3. XML en el *Corpus Histórico del Español en México*

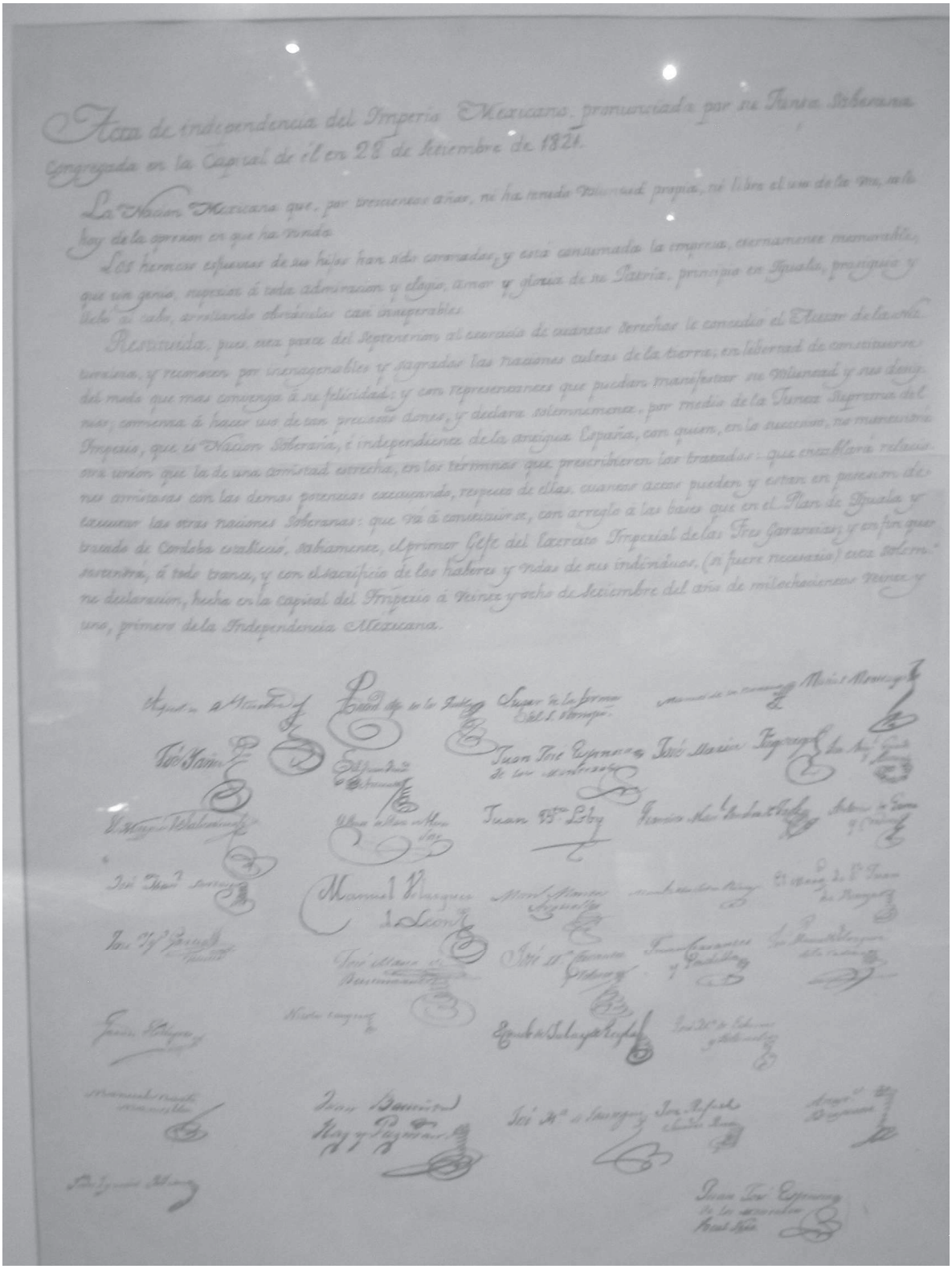
En este capítulo se muestran aspectos básicos de la utilización de XML en la codificación del *Corpus Histórico del Español en México*, al que esta tesis se enfoca. Esto es importante para describir el contexto en el que se desarrolla el constructor automático de la base bibliográfica de dicho corpus. En esencia, se presentará un documento XML, el esquema XML que lo valida y que el constructor analizará para construir la base de datos y la porción del documento XML que servirá para poblar la base de datos creada.

Los documentos XML, como se dijo arriba, están formados de etiquetas con picoparéntesis y se ven como en la figura 3.1, que contiene el Acta de Independencia del Imperio Mexicano (1821) ya codificada. El documento original se muestra en la figura 3.2.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text/xsl" href="estilo.xsl"?>

<documento      xmlns="http://www.ii.unam.mx"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://www.ii.unam.mx esquema.xsd">

    <encabezado id="0" permisos="todos">

        <titulo>Acta de independencia del Imperio Mexicano</titulo>

        <parametros     generoLiterario="prosa" registro="estándar">
            <corpus><CHEM   zonaGeografica="Altiplano Central" areaTematica="derecho"/></corpus>
            <hablante   genero="grupo masculino"/>
        </parametros>

        <institucion>Junta Soberana del Imperio Mexicano</institucion>
        <referencia><otro tipo="acta" pgs="1"/></referencia>
        <ciudadPublicacion>ciudad de México</ciudadPublicacion>
        <fechaPublicacion>1821</fechaPublicacion>
        <imagen origen="internet" ancho="400" alto="532">Independencia-mx-acta.jpg</imagen>

        <responsables>
            <transcriptor nombres="" apellidos="" fecha="noviembre 2007"/>
            <etiquetador nombres="" apellidos="" fecha="noviembre 2007"/>
            <revisor nombres="" apellidos="" fecha="noviembre 2007"/>
        </responsables>

    </encabezado>

<cuerpo tipoFuente="Espinosa">
<bastardillas>
<seccion>
<tituloSecc>
    <g>Acta</g><e/>
    <g>de</g><e/>
    <g>independencia</g><e/>
    <g>del</g><e/>
    <g>Imperio</g><e/>
    <g>Mexicano</g><r c="Fc">,</r><e/>
    <g>pronunciada</g><e/>
    <g>por</g><e/>
```

Figura 3.1. Acta de Independencia codificada en XML

Como se observa en la figura 3.1, el documento está formado por dos elementos principales: un encabezado y un cuerpo. Por un lado, el encabezado contiene generalidades del documento, como sus características de clasificación en el corpus (su género literario, registro dialectal, tipo de hablantes, etc.) la información bibliográfica (referencias, lugar y fecha de publicación, etc.) que servirá para poblar la base de datos que generará el constructor de esta tesis y datos sobre los diversos responsables del documento en cuestión. Por el otro, el



Acta de independencia del Imperio Mexicano, pronunciada por su Junta Soberana congregada en la Capital de él en 28 de Setiembre de 1821.

La Nacion Mexicana que, por trescientos años, no ha tenido voluntad propia, ni libre el uso de la voz, sale hoy de la opresion en que ha vivido.

Los heroicos esfuerzos de sus hijos han sido coronados, y está consumada la empresa, eternamente memorable, que un genio, superior á toda admiracion y elogio, amor y gloria de su Patria, principió en Iguala, prosiguió y llevó al cabo, arrollando obstáculos casi insuperables.

Restituida, pues, esta parte del Septentrion al exercicio de cuantos derechos le concedió el Eterno en la sociedad humana, y reconoce por inenagenables y sagrados las naciones cultas de la tierra; en libertad de constituirse del modo que mas convenga á su felicidad; y con representantes que puedan manifestar su voluntad y sus designios; comienza á hacer uso de tan preciosos dones, y declara solemnemente, por medio de la Junta Suprema del Imperio, que es Nacion Soberana, é independiente de la antigua España, con quien, en la sucesion, no mantendrá otra union que la de una amistad estrecha, en los términos que prescribieren los tratados: que entablará relaciones amistosas con las demas potencias executando, respecto de ellas, cuantos actos pueden y estan en practica de executar las otras naciones Soberanas: que va á constituirse, con arreglo á las bases que en el Plan de Iguala y tratado de Cordoba acordó, sabiamente, el primer Gefe del Exército Imperial de las Tres Garantías; y en fin que sostendrá, á todo trance, y con el sacrificio de los haberes y vidas de sus individuos, (si fuere necesario) esta solemne declaracion, hecha en la capital del Imperio á veinte y ocho de Setiembre del año de mil ochocientos veinte y uno, primero de la Independencia Mexicana.

Figura 3.2. Imagen del Acta de Independencia del Imperio Mexicano (1821)

cuerpo contiene el documento mismo, es decir la secuencia específica de “tokens” u ocurrencias de palabras con las particularidades tipográficas del original (tipo de fuente, letras cursivas, bastardillas, versales, negritas, etc.), así como títulos, subtítulos, firmas y errores del original, todo codificado como ya se dijo en XML mediante elementos y atributos diseñados específicamente para este corpus. Todos estos elementos y atributos se validan mediante el esquema XML, parte del cual aparece en la figura 3.3.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <xsd:element name="documento">
        <xsd:complexType>
            <xsd:sequence>
                <xsd:element ref="encabezado" minOccurs="1" maxOccurs="1"/>
                <xsd:element ref="cuerpo" minOccurs="1" maxOccurs="1"/>
            </xsd:sequence>
        </xsd:complexType>
    </xsd:element>

    <xsd:element name="corpus">
        <xsd:complexType>
            <xsd:choice minOccurs="1" maxOccurs="1">
                <xsd:element ref="CHEM" minOccurs="1" maxOccurs="1"/>
                <xsd:element ref="COCIEM" minOccurs="1" maxOccurs="1"/>
                <xsd:element ref="CSMX" minOccurs="1" maxOccurs="1"/>
                <xsd:element ref="CLI" minOccurs="1" maxOccurs="1"/>
            </xsd:choice>
        </xsd:complexType>
    </xsd:element>
    <xsd:element name="CHEM">
        <xsd:complexType>
            <xsd:attribute name="zonaGeografica" use="required">
                <xsd:simpleType>
                    <xsd:restriction base="xsd:token">
                        <xsd:enumeration value="Altiplano Central"/>
                        <xsd:enumeration value="Noroeste"/>
                        <xsd:enumeration value="Noreste"/>
                        <xsd:enumeration value="Golfo"/>
                        <xsd:enumeration value="Sureste"/>
                        <xsd:enumeration value="Oaxaca"/>
                        <xsd:enumeration value="España"/>
                    </xsd:restriction>
                </xsd:simpleType>
            </xsd:attribute>
            <xsd:attribute name="areaTematica" use="required">
                <xsd:simpleType>
                    <xsd:restriction base="xsd:token">
                        <xsd:enumeration value="literatura"/>
                        <xsd:enumeration value="derecho"/>
```

Figura 3.3. Esquema XML para los documentos del Corpus Histórico

En este esquema se puede ver cómo el documento está definido mediante dos elementos, el encabezado y el cuerpo. De hecho, más debajo de las definiciones de los elementos “corpus” y “CHEM” se encuentran las definiciones de “encabezado” y “cuerpo” y todos los elementos que contienen.

Para visualizar o presentar la información de los documentos XML de tal manera que puedan ser leídos como un texto de usuario final, se usan las hojas de estilo *XSLT*. Una hoja de estilo permite que el autor o el receptor del documento XML puedan definir cómo se tiene que mostrar el contenido del documento. Se puede decir que la hoja de estilo contiene una serie de instrucciones para reescribir el documento y presentarlo de una forma amigable a la vista de la persona que quiera leerlo. Con la hoja de estilo diseñada específicamente para este corpus el Acta de Independencia se visualiza como en la figura 3.4.

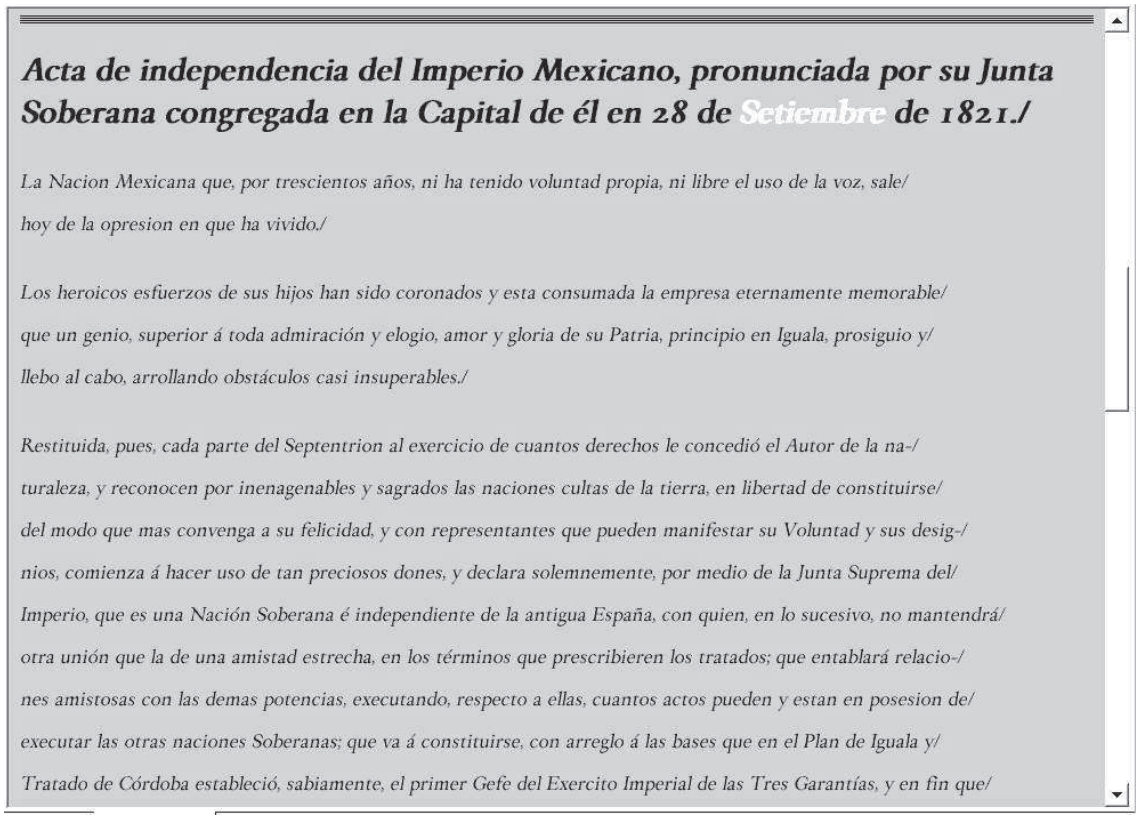
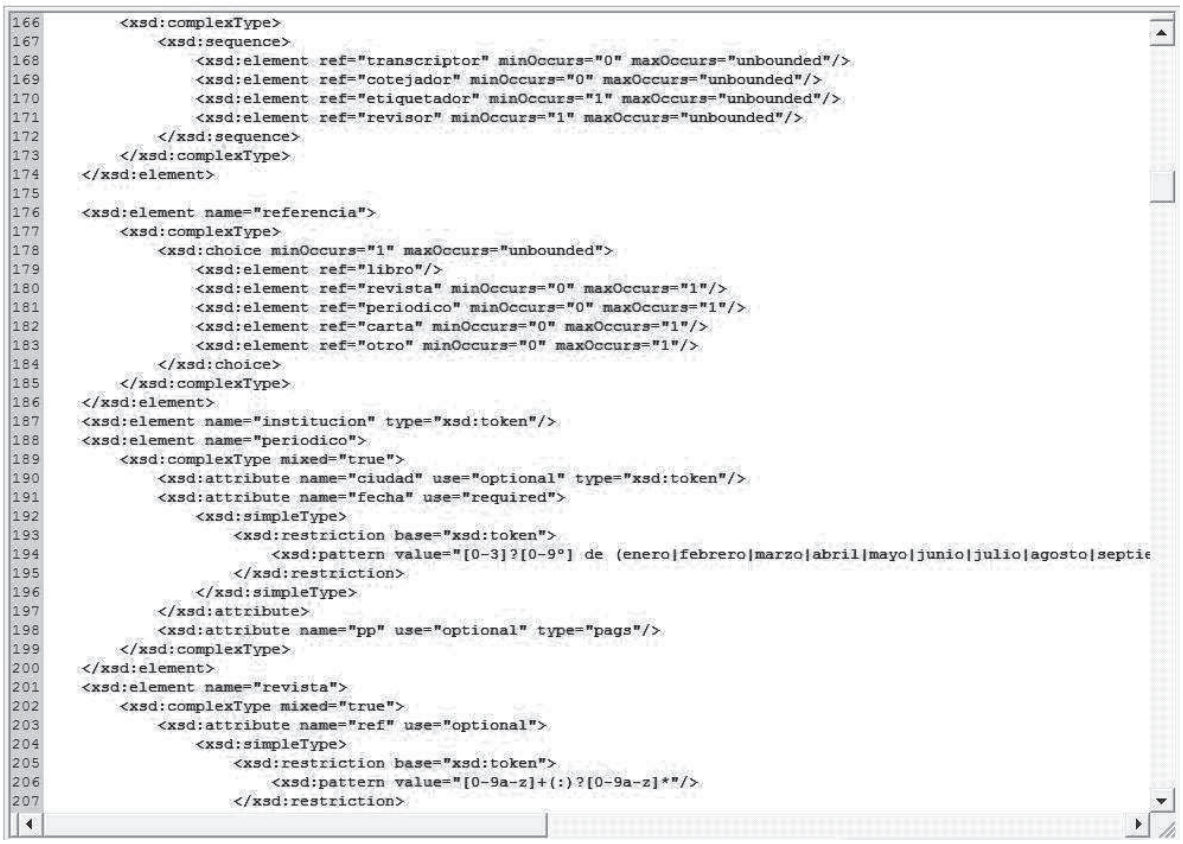


Figura 3.4. Visualización de documento XML del CHEM (Acta de Independencia)

En este contexto, la tarea del constructor automático de la base de datos de esta tesis es analizar la parte del esquema XML que valida la información bibliográfica de cada documento del corpus. La idea es generar las instrucciones para crear la tabla SQL a partir de este análisis. En el siguiente capítulo, veremos las particularidades de SQL para llevar esto a cabo. Finalmente, esta tabla se puebla con la información específica de cada documento del corpus (lo que se verá en la segunda parte del capítulo 6). En la figura 3.5 se muestra la porción del esquema XML que define el elemento “referencia”, que contiene la estructura de las referencias bibliográficas del corpus.



```
166         <xsd:complexType>
167             <xsd:sequence>
168                 <xsd:element ref="transcriptor" minOccurs="0" maxOccurs="unbounded"/>
169                 <xsd:element ref="cotejador" minOccurs="0" maxOccurs="unbounded"/>
170                 <xsd:element ref="etiquetador" minOccurs="1" maxOccurs="unbounded"/>
171                 <xsd:element ref="revisor" minOccurs="1" maxOccurs="unbounded"/>
172             </xsd:sequence>
173         </xsd:complexType>
174     </xsd:element>
175
176     <xsd:element name="referencia">
177         <xsd:complexType>
178             <xsd:choice minOccurs="1" maxOccurs="unbounded">
179                 <xsd:element ref="libro"/>
180                 <xsd:element ref="revista" minOccurs="0" maxOccurs="1"/>
181                 <xsd:element ref="periodico" minOccurs="0" maxOccurs="1"/>
182                 <xsd:element ref="carta" minOccurs="0" maxOccurs="1"/>
183                 <xsd:element ref="otro" minOccurs="0" maxOccurs="1"/>
184             </xsd:choice>
185         </xsd:complexType>
186     </xsd:element>
187     <xsd:element name="institucion" type="xsd:token"/>
188     <xsd:element name="periodico">
189         <xsd:complexType mixed="true">
190             <xsd:attribute name="ciudad" use="optional" type="xsd:token"/>
191             <xsd:attribute name="fecha" use="required">
192                 <xsd:simpleType>
193                     <xsd:restriction base="xsd:token">
194                         <xsd:pattern value="[0-3]?[0-9°] de (enero|febrero|marzo|abril|mayo|junio|julio|agosto|septie
195                     </xsd:restriction>
196                 </xsd:simpleType>
197             </xsd:attribute>
198             <xsd:attribute name="pp" use="optional" type="pags"/>
199         </xsd:complexType>
200     </xsd:element>
201     <xsd:element name="revista">
202         <xsd:complexType mixed="true">
203             <xsd:attribute name="ref" use="optional">
204                 <xsd:simpleType>
205                     <xsd:restriction base="xsd:token">
206                         <xsd:pattern value="[0-9a-z]+(:)?[0-9a-z]*"/>
207                     </xsd:restriction>
```

Figura 3.5. Porción del esquema XML para los documentos del corpus que define el elemento "referencia", objetivo principal del constructor.