Temario de Ayudante de Biblioteca de la Administración General del Estado

Este temario ha sido elaborado por un opositor, para presentarse al proceso selectivo de Ayudante de Bibliotecas de la Administración General del Estado en la convocatoria de 2021.

Incluye todos los temas, de legislación y específicos de bibliotecas, del programa correspondiente a la convocatoria de la Administración General del Estado para cubrir plazas de Ayudante de Bibliotecas en el Ministerios de Cultura y Deporte, Ministerio de Defensa, Ministerio de Asuntos Exteriores, Unión Europea y Cooperación y Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática. «BOE» núm. 149, de 23 de junio de 2021.

Temario completo disponible en:

https://www.bibliopos.es/

B12 Lenguajes de marcado y su aplicación en bibliotecas

Los lenguajes de marcado

Los lenguajes de marcado, también denominados lenguajes de marcas o lenguajes de descripción de documentos (*markup languages*), son modos de codificar un documento electrónico mediante la incorporación, junto al texto sin formato o plano, de etiquetas provistas de información adicional sobre la estructura, semántica o presentación que ha de tener el mismo. Dichas etiquetas serán posteriormente interpretadas por los intérpretes del lenguaje (como los navegadores) y ayudan al procesado del documento. Los lenguajes de marcas no son lenguajes de formato similares a los lenguajes que se usan en Internet como los de descripción de páginas (como los archivos .pdf) ni son lenguajes de programación (Java, Perl, C++...). En realidad, se podría hablar de metalenguajes o sistemas formales mediante los cuales se añade información o codificación a la forma digital de un documento bien para controlar su procesamiento, bien para representar su significado.

En general, siguen una **sintaxis** basada en el uso de dos marcas, tags o etiquetas, en minúsculas, encerradas entre ángulos (<>): una de inicio y otra de fin para indicar que ha terminado el atributo deseado (la de final se diferencia por una barra inclinada "/" antes del código). A través de las etiquetas se van definiendo los elementos del documento, como enlaces, párrafos, imágenes, etc. Esto servirá al navegador para saber cómo presentar el texto y otros elementos en la página.

Originalmente reciben este nombre por la práctica tradicional de marcar los manuscritos con instrucciones de impresión en los márgenes acerca de cómo se deben utilizar los tipos de letra, los tamaños, los espacios, el sangrado, etc., para definir la forma que tendrá el documento impreso. En los documentos digitales, se utiliza el término *marcado* para describir los códigos y etiquetas añadidos al texto, cuya finalidad será la de definir su estructura o su formato de presentación.

Cuando se habla de lenguajes de marcado, es importante distinguir entre la estructura física del documento (indica la apariencia del documento, incluyendo sus componentes físicos, el posicionamiento de los elementos y la tipografía empleada) y la estructura lógica (formada por las partes que lo componen y por sus relaciones). Esto es, en un documento existen distintos niveles de información: por un lado, los datos que conforman el contenido de un documento (caracteres de contenido) y, por otro, una información superpuesta al contenido, que es lo que constituye el etiquetado, marcado o "markup" (caracteres de etiquetado).

Un lenguaje de marcado cumple con dos **objetivos esenciales** a la hora de diseñar y procesar un documento digital:

- Especifica las operaciones tipográficas y las funciones que debe ejecutar el programa, navegador o visualizador sobre dichos elementos. Las operaciones tipográficas son instrucciones de formato que se aplican a cada uno de los elementos de un documento digital como, por ejemplo, imprimir un título en negrita y a un determinado tamaño.
- Separa un texto en los elementos de los que se compone: párrafos, capítulos, etc.

Por lo general, se pueden distinguir tres **tipos básicos** de lenguajes de marcado:

- **De presentación**. Es aquel que indica el formato del texto. Éstos suelen ocultar las etiquetas y mostrar al usuario solamente el texto con su formato. Es útil para la presentación de un documento, pero resulta insuficiente para el procesamiento automático de la información.
- **De procedimiento o procesado**. Está enfocado hacia la presentación del texto, sin embargo, también es visible para el usuario que lo edita. Las anotaciones o marcas de los lenguajes de procedimiento describen la forma y el significado de las operaciones tipográficas que van a ser aplicadas a cada uno de los elementos del documento. El programa que representa el

documento debe interpretar el código en el mismo orden en que aparece. Por ejemplo, para formatear un título, debe haber una serie de directivas inmediatamente antes y después del texto en cuestión, indicándole al software instrucciones tales como centrar, aumentar el tamaño de la fuente, o cambiar a negrita.

- **De marcado estructural o descriptivo**. Las marcas o anotaciones únicamente describen la estructura lógica del documento digital y/o la descripción del contenido, no su tipografía. Utiliza etiquetas para describir los fragmentos de texto, pero sin especificar cómo deben ser representados, o en qué orden.

El **concepto de lenguaje de marcas** fue expuesto por vez primera por William W. Tunnicliffe en 1967, sin embargo, quien es considerado el padre de los lenguajes de marcas es Charles Goldfarb (investigador de IBM), que participó en la creación del lenguaje **GML** (Generalized Markup Language), y posteriormente dirigió el comité que elaboró en 1986 el estándar **SGML** (Standard Generalized Markup Languaje, la norma ISO 8879:1986), la piedra angular de los lenguajes de marcas. El SGML especifica la sintaxis para la inclusión de marcas en los textos, así como la sintaxis del documento que especifica qué etiquetas están permitidas y dónde: el Document Type Definition (DTD, definición de tipo de documento, es la abstracción de todos los documentos que recogen el mismo tipo de información y que comparten características comunes) o *schema*.

En 1991, Tim Berners-Lee utilizó la sintaxis SGML para describir los 18 elementos que incluyen el diseño inicial de **HTML** (HyperText Markup Language, lenguaje de marcas de hipertexto). La característica que dota a HTML de un poder extraordinario frente a otros lenguajes es su capacidad hipertextual, es decir, su capacidad para establecer vínculos con otros documentos electrónicos. Los documentos web escritos con HTML tienen la gran ventaja de poder mostrar texto con enlaces a otras partes de la misma página (anclas), a otras páginas (hiperenlaces), o incluso a servidores distintos. Su invención se considera crucial en la aparición, desarrollo y expansión de la World Wide Web (WWW). La flexibilidad y escalabilidad del marcado HTML fue uno de los principales factores, junto con el empleo de URL y la distribución libre de navegadores, del éxito de la Web. En la actualidad, la construcción de la mayoría de los sitios web continúa basándose en lenguaje HTML con marcas o etiquetas que se muestran cuando se visualiza el código fuente, pero que permanecen ocultas en la visualización normal de los navegadores y que contienen información sobre el contenido de la página, enlaces hacia otras páginas, formatos de letra, imágenes, etc.

Sin embargo, el problema de este lenguaje es que se encuentra destinado principalmente a la visualización y no a la estructura de la información. La respuesta a los problemas surgidos en torno al HTML vino de la mano del **XML** (eXtensible Markup Language): un lenguaje de marcado desarrollado por el World Wide Web Consortium (W3C) y derivado del lenguaje SGML, que proporciona una sintaxis superficial para documentos estructurados, pero que no impone restricciones semánticas sobre el significado de los mismos. Es un lenguaje de marcas para la codificación de información en formato electrónico y su transferencia e intercambio a través de la Red. El lenguaje ofrece un modelo normalizado para representar la información y estructurar los documentos, de forma que sean fácilmente procesables por aplicaciones informáticas. Su caracterización como “extensible” se deriva de la no limitación en el número de marcas o etiquetas, pues permite definir todas aquellas que sean necesarias, así como crear tipos de documentos adecuados a la resolución de un problema particular. Otra de sus características principales es que permite enlaces multidireccionales (esto es, que apuntan a varios documentos). XML no es una nueva versión de HTML, aunque ambos proceden de un mismo metalenguaje, el SGML, sino que surgió para superar las limitaciones del formato HTML.

El **XHTML** (eXtensible HyperText Markup Language o Lenguaje de Etiquetado Hipertextual Extensible) es una reformulación del lenguaje HTML como aplicación XML. Su objetivo es avanzar en el proyecto del W3C de lograr una Web semántica. Los lenguajes de marcado son la herramienta fundamental en el diseño de la **Web semántica**, aquella que no sólo permite acceder a la información, sino que además define su significado, de forma que sea más fácil su procesamiento automático y se pueda reutilizar para distintas aplicaciones. Esto se consigue

añadiendo datos adicionales a los documentos, por medio de dos lenguajes expresamente creados: el **RDF** (Resource Description Framework, Marco de descripción de recursos) y **OWL** (Web Ontology Language, Lenguaje de ontologías para la web), ambos basados en XML. RDF y OWL se han convertido en estándares semánticos de la Web para proveer un marco de trabajo que asegure la gestión y la integración de iniciativas para compartir y reutilizar los datos sobre la Web.

Su aplicación en bibliotecas

La gestión de recursos de información se ha visto impactada por la generalización de los documentos en formato electrónico. Este cambio ha obligado a las bibliotecas a gestionar un nuevo tipo de materiales y a familiarizarse con los nuevos formatos y con las herramientas y tecnologías necesarias para su tratamiento. La adopción y uso de los lenguajes de marcas, y en particular XML, ofrecen posibilidades para la resolución de distintos problemas relacionados con la gestión de documentos electrónicos en bibliotecas, archivos y centros de documentación: descripción de recursos bibliográficos, codificación de documentos digitales, recuperación de información, recolección y agregación de metadatos y preservación de documentos digitales.

La proliferación de documentos publicados en la Web hizo que la comunidad bibliotecaria se plantease la necesidad de aplicar técnicas similares a la catalogación tradicional para facilitar la identificación, localización y posterior recuperación de estos recursos. Los sistemas para la creación de descripciones de recursos web debían resultar más fácil de usar que los que se venían aplicando tradicionalmente en bibliotecas. Se plantearon que fuesen los propios autores de los documentos web los que facilitasen unos **metadatos** básicos para sus documentos, ante la imposibilidad de que las bibliotecas asumiesen la responsabilidad de describir los recursos publicados en la Web, dado su incremento exponencial.

Surgió una línea de trabajo multidisciplinar que planteó sistemas alternativos para la descripción y catalogación de los recursos de información electrónicos. Su principal resultado fue la publicación en 1998 del llamado conjunto de metadatos **Dublin Core** (o DCMES, Dublin Core Metadata Element Set), de la DCMI (Dublin Core Metadata Initiative, organización dedicada a fomentar la adopción extensa de los estándares interoperables de los metadatos), con el patrocinio de OCLC (Online Computer Library Center). DCMES estableció un modelo de descripción básico formado por quince campos o metadatos, independientes de cualquier método de codificación (no obligaba a codificar de una forma particular). Para facilitar el uso de las distintas alternativas, la DCMI ha publicado recomendaciones sobre cómo codificar los metadatos Dublin Core en HTML y en XML. Si bien las pretensiones iniciales fueron bastante ambiciosas y se afirmaba que los buscadores en Internet serían capaces de identificar, extraer y recuperar documentos web que usasen esos metadatos, la evolución real no se correspondió con las expectativas iniciales. En los últimos años su uso se ha generalizado, concretamente en los repositorios institucionales y como medio para la agregación de metadatos en los archivos abiertos.

Pero Dublin Core no es el único sistema de metadatos surgido en torno a la Web. Otra iniciativa similar es **MODS** (Metadata Object Description Schema, Esquema para la Descripción de Objetos de Metadatos), elaborado en 2002 por la Network Development and MARC Standards Office (NDMSO, Oficina de Desarrollo de Redes y Normas MARC) de la Library of Congress. Frente a las críticas que recibía Dublin Core por la extrema simplicidad de su sistema, el sistema MODS ofrece un conjunto de metadatos más amplio (20 elementos), compatible en mayor medida con el formato MARC (MAchine Readable Cataloging, norma ISO 2709, utilizado en bibliotecas para la codificación de distintos tipos de registros). Establece una representación de los metadatos en forma de documentos XML, es decir, además de la información que debe indicarse en las descripciones, especifica la forma en la que éstas deben codificarse en XML.

Los componentes básicos de un catálogo son los registros bibliográficos y los registros de autoridades. Al igual que los registros bibliográficos se intercambian entre distintos centros, los registros de autoridades también se comparten en iniciativas de catalogación cooperativa. El sistema

de metadatos MODS fue complementado con un segundo esquema llamado **MADS** (Metadata Authority Description Schema), cuya primera versión es de 2004. Su propósito es ofrecer un esquema XML para codificar registros de autoridades que puedan vincularse a las descripciones bibliográficas basadas en MODS.

Además de MODS, para la adaptación de MARC a SGML, la NDMSO durante los años 1990 desarrolló dos DTD, uno para registros bibliográficos, de información de la comunidad y de fondos y otro para registros de autoridad y clasificación, capaces de convertir registros MARC a SGML y viceversa, sin pérdida de información, que pronto migrarían a XML para adecuarse a las nuevas necesidades tecnológicas: **MARC XML DTD**.

También publicó en 2002 un esquema XML, yendo más allá de un mero mecanismo de conversión y facilitando la representación de registros MARC en formato XML, para eliminar complejidades innecesarias y evitar que MARC quedara relegado frente a otras propuestas, en el marco de la biblioteca electrónica. Con esta especificación se adaptó el formato MARC a los lenguajes de marcas. **MARCXML** estableció una forma de codificar registros MARC como documentos XML, que fue convertida en 2008 en la actual norma ISO 25577:2013, que especifica los requisitos para un formato de intercambio generalizado basado en XML para los registros bibliográficos, así como otros tipos de metadatos. Se mantuvo para ello la estructura del registro y su organización en campos y subcampos y otras características del MARC (como los indicadores, la cabecera del registro, etc.). También se mantuvieron los códigos usados para nombrar campos y subcampos, así como su significado. El cambio consistió en una nueva forma de representar registros MARC para facilitar su procesamiento. Además se evita que las bibliotecas tengan que mantener dos tipos de repositorios: los catálogos tradicionales en los que se usa MARC y los basados en otros sistemas de metadatos. Con la especificación MARCXML los centros disponen de una forma de intercambiar datos bibliográficos con otros catálogos y bases de datos que no usan el MARC tradicional. La iniciativa llamada *MARCXML Framework* incluye utilidades y programas para la conversión de registros MARC a MARCXML, la conversión entre registros MARCXML, Dublin Core y MODS y la presentación y visualización de registros MARCXML mediante hojas de estilo XSLT (eXtensible Stylesheet Language Transformations, estándar de la organización W3C que presenta una forma de transformar documentos XML en otros).

Una aplicación de los lenguajes de marca sumamente interesante para bibliotecas es la codificación de documentos en texto completo (y no sólo de sus descripciones y metadatos). Las primeras iniciativas orientadas a este fin son anteriores a la formulación de XML. El uso de lenguajes de marcado para codificar documentos electrónicos resulta clave en proyectos de biblioteca digital, donde el intercambio de materiales es una práctica casi obligatoria, ya que es improbable que un centro disponga de todos los materiales a los que quiera dar acceso. Por ello, disponer de documentos codificados de forma uniforme es una garantía para el intercambio de recursos. La especificación más relevante en esta área es **TEI** (Text Encoding Initiative), que surgió para codificar *corpus* textuales en formato electrónico, universalizar el acceso a los mismos y facilitar su preservación en formato digital. Basado en XML, se trata de un etiquetado semántico y no presentacional, que establece el significado de cada elemento y atributo. Es un modelo de metadatos para codificar cualquier tipo de estructura textual de contenido literario y humanístico.

Pero no sólo gestionan documentos en formato texto, sino que parte de los materiales disponibles en sus colecciones consisten en documentos obtenidos por un proceso de digitalización, formados por una serie de archivos de imágenes (TIFF, PDF, etc.). Además es preciso mantener información sobre cómo se organizan esos ficheros para poder presentarlos al lector en una secuencia correcta y mostrar su organización en capítulos, partes, etc. Para codificar y representar la estructura de documentos digitales complejos que reúnen distintos archivos, se emplean los *metadatos estructurales*, que pueden ser codificados a través de la especificación **METS** (Metadata Encoding and Transmission Standard), cuya evolución también se encarga la Library of Congress.

El lenguaje XML también ha adquirido un gran protagonismo en los sistemas informáticos para la recuperación textual y en la normalización de los métodos utilizados para integrar sistemas de información heterogéneos. En este sentido, las bibliotecas cuentan con una norma de uso

universal, la Z39.50 (ISO 23950), ampliamente adoptada por los fabricantes de los sistemas integrados de gestión bibliotecaria. Este estándar permite la interrogación y búsqueda simultánea en múltiples catálogos bibliográficos remotos para obtener una lista unificada de registros, descargarlos en formato MARC e integrarlos en el catálogo propio. Pero con la generalización de la web y la aparición de nuevas tecnologías y protocolos, se han abierto nuevas posibilidades para la integración de sistemas de recuperación. Así surge en 2001 la iniciativa **ZING** (Z39.50 International Next Generation), con el objetivo de ofrecer un protocolo que resulte funcionalmente equivalente a Z39.50, pero que se base en las tecnologías de la Web. Incluyó tres líneas de trabajo: SRU (Search/Retrieve via URL) y SRW (Search/Retrieve Web service), que son protocolos que describen y normalizan el proceso de recuperación de información, los servicios y los mensajes que se deben intercambiar; y CQL (Contextual Query Language), lenguaje para la definición de consultas. La aplicación de XML en los procesos de recuperación de información también se ha hecho manifestar en las **pasarelas XML** que ofrecen distintos proveedores de servicios de información para exponer sus contenidos a los llamados metabuscadores o sistemas de búsqueda federada (aplicaciones donde el usuario puede buscar simultáneamente en múltiples servicios de información, como bases de datos, no sólo en catálogos). El lenguaje XML es el formato que permite distribuir las peticiones de búsqueda y transferir los resultados recuperados.

Otra aplicación donde se ha hecho un uso extensivo de XML son los llamados **archivos abiertos**: bases de datos centralizadas que reúnen descripciones de recursos creadas por distintos centros. Al poner en común estas descripciones en un único sitio de la Red, los centros dan una mayor visibilidad a sus recursos de información. Además de esa visibilidad, los centros no necesitan disponer ni mantener una infraestructura técnica compleja y únicamente tienen que facilitar las descripciones de sus recursos al centro responsable de agregarlos. Esto lleva a la necesidad de contar con un formato de intercambio de datos normalizado, común para todos los centros participantes, y que sea fácilmente procesable, lo que apunta nuevamente a los lenguajes de marcas. XML se ha convertido en el formato escogido para transferir los metadatos entre los centros participantes y los repositorios donde se centralizan, usándose para ello un protocolo técnico llamado **OAI-PMH** (Open Archive Initiative-Protocol for Metadata Harvesting), que regula la interacción entre un proveedor de datos y los programas de recolección. Por tanto, permite la recolección de los metadatos mediante un proceso automático que se ejecuta periódicamente. OAI-PMH utiliza XML para codificar y transferir las respuestas que genera el proveedor de datos a las peticiones cursadas por el recolector. Además, también se dispone de otros protocolos basados en XML para la agregación de información, como RSS (Really Simple Syndication, en español, sindicación realmente simple) o Atom, formatos XML para compartir y difundir información actualizada frecuentemente a usuarios suscritos a la fuente de contenidos.

Pero, los documentos electrónicos presentan un reto adicional: su preservación. El papel de los lenguajes de marcado en la preservación digital puede resumirse en dos líneas de actuación. Por una parte, el hecho de crear y almacenar documentos en un formato estable, soportado por múltiples fabricantes, ofrece mayores garantías que el uso de formatos propietarios. Por otra, una estrategia de preservación exige gestionar metadatos sobre los documentos digitales adicionales: información sobre el programa informático utilizado para generar un documento, la versión del software necesaria para leerlo o el método aplicado al generar una firma digital que garantice su autenticidad, son metadatos que se deben registrar y mantener para su preservación. En los últimos años se han lanzado distintas iniciativas para establecer el conjunto de metadatos necesarios para la preservación de objetos digital. De todas estas iniciativas, destaca el proyecto **PREMIS** (Preservation Metadata: Implementation Strategies, Metadatos de preservación: estrategias de ejecución), en el que se estableció un detallado diccionario de datos y una forma de codificarlos en forma de documentos XML.

Marzo 2021