DCT Schema Overview

I found the config file and consequently the folder containing information about DTD within prod. I shared that with you so please look inside for relevant information. Below is some info that our Bookloader does.

This is generated from AI so take it with a grain of salt!

1) Processing Flow (Functions & Order)

prepareBookXml(baseDestDirName) // parses metadata, validates against DTD, prepares book XML

processRules() // applies rules (addRISInfo, addRISIndex*, linkDisease, linkDrug, addPMID, loadContent) and writes final output

2) Core Classes & Methods (Explicit)

Class / File	Method(s)	Role in Conversion
Main.java	prepareContentFiles()	Detects input types; calls EPUBParser for .epub; stages resources.
Main.java	prepareBookXml()	Parses/collects book metadata (ISBN, title, authors); validates against DTD; preps for rules.
Main.java	processRules()	Applies ris_rules.xml: tagging, medical term linking, RIS index, PMID; generates final XML output.

3) Primary RIS Book XML (Sample Output of processRules())

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE book SYSTEM "RittDocBook.dtd">
<book xmlns="http://www.rittenhouse.com/dtd/v1.1">
  <!-- Book Metadata -->
  <title>Clinical Medicine Handbook</title>
  <authors>
    <author>Dr. John Smith, MD</author>
    <author>Dr. Jane Doe, PhD</author>
  </authors>
```

```
<isbn>9781234567890</isbn>

<!-- Book Information Block -->
<bookinfo>
  <title>Clinical Medicine Handbook</title>
  <isbn>9781234567890</isbn>
  <authorgroup>
    <author>
      <personname>
        <firstname>John</firstname>
        <surname>Smith</surname>
        <degree>MD</degree>
      </personname>
    </author>
  </authorgroup>
  <publisher>
    <publishername>Medical Publishers Inc</publishername>
  </publisher>
  <pubdate>October 2023</pubdate>
  <edition>Third Edition</edition>
  <copyright>
    <year>2023</year>
    <holder>Medical Publishers Inc</holder>
  </copyright>
</bookinfo>

<!-- Chapter Content -->
<chapters>
  <chapter id="chapter-001">
    <title>Introduction to Clinical Medicine</title>
    <content>
      This chapter covers the fundamental principles of clinical medicine,
      including diagnostic approaches and treatment methodologies...
    </content>
    <!-- RIS Index Tags (added by processing rules) -->
    <risindex>
      <risrule>addRISIndex1</risrule>
      <risterm>clinical medicine</risterm>
      <risweight>1.0</risweight>
    </risindex>
  </chapter>

  <chapter id="chapter-002">
```

```
        <title>Cardiovascular Diseases</title>
        <content>
          Cardiovascular diseases are among the leading causes of mortality.
          Common conditions include <ulink type="disease" id="123">hypertension</ulink>
          and <ulink type="disease" id="456">myocardial infarction</ulink>.
          Treatment often involves <ulink type="drug" id="789">aspirin</ulink> therapy.
        </content>
        <!-- Medical Term Links (added by linking rules) -->
        <risindex>
          <risrule>linkDisease</risrule>
          <risterm>hypertension</risterm>
          <risid>123</risid>
        </risindex>
      </chapter>
    </chapters>
</book>
```

4) DTD Definition (RittDocBook.dtd excerpt)

```
<!-- Root Elements -->
<!ELEMENT book (title?, authors?, isbn?, bookinfo?, chapters?)>
<!ATTLIST book
   xmlns CDATA #FIXED "http://www.rittenhouse.com/dtd/v1.1">

<!-- Basic Metadata -->
<!ELEMENT title (#PCDATA)>
<!ELEMENT authors (author+)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT isbn (#PCDATA)>

<!-- Detailed Book Information -->
<!ELEMENT bookinfo (title?, isbn?, authorgroup?, publisher?, pubdate?, edition?, copyright?,
legalnotice?)>

<!ELEMENT authorgroup (author+ | editor+)>
<!ELEMENT author (personname)>
<!ELEMENT editor (personname)>
<!ELEMENT personname (firstname?, surname?, degree?)>
<!ELEMENT firstname (#PCDATA)>
<!ELEMENT surname (#PCDATA)>
<!ELEMENT degree (#PCDATA)>

<!ELEMENT publisher (publishername)>
<!ELEMENT publishername (#PCDATA)>
```

```
<!ELEMENT pubdate (#PCDATA)>
<!ELEMENT edition (#PCDATA)>

<!ELEMENT copyright (year+, holder?)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT holder (#PCDATA)>

<!ELEMENT legalnotice (para+)>
<!ELEMENT para (#PCDATA | emphasis | ulink | biblioid)*>

<!-- Chapter Structure -->
<!ELEMENT chapters (chapter*)>
<!ELEMENT chapter (title?, content?, risindex*, sect1*)>
<!ATTLIST chapter id CDATA #REQUIRED>

<!ELEMENT content (#PCDATA | emphasis | ulink)*>

<!-- Section Hierarchy -->
<!ELEMENT sect1 (title?, para*, sect2*, risindex*)>
<!ATTLIST sect1 id CDATA #IMPLIED>

<!ELEMENT sect2 (title?, para*, sect3*)>
<!ELEMENT sect3 (title?, para*, sect4*)>
<!ELEMENT sect4 (title?, para*)>

<!-- Inline Elements -->
<!ELEMENT emphasis (#PCDATA | emphasis)*>

<!-- Medical Term Links -->
<!ELEMENT ulink (#PCDATA)>
<!ATTLIST ulink
  type (disease|drug|keyword) #REQUIRED
  id CDATA #REQUIRED
  url CDATA #IMPLIED>

<!-- RIS Index Tags -->
<!ELEMENT risindex (risrule, risterm?, risid?, risweight?)>
<!ELEMENT risrule (#PCDATA)>
<!ELEMENT risterm (#PCDATA)>
<!ELEMENT risid (#PCDATA)>
<!ELEMENT risweight (#PCDATA)>

<!-- Bibliography Support -->
```

```
<!ELEMENT bibliomixed (#PCDATA | title | volumenum | artpagenums | authorgroup |
biblioid)*>
<!ELEMENT volumenum (#PCDATA)>
<!ELEMENT artpagenums (#PCDATA)>

<!ELEMENT biblioid (#PCDATA)>
<!ATTLIST biblioid
  class CDATA #IMPLIED
  otherclass CDATA #IMPLIED>
```

5) XSD (Alternative to DTD)

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
        targetNamespace="http://www.rittenhouse.com/dtd/v1.1"
        xmlns:rh="http://www.rittenhouse.com/dtd/v1.1"
        elementFormDefault="qualified">

  <xs:element name="book" type="rh:BookType"/>

  <xs:complexType name="BookType">
    <xs:sequence>
      <xs:element name="title" type="xs:string" minOccurs="0"/>
      <xs:element name="authors" type="rh:AuthorsType" minOccurs="0"/>
      <xs:element name="isbn" type="xs:string" minOccurs="0"/>
      <xs:element name="bookinfo" type="rh:BookInfoType" minOccurs="0"/>
      <xs:element name="chapters" type="rh:ChaptersType" minOccurs="0"/>
    </xs:sequence>
    <xs:attribute name="xmlns" type="xs:string"
fixed="http://www.rittenhouse.com/dtd/v1.1"/>
  </xs:complexType>

  <xs:complexType name="AuthorsType">
    <xs:sequence>
      <xs:element name="author" type="xs:string" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="BookInfoType">
    <xs:sequence>
      <xs:element name="title" type="xs:string" minOccurs="0"/>
      <xs:element name="isbn" type="xs:string" minOccurs="0"/>
      <xs:element name="authorgroup" type="rh:AuthorGroupType" minOccurs="0"/>
      <xs:element name="publisher" type="rh:PublisherType" minOccurs="0"/>
```

```
      <xs:element name="pubdate" type="xs:string" minOccurs="0"/>
      <xs:element name="edition" type="xs:string" minOccurs="0"/>
      <xs:element name="copyright" type="rh:CopyrightType" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="AuthorGroupType">
    <xs:choice maxOccurs="unbounded">
      <xs:element name="author" type="rh:PersonType"/>
      <xs:element name="editor" type="rh:PersonType"/>
    </xs:choice>
  </xs:complexType>

  <xs:complexType name="PersonType">
    <xs:sequence>
      <xs:element name="personname" type="rh:PersonNameType"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="PersonNameType">
    <xs:sequence>
      <xs:element name="firstname" type="xs:string" minOccurs="0"/>
      <xs:element name="surname" type="xs:string" minOccurs="0"/>
      <xs:element name="degree" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="PublisherType">
    <xs:sequence>
      <xs:element name="publishername" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="CopyrightType">
    <xs:sequence>
      <xs:element name="year" type="xs:string" maxOccurs="unbounded"/>
      <xs:element name="holder" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="ChaptersType">
    <xs:sequence>
      <xs:element name="chapter" type="rh:ChapterType" maxOccurs="unbounded"/>
```

```
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="ChapterType">
    <xs:sequence>
      <xs:element name="title" type="xs:string" minOccurs="0"/>
      <xs:element name="content" type="rh:MixedContentType" minOccurs="0"/>
      <xs:element name="risindex" type="rh:RISIndexType" minOccurs="0"
maxOccurs="unbounded"/>
      <xs:element name="sect1" type="rh:SectionType" minOccurs="0"
maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="id" type="xs:string" use="required"/>
  </xs:complexType>

  <xs:complexType name="MixedContentType" mixed="true">
    <xs:choice minOccurs="0" maxOccurs="unbounded">
      <xs:element name="emphasis" type="rh:EmphasisType"/>
      <xs:element name="ulink" type="rh:ULinkType"/>
    </xs:choice>
  </xs:complexType>

  <xs:complexType name="ULinkType" mixed="true">
    <xs:attribute name="type" use="required">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="disease"/>
          <xs:enumeration value="drug"/>
          <xs:enumeration value="keyword"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="id" type="xs:string" use="required"/>
    <xs:attribute name="url" type="xs:string"/>
  </xs:complexType>

  <xs:complexType name="RISIndexType">
    <xs:sequence>
      <xs:element name="risrule" type="xs:string"/>
      <xs:element name="risterm" type="xs:string" minOccurs="0"/>
      <xs:element name="risid" type="xs:string" minOccurs="0"/>
      <xs:element name="risweight" type="xs:decimal" minOccurs="0"/>
    </xs:sequence>
```

```
  </xs:complexType>

  <xs:complexType name="EmphasisType" mixed="true">
    <xs:sequence>
      <xs:element name="emphasis" type="rh:EmphasisType" minOccurs="0"
maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="SectionType">
    <xs:sequence>
      <xs:element name="title" type="xs:string" minOccurs="0"/>
      <xs:element name="para" type="rh:MixedContentType" minOccurs="0"
maxOccurs="unbounded"/>
      <xs:element name="risindex" type="rh:RISIndexType" minOccurs="0"
maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="id" type="xs:string"/>
  </xs:complexType>
</xs:schema>
```

6) Processing Rules (ris_rules.xml)

```
<?xml version="1.0" encoding="UTF-8"?>
<rules>
  <!-- Content Enhancement Rules -->
  <rule type="addRISInfo" action="apply"/>
  <rule type="addRISIndex1" action="apply"/>
  <rule type="addRISIndex2" action="apply"/>
  <rule type="addRISIndex3" action="apply"/>

  <!-- Medical Term Linking -->
  <rule type="linkDisease" action="apply"/>
  <rule type="linkDrug" action="apply"/>
  <rule type="linkKeyword" action="apply"/>

  <!-- Bibliography Enhancement -->
  <rule type="addPMID" action="apply"/>

  <!-- Content Loading -->
  <rule type="loadContent" action="apply"/>
</rules>
```

7) Configuration Schema (RISBackend.cfg)

```
# Input/Output Directories
RIS.CONTENT_IN=/input/epub/books/
RIS.CONTENT_TEMP=/temp/processing/
RIS.CONTENT_OUT=/output/processed/xml/
RIS.CONTENT_MEDIA=/web/media/files/

# DTD Configuration
RIS.RITTENHOUSE_DTD_PATH=/dtd/schemas/

# Database Configuration
RISDB.URL=jdbc:sqlserver://server:1433;database=RIS
RISDB.USERNAME=risuser
RISDB.PASSWORD=rispass

# XML Database (TextML)
XMLDB.DEFAULT_COLLECTION_NAME=medical_books
RIS.SKIP_TEXTML=false

# Alternative Output Path
RIS.LOAD_CONTENT_TO_NON_TEXTML_PATH=true
RIS.DEST_NON_TEXTML_CONTENT_PATH=/alternative/output/path/

# Threading Configuration
RIS.THREAD_COUNT_MAXIMUM=8
RIS.LINKING_THREAD_COUNT_MAXIMUM=4

# External APIs
NCBI.EUTILITIES_URL=https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi
NCBI.API_KEY=your_ncbi_api_key_here
```

8) Data Flow Structures

EPUB Package Structure

```
medical_book.epub (ZIP format)
├── META-INF/
│   └── container.xml
├── OEBPS/
│   ├── content.opf
│   ├── chapter01.xhtml
│   ├── chapter02.xhtml
│   └── images/
└── mimetype
```

/output/processed/xml/9781234567890/ (Processed Output)
├── book.9781234567890.xml
├── sect1.chapter001.xml
├── sect1.chapter002.xml
├── toc.9781234567890.xml
└── images/
  ├── figure01.jpg
  └── diagram02.png

Database Integration (illustrative tables)

```
CREATE TABLE resources (
  resource_id INT PRIMARY KEY,
  isbn VARCHAR(20),
  title VARCHAR(500),
  authors TEXT,
  publisher_id INT,
  pub_date DATE,
  edition VARCHAR(50),
  copyright_info TEXT
);

CREATE TABLE keyword_resources (
  keyword_id INT,
  resource_id INT,
  chapter_id VARCHAR(100),
  weight DECIMAL(3,2)
);
```

9) End-to-End Workflow (As Implemented)

Phase 1: EPUB Ingestion — EPUB file → EPUBParser.extractEPUB() → temp directory

Phase 2: Content Conversion — EPUB structure → EPUBParser.parseContentOPF() → base RIS XML

Phase 3: Content Enhancement — base XML → processRules() → medical term linking & RIS index tags

Phase 4: Output & Integration — enhanced XML → file system output and (where applicable) DB writes