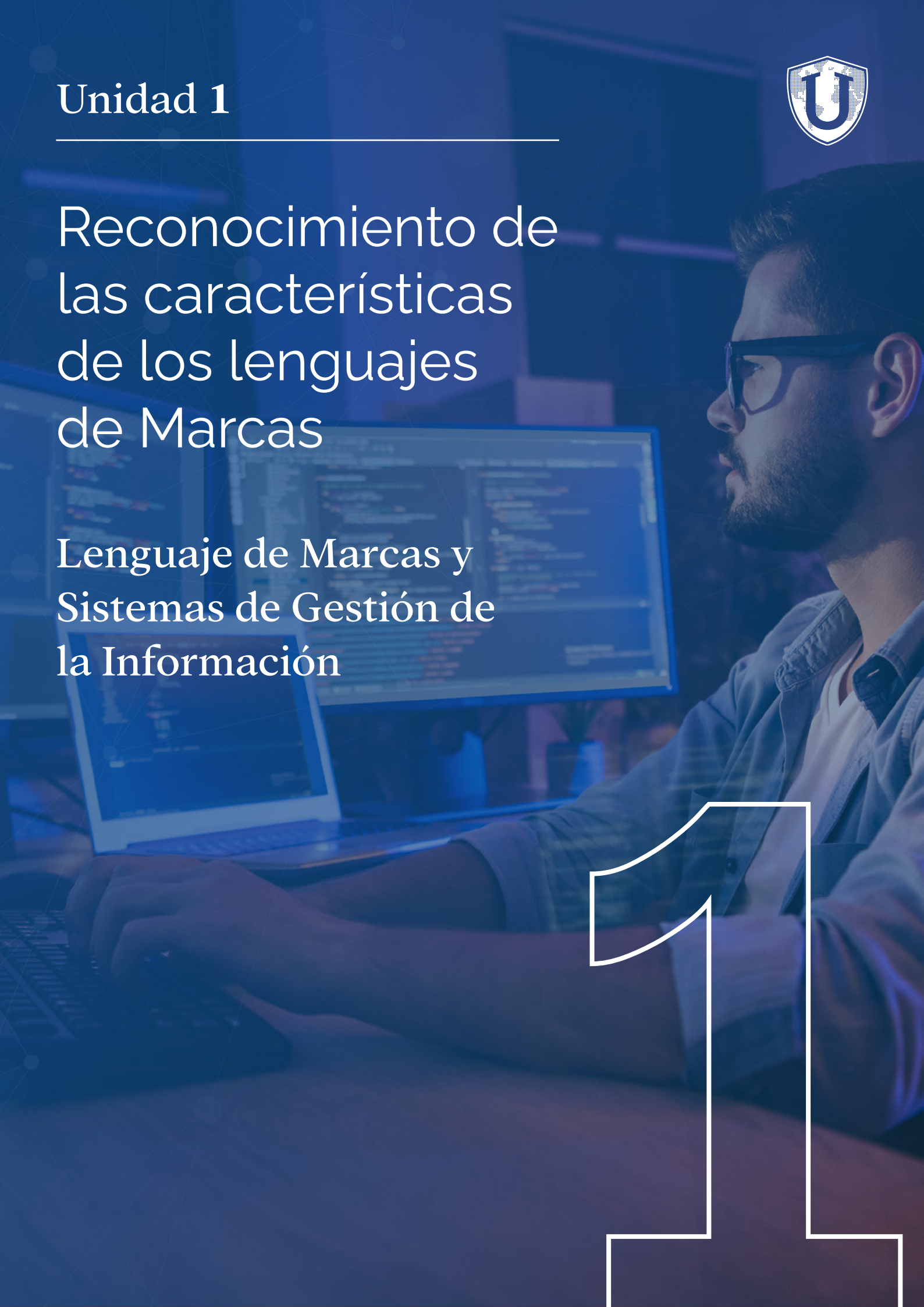
Unidad 1

Reconocimiento de las características de los lenguajes de Marcas

Lenguaje de Marcas y Sistemas de Gestión de la Información



1

Índice

Introducción

En los últimos años Internet ha dominado el mundo hasta el punto en que cualquier persona usa internet para todo en estos días. Y gran parte de los internautas de hoy en día usan páginas webs. Pues bien, detrás de estas páginas web están los lenguajes de marcas dándole sentido y forma a la información para que se nos presente del modo tan visual que lo hace hoy en día. Este proceso de transformación es invisible para el usuario final.

Además, hay muchos programas que podemos tener instalados en el ordenador de nuestra casa o nuestra oficina que también llevan por debajo un código escrito en lenguaje de marcas ya sea para alguna configuración de la aplicación o demás propósitos.

Hay que tener en cuenta que un lenguaje de marcas no es un lenguaje de programación, aunque se puede combinar con estos dentro del mismo código de un programa.

Para terminar, hay que tener en cuenta que para visualizar los lenguajes de marcas bien formados se usará casi siempre el navegador web, que no es lo mismo que un buscador, y debemos de saber diferenciarlos.

Al finalizar esta unidad

- Iniciaremos en el mundo de los lenguajes de marcas, identificando sus características más generales, reconociendo las ventajas que proporcionan, estudiando su clasificación y estructura e identificando los más relevantes,
- Conoceremos los orígenes y evolución de los lenguajes de marcas.
- Conoceremos las organizaciones desarrolladoras de los lenguajes de marcas
- Conoceremos las gramáticas de los lenguajes de marcas.

1.1. Clasificación y características comunes de los lenguajes de marcas

Atendiendo a tres tipos distintos, clasificaremos los lenguajes de marcas poniendo de manifiesto sus principales características.

1.1.1. Clasificación de los lenguajes de marcas

Los lenguajes de marcas tienen unos usos muy diversos, desde mensajería instantánea con XMPP hasta documentación electrónica con RTF, pasando por páginas web y sindicación de contenidos con HTML y RSS respectivamente.

Como hemos comentado antes, los clasificamos en tres tipos:

- **Tipo 1 o de procedimiento.** Este tipo de lenguaje suele ser visible para el usuario y se usa para la presentación del texto, Es común que se use etiquetas para por ejemplo resaltar un texto en negrita o alinear un texto a la izquierda. Un ejemplo de este tipo de lenguaje de marcas es Nroff.
- **Tipo 2 o de presentación.** Este tipo de lenguaje es el usado para dar formato al texto, es decir, dictaminar los espaciados, los saltos de línea etc. Resultan invisibles para el usuario y, además, aunque no es complicado de implementar, si lo es a la hora de modificarlo o mantenerlo. Un uso de este tipo de lenguaje es por ejemplo la maquetación de un documento de cara al lector de Microsoft Word.
- **Tipo 3 o descriptivos o semánticos.** Se trata del más flexible ya que sus Etiquetas se usan sin necesidad de que indiquemos un orden específico o su manera de representación. Estas marcas se van dando forma entre sí junto con el contenido a medida que se desarrolla el recurso en cuestión. En este grupo se incluyen los más famosos como XML y HTML.

1.1.2. Características comunes

Como hemos podido apreciar, hay gran cantidad de lenguajes de marcas en el mercado actual y es por esto por lo que no todos tienen las mismas características, pero sí que comparten muchas que vamos a detallar a continuación.

Texto plano (plain text)

Este tipo de documentos, también llamado texto sin formato o simple no permite almacenar información con formato (color, tipo de letra, negrita, tamaño, subrayados, etc.) tal y como CAPÍTULO 1 Abre el Bloc de notas de Windows y escribe la siguiente frase: "alumno LM". ¿Cuántos byt 6lmes ocupa? Solución CAPÍTULO 1 W ' RECONOCIMIENTO DE LAS CARACTERÍSTICAS DE LENGUAJES DE MARCAS puede hacerlo un procesador de texto avanzado. Están formados exclusivamente por caracteres (letras, números, caracteres especiales y de control). SABÍAS QUE... El espacio en blanco, el retorno de carro y las tabulaciones se consideran caracteres. Uno de los editores que puede encontrarse en Windows para crear un documento de texto plano es el Bloc de notas (Notepad), cuya extensión es .txt, mientras que, si se usa Linux, Gedit, Vi o Nano pueden ser los más usados. Tal y como se ha mencionado anteriormente, al estar formado por caracteres, se emplea una codificación específica que suele ser ASCII, ISO 8859-1, Unicode o UTF-8. Estos estándares de codificación permiten representar los distintos caracteres en todos los idiomas. ISO 8859-1 se usa para la codificación de alfabeto latino (fi, letras acentuadas, etc.) En el ejercicio resuelto 1.1, se verá que cada carácter que se escribe en un documento de texto plano ocupa un byte o, lo que es lo mismo, 8 bits. Hay que recordar que el espacio en blanco también se considera un carácter.

Interoperabilidad o independencia

Se considera el texto plano como formato universal, ya que puede abrirse y editarse desde cualquier máquina, aunque ha de tenerse en cuenta la codificación empleada. Son independientes a la plataforma usada y a cualquier sistema de representación.

Flexibles y fáciles de crear

Simplemente, se necesita un editor de texto para poder crearlos y guardarlos en la extensión que se desee. Algunos de ellos permiten combinarse con otro lenguaje para darle mayor funcionalidad.

1.2. Identificación de ámbitos de aplicación

El ámbito de aplicación de los lenguajes de marcas es muy diverso, ya que permite el intercambio de datos entre distintas aplicaciones, independientemente de la plataforma usada y la tecnología en la que estén creadas. Desde un fichero XML, pueden generarse vistas como HTML, WML o PDF. Java EE usa ficheros XML para poder especificar datos de configuración. Visual Studio, a la hora de crear servicios web, genera varios documentos con estructura XML. Tanto Windows Phone como Android Studio usan esta estructura para guardar las vistas. En general, existe una gran cantidad de software que usa datos en formato XML para configuración o guardar información. En el mundo real, puede usarse para bases de datos, frameworks de desarrollo, sistemas de publicación de contenidos, definición de interfaces gráficas, etc.

1.3. Evolución de los lenguajes de marcas

Los lenguajes de marcas comenzaron a usarse a finales de la década de los 60 para poder introducir anotaciones dentro de documentos electrónicos, de la misma forma que se hacía cuando la documentación estaba en papel. De esta posibilidad de incorporar marcas es de donde reciben su nombre. Es en esas fechas cuando se estandariza el lenguaje **SGML** (Standard Generalized Markup Language), que es un descendiente directo del lenguaje **GML** propuesto por IBM. Este lenguaje surgió para permitir compartir información por parte de sistemas informáticos. Este estándar tuvo una gran aceptación, pero no consiguió asentarse del todo debido principalmente a su complejidad lo que provocaba que el software que usará SGML terminaba siendo excesivamente extenso y complejo.

A finales de los 80 dentro del CERN (Conseil Européen pour la Recherche Nucléaire) se creó un lenguaje de marcado pensado para compartir información usando las redes de computadores y, de forma más general, a través de Internet. Este lenguaje se basaba en algunos principios de SGML y lo denominaron HTML (Hyper-text Markup Language). La aparición de este lenguaje supuso de alguna manera una revolución en la forma de compartir información, gracias principalmente a la sencillez de sus sintaxis y del software necesario para interpretarlo. En poco tiempo el lenguaje HTML se extendió y empezó a crecer de forma en ocasiones descontrolada y casi siempre influenciado por razones meramente comerciales.

A mediados de los años 90 el consorcio W3C (World Wide Web Consortium) comenzó una iniciativa para intentar dotar a la web de un lenguaje más potente y que pudiera dar una estructurar semántica a la misma. Para ello se marcaron el objetivo de crear un nuevo lenguaje de marcas basado en SGML y que fuera sencillo como HTML. Finalmente, en el 1998, W3C hizo público un nuevo estándar que denominaron XML (eXtended Markup Language), más sencillo que SGML y más potente que HTML.

1.4. Etiquetas, elementos y atributos

Existen tres términos comúnmente usados para describir las partes de un documento de lenguajes de marcas: etiquetas, elementos y atributos.

Una **etiqueta** (tag) es un texto que va entre el símbolo menor que (<) y el símbolo mayor que (>). Existen etiquetas de inicio (como <nombre>) y etiquetas de fin (como </nombre>).

Los **elementos** representan estructuras mediante las que se organizará el contenido del documento o acciones que se desencadenan cuando el programa navegador interpreta el documento. Constan de la etiqueta de inicio, la etiqueta de fin y de todo aquello que se encuentra entre ambas.

Algunos elementos no tienen contenido. Se les denomina elementos vacíos y no deben llevar etiqueta de fin.

Un **atributo** es un par nombre-valor que se encuentra dentro de la etiqueta de inicio de un elemento e indican las propiedades que pueden llevar asociadas los elementos.

1.5. Organizaciones desarrolladoras

Dentro de las organizaciones que se han encargado de desarrollar los lenguajes de marcas se encuentran:

- **Organización Internacional para la Estandarización (ISO, International Organization for Standardization).** Se formó después de la Segunda Guerra Mundial (23 de febrero de 1947) y es el organismo encargado de promover el desarrollo de normas internacionales de fabricación, comercio y comunicación para todas las ramas industriales a excepción de la eléctrica y la electrónica. Su función principal es la de buscar la estandarización de normas de productos y seguridad para las empresas u organizaciones a nivel internacional.

  Es una red de los institutos de normas nacionales de 163 países, sobre la base de un miembro por país, con una Secretaría Central en Ginebra (Suiza) que coordina el sistema.

  Las normas desarrolladas por ISO son voluntarias, ya que es un organismo no gubernamental y no depende de ningún otro organismo internacional, por tanto, no tiene autoridad para imponer sus normas a ningún país. El contenido de los estándares está protegido por derechos de copyright y para acceder a ellos el público en general ha de comprar cada documento.

  Esta organización después del éxito que tuvo GML y, después de un largo proceso, publicó en 1986 el Standard Generalized Markup Language (SGML) con rango de Estándar Internacional con el código ISO 8879.

- **World Wide Web Consortium (W3C).** El W3C se creó en 1994 por Tim Berners-Lee en el MIT, actual sede central del consorcio. Posteriormente se unió, en abril de 1995, el INRIA en Francia, reemplazado por el ERCIM en el 2003 como el huésped europeo del consorcio y la Universidad de Keiō (Shonan Fujisawa Campus) en Japón en septiembre de 1996 como huésped asiático. Su función principal es tutelar el crecimiento y organización de la web.

  Su primer trabajo fue normalizar el lenguaje HTML, el lenguaje de marcas con el que se escriben las páginas web. Al crecer el uso de la web, crecieron las presiones para ampliar el HTML. El W3C decidió que la solución no era ampliar el HTML, sino crear unas reglas para que cualquiera pudiera crear lenguajes de marcas adecuados a sus necesidades, pero manteniendo unas estructuras y sintaxis comunes que permitieran compatibilizarlos y tratarlos con las mismas herramientas. Ese conjunto de reglas es el XML, cuya primera versión se publicó en 1998.

1.6. Gramáticas

Todo documento de un lenguaje de marcas tiene en común una gramática que define el marcado permitido en esa clase, el marcado requerido y cómo debe ser utilizado dicho marcado en la instancia del documento.

1.6.1. DTD

El estándar define esta gramática mediante la **DTD (Definición de Tipo de Documento)** que establece las reglas de formación del lenguaje formal, es decir, qué combinaciones de símbolos elementales son sintácticamente correctas.

En la DTD se identifica la estructura del documento, es decir, aquellos elementos que son necesarios en la elaboración de un documento o un grupo de documentos estructurados de manera similar. Contiene las reglas de dichos elementos: el nombre, su significado, dónde pueden ser utilizados y qué pueden contener.

La especificación del W3C para HTML 4.0 contempla tres DTD:

- **DTD estricta (HTML 4.0 Strict DTD):** incluye todos los elementos y atributos que no han sido declarados "desaprobados" (deprecated), interpretando la expresión en el sentido de que no se recomienda ya su uso proponiéndose nuevos y mejores recursos para hacer lo mismo.
- **DTD transicional o flexible (HTML 4.0 Transitional DTD):** incluye todo lo que la anterior más los elementos y atributos desaprobados (deprecated).
- **DTD para documentos con marcos (HTML 4.0 Frameset DTD):** engloba todo lo incluido en la transicional más lo relativo a la creación de documentos con marcos (frames).

Recuerde que, aunque la especificación recomienda ceñirse a los recursos de la DTD estricta, utilizar el resto de los elementos y atributos no es incorrecto.

La DTD es el formato de esquema nativo (y el más antiguo) para validar documentos XML, heredado de SGML. Utiliza una sintaxis no-XML para definir la estructura o modelo de contenido de un documento XML válido:

- Define todos los elementos.
- Define las relaciones entre los distintos elementos.
- Proporciona información adicional que puede ser incluida en el documento (atributos, entidades, notaciones).
- Aporta comentarios e instrucciones para su procesamiento y representación de los formatos de datos.

Es el método más sencillo usado para validar, y por esta razón presenta varias limitaciones, ya que no soporta nuevas ampliaciones de XML y no es capaz de describir ciertos aspectos formales de un documento a nivel expresivo.

Las DTD pueden ser internas o externas a un documento, o ambas cosas a la vez.

1.6.2. Esquema XML

XML Schema es la evolución de la DTD descrita por el W3C, también denominado XSD (XML Schema Definition). Es un lenguaje de esquema más complejo y potente, basado en la gramática para proporcionar una potencia expresiva mayor que la DTD. Utiliza sintaxis XML, cosa que le permite especificar de forma más detallada un extenso sistema de tipos de datos. A diferencia de las DTD, soporta la extensión del documento sin problemas.

A la hora de la validación del documento, la utilización de XSD supone un gran consumo en recursos y tiempo debido a su gran especificación y complejidad en la sintaxis (los esquemas son más difíciles de leer y de escribir).

Después de validar el documento con XML Schema, es posible expresar su estructura y contenido en términos del modelo de datos usado por el esquema de validación. Esta funcionalidad, conocida como Post-Schema-Validation Infoset (PSVI), se puede utilizar para transformar el documento en una jerarquía de objetos, a los cuales se puede acceder a través de un lenguaje de programación orientada a objetos (OOP).

1.6.3. Relax NG

RELAX NG es un lenguaje de esquema basado en la gramática, muy intuitivo y más fácil de entender que el XML Schema. Tiene un alto poder expresivo, ya que, por ejemplo, permite validar elementos intercalados que pueden aparecer en cualquier orden.

Las aplicaciones de definición de documentos y validación para RELAX NG son más sencillas que las de XML Schema, haciéndolo más fácil de utilizar e implementar.

RELAX NG se ha convertido recientemente en un estándar ISO como la parte 2 de DSDL (Document Schema Definition Language).

www.universae.com