos perdurarán y permanecerán accesibles en el futuro, permitiendo ser interpretados por humanos y por ordenadores, fomentando la interoperabilidad para el intercambio de datos entre las distintas plataformas, estructuras de datos e interfaces.

5.2 Conclusiones del caso estudiado

A pesar de lo anteriormente comentado y mediante el uso de herramientas avanzadas se ha detectado que el nivel actual de aplicación de metadatos en las webs institucionales oficiales españolas es ínfimo y, a menudo, su calidad es errónea e inexacta.

Los problemas que habitualmente nos encontramos al analizar los metadatos y su implementación son:

- Datos incompletos o insuficientes. En la gran mayoría de los casos se dejan sin especificar aspectos que se consideran obvios o sin utilidad para la institución que crea los metadatos (por ejemplo, porque el contenido de un elemento sería exactamente el mismo en todos los registros); o porque los metadatos proceden de aplicaciones ya existentes en las que simplemente dichos elementos no estaban definidos.

- Datos incorrectos. Por ejemplo, por la utilización errónea de elementos tales como la inclusión en un solo elemento contenidos que deberían repetirse en varios.

- Datos confusos. Por ejemplo el uso inconsistente de cadenas de caracteres (palabras clave separadas unas veces por ” ; ” y otras por ” , ”),inclusión de caracteres erróneos o mal codificados (UTF-8 incorrecto en XML), o entidades codificadas doblemente en XML, etc.

Tal vez las instituciones estudiadas podrán trabajar sin problemas a pesar de los citados errores en el contenido de los metadatos, pero cuando éstos intenten ser integrados en agregadores temáticos será el momento en el que surjan los problemas al perderse el contexto original.

Excepto el caso de la web institucional del País Vasco con Dublin Core, ninguna se rige claramente por un estándar o esquema de metadatos. Algunas de ellas incluso prácticamente carecen de metadatos (es el caso de los portales de Aragón, Cantabria y La Rioja). La media nacional se encuentra en apenas cinco metadatos por página y la mayoría de websites se basan en tres etiquetas meta básicas de HTML:

- *description*: ofrece la información general sobre la página: de qué trata, qué tipo de información contiene, qué institución u organización es la responsable, etc.

- *keywords*: las palabras clave o descriptores del contenido de la página.

- *viewport*: referente a la visualización en dispositivos móviles.

Es reseñable indicar el gran número de metatags de redes sociales (especialmente Twitter). Y de otros enfocados al manejo de los portales con distintos dispositivos (en cuanto al tamaño o color). Hay que indicar, que debido al bilingüismo de algunas comunidades el metadato *language* se manifiesta en alguna ocasión.

El portal con mayor cantidad de metadatos recuperados es, de forma destacada, Castilla y León y mezcla los metatags clásicos con DCterms, pero la comunidad que más firmemente ha apostado por un sistema o esquema de metadatos (usa siete elementos DC) es el País Vasco.

Por tanto, ante la disparidad y pobreza de los datos recopilados se recomendaría seguir una serie de mejoras elaborando un de un programa de actualización y ampliación progresiva del modelo de metadatos, estableciendo mecanismos de control de la calidad en la asignación y uso de metadatos y diseñando un programa de formación del personal que asigna los metadatos en los recursos de información.

6. BIBLIOGRAFÍA

Alonso Berrocal, J.L; Figuerola, C.G; Zazo , Á. F. (2006). SACARINO (Sonda AutomátiCA para la Recuperación de INformaciÓn en la Web): un robot para recorrer y procesar la Web. // Scire: representación y organización del conocimiento. 12:1 (en.-jun. 2006) 211-224. Disponible en: https://www.ibersid.eu/ojs/index.php/scire/article/view/1160 [recuperado el 3 de julio de 2019]

Alonso Vega, A. (2017). Developer Tools #2: Construyendo expresiones regulares con Regex101. Disponible en: https://medium.com/@alonsus91/developer-tools-2-construyendo-expresiones-regulares-con-regex101-50d4f531b366 [recuperado el 3 de julio de 2019]

Álvarez Díaz, M. (2007). Arquitectura para Crawling Dirigido de Información Contenida en la Web Oculta. [Tesis doctoral]. A Coruña: Departamento de Tecnoloxías da Información e as Comunicacións. Universidade da Coruña. Disponible en: http://hdl.handle.net/2183/1000 [recuperado el 3 de julio de 2019]

Alvite Díaz, M.L. (2014). Metadatos en el contexto archivístico. El reto de la gestión y conservación de documentos electrónicos. Jornadas Archivando: la nueva gestión de archivos. León, 6 y 7 de noviembre 2014. Actas de las jornadas. Disponible en: https://www.researchgate.net/publication/272678911_Metadatos_en_el_contexto_archivistico_El_reto_de_la_gestion_y_conservacion_de_documentos_electronicos [recuperado el 3 de julio de 2019]

Americati (2007) Tutorial y guía breve del Comando Sed para Unix y Linux. Disponible en: https://www.americati.com/doc/sed/sed.html [recuperado el 3 de julio de 2019]

Barker, P. (2005) What is IEEE Learning Object Metadata / IMS Learning Resource Metadata? Cetis. Standards briefing series. JISC. Disponible en: http://www.dia.uniroma3.it/~sciarro/e-learning/WhatIsLOMscreen.pdf [recuperado el 3 de julio de 2019]

Barrueco, J.M. (2011). Introducción a los metadatos para las colecciones digitales. Servicio de bibliotecas y documentación. Universidad de Valencia. Disponible en: https://www.uv.es/=barrueco/00/curso.pdf [recuperado el 3 de julio de 2019]

Beard, K. (1996). A Structure for Organizing Metadata Collection. Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, New Mexico, USA, January 21-25.

Benítez Andrades, J. A. (2010). Resumen Tema 2: Crawling. Disponible en: https://www.jabenitez.com/personal/MASTER/MINERIA_DE_LA_WEB/TAREAS/MDW-JoseAlbertoBenitezAndrades-Tema2.pdf [recuperado el 3 de julio de 2019]

Boldi, B. Codenotti, M. Santini y S. Vigna (2004), Ubicrawler: A scalable fully distributed web crawler, Software: Practice and Experience, vol. 34, pp. 711-726, 2004. Disponible en: https://doi.org/10.1002/spe.587 [recuperado el 3 de julio de 2019]

Camargo, F. y Ordóñez, S. (2013). Evolución y tendencias actuales de los Web crawlers. Ingeniería, Vol. 18, No. 2, pp. 19–35. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=4797426 [recuperado el 3 de julio de 2019]

Caplan, P. (1995). You call it corn, we call it syntax-independent metadata for document like objects. Public Access Computer Systems Review, 4, pp. 19-23. Disponible en: https://journals.tdl.org/pacsr/index.php/pacsr/article/view/5992/5621 [recuperado el 3 de julio de 2019]

Carbajal Sardá, P. J. y Moreno Coatzozón, G. (2011). Metadatos: Estándares, lenguajes de ontología, web semántica. Universidad Veracruzana. Disponible en:

https://es.slideshare.net/Coatzozon20/capitulo-8-metadatos [recuperado el 3 de julio de 2019]

Castro-Ponce, S. (2013). Metadatos como herramienta en la recuperación de la información. Disponible en: https://www.infotecarios.com/metadatos-y-recuperacion-de-la-informacion/#.XR2rYP7tbDc [recuperado el 3 de julio de 2019]

Ceweb.br (2008) Guía de la web semántica. Disponible en: https://ceweb.br/guias/web-semantica/es/capitulo-2/ [recuperado el 3 de julio de 2019]

Chacón Beltrán, M. R. (2008). El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical. El valor de la diversidad (meta)lingüística. Actas del VIII Congreso de Lingüística General, Madrid, UAM, pp. 27-34. Disponible en: https://www.researchgate.net/publication/258019337_El_uso_de_expresiones_regulares_en_la_deteccion_de_errores_escritos_implicaciones_para_el_diseno_de_un_corrector_gramatical [recuperado el 3 de julio de 2019]

Cho J., H. Garcia-Molina, y P. Lawrence, (1998). Efficient crawling through URL ordering, Proceedings of the seventh international conference on World Wide Web 7 (WWW7), Amsterdam, The Netherlands, pp. 161-172, 1998. Disponible en: https://dl.acm.org/citation.cfm?id=297835 [recuperado el 3 de julio de 2019]

Collado, M. (2013). Lenguaje AWK. Material de estudio de Entornos de programación. Ingeniería informática. UPM. Disponible en: http://lml.ls.fi.upm.es/ep/awk.html [recuperado el 3 de julio de 2019]

Delgado Gómez, A. (2005). Archivos y metadatos de conservación: estado del arte y propuesta metodológica. Scire. 11: 1 (en.-jun. 2005) 83-101. ISSN 1135-3761. Disponible en: https://www.ibersid.eu/ojs/index.php/scire/article/view/1509/1487 [recuperado el 3 de julio de 2019]

Dempsey, L. (2005). Metadata: practice and practice. CLIR/DLF. Managing Digital Assets: A Primer for Library and Information Technology Administrators. Disponible en: https://www.clir.org/wp-content/uploads/sites/6/dempsey.ppt [recuperado el 3 de julio de 2019]

Díaz, L., Granell, C., Beltrán Fonollosa, A., Llaves, A., y Gould, M. (2008). Extracción semiautomática de metadatos: hacia los metadatos implícitos. Actas de las II Jornadas de SIG Libre. Girona, España. Disponible en: http://hdl.handle.net/10256/1160 [recuperado el 3 de julio de 2019]

Duval, E.; Hodgins, W.; Sutton, S.; Weibel, S. (2002). Metadata Principles and Practicalities. D-Lib Magazine, v.8, n.4, abril de 2002. Disponible en: http://www.dlib.org/dlib/april02/weibel/04weibel.html [recuperado el 3 de julio de 2019]

ECNBII (2003). FGDC Biological Data Profile As it maps Dublin Core. GeoConnection.

Ebner, H. (2009). Introduction to Dublin Core Metadata. Knowledge Management Research group of Royal Institute of Technology (KTH), Sweden. Disponible en: https://www.slideshare.net/ebner/introduction-to-dublin-core-metadata [recuperado el 3 de julio de 2019]

Ercegovac, Z. (1999). Introduction. Special Issue: Integrating Multiple Overlapping Metadata Standards. Vol 50, Issue 3. Disponible en: https://doi.org/10.1002/(SICI)1097-4571(1999)50:13<1165::AID-ASI2>3.0.CO;2-I [recuperado el 3 de julio de 2019]

España (2010). Real Decreto 1671/2009, de 6 de noviembre, por el que se desarrolla parcialmente la Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los servicios públicos. BOE núm. 278, de 18 de noviembre de 2009.

Disponible en: https://www.boe.es/buscar/pdf/2009/BOE-A-2009-18358-consolidado.pdf [recuperado el 3 de julio de 2019]

España (2010). Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica. BOE núm. 25, de 29 de enero de 2010. Disponible en: http://www.boe.es/boe/dias/2010/01/29/pdfs/BOE-A-2010-1331.pdf [recuperado el 3 de julio de 2019]

Friedl, J.E.F. (2006). Mastering. Regular expressions. Disponible en: http://xlb.es/Mastering%20Regular%20Expressions%20(Friedl-2006).pdf [recuperado el 3 de julio de 2019]

Gayatri y Ramachandran, S. (2007). Understanding Metadata. The Icfai Journal of InformationTechnology, March 2007

Gilliland-Swetland, A. (1998). An Exploration of K-12 User Needs for Digital Primary Source Materials. The American Archivist 1998 61:1, 136-157. Disponible en: https://doi.org/10.17723/aarc.61.1.w851770151576l03 [recuperado el 3 de julio de 2019]

Gilliland-Swetland, A. (2000). Introduction to Metadata: Pathways to Digital Information. Getty Publications. Disponible en: https://www.researchgate.net/publication/234802161_Introduction_to_Metadata_Pathways_to_Digital_Information [recuperado el 3 de julio de 2019]

Gmelendez (2013). Wikimedia Commons. Disponible en: https://commons.wikimedia.org/wiki/File:SedDiagram.jpg [recuperado el 3 de julio de 2019]

GNU Wget (1996). GNU Wget 1.20 Manual. Disponible en: www.gnu.org/software/wget/manual/wget.html [recuperado el 3 de julio de 2019]

Heery R. (1996) Review of metadata formats. Program, Vol. 30 Issue: 4, pp.345-373, Disponible en: https://doi.org/10.1108/eb047236 [recuperado el 3 de julio de 2019]

Hípola, P., Varga-Quesada, B. y Senso, J. A. (2000). Bibliotecas digitales: situación actual y problemas. El profesional de la información, 2000, vol. 9, n. 4, pp. 4-13. Disponible en: http://eprints.rclis.org/14479/ [recuperado el 3 de julio de 2019]

Howe, D. (1993). Free On-Line Dictionary of Computing (FOLDOC) Disponible en: http://foldoc.org/ [recuperado el 3 de julio de 2019]

Iqbal, M.; Muneeb Abid, M.; Khurshid, F. (2015). Information Retrieval Process on the Web: A survey on Web Crawler Types & Algorithms. Vol. 2. Disponible en: https://www.researchgate.net/publication/282245933_Information_Retrieval_Process_on_the_Web_A_survey_on_Web_Crawler_Types_Algorithms [recuperado el 3 de julio de 2019]

IONOS (2019). Los metatags más importantes. Disponible en: https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/los-meta-tags-mas-importantes-y-su-funcion/ [recuperado el 3 de julio de 2019]

Johnston, P. y Powell, A. (2005). A simple text-based format for DC metadata. Disponible en: http://dcpapers.dublincore.org/pubs/article/view/836 [recuperado el 3 de julio de 2019]

Kerrisk, M. (2010). The Linux Programming Interface: A Linux and UNIX System Programming Handbook. No Start Press.

Lamarca Lapuente, M. J. (2006). Hipertexto: El nuevo concepto de documento en la cultura de la imagen. [Tesis doctoral]. Disponible en: http://www.hipertexto.info/index.htm [recuperado el 3 de julio de 2019]

Lange, H. R.; Winkler, B. J. (1997). Taming the Internet: metadata, a work in progress. Advances in Librarianship, v. 21, p. 47-72, 1997. Disponible en: https://www.emeraldinsight.com/doi/abs/10.1108/S0065283019970000021005 [recuperado el 3 de julio de 2019]

Lara Carrascal, J. L. (2008). Gestores de descargas - GNU Wget. Manualinux: Documentación en Español del Sistema Operativo GNU/Linux. Disponible en: http://manualinux.eu/wget.html [recuperado el 3 de julio de 2019]

Llamaret Heredia, M. (2019). Explotando las potencialidades de Wget. Blog SWL-X. Disponible en: https://swlx.info/blog/explotando-potencialidades-wget [recuperado el 3 de julio de 2019]

Lynch, C. (1999). Canonicalization: a fundamental tool to facilitate preservation and management of digital information, D-Lib Magazine, 5 (9), Disponible en: www.dlib.org/dlib/september99/09lynch.html [recuperado el 3 de julio de 2019]

Méndez Ortega, J. (2018). Expresiones regulares. Disponible en: https://medium.com/@jmz12/expresiones-regulares-215af64acab1 [recuperado el 3 de julio de 2019]

Méndez Rodríguez, E. M. (2002). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón, Ediciones Trea.

Méndez Rodríguez, E. M. (2003a). La descripción de documentos electrónicos a través de los metadatos: una visión para la archivística desde la nueva e-administración. Revista d'arxius (2), 47-84. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=3209501 [recuperado el 3 de julio de 2019]

Méndez Rodríguez, E. M. (2003b). Tratamiento de los objetos de información en los archivos: retos y estándares para la descripción basada en metadatos. Disponible en: https://www.researchgate.net/publication/28809834_Tratamiento_de_los_objetos_de_informacion_en_los_archivos_retos_y_estandares_para_la_descripcion_basada_en_metadatos [recuperado el 3 de julio de 2019]

Méndez Rodríguez, E. M. y Senso J. A. (2004). Introducción a los metadatos: estándares y aplicación. Unidad de Autoformación. SEDIC. Disponible en: https://www.sedic.es/autoformacion/metadatos/ [recuperado el 3 de julio de 2019]

Moellering, H., Brodeur, J., Danko, D. y Shin, S. (2006). Towards a North American Profile of the ISO 19115 World Spatial Metadata Standard. Disponible en: https://www.researchgate.net/publication/228752938_Towards_a_North_American_Profile_of_the_ISO_19115_World_Spatial_Metadata_Standard [recuperado el 3 de julio de 2019]

Nebert, D., (2004). Developing spatial data infrastructures: The sdicookbook v.1.1. Global Spatial Data Infrastructure.

Ortiz-Repiso Jiménez, V. (1999). Nuevas perspectivas para la catalogación: metadatos versus MARC. Revista Española de Documentación Científica, 1999, v. 22, n. 2, pp. 198-219. Disponible en: http://redc.revistas.csic.es/index.php/redc/article/view/338 [recuperado el 3 de julio de 2019]

Oosterom, P.J.M van (2004). NCG/GIN Farewell Seminar Henri J.G.L. Aalders, Delft, 17 de noviembre 2004. Disponible en: https://www.ncgeo.nl/index.php/en/publicatiesgb/green-series/item/2362-gs-42-p-

van-oosterom-geo-information-standards-in-action [recuperado el 3 de julio de 2019]

Pasquinelli, A. (1997). Information technology directions in libraries: a sun microsystems white paper.

Ramos, J. (2011). SEO: Guía práctica de posicionamiento en buscadores. Disponible en: https://books.google.es/books?id=j5E2DgAAQBAJ&pg=PT6&lpg=PT6&dq= [recuperado el 3 de julio de 2019]

Riley, J. (2010). Seeing Standards: A Visualization of the Metadata Universe. Disponible en: http://jennriley.com/metadatamap/ [recuperado el 3 de julio de 2019]

Senso, J. A., de la Rosa Piñero, A. (2002) Evolución del Dublin Core Metadata Initiative. Disponible en: https://www.ugr.es/~jsenso/curriculum/dcmi.pdf [recuperado el 3 de julio de 2019]

Senso, J. A., de la Rosa Piñero A. (2003), El concepto de metadato: algo más que descripción de recursos electrónicos, Ciência da Informação, 32, 2, (95-106). Disponible en: http://www.scielo.br/pdf/ci/v32n2/17038.pdf [recuperado el 3 de julio de 2019]

Testa, P. M. y Degiorgi, E. H. (2013) Esquemas de metadatos para los repositorios institucionales de las universidades argentinas Tesis. Universidad Nacional de Cuyo. Disponible en: http://bdigital.uncu.edu.ar/objetos_digitales/5881/tesisdegiorgitesta.pdf [recuperado el 3 de julio de 2019]

UNE-ISO 19115. Información geográfica. Metadatos. Parte 1: Fundamentos. Modificación 1. (ISO 19115-1:2014/Amd 1:2018). Madrid: Asociación Española de Normalización y Certificación (AENOR).

UNE-ISO 15489-1:2016 Información y documentación. Gestión de documentos. Parte 1: Conceptos y principios (2016). Madrid: Asociación Española de Normalización y Certificación (AENOR)

UNE-ISO 15489-2 Información y documentación. Gestión de documentos. Parte 2: Directrices. (ISO/TR 15489-2:2001). Madrid: Asociación Española de Normalización y Certificación (AENOR)

UNE-ISO 15836: Información y documentación. Conjunto de elementos de metadatos Dublin Core (2011). Madrid: Asociación Española de Normalización y Certificación (AENOR).

UNE-ISO 23081-1: Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 1: Principios. (2008). Madrid: Asociación Española de Normalización y Certificación (AENOR).

UNE-ISO 23081-2: Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 2: Elementos de implementación y conceptuales. (2011). Madrid: Asociación Española de Normalización y Certificación (AENOR).

Weibel, S. (1995) Metadata: the foundations of resource description. D-lib magazine, julio 1995. Disponible en: http://www.diglib.org/dlib/July95/07weibel.html [recuperado el 3 de julio de 2019]

W3C Consortium (1999). Especificaciones oficiales elaboradas por W3C Consortium. Disponible en: https://www.w3schools.com/tags/tag_meta.asp [recuperado el 3 de julio de 2019]

7. ENLACES DE INTERÉS

[ASGF] Australian Subject Gateways Forum	http://www.nla.gov.au/initiatives/sg/gateways.html
[AGLS] Australian Government Locator Service	http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html
[CERIF] Common European Research Information Format	http://www.cordis.lu/cerif
[DC-Lib] Dublin Core Library Application Profile	http://dublincore.org/documents/library-application-profile
[DCMI] Dublin Core Metadata Initiative	http://www.dublincore.org http://es.dublincore.org
[DCMES] Dublin Core Metadata Element Set (ISO 15836)	http://www.dublincore.org/documents/dces/ http://www.niso.org/international/SC4/n515.pdf
[DIG35]	http://www.i3a.org/i_dig35.html
[DOI] Digital Object Identifier	http://www.doi.org
[EAD] Encoded Archival Description	http://www.loc.gov/ead
[EdNA] Educational Network Australia	http://www.edna.edu.au/metadata
[FGDC] Federal Geographic Data Committee	http://www.fgdc.gov
[FRBR] Functional Requirements for Bibliographic Records	http://www.ifla.org/VII/s13/frbr/frbr.htm
[GEM] Gateway to Educational Materials	http://www.geminfo.org
[GILS] Global Information Locator Service	http://www.gils.net
[ICPSR] Inter-University Consortium for Political and Social Research	http://www.icpsr.umich.edu
[IEEE/LOM] Learning Object Metadata	http://ltsc.ieee.org/wg12
[MARC] Standards Website	http://lcweb.loc.gov/marc
[MARC-ES] MARC en español	http://www.loc.gov/marc/marcspa.html
[MARCXML] MARC XML schema	http://www.loc.gov/standards/marcxml
[METS] Metadata Encoding and Transmission Standard	http://www.loc.gov/standards/mets
[MOA] Making Of America	http://sunsite.berkeley.edu/MOA2
[MODS] Metadata Object Description Schema	http://www.loc.gov/standards/mods

[NDLTD] Networked Digital Library of Theses and Dissertations	http://www.theses.org
[NZGLS] New Zealand Government Locator Service	http://www.e-government.govt.nz/nzgls
[OAI] Open Archives Initiative	http://www.openarchives.org
[OMNI] UK's Gateway to High Quality Internet Resources in Health and Medicine	http://omni.ac.uk
[ONIX] Online Information eXchange	http://www.editeur.org/onix.html
[PADI] Preserving Access to Digital Information	http://www.nla.gov.au/padi
[RDF] Resource Description Framework	http://www.w3.org/RDF
[RDFS] RDF Schema: RDF Vocabulary Description Language	http://www.w3.org/TR/rdf-schema
[RSS] Really Simple Syndication Rich/RDF Site Summary	http://web.resource.org/rss/1.0/
[SGML] Standard Generalized Mark-up Language	http://xml.coverpages.org/sgml.html http://www.w3.org/MarkUp/SGML/
[SVG] Scalable Vector Graphics	http://www.w3.org/Graphics/SVG
[SW] Semantic Web (W3C)	http://www.w3.org/2001/sw
[TEI] Text Encoding Initiative	http://www.tei-c.org
[ViDe] Video Development Initiative (DC application profile)	http://www.vide.net/workgroups/videoaccess/resources.shtml
[VRA] Virtual Resources Association Core Categories	http://www.vraweb.org/vracore3.htm
[W3C] World-Wide Web Consortium	http://www.w3.org
[XML] eXtensible Markup Language	http://www.w3.org/XML
[XMP] eXtensible Metadata Platform (Adobe)	http://www.adobe.com/products/xmp/main.html
[Z39.50] Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ISO23950)	http://lcweb.loc.gov/z3950/agency
[ZING] Z39.50 International Next Generation	http://www.loc.gov/z3950/agency/zing