CS 501- Software Engineering

Legal Data Markup Software
DTD Design Document

Version 1.0

Document Revision History

Date	Version	Description	Author
11/27/00	1.0	Draft for Delivery	LDMS Team

Table of Contents

1. Introduction 4
- 1.1 Purpose 4
- 1.2 Scope 4
- 1.3 Definitions, Acronyms and Abbreviations 4
- 1.4 References 5
- 1.5 Overview 5
- 1.6 Roles and Responsibilities 5

2. The Problem 6
- 2.1 Example of Variation in Input ASCII 6
- 2.2 Consistencies in input ASCII - Dashlines 7

3. DTD Design 7
- 3.1 XML Tag Descriptions 7
  - 3.1.1 LDMS 7
  - 3.1.2 STRUCTDIV 8
  - 3.1.3 TITLEDATA 8
  - 3.1.4 NAVGROUP 8
  - 3.1.5 CITE and DATE 9
  - 3.1.6 EXPCITE, DIVEXPCITE, and HEAD 9
  - 3.1.7 STATGROUP, STATUTE, SOURCE, DIVSOURCE, and STATAMEND 9
  - 3.1.8 MISC1, REFTEXT, MISC2, COD, MISC3, CHANGE, MISC4, TRANS, MISC5, EXEC, MISC6, CROSS, MISC7, SECREF, and MISC8 10
  - 3.1.9 DATATEXT, DATATEXTNAME, DIVDATATEXT, PRE 11
  - 3.1.10 XREF 12
  - 3.1.11 FOOTNOTE and FOOTREF 12
  - 3.1.12 TABLE, TABLENAME, and FIELD 12
- 3.2 UML Diagram 15

4. LDMS DTD 16

DTD Design Document

1. Introduction

The intent of this project is to create a software tool that will convert the US Code of law from its distribution ASCII format into well-formed, valid XML. The XML output would subsequently be utilized by our client, the Legal Information Institute, in next-generation applications that will make the U.S. Code available in a variety of different formats to the general public. Examples of such use include the electronic publication of the code on the Internet and downloadable versions in Folio Views format.

1.1 Purpose

The purpose of this document is to describe the design of the Document Type Definition (DTD). The DTD is crucial to the LDMS project, and its structure is hard coded into the LDMS software.

1.2 Scope

This document applies only to the LDMS DTD.

1.3 Definitions, Acronyms and Abbreviations

DTD	Document Type Definition
LDMS	Legal Data Markup Software
LII	Legal Information Institute
HTML	Hyper Text Markup Language
W3C	World Wide Web Consortium
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
PDD	Program Design Document
DDD	DTD Design Document

1.4 References

Developers may find the following documents useful:

- http://uscode.house.gov/download.htm – U.S. Code related input formats.
- http://www.w3.org/TR/REC-xml – The W3C XML 1.0 Draft Specification (2nd Edition).
- http://uscode.house.gov/uschelp.htm dashline explanations.

1.5 Overview

This document is aimed primarily at developers working directly on the LDMS DTD. To that end, it shall document the structure of the DTD, and will provide a high-level overview of how the DTD represents the US Code.

1.6 Roles and Responsibilities

Name	Department	Responsibility
Thomas Bruce	Legal Information Institute	Project Sponsor
William Arms	Computer Science Department	Project Sponsor
Amy Siu	Computer Science Department	Project Reviewer
Ju Joh	Computer Science Department	Student Developer
Sylvia Kwakye	Computer Science Department	Student Developer
Jason Lee	Computer Science Department	Student Developer
Nidhi Loyalka	Computer Science Department	Student Developer
Omar Mehmood	Computer Science Department	Student Developer
Charles Shagong	Computer Science Department	Student Developer
Brian Williams	Computer Science Department	Student Developer

2. The Problem

The LII wishes to be able to convert the existing US Code from the House of Representatives into XML. However, there are many variations in structure from Title to Title, and the LII wishes to keep most, if not all of the functionality present in the current HTML version.

2.1 Example of Variation in Input ASCII

In Title 11 we see:

```
-CITE-
    11 USC Sec. 506                                  01/23/00
-EXPCITE-
    TITLE 11 – BANKRUPTCY
    CHAPTER 5 - CREDITORS, THE DEBTOR, AND THE ESTATE
    SUBCHAPTER I - CREDITORS AND CLAIMS
-HEAD-
    Sec. 506. Determination of secured status
```

In Title 46 we see:

```
-CITE-
    46 USC Sec. 13102                                01/05/99
-EXPCITE-
    TITLE 46 – SHIPPING
    Subtitle II - Vessels and Seamen
    Part I - State Boating Safety Programs
    CHAPTER 131 - RECREATIONAL BOATING SAFETY
-HEAD-    Sec. 13102. Program acceptance
```

Notice that Titles vary in their internal structures. Title 11 is divided into Chapters, then Subchapters, then Sections whereas Title 46 is divided into Subtitles, Parts, Chapters, then Sections.

2.2 Consistencies in input ASCII - Dashlines

The words surrounded by dashes, in this example –CITE-, -EXPCITE-, and –HEAD-, although not part of the actual legal data, are consistent throughout all titles in the ASCII input. The LDMS team calls these words dashlines. The meaning for each dashline is defined at http://uscode.house.gov/uschelp.htm The dashlines always appear in the specific order CITE, EXPCITE, HEAD, STATUTE, SOURCE, STATAMEND, TEXT, MISC1, REFTEXT, MISC2, COD, MISC3, CHANGE, MISC4, TRANS, MISC5, EXEC, MISC6, CROSS, MISC7, SECREF, MISC8, NOTES; however, not all dashlines are required in a given sequence.

3. DTD Design

3.1 XML Tag Descriptions

3.1.1 LDMS

LDMS is the outer most tag that denotes a singular LDMS produced XML file. LDMS elements contains one or more STRUCTDIV elements.

3.1.2 STRUCTDIV

STRUCTDIV is a generic tag that marks up a logical divisions of the US Code, e.g. a Title, a Chapter, a Part, a Section, etc. STRUCTDIV elements can contain parsed character data (#PCDATA), TITLEDATA, or nested STRUCTDIV elements. The attributes for STRUCTDIV are NAME, VLEVEL, and HLEVEL. The NAME denotes the type of division it is, Title, Chapter, Part, etc. VLEVEL denotes the verticle level within the hierarchy and HLEVEL denotes the horizontal level within the hierarchy. In our example above from title 46, the NAME would be Section with VLEVEL = 5 and HLEVEL = 13102. EID is the unique identifier for cross referencing purposes.

3.1.3 TITLEDATA

Each TITLEDATA element represents a single sequence of the dashline markers from the input ASCII. It contains one NAVGROUP element, and zero or one each of STATGROUP, MISC1, REFTEXT, MISC2, COD, MISC3, CHANGE, MISC4, TRANS, MISC5, EXEC, MISC6, CROSS, MISC7, SECREF, MISC8, and NOTES elements.

3.1.4 NAVGROUP

The NAVGROUP element represents navigational information containing one each of the CITE, EXPCITE, and HEAD elements. The attribute for NAVGROUP is MAGICWORD which could be the entities RESERVED, REPEALED, TRANSFERRED, or OMITTED, having special legal meaning each.

3.1.5 CITE and DATE

The CITE tag directly corresponds to the CITE dashline in the input ASCII. A CITE element contains PCDATA and a DATE element. The attribute for CITE is TITLENUMBER, marking the title that contains this CITE.

The DATE element marks the date contained in the CITE element. It contains the PCDATA version of the date, and the attribute represents that date in the ISO standard format.

3.1.6 EXPCITE, DIVEXPCITE, and HEAD

The EXPCITE tag directly corresponds to the EXPCITE dashline in the input ASCII. It contains one or more DIVEXPCITE elements. The attribute is simply the level it represents.

The DIVEXPCITE tag divides the EXPCITE into individual catchlines. In our Title 46 example above, the EXPCITE would be divided into Title, Subtitle, Part, and Chapter DIVEXPCITE elements. It contains PCDATA.

The HEAD tag corresponds directly to the HEAD dashline and the element simply contains the PCDATA naming the current structural division.

3.1.7 STATGROUP, STATUTE, SOURCE, DIVSOURCE, and STATAMEND

STATGROUP is an entity that contains one or more DATATEXT elements, one STATUTE element, zero or one SOURCE element, and one STATAMEND. The STATGROUP entity shows the logical reletionship of these three dashlines which mark the location of the actual law as opposed to navigational information and notes.

The STATUTE element corresponds directly to the STATUTE dashline and contains one or more DATATEXT elements representing the law as passed by the US House of Representatives.

The SOURCE element corresponds directly to the SOURCE dashline and contains one or more DIVSOURCE elements.

The DIVSOURCE element contains PCDATA representing individual sources used by the US House of Representatives to create the preceding statute.

The STATAMEND element corresponds directly to the STATAMEND dashline and contains one or more DATATEXT elements representing amendments made to the preceding statute.

3.1.8 MISC1, REFTEXT, MISC2, COD, MISC3, CHANGE, MISC4, TRANS, MISC5, EXEC, MISC6, CROSS, MISC7, SECREF, and MISC8

All of these elements correspond directly to a dashline with the same name and contain one or more DATATEXT elements.

3.1.9 DATATEXT, DATATEXTNAME, DIVDATATEXT, PRE

The DATATEXT tag marks up various structures that are not represented by dashline markers in the ASCII input including numbered lists, cross references, and tables. It contains zero or one DATATEXTNAME elements, one or more DIVDATATEXT elements, and zero or more PRE elements. It has an EID attribute for crossreferencing and an INDENTLEVEL attribute to show how far it is indented relative to other DATATEXT elements.

The DATATEXTNAME element contains the PCDATA that is centered above the DATATEXT, naming this section of text.

The DIVDATATEXT element represent a logical section of the DATATEXT, such as all the text under (b) in an ordered list (a), (b), (c)… It contains one or more of any of DATATEXT, XREF, FOOTNOTE, TABLE, or PCDATA elements. The attributes NAME, VLEVEL and HLEVEL. NAME corresponds to the label of the section, which is b in the previous example, and VLEVEL and HLEVEL represent the location within the structure of the enclosing DATATEXT element.

The PRE tag marks a graceful failure: when the LDMS script fails to parse the data. The text contained in a PRE element has not been parsed and original format has been preserved.

3.1.10 XREF

The XREF tag marks a cross reference in the input. An XREF element contains one or more of either PCDATA or FOOTREF. The attribute TARGET shows where the XREF is pointing.

3.1.11 FOOTNOTE and FOOTREF

The FOOTNOTE tag marks the actual footnote at the end of a DATATEXT. A FOOTNOTE element contains one or more of either XREF or PCDATA. The attribute FNUMBER denotes what number footnote this is and the EID is the location used by XREF or FOOTREF to point to this footnote.

The FOOTREF element is a reference to FOOTNOTE. It contains the PCDATA where the text refers to a footnote, e.g. (Footnote 6). The attribute TARGET is the ID that points to the actual footnote text contained in a FOOTNOTE element.

3.1.12 TABLE, TABLENAME, and FIELD

The TABLE element represents a table found in the input. It contains zero or one TABLENAME elements and one or more FIELDS.

The TABLENAME element simply contains the PCDATA that names the table.

The FIELD tag marks each column on the table. It contains zero or one FIELDNAME elements and one or more DIVFIELD elements.

The FIELDNAME element contains PCDATA naming the FIELD that contains it.

The DIVFIELD element contains exactly one DATATEXT element representing the entry at that position in the table.

3.2 EID/Target Naming Conventions

3.2.1 For XREF purposes

The TARGET attribute in crossreferences follows the scheme:

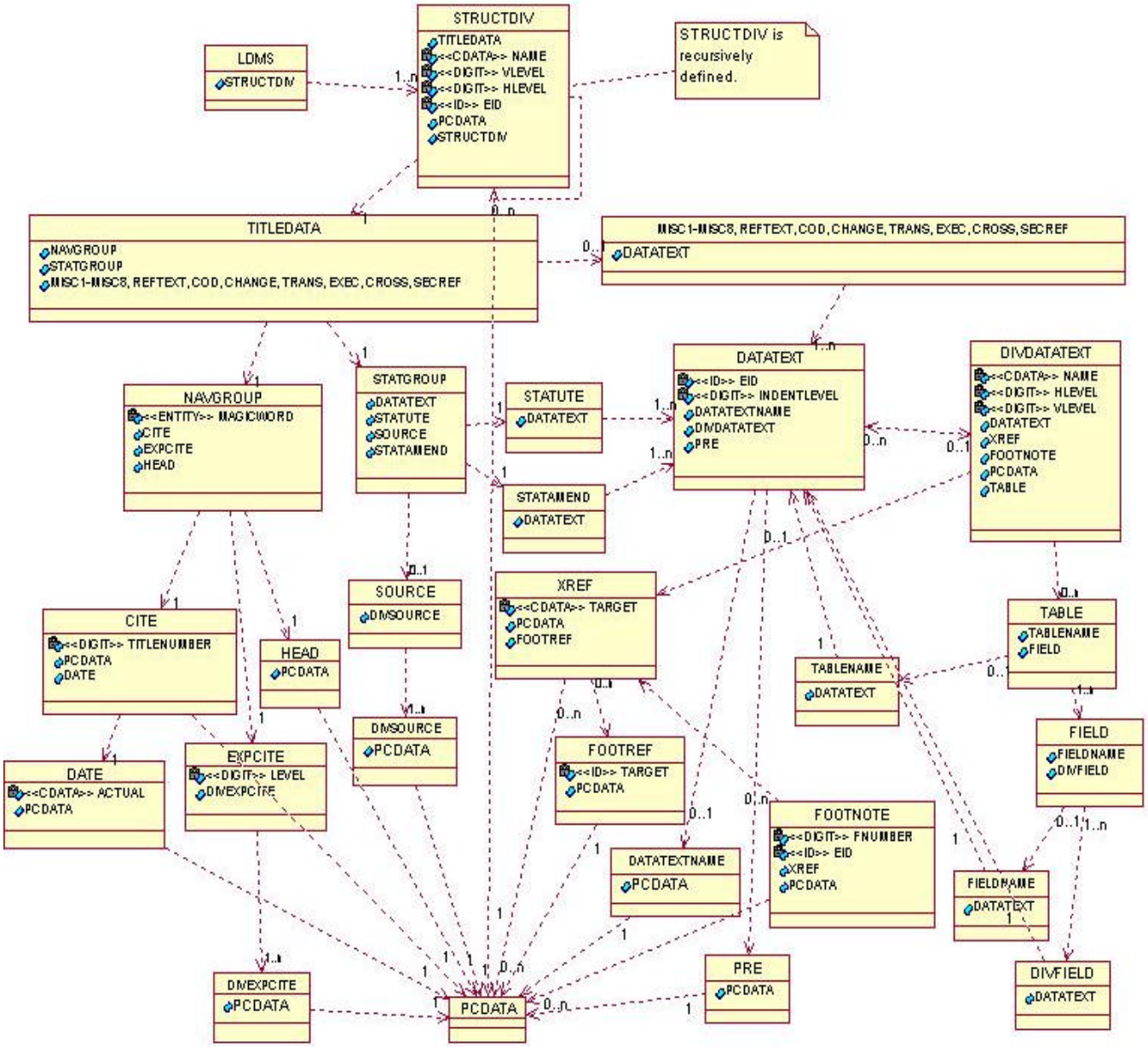
usc:section_chapter_number. eg. Title 20, chapter 5, section 117 is usc:117_5_20. If any field is 0, it means the information was not available from the xref. In a lot of the cases the chapter number is not supplied. This is not a problem since sections are unique within titles. so my target for something like title 50 section 101 is usc:101_0_50. If there is more than 1 element in any field, they are separated by ";". eg. sections 1,2,3b,4,5 of Title 27 will be usc:1;2;3b;4;5_0_27. For crossreferences like 1 USC 119, the convention is the same. In this case the target is usc:119_0_1. For public laws matched, the convention is the same except usc is replaced by pl. so Pub 211-459 sec. 12 will be pl:12_0_211-459.

3.2.2 For FOOTNOTE/FOOTREF purposes

The TARGET attribute for footreferences uses the following scheme: The attribute value is formatted as a dash delimited sequence of numbers. The first number is a period separated pair where the former number indicates the footnote number and the latter number indicates which particular instance of that footnote number is linked. The latter number handles cases where two footnote definitions with the same number appear in

the text. The subsequent numbers in the attribute value indicate the hierarchical division in the text in ascending hierarchical order, such as: subchapter, chapter, title.

3.3 UML Diagram



4. LDMS DTD

```
<!-- XML DTD LII Ver. 1.00-->
<!--LDMS is the outer most tag to denote XML output of LDMS-->
<!ELEMENT LDMS(STRUCTDIV+)>

<!--PRE is the tag to mark graceful failures.-->
<!--It indicates that there is no structural tags within the marked text, and any structural relationships are preserved through preserving the formatting.-->
<!ELEMENT PRE (#PCDATA)>


<!--XREF is the tag for cross reference-->
<!ELEMENT XREF ((#PCDATA|FOOTREF)+)>
  <!ATTLIST XREF
    TARGET CDATA #REQUIRED>
<!--FOOTNOTE is a footnote uniquely identified within TITLEDATA scope.-->
<!--It is the very last mention of footnote within a TITLEDATA.-->
<!--It is marked by structure (FOOTNOTE 1) at the beginning of a differently indented new line.-->
<!--It is ended by a beginning of a new DATATEXT|DIVDATATEXT or an end of super-tag-->
<!ELEMENT FOOTNOTE ((XREF|#PCDATA)+)
  <!--FNUMBER is the footnote number.-->
  <!--EID is the globally unique string.-->
    FNUMBER DIGIT #IMPLIED
    EID ID #REQUIRED>
<!--FOOTREF is a reference to FOOTNOTE.-->
<!--They are all non-ultimate mentions of (FOOTNOTE #).-->
<!--This tag will mark up the text (FOOTNOTE #).-->
<!ELEMENT FOOTREF (#PCDATA)>
  <!--TARGET is the ID-REF to FOOTNOTE's EID-->
  <!ATTLIST FOOTREF
    TARGET ID-REF #IMPLIED>

<!--TABLE is the tag for marking tables.-->
<!--Content of the tables are delimited by lines of dashes.-->
<!ELEMENT TABLE (TABLENAME?, FIELD+)>
  <!--TABLENAME is the title of the table-->
  <!ELEMENT TABLENAME (DATATEXT)>
  <!--FIELD corresponds to the fields of the table-->
  <!ELEMENT FIELD(FIELDNAME?, DIVFIELD+)>
    <!ELEMENT FIELDNAME (DATATEXT)>
    <!ELEMENT DIVFIELD(DATATEXT)>

<!--DATATEXT is the generic tag for the texts-->
<!--DATATEXT may include various other elements that can be found within the text, e.g. crossreferences, tables, etc.-->
<!--DATATEXTNAME marks the preceding header-esque line-->
<!ELEMENT DATATEXT (DATATEXTNAME?, (DIVDATATEXT), PRE+)>
  <!--INDENTLEVEL is the count of indentation depth, i.e. how many levels of white-space delimited texts there exists at the current DATATEXT.-->
  <!ATTLIST DATATEXT
    INDENTLEVEL DIGIT #IMPLIED
```

```
    EID ID #IMPLIED>
  <!ELEMENT DATATEXTNAME (#PCDATA)>
  <!--DIVDATATEXT marks the ordered division within DATATEXT-->
  <!--These will include things such as ordered list (a) to (z).-->
  <!--These are marked by NAME that contains (). e.g. (a), (i), (1), etc.-->

  <!ELEMENT DIVDATATEXT ((DATATEXT|XREF|FOOTNOTE|TABLE|#PCDATA)+)>
  <!--NAME is the label of the division, e.g. (a), (i), (1), etc.-->
  <!--VLEVEL denotes the vertical depth of the tag. Starts at 0.-->
  <!ATTLIST DIVDATATEXT
  <!--HLEVEL denotes the sequencial (horizontal) order of the tag.  Starts at 0.-->
  <!--VLEVEL denotes the vertical depth of the tag. Starts at 0.-->
  <!ATTLIST DIVDATATEXT
    NAME CDATA #REQUIRED
    VLEVEL DIGIT #REQUIRED
    HLEVEL DIGIT #REQUIRED>

<!--STRUCTDIV is the generic structural division tag.-->
<!ELEMENT STRUCTDIV (#PCDATA, TITLEDATA, STRUCTDIV*)>
  <!--NAME denotes the label of the division, e.g. Title, Subtitle, Chapter, etc.-->
  <!--VLEVEL denotes the vertical depth of the tag. Starts at 0.-->
  <!--HLEVEL denotes the sequential (horizontal) order of the tag.  Starts at 0.-->
  <!ATTLIST STRUCTDIV
    NAME CDATA #REQUIRED
    VLEVEL DIGIT #REQUIRED
    HLEVEL DIGIT #REQUIRED
    EID ID #REQUIRED>


<!--TITLEDATA contains the sequence of ordered dashline tags. Within each title there will be several sets of
TITLEDATA-->
<!ELEMENT TITLEDATA (NAVGROUP, STATGROUP?, MISC1?, REFTEXT?, MISC2?, COD?, MISC3?,
CHANGE?, MISC4?, TRANS?, MISC5?, EXEC?, MISC6?, CROSS?, MISC7?, SECREF?, MISC8?, NOTES?)>

<!--NAVGROUP contains the navigational dashline information of the title, i.e. the current section within the table
of content.  It is required in the beginning of each TITLEDATA-->
<!ELEMENT NAVGROUP (CITE, EXPCITE, HEAD)>
  <!ATTLIST NAVGROUP
        MAGICWORD ENTITY #IMPLIED>
  <!--MAGICWORD is one of the special words, e.g. RESERVED, REPEALED, etc.-->
  <!ENTITY % MAGICWORD "(RESERVED | REPEALED | TRANSFERRED | OMITTED)">

  <!ELEMENT CITE (#PCDATA, DATE)>
  <!--TITLENUMBER is the current title number-->
    <!ATTLIST CITE
      TITLENUMBER DIGIT #REQUIRED>
    <!--DATE is the tag to mark up the date of the document.-->
    <!ELEMENT DATE (#PCDATA)>
    <!--ACTUAL is the ISO standard formatted date.-->
    <!ATTLIST DATE
      ACTUAL CDATA #REQUIRED>
    <!ELEMENT EXPCITE (DIVEXPCITE+)>
    <!ATTLIST EXPCITE
```

```
      LEVEL DIGIT #IMPLIED>
    <!--DIVEXPCITE is a divider within EXPCITE.  Each DIVEXPCITE corresponds to a catchline.  It is divided by specific pattern, e.g.  TITLE 27 --->
  <!ELEMENT DIVEXPCITE (#PCDATA)>
  <!ELEMENT HEAD (#PCDATA)>

<!--STATGROUP is the entity group that shows the abstract relationship of STATUTE, SOURCE, STATAMEND-->

<!ENTITY STATGROUP "DATATEXT+, STATUTE, SOURCE?, STATAMEND">
  <!--STATUTE dashline contains the actual law text.  SOURCE and STATAMEND dashlines must be matched to a statute, i.e. included within the STATUTE element.-->
  <!ELEMENT STATUTE (DATATEXT+)>
  <!ELEMENT SOURCE (DIVSOURCE+)>
  <!--DIVSOURCE is separated by semicolon-->
    <!ELEMENT DIVSOURCE (#PCDATA)>

  <!ELEMENT STATAMEND (DATATEXT+)>

<!--MISC1 through MISC8 are identical texts, except in terms of physical location. e.g. MISC1 is found between REFTEXT and STATAMEND, etc.-->
<!ELEMENT MISC1 (DATATEXT+)>
<!ELEMENT REFTEXT (DATATEXT+)>

<!ELEMENT MISC2 (DATATEXT+)>
<!ELEMENT COD (DATATEXT+)>

<!ELEMENT MISC3 (DATATEXT+)>
<!ELEMENT CHANGE (DATATEXT+)>

<!ELEMENT MISC4 (DATATEXT+)>
<!ELEMENT TRANS (DATATEXT+)>


<!--TRANS contains DATATEXT elements only -->
<!ELEMENT MISC5 (DATATEXT+)>
<!ELEMENT EXEC (DATATEXT+)>

<!ELEMENT MISC6 (DATATEXT+)>

<!--CROSS dashlines denote the various crossreferences from this portion of the document to other titles and/or sections.-->
<!ELEMENT CROSS (DATATEXT+)>

<!--SECREF lists the various sections that have this section listed in their crossreference-->
<!ELEMENT MISC7 (DATATEXT+)>
<!ELEMENT SECREF (DATATEXT+)>

<!ELEMENT MISC8 (DATATEXT+)>

<!ENTITY NOTES "">
<!ENTITY TEXT "">
```