U.S. Government Printing Office Federal Digital System User Guide Document Code of Federal Regulations XML Rendition

Prepared by: Program Management Office

Office of the Chief Information Officer U.S. Government Printing Office

December 17, 2009

Revision History

Revision	Date	Description
1.0	December 17, 2009	Version 1.0

Table of Contents

1. Int	oduction	1
1.1.	Purpose	1
1.2.	Legal Status & Authenticity of Code of Federal Regulations files via Data.gov	1
2. Sch	ema Description	5
2.1.	Sections Available in XML	7
2.2. 2.2.1	Sections, Parts, and ArticlesXPath Examples	7 8
2.3.	The TITLE, CHAPTER, PART and SUBPART Section	
2.4.	Cite This Code Section	10
2.5.	The Contents of Section	12
2.6.	Chapter Tag	13
2.7.	ALPHLIST Tag	14
3. Res	ources Directory	15

1. Introduction

The U.S. Government Printing Office (GPO) and the National Archives' Office of the Federal Register (OFR) partnership is offering bulk data downloads of Code of Federal Regulations (CFR) files to the general public via Data.gov and FDsys. This effort began when the President challenged Federal agencies to create a more open and transparent government, promote accountability, and provide information to citizens about what their Government is doing (see 74 FR 4685, January 26, 2009 at http://www.gpo.gov/fdsys/pkg/FR-2009-01-26/pdf/E9-1777.pdf). The Public Printer's letter of March 9, 2009 pledged to provide trusted information in whatever form is required to meet the President's objectives http://www.gpo.gov/pdfs/news-media/letter_030909.pdf.

In addition, the Office of the Federal Register coordinates with the Office of Science Technology Policy to ensure that the OFR/GPO partnership meets customer expectations. To follow through on our commitment, we are expanding and accelerating the development of FDsys to provide XML-structured content as rendered output. This will give users access to masses of data to reconfigure and redistribute as they wish to meet the specialized needs of their constituencies.

1.1. Purpose

The purpose of this document is to provide an overview of Code of Federal Regulations XML files and associated schema. The FDsys Bulk Data repository at http://www.gpo.gov/fdsys/bulkdata/ contains the Code of Federal Regulations in XML from 2007 to the present and additional years will be added as they become available. Please see FDsys at www.fdsys.gov for access to the Code of Federal Regulations in PDF and HTML formats.

1.2. Legal Status & Authenticity of Code of Federal Regulations files via Data.gov

Q. What is the data set available for the Code of Federal Regulations in XML?

A. Code of Federal Regulations files in XML have been converted and simplified from the SGML rendition used for the printed publication. The Code of Federal Regulations is available in XML starting with the year 2000, when GPO began composing the Code of Federal Regulations in SGML. OFR/GPO plan to convert the remaining electronic data set at a future date.

Q. Are bulk data downloads of XML files offered on FDsys via Data.gov and FDsys part of the official version of the Code of Federal Regulations?

A. No, the XML-structured files offered for bulk download are not part of the official on-line format of the *Code of Federal Regulations*. While GPO's XML files are based on the original source

data submitted by Federal agencies, OFR markup, and GPO typesetting and composition markup in SGML, only the PDF and Text versions of *Code of Federal Regulations* content on GPO Access and FDsys have legal status as parts of the official online format of the *Code of Federal Regulations*. Additional development will be required before OFR/GPO can specify that XML files are a part of the official online edition of the *Code of Federal Regulations*.

Q. Do OFR/GPO plan to include XML files as part of the official format of the Code of Federal Regulations?

A. Yes. The current set of XML-structured material (minus graphics) is not yet characterized as part of the official online *Code of Federal Regulations* format because the underlying data used to create a viewable rendition of the material must be thoroughly scrutinized and tested. The XML-structured files are derived from SGML-tagged data and printing codes, which may produce anomalies in display. For example, complex tabular material in XML files may not display as correctly composed objects equivalent to the tables that appear in the Text and PDF files. Tabular material that displays in an ambiguous or distorted manner could affect the substantive meaning of regulations and other documents.

Our goal is to clean up the XML data and develop associated style sheets to the point that the XML rendition can be characterized as a display format of the official online edition. We are also developing the means to embed graphics in the files, so that an XML version may be deemed both official and complete. The OFR will issue a *Federal Register Bulletin* to alert users when the XML-structured files are ready to be included as part of the official online edition.

Q. What is the legal basis for making determinations about official status?

A. The Federal Register Act established the Administrative Committee of the Federal Register (Administrative Committee or ACFR) as the regulatory body with authority to determine the format(s) of its official publications, including the Code of Federal Regulations (44 U.S.C. Ch. 15). Under 1 CFR 8.6(a), the official formats approved by the ACFR include a paper edition, a microfiche edition, and an online edition. OFR and GPO carry out ACFR regulations by developing appropriate means to display *Code of Federal Regulations* material in the official formats.

Q. When did the Code of Federal Regulations first appear online?

A. The first official volumes were published online in 1996. The first full online edition dates to 1997 as authorized by resolution of the Administrative Committee and its regulations at 1 CFR 8.6, and by Public Law 103-40 (the Government Printing Office Electronic Information Enhancement Act of 1993; codified at 44 U.S.C. 4101).

Q. Are Code of Federal Regulations XML bulk download files digitally signed?

No, XML files available for download are not digitally signed. They can be manipulated and enriched to operate in the various applications that users may devise. GPO is evaluating technology that could be used to digitally sign XML files for future official editions posted on FDsys. Adding signed non-PDF files to FDsys would be an enhancement for FDsys users, but would not be used to restrict or adversely affect the XML bulk data downloads available to our customers.

O. What does the term "digitally signed" mean?

A. Currently, GPO uses digital signature technology on PDF documents to add a visible Seal of Authenticity (a graphic of an eagle) to authenticated and certified documents. The technology allows GPO to secure data integrity, and provide users with assurance that the content is unchanged since it was disseminated by GPO. A signed and certified document also displays a blue ribbon icon to the left of the Seal of Authenticity and in the Signatures tab within Adobe Acrobat or Reader. When users print a document that has been signed and certified by GPO, the Seal of Authenticity will automatically print on the document, but the blue ribbon will not print.

Q. How reliable is the metadata and the underlying tagging in bulk data files?

A. The document markup and metadata found within bulk data files are generally reliable and complete. However, there are variations in this underlying data due to inconsistencies in the composition and typesetting process. As a result, some data-mining applications and user aids may not produce 100 per cent accurate results.

Q. What is the legal status of Code of Federal Regulations user aids?

A.: Code of Federal Regulations user aids, including, finding aids, indexes, search tools, metadata associations, and tagging schemes are not part of the legal text of the *Code of Federal Regulations*. These ancillary features help users explore and extract data, but the official legal text stands on its own. No person should form absolute legal conclusions based on search results, finding aids, metadata associations, extractions of data, and the like. Ultimately, only the official text of the *Code of Federal Regulations* may be relied upon as evidence in a court of law.

Q. What is the authenticity of Code of Federal Regulations bulk data files after they have been downloaded to another site?

A. We cannot vouch for the authenticity of data that is not under OFR/GPO control. OFR and GPO are providing free access to Code of Federal Regulations data via XML for display in various applications and mash-ups outside the FDsys domain. The OFR/GPO partnership does not endorse third party applications, and does not evaluate how our original legal content is displayed on other sites. Consumers should form their own conclusions as to whether the downloaded data can be relied upon within an application or mash-up. An application may link to the official *Code of Federal Regulations* on FDsys to provide users with additional assurance.

Q. Do OFR and GPO assert any control over downstream uses of bulk data?

A. In general, there are no restrictions on re-use of information in Code of Federal Regulations material because U.S. Government works are not subject to copyright. OFR and GPO do not restrict downstream uses of Code of Federal Regulations data, except that independent providers should be aware that only the OFR and GPO are entitled to represent that they are the providers of the official versions of the *Code of Federal Regulations* and related Federal Register publications.

Q. How can re-publishers indicate the source of Code of Federal Regulations data?

A. Re-publishers of Code of Federal Regulations data may cite FDsys and OFR/GPO as the source of their data, and they are free to characterize the quality of data as it appears on their site. But private sector re-publishers are prohibited from using the seal of the National Archives and Records Administration (NARA) or stylized Code of Federal Regulations logos identified in NARA regulations (36 CFR part 1200) on their products because that would unlawfully misrepresent the legal status of the material, and could falsely identify private organizations as entities of the Federal Government.

2. Schema Description

The schema chosen to represent the Code of Federal Regulations is a simplified version of the SGML schema that is used as part of the print production process, with some presentation and print specific tags removed or collapsed, and then converted to well-formed XML. This schema was chosen for the following reasons:

- 1. It is a complete and faithful representation of the Code of Federal Regulations, which matches most closely to the author's original intent.
- 2. It describes the data using semantic tags in a way that is appropriate to the Code of Federal Regulations Domain.
- 3. It fully describes the structure of the Code of Federal Regulations, including the large structure (chapters, parts, sections, etc.), the document structure (paragraphs, etc.), and semantic structure.

Since the schema is not an authoring schema and the SGML to XML conversion process maintains the order of the tags, this allows the XML schema to be more permissive than if it were used for checking authored content.

The schema being produced for this effort describes the data as it actually occurs from the OFR. Documents are not being cleaned up because they do not match the schema; instead, the schema was selectively relaxed. Such an approach maintains 100% fidelity to the original data, and eliminates any errors that might occur in schema interpretation or further data manipulation.

The following table lists the SGML tags that were removed or collapsed into a source attribute in the Code of Federal Regulations XML.

Purposed Fields to be Removed	Fields to be Removed or Collapsed	Description
FNC	Removed	Used to generate a new column.
Q	Removed	Insert vertical space.
FNP	Removed	Used to generate a new even page.
FP	Collapse to FP	Flush paragraph entries.
FP-1	Collapse to FP	Flush paragraph, turnovers indented one em.
FP-2	Collapse to FP	Flush paragraph, turnovers indented two ems.
FP-DASH	Collapse to FP	Used for flush paragraphs that fill with dashesu.
FP1	Collapse to FP	First level indented flush paragraph.
FP1-2	Collapse to FP	Paragraph, first line indented one em and turnovers indented two ems.
FP2	Collapse to FP	Second level flush paragraph, all lines indented two ems.
FP2-2	Collapse to FP	Second level indented paragraph with turnovers indented two .levels also.
FP2-3	Collapse to FP	Second level indented paragraph with turnovers indented three

		levels.
FP3	Collapse to FP	Third level indented flush paragraph.
FP4	Collapse to FP	Fourth level indented flush paragraph.
FRP	Collapse to FP	Flush right material, actually held in 1 em from right margin.
FRP0	Collapse to FP	True flush right material.
HD	Collapse to HD	First level head in the following sections.
HD1	Collapse to HD	Usually used for first level (bold?) head.
HD2	Collapse to HD	Usually used for second level (caps and small caps?) head.
HD3	Collapse to HD	Usually used for third level (italic?) head.
HD4	Collapse to HD	Usually used for fourth level (roman?) heads.
HD5	Collapse to HD	Fifth level head in the following sections.
HD6	Collapse to HD	Sixth level head in the following sections.
HD8	Collapse to HD	Lowest level head in the following sections.
HED	Collapse to HD	Usually the first head within a group or container where the group or container is designated for search and retrieval purposes.
HED1	Collapse to HD	Special first level head in text of Section.
THED	Collapse to HD	Head that turns sideways on the page.
P	Collapse to P	Normally used for paragraph. A paragraph here has the first line indented.
NPAR	Collapse to P	Used to generate a new paragraph where the <p> tag would create a run in entry.</p>
P-1	Collapse to P	Paragraph, turnovers indent one level.
P-2	Collapse to P	Paragraph, turnovers indented two ems on left.
P-3	Collapse to P	Paragraph, turnovers indent third level.
P-DASH	Collapse to P	Used for paragraphs that fill with dashes.
P1	Collapse to P	First level indent paragraph.
P2	Collapse to P	Second level indent paragraph.
P2-3	Collapse to P	Second level indent paragraph, turnovers indent third level.
P4-5	Collapse to P	Special indents for BLM index.
P6-7	Collapse to P	Special indents for BLM index.
P8-9	Collapse to P	Special indents for BLM index.
SUBJECT	Collapse to SUBJECT	Subject of Section, usually follows Section Number.
SUBJECT1	Collapse to SUBJECT	First level indented subject in special indexes.
SUBJECT2	Collapse to SUBJECT	Second level indented subject in special indexes.
SUBJECT3	Collapse to SUBJECT	Third level indented subject in special indexes.

SUBJL	Collapse to SUBJL	Lead subject in special indexes.
SUBJ1L	Collapse to SUBJL	First level lead subject in special indexes.
SUBJ2L	Collapse to SUBJL	Second level lead subject in special indexes.
SUBJ3L	Collapse to SUBJL	Third level lead subject in special indexes.
SUBJ4L	Collapse to SUBJL	Fourth level lead subject in special indexes.

2.1. Sections Available in XML

The following are currently available in Code of Federal Regulations XML:

- Table of Contents
- Title
- Chapter
- Alphabetical list of agencies

Note: The sections above are not a complete set of what is available in Code of Federal Regulations XML.

2.2. Sections, Parts, and Articles

This section describes the top-level structure of Code of the Federal Regulations XML file.

The XML schema, being a translated reproduction of the SGML schema, contains tags and content in the same order as they appear in the printed document. Major sections are grouped appropriately (<TOC>, <PART>, <SUBPART>, <CITE>, <EXPLA>, <SECTION>, <CHAPTER>, <ALPHLIST>), and all data and tags are represented – with the exception of the reduced tags from the previous section.

The XML tags and their descriptions of the TOC schema above is shown below:

XML Tag	Description
CFRDOC	The root tag for all document types in Code of Federal Regulations publications. This tag includes children such as TOC, CITE, EXPLA, SECTION, CHAPTER and ALPHLIST. Note: These children are not a complete set of tags. See the CFRDOC.XSD for more details.
TOC	Contains the main table of contents in major divisions of publication.
TITLENO	Contains title number in tables of contents.
SUBTI	Contains subtitle head in table of contents.
PGHD	Usually, it contains the word "Page" when it appears as a heading.
EXPL	Used for explanatory data following the "Cite this Code" page.
SUBJECT	Contains subject of section, usually follows section number.
PG	Used for page number data in lists.

СНАРТІ	Contains chapter division in tables of contents sections.
FAIDS	Finding Aids entries

An abbreviated example of the overall section and structure is below:

```
<CFRDOC>
   <TOC>
       <HD SOURCE="HED">Table of Contents/HD>
        <PGHD>Page</PGHD>
        <EXPL>
            <SUBJECT>Explanation</SUBJECT>
           <PG>v</PG>
        </EXPL>
        <TITLENO>
           <HD SOURCE="HED">Title 7:</hd>
                <HD SOURCE="HED">Subtitle B-Regulations of the.../HD>
           </SUBTI>
            <CHAPTI>
                <SUBJECT>Chapter X-Agricultural Marketing Service... </SUBJECT>
                <PG>5</PG>
            </CHAPTI>
        </TITLENO>
        <FAIDS>
            <HD SOURCE="HED">Finding Aids:
            <SUBJECT>Table of CFR Titles and Chapters/SUBJECT>
           <PG>225</PG>
           <SUBJECT>Alphabetical List of Agencies Appearing in.../SUBJECT>
           <PG>245</PG>
            <SUBJECT>List of CFR Sections Affected</SUBJECT>
           <PG>255</PG>
        </FAIDS>
    </TOC>
</CFRDOC>
```

2.2.1.XPath Examples

The schema allows for a wide variety XPath commands for extracting items:

```
//TOC → Output table of contents

(//PRTPAGE/@P)[0] → Get the first page number

(//PRTPAGE/@P)[last()] → Get the last page number

//TOC[descendant::PRTPAGE/@P = 'iii'] → Get the table of contents which contains page iii.
```

//TOC[TITLENO/CHAPTI

→ Get all chapter titles in table of contents section.

2.3. The TITLE, CHAPTER, PART and SUBPART Section

The TITLE, CHAPTER, PART and SUBPART tags are organized in an xml structure defined below and contains logical divisions in the Code of Federal Regulations. For example, each title is divided into chapters, which usually bear the name of the issuing agency. Each chapter is further subdivided into parts that cover specific regulatory areas. Large parts may be subdivided into subparts. All parts are organized in sections, and most citations in the CFR are provided at the section level.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	Description
TITLE	Contains a major group of the publication.
SUBCHAP	Contains a major division within a Chapter.
CONTENTS	Contains table of contents entries.
SECHD	Usually used for the words "Section" or "Sec." when it appears as heading.
SECTNO	Used for the CFR section number.
SUBJECT	Subject of Section, usually follows Section Number.
APP	Used for Appendix head entries in tables of contents.
CHAPTER	Contains a major division within Title and Subtitle.
PART	Contains a major division within Chapters and Subchapters.
EAR	Used to override default entry for first or last section number in running head.
SUBPART	Contains a major division within Part.
RESERVED	Used mostly for units that are reserved for future use, and therefore contain no data at this time.

The abbreviated example of the TITLE, CHAPTER, PART and SUBPART tags are shown below.

```
<SECTNO>53.1</SECTNO>
                            <SUBJECT>Definitions.</SUBJECT>
                            <SECTNO>53.2</SECTNO>
                            <SUBJECT>General requirements for a...</SUBJECT>
                            <SECTNO>53.3</SECTNO>
                            <SUBJECT>General requirements for an...
                        </SUBPART>
                        <SUBPART>
                            <HD SOURCE="HED">Subpart B-Procedures for...</HD>
                            <SECTNO>53.20</SECTNO>
                            <SUBJECT>General provisions.</SUBJECT>
                            <SECTNO>53.21</SECTNO>
                            <SUBJECT>Test conditions.</SUBJECT>
                        </SUBPART>
                       <APP> <E T="04">Appendix A to Subpart...</E></APP>
                  </CONTENTS>
               </PART>
           </SUBCHAP>
       </CHAPTER>
   </TITLE>
</CFRDOC>
```

2.4. Cite This Code Section

Cite this code section holds the first text line of site this code pages followed by explanatory data. The contents of a CITE always includes a CITEP paragraph on the regulation. Cite this code section is followed by EXPLA section which explains the data of Code of Federal Regulations. The signature block is often followed by EXPLA section and may contain empty xml values.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	Description
CITE	Contains the first text line of "cite this code" page.
CITEP	Contains "To Cite" paragraph on cite this code page in front matter
EXPLA	Used for explanatory data following the "Cite this Code" page in the front matter
IPAR	Contains lead material on first page of explanation section of the front matter.
STUB	Contains the back part of lead material.

SIDEHED	Used within explanatory data following the "Cite this Code" page in the front matter.
SIG	Used to designate a signature block.
NAME	Used for person's name.
POSITION	Used to denote the job title only in signature blocks.
OFFICE	Used for the "Office of the Federal Regulations" line in the signature block of the explanation pages of the Front matter.
DATE	Used for date entry.

An abbreviated example of the CITE and EXPLA section and structure is below:

```
<CFRDOC>
   <CITE>
       <P>Cite this Code:<E T="01">CFR</E></P>
       <CITEP>To cite the regulations in this volume use title...</CITEP>
   </CITE>
    <EXPLA>
        <HD SOURCE="HED">Explanation</HD>
        <P>The Code of Federal Regulations is a codification of the...
       <P>Each volume of the Code is revised at least once each...</P>
        <IPAR>
           <P SOURCE="P1">Title 1 through Title 16 </P>
           <STUB>as of January 1</STUB>
           <P SOURCE="P1">Title 17 through Title 27 
           <STUB>as of April 1</STUB>
        </IPAR>
        <P>The appropriate revision date is printed on the cover of each...</P>
        <SIDEHED>
            <HD SOURCE="HED">LEGAL STATUS</HD>
            <P>The contents of the Federal Register are required to be...</P>
        </SIDEHED>
        <SIDEHED>
           <HD SOURCE="HED">HOW TO USE THE CODE OF FEDERAL REGULATIONS</HD>
            <P>The Code of Federal Regulations is kept up to date by the...</P>
            <P>To determine whether a Code volume has been amended since...</P>
        </SIDEHED>
        <SIG>
           <NAME/>
            <POSITION/>
            <OFFICE/>
        </SIG>
        <DATE/>
   </EXPLA>
```

```
</CFRDOC>
```

2.5. The Contents of Section

The contents of section tags are the ones that usually describe the majority of the data in the Code of Federal Regulations. In some cases, the section includes XML GPOTABLE tags which describe elements of tables and their data. Section tag always contains a section number which is usually followed by a subject of that section. A paragraph or XML tag P is used to denote large or small text in the section. In some cases, a section may also contain a citation.

XML Tag	Description
SECTION	Contains group or container tag for search and retrieval purposes.
SECTNO	Used for the code of Federal Regulations section number.
SUBJECT	Contains subject of section, usually follows section number.
GPOTABLE	Defines the elements necessary to produce the structure of GPO tables.
BOXHD	Generation of box heads delimiter. It is used only if there are actual CHEDs in the table.
CHED	Generation of column boxheads. Do not use if there are no actual CHEDs in the table.
ROW	Generation row delimiter.
ENT	Generation table body cell.
CITA	Used for citations.

The contents will roughly have the same structure

```
<CFRDOC>
 <SECTION>
   <SECTNO>§ 1007.51
   <SUBJECT>Class I differential, adjustments to Class I prices...</SUBJECT>
   <P>(a) The Class I differential shall be the differential...</P>
   <P>(b) Adjustment to Class I prices. Class I prices shall...</P>
   <GPOTABLE CDEF="xs60,r100,13,13," COLS="4" OPTS="L2">
       <CHED H="1">State</CHED>
       <CHED H="1">Country/parish</CHED>
       <CHED H="1">FIPS</CHED>
       <CHED H="1">Class I price adjustment</CHED>
     </BOXHD>
       <ENT I="01">AL</ENT>
       <ENT>AUTAUGA</ENT>
       <ENT>01001</ENT>
       <ENT>0.50</ENT>
     </ROW>
```

2.6. Chapter Tag

The CHAPTER tag holds contents that list agency names along with their code of Federal Regulations data. It contains chapter number entries and repeatable Code of Federal Regulations numbers heads.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	Description	
CHAPTER	Used for group or container tag for search and retrieval purposes.	
CHAPNO	Used for chapter number entries in the "List of Code of Federal Regulations Sections Affected" section in Back matter and elsewhere.	
AGENCY	Used Agency name.	
CFRHD	Contains repeatable "CFR, Title." Head in Alphabetical Listing of Agencies in Back matter.	
ONOTE	Contains "Material Approved for Incorporation by Reference" section of Back matter.	
OWNER	Contains "Material Approved for Incorporation by Reference" section of Back matter.	
ADDR	Contains Address of organization publishing standards that follow. Used in "Material Approved for Incorporation by Reference" section of back matter.	
PUBLI	Used for publications.	
CFRNO	Repeatable CFR Number heads in Material Incorporated for Reference and List of CFR Sections Affected section of Back matter.	

An abbreviated example of the CHAPTER section and structure is below:

2.7. ALPHLIST Tag

The ALPHLIST tag contains list of agencies appearing in the Code of Federal Regulations documents. This information may include revised tag with a date along with repeatable agency tags and CFRID values. The XML tags and their descriptions of the ALPHLIST schema are shown below.

XML Tag	Description
ALPHLIST	Alphabetical List of Agencies Appearing in the code of Federal Regulations.
REV	Contains "Revised as of" line under Head.
AGHD	Contains Repeatable "Agency" head. It is used only in alphabetical list of agencies in Back matter.
CFRID	Contains CFR Title, Subtitle, or Chapter data in second column of Alphabetical Listing of Agencies in Back matter.

An abbreviated example of the ALPHLIST section and structure is below:

3. Resources Directory

The resources directory in the Code of Federal Regulations bulk data repository at http://www.gpo.gov/fdsys/bulkdata/CFR/resources contains the current version of the XML schema, the XML stylesheet used to display the XML files in a browser on the FDsys website, and this user guide.