Resistencia, Chaco

UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL RESISTENCIA

Trabajo Práctico Integrador

Diseño e implementación de Lexer y Parser
Lenguaje DocBook/XML

Carrera: Ingeniería en Sistemas de Información

Asignatura: Sintaxis y Semántica de los Lenguajes

Curso académico: 2023

Primer cuatrimestre

Grupo Nº: 12

Integrantes:

- Sotelo, Julián
- Beron de Astrada, Santiago Agustín
- Bregant, Joaquín
- Niveiro, Gianfranco

Primera entrega: 30/04/2023
Versión de la documentación: 1.0

ÍNDICE

Introducción

Este es un trabajo realizado durante el cursado de la asignatura *Sintaxis y Semántica de los Lenguajes* correspondiente a la carrera de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional, Facultad Regional Resistencia. El mismo tiene como objetivo el diseño, creación e implementación de un Lexer y un Parser que pueda interpretar correctamente el lenguaje *Docbook/XML*, para luego traducirlo al lenguaje *HTML*.

Para la realización de esta tarea se consideró adecuado utilizar el lenguaje de programación de alto nivel *Python* debido a su versatilidad y al hecho de que todos los integrantes del grupo poseen conocimiento sobre el mismo. Adicionalmente se utilizará la dependencia PLY, la cual permitirá una fácil integración con la gramática definida.

Lenguaje Docbook/XML

Es un lenguaje de marcas que se utiliza para definir solo la estructura de un documento, y no su formato. Los documentos DocBook no describen ni la apariencia ni la presentación de sus contenidos, sino únicamente el sentido de dichos contenidos. Contiene básicamente texto. Ahora bien, hay texto que sirve para indicar elementos, describir el contenido y propiedades de la página. Es decir, un documento contiene lo que se denominan etiquetas o *tags*. Las etiquetas sirven para delimitar y determinar el tipo de elemento que representa un fragmento de texto en la página; por ejemplo, un párrafo o una tabla son elementos que pueden conformar una página. Incluso hay elementos que contienen otros elementos; por ej, las tablas constan de filas y las filas de celdas.

Componentes léxicos o tokens

<u>Observaciones</u>

- El texto dentro de las páginas webs no respeta los espacios en blanco ni tabulaciones que se coloquen en el código a la hora de mostrar el contenido por pantalla. Solo se considera el primer espacio en blanco, el resto se elimina.
- Las etiquetas distinguen entre mayúsculas y minúsculas.

Tipos de Datos:

El tipo de dato Cadena y URL se tratarán de manera especial

- Cadena: Estará compuesta por letras, números, signos de puntuación, caracteres especiales,
- URL: Los únicos caracteres permitidos en URL son letras, números, guión medio, guión bajo y punto. además de los caracteres reservados: # , /, :, &, ?, =

Reglas generales de las Etiquetas

- La mayoría de etiquetas reservadas requieren una apertura y un cierre de la misma. Al inicializarse estarán encerradas por los símbolos *<nombreDeEtiqueta>* y al finalizar su uso deberán cerrarse con *</nombreDeEtiqueta>*. En caso de que la etiqueta no requiera cierre, se colocará solo la de apertura.
- Cada etiqueta podrá contener determinados atributos, en caso de que los tenga, será con el formato:
  nombreAtributo="valorAtributo"
- Las etiquetas serán utilizadas en minúscula.

Etiqueta **DOCTYPE:** Esta indica el tipo de XML que estamos utilizando, en este caso sería Article.

- **Article**: Todo artículo está encerrado dentro de la apertura y el cierre de esa etiqueta. Marca el principio y fin del mismo (marca el elemento raíz de un documento). Aparece por única vez, puede o no poseer información o título, y a su vez debe obligatoriamente contener alguna otra etiqueta de texto.

- **Info:** Lo usamos para agregar si necesitamos la información de "encabezado". Tiene que contener como mínimo alguna información sobre el autor

- **Title**: Se utiliza para indicar los títulos de una sección en un documento o de un bloque dentro del mismo.
- **MediaObject:** Imágenes y multimedia
- **ItemizedList:** crea una lista no ordenada, la cual agrupa párrafos alineados a través de viñetas.
- **Important:** indica que un fragmento de texto es importante.

Gramática

Símbolos NO terminales (N)

NT_ARTICLE: sentencia de tipo *article.*
NT_SECTION: sentencia de tipo *section.*
NT_SIMPLESEC: sentencia de tipo *simplesec.*
CONT_A_S: elementos opcionales y obligatorios para *article* y *section.*
CONT_SS: elementos opcionales y obligatorias para *simplesec.*
SECTIONS: permite agregar *section* y *simplesec* recursivamente.
CONT_1: conjunto de etiquetas permitidas para *article*, *section*, *simplesec*, *itemizedlist* e *important*.
NT_INFO: sentencia de tipo *info*.
CONT_INFO: permite agregar etiquetas recursivamente para *info*.
ELEM_INFO: etiquetas permitidas para *info*.
NT_TITLE: sentencia de tipo *title.*
CONT_TITLE: permite agregar etiquetas recursivamente para *title.*
ELEM_TITLE: etiquetas permitidas para *title.*
PARAS: permite agregar recursivamente uno o más *para* y *simpara.*
NT_PARA: sentencia de tipo *para.*
CONT_PARA: permite agregar etiquetas recursivamente para la sentencia de tipo *para.*
ELEM_PARA: etiquetas permitidas para la sentencia de tipo *para.*
NT_SIMPARA: sentencia de tipo *simpara.*
NT_ABSTRACT: sentencia de tipo *abstract.*
NT_ITEMIZEDLIST: sentencia de tipo *itemizedlist.*
LISTITEM: permite agregar recursivamente etiquetas permitidas para *itemizedlist.*
NT_MEDIAOBJECT: sentencia de tipo *mediaobject*.
CONT_MEDIAOBJECT: etiquetas permitidas para *mediaobject*.
NT_IMAGEOBJECT: sentencia de tipo *imageobject*.
NT_VIDEOOBJECT: sentencia de tipo *videoobject*.
NT_AUTHOR: sentencia de tipo *author.*
CONT_AUTHOR: etiquetas permitidas para *author.*
NT_ADDRESS: sentencia de tipo *address.*
CONT_ADDRESS: etiquetas permitidas para *address*.
NT_COPYRIGHT: sentencia de tipo *copyright.*

CONT_2: conjunto de etiquetas permitidas para *simpara*, *emphasis*, *link* y *comment*.
CONT_SECL: permite agregar recursivamente etiquetas permitidas para *emphasis*, *comment*, *simpara* y *link.*
NT_SIMPARA: sentencia de tipo *simpara.*
NT_EMPHASIS: sentencia de tipo *emphasis.*
NT_COMMENT: sentencia de tipo *comment.*
NT_LINK: sentencia de tipo *link*.
NT_IMPORTANT: sentencia de tipo *important.*
CONT_IMPORTANT: permite agregar recursivamente etiquetas permitidas para *important*.
CONT_3: conjunto de etiquetas permitidas para *firstname*, *surname*, *street*, *city*, *state*, *phone*, *email*, *date, year* y *holder*.
CONT_VAR: permite agregar recursivamente etiquetas que pertenecen al conjunto CONT_3.
NT_FIRSTNAME: sentencia de tipo *firstname*.
NT_SURNAME: sentencia de tipo *surname*.
NT_STREET: sentencia de tipo *street*.
NT_CITY: sentencia de tipo *city*.
NT_STATE: sentencia de tipo *state*.
NT_PHONE: sentencia de tipo *phone*.
NT_EMAIL: sentencia de tipo *email*.
NT_DATE: sentencia de tipo *date.*
NT_YEAR: sentencia de tipo *year.*
NT_HOLDER: sentencia de tipo *holder.*
NT_INFORMALTABLE: sentencia de tipo *informaltable*.
TABLE_MEDIA: permite agregar recursivamente etiquetas *mediaobject*.
TABLE_GROUP: permite agregar recursivamente etiquetas *tgroup*.
NT_TGROUP: sentencia de tipo *tgroup*.
NT_THEAD: sentencia de tipo *thead*.
NT_TFOOT: sentencia de tipo *tfoot*.
NT_TBODY: sentencia de tipo *tbody*.
CONT_T: permite agregar recursivamente etiquetas *row* para las sentencias de tipo *thead*, *tfoot* y *tbody*.
NT_ROW: sentencia de tipo *row*.
CONT_ROW_1: permite agregar recursivamente etiquetas *entry* dentro de etiquetas *row*.

CONT_ROW_2: permite agregar recursivamente etiquetas *entrytbl* dentro de etiquetas *row*.
NT_ENTRY: sentencia de tipo *entry*.
NT_ENTRYTBL: sentencia de tipo *entrytbl*.
CONT_ENTRY: permite agregar recursivamente las etiquetas permitidas para *entry*.
NT_URL: sentencia de tipo *url*.
PROTOCOL: tipos de protocolos web aceptados para un enlace.
DOM: dominio del enlace.
PORT: puerto del enlace.
ROUTE: ruta del enlace.
FRAGMENT: localizador interno del enlace.

Símbolos terminales (T)

- <
- >
- /
- :
- #
- =
- !DOCTYPE
- article
- section
- simplesec
- info
- title
- para
- simpara
- itemizedlist
- author
- firstname
- surname
- address
- copyright
- emphasis
- comment
- link
- important
- street
- city
- state
- phone
- email
- date
- year
- holder
- https
- http
- ftp
- ftps
- fileref
- xlink
- href
- informaltable
- tgroup
- thead
- tfoot
- tbody
- row
- entrytbl
- entry

Producciones (P)

Los símbolos escritos en mayúsculas se corresponden con símbolos no terminales de la gramática, y los escritos en minúsculas, con símbolos terminales, excepto por *DOCTYPE* que también es un terminal.

Σ → <!DOCTYPE article>**NT_ARTICLE**

NT_ARTICLE → <article>**NT_INFO NT_TITLE CONTENT_A_S**</article> |
<article>**NT_TITLE CONTENT_A_S**</article> |
<article>**NT_INFO CONTENT_A_S**</article> |
<article>**CONTENT_A_S**</article>

NT_SECTION → <section>**CONTENT_A_S**</section> |
<section>**NT_INFO CONTENT_A_S**</section> |
<section>**NT_TITLE CONTENT_A_S**</section> |
<section>**NT_INFO NT_TITLE CONTENT_A_S**</section>

NT_SIMPLESEC → <simplesec>**CONT_SS**</simplesec> |
<simplesec>**NT_INFO CONT_SS**</simplesec> |
<simplesec>**NT_TITLE CONT_SS**</simplesec> |
<simplesec>**NT_INFO NT_TITLE CONT_SS**</simplesec>

CONT_A_S → **CONT_1** | **CONT_1 CONT_A_S** | **CONT_1 SECTIONS** |
CONT_1 CONT_A_S SECTIONS

SECTIONS → **NT_SECTION** | **NT_SECTION SECTIONS** | **NT_SIMPLESEC** |
NT_SIMPLESEC SECTIONS

CONT_1 → **NT_ITEMIZEDLIST** | **NT_IMPORTANT** | **NT_PARA** | **NT_SIMPARA** |
NT_ADDRESS | **NT_MEDIAOBJECT** | **NT_INFORMALTABLE** | **NT_COMMENT** |
NT_ABSTRACT

NT_INFO → <info>**CONT_INFO**</info>

CONT_INFO → **ELEM_INFO** | **ELEM_INFO CONT_INFO**

ELEM_INFO → **NT_MEDIAOBJECT** | **NT_ABSTRACT** | **NT_ADDRESS** | **NT_AUTHOR** |
NT_DATE | **NT_ABSTRACT** | **NT_COPYRIGHT** | **NT_TITLE**

NT_ABSTRACT → **NT_TITLE** | **NT_TITLE PARAS**

NT_TITLE → <title>**CONT_TITLE**</title>

CONT_TITLE → **ELEM_TITLE** | **ELEM_TITLE CONT_TITLE**

ELEM_TITLE → Cadena | **NT_EMPHASIS** | **NT_LINK** | **NT_EMAIL**

PARAS → **NT_PARA** | **NT_SIMPARA** | **NT_PARA PARAS** | **NT_SIMPARA PARAS**

NT_PARA → <para>**CONT_PARA**</para>

CONT_PARA → **ELEM_PARA** | **ELEM_PARA CONT_PARA**

ELEM_PARA → Cadena | NT_EMPHASIS | NT_LINK | NT_EMAIL | NT_AUTHOR | NT_COMMENT | NT_ITEMIZEDLIST | NT_IMPORTANT | NT_ADDRESS | NT_MEDIAOBJECT | NT_INFORMALTABLE

NT_ITEMIZEDLIST → <itemizedlist>LISTITEM</itemizedlist>

NT_MEDIAOBJECT → <mediaobject>NT_INFO CONT_MEDIAOBJECT</mediaobject>| <mediaobject>CONT_MEDIAOBJECT</mediaobject>

CONT_MEDIAOBJECT → NT_IMAGEOBJECT | NT_VIDEOOBJECT | NT_IMAGEOBJECT CONT_MEDIAOBJECT | NT_VIDEOOBJECT CONT_MEDIAOBJECT

NT_IMAGEOBJECT → <imageobject>NT_INFO<imagedata fileref=”URL”/> | <imagedata fileref=”URL” /> </imageobject>

NT_VIDEOOBJECT → <videoobject>NT_INFO<videodata fileref=”URL”/> | <videodata fileref=”URL” /></videoobject>

LISTITEM → CONT_1 | CONT_1 LISTITEM

NT_AUTHOR → <author>CONT_AUTHOR</author>

CONT_AUTHOR → NT_FIRSTNAME | NT_SURNAME | NT_EMAIL | NT_FIRSTNAME NT_SURNAME | NT_FIRSTNAME NT_EMAIL | NT_SURNAME NT_EMAIL | NT_FIRSTNAME NT_SURNAME NT_EMAIL

CONT_SS → CONT_1 | CONT_1 CONT_SS

NT_ADDRESS → <address></address> | <address>(Cadena | CONT_ADDRESS)</address>

CONT_ADRESS → NT_STREET CONT_ADRESS | NT_CITY CONT_ADRESS | NT_STATE CONT_ADRESS | NT_PHONE CONT_ADRESS | NT_EMAIL CONT_ADRESS | lambda

NT_COPYRIGHT → <copyright>NT_YEAR</copyright> | <copyright>NT_YEAR NT_HOLDER</copyright>

CONT_2 → Cadena | NT_EMPHASIS | NT_LINK | NT_EMAIL | NT_AUTHOR | NT_COMMENT

NT_SIMPARA → <simpara>CONT_SECL</simpara>

NT_EMPHASIS → <emphasis>CONT_SECL</emphasis>

NT_COMMENT → <comment>CONT_SECL</comment>

NT_LINK → <link xlink:href=”NT_URL”>CONT_SECL</link>

CONT_SECL → CONT_2 | CONT_2 CONT_SECL

NT_IMPORTANT → <important>NT_TITLE CONT_IMPORTANT</important> | <important>CONT_IMPORTANT</important>

CONT_IMPORTANT → CONT_1 | CONT_1 CONT_IMPORTANT

CONT_3 → Cadena | NT_LINK | NT_EMPHASIS | NT_COMMENT

CONT_VAR → CONT_3 | CONT_3 CONT_VAR

NT_FIRSTNAME → <firstname>CONT_VAR</firstname>

NT_SURNAME → <surname>CONT_VAR</surname>

NT_STREET → <street>CONT_VAR</street>

NT_CITY → <city>CONT_VAR</city>

NT_STATE → <state>CONT_VAR</state>

NT_PHONE → <phone>CONT_VAR</phone>

NT_EMAIL → <email>CONT_VAR</email>

NT_DATE → <date>CONT_VAR</date>

NT_YEAR → <year>CONT_VAR</year>

NT_HOLDER → <holder>CONT_VAR</holder>

NT_INFORMALTABLE → <informaltable>TABLE_MEDIA</informaltable> |
<informaltable>TABLE_GROUP</informaltable>

TABLE_MEDIA → NT_MEDIAOBJECT | NT_MEDIAOBJECT TABLE_MEDIA

TABLE_GROUP → NT_TGROUP | NT_TGROUP TABLE_GROUP

NT_TGROUP → <tgroup>NT_THEAD NT_TFOOT NT_TBODY</tgroup> |
<tgroup>NT_THEAD NT_TBODY</tgroup> |
<tgroup>NT_TFOOT NT_BODY</tgroup> | <tgroup>NT_TBODY</tgroup>

CONT_T → NT_ROW | NT_ROW CONT_T

NT_THEAD → <thead>CONT_T</thead>

NT_TFOOT → <tfoot>CONT_T</tfoot>

NT_TBODY → <tbody>CONT_T</tbody>

NT_ROW → <row>CONT_ROW_1</row> | <row>CONT_ROW_2</row>

CONT_ROW_1 → NT_ENTRY | NT_ENTRY CONT_ROW_1

CONT_ROW_2 → NT_ENTRYTBL | NT_ENTRYTBL CONT_ROW_2

NT_ENTRYTBL → <entrytbl>NT_THEAD NT_TBODY</entrytbl> |
<entrytbl>NT_TBODY</entrytbl>

NT_ENTRY → <entry>CONT_ENTRY</entry>

CONT_ENTRY → Cadena | Cadena CONT_ENTRY | NT_ITEMIZEDLIST |
NT_ITEMIZEDLIST CONT_ENTRY | NT_IMPORTANT | NT_PARA | NT_SIMPARA |
NT_IMPORTANT CONT_ENTRY | NT_PARA CONT_ENTRY | NT_COMMENT |
NT_SIMPARA CONT_ENTRY | NT_COMMENT CONT_ENTRY | NT_ABSTRACT |
NT_ABSTRACT CONT_ENTRY | NT_MEDIAOBJECT CONT_ENTRY |
NT_MEDIAOBJECT

NT_URL → PROTOCOL://DOM:PORT/ROUTE#FRAGMENT | PROTOCOL://DOM:PORT |
PROTOCOL://DOM/ROUTE | PROTOCOL://DOM#FRAGMENT |

PROTOCOL://DOM:PORT/ROUTE | **PROTOCOL://DOM:PORT#FRAGMENT** | **PROTOCOL://DOM/ROUTE#FRAGMENT** | Url

PROTOCOL → http | https | ftp | ftps

DOM → Url

PORT → Url

ROUTE → Url

FRAGMENT → Url

Conclusión

Bibliografía

https://tdg.docbook.org/tdg/4.5/docbook.html