

XML 101

It's Not Just Markup Anymore

Bill Kasdorf

Vice President, Apex Publishing, LLC

General Editor, *The Columbia Guide to Digital Publishing*

XML: Extensible Markup Language



Extensible:

Designed to adapt to various

- Kinds of documents
- Modes of publication
- Styles of presentation
- Patterns of access and use

XML: Extensible Markup Language



X **M** **L**

Markup:

Tagging a document to provide

- Semantic information
- Structural information
- Formatting information
- Supplemental information

XML: Extensible Markup Language



X M L

Language:

A formal way to express markup

Not a set of tags or a vocabulary,
but an agreed-upon *way to express*
a given vocabulary or tag set

1

Chapter Title

*Here's some text
with no markup.*

Chapter Author
Author Identification

Here's some text at the beginning of this chapter.
Let's make one more line's worth.

Level One Subhead

Here's some more text. This author's a pretty nice
girl, but she doesn't have a lot to say.

Level Two Subhead

The end.

*A typical
MS implies
markup . . .*

1

Chapter Title

Chapter Author
Author Identification

Here's some text at the beginning of this chapter. Let's make one more line's worth.

Level One Subhead

Here's some more text. This author's *a pretty nice girl*, but she doesn't have a lot to say.

Level Two Subhead

The end.

(CN) 1

(CT) Chapter Title

(CA) Chapter Author
Author Identification

(NI) Here's some text at the beginning of this chapter. Let's make one more line's worth.

(H1) **Level One Subhead**

(NI) Here's some more text. This author's *a pretty nice girl*, but she doesn't have a lot to say.

(H2) *Level Two Subhead*

“Editorial” markup . . .

(IT) The end.

“Well Formed” (but not very good) XML

CN

<CN>1</CN>

CT

<CT>Chapter Title</CT>

CA

<CA>Chapter Author

NI

<ITAL>Author Identification</ITAL></CA>

NI

<NI>Here’s some text at the beginning of this chapter. Let’s make one more line’s worth.</NI>

H1

<H1>Level One Subhead</H1>

NI

<NI>Here’s some more text. This author’s <ITAL>a pretty nice girl</ITAL>, but she doesn’t have a lot to say.</NI>

H2

<H2>Level Two Subhead</H2>

IT

<IT>The end.</IT>

<CN>1</CN>

<CT>Chapter Title</CT>

<CA>Chapter Author

<ITAL>Author Identification</ITAL></CA>

<NI>Here's some text at the beginning of this chapter.
Let's make one more line's worth.

</NI>

<H1>Level One Subhead</H1>

<NI>Here's some more text. This author's

<ITAL>a pretty nice girl</ITAL>, but she doesn't have a
lot to say.</NI>

<H2>Level Two Subhead</H2>

<IT>The end.</IT>

*Structural
markup*

Presentational markup

<CN>1</CN>

<CT>Chapter Title</CT>

<CA>Chapter Author

<ITAL>Author Identification</ITAL></CA>

<NI>Here's some text at the beginning of this chapter.
Let's make one more line's worth.

</NI>

<H1>Level One Subhead</H1>

<NI>Here's some more text. This author's

<ITAL>a pretty nice girl</ITAL>, but she doesn't have a
lot to say.</NI>

<H2>Level Two Subhead</H2>

<IT>The end.</IT>

Semantic markup

<CN>1</CN>

<CT>Chapter Title</CT>

<CA>Chapter Author

<ITAL>Author Identification</ITAL></CA>

<NI>Here’s some text at the beginning of this chapter.
Let’s make one more line’s worth.

</NI>

<H1>Level One Subhead</H1>

<NI>Here’s some more text. This author’s

<ITAL>a pretty nice girl</ITAL>, but she doesn’t have a
lot to say.</NI>

<H2>Level Two Subhead</H2>

<IT>The end.</IT>

XML: Extensible Markup Language



- Tells what each element **is**
—*E.g., “chapter title,” not “18’ Bulmer caps”*
- Tags where each element **starts & ends**
—*Unambiguous; must be properly nested*
- Defines **attributes** of each element
—*E.g., brand name vs. generic drug names*
- Defines **relationship** between elements
—*E.g., “H1 subhead must precede H2 subhead”*

XML: Not Just for Markup Anymore



Metadata—*Making content discoverable*

- Sales, marketing, and rights information
- ONIX metadata for booksellers
- CrossRef for journals
- DOI (for both books and journals)
- Subject classifications, metadata

Beginning to be integrated into ed/prod workflow

XML: Not Just for Markup Anymore



Metadata—*Making your content work*

- Links are a given (internal and external) to references, tables, figures, etc.
- Keywords, taxonomies/controlled vocabularies
- Metadata to track revision history, administrative information, editorial and production notes, etc.

Semantic markup and rich metadata are essential for optimizing use of your content

Example of a header for a scholarly book

```
<cup_document>
<cup_header>
<cupid>segel-_811404</cupid>
<author>Segel, Harold B</author>
<author_last_name>Segel</author_last_name>
<title>The Columbia Guide to the Literatures of Eastern Europe Since 1945</title>
<subtitle></subtitle> <edition></edition> <isbn>978-0-231-11404-2</isbn>
<pub_date>4/15/2003</pub_date> <cup_subject_1>European</cup_subject_1>
<cup_subject_2></cup_subject_2>
<cup_subject_code_1>233</cup_subject_code_1>
<cup_subject_code_2></cup_subject_code_2>
<bisac_subject_1>LITERARY CRITICISM / European / General</bisac_subject_1>
<bisac_subject_2></bisac_subject_2>
<bisac_code_1>LIT004130</bisac_code_1>
<bisac_code_2>LIT004130</bisac_code_2>
<bisac_codes>LIT004130; LIT004130</bisac_codes>
<price_dollars>105</price_dollars>
<page_count>512</page_count>
<market>World</market>
<season>Spring 2008</season>
<press>CUP</press>
<publisher>Columbia University Press</publisher> <figures></figures>
<rights>All rights in all media (now or hereafter known): CUP</rights>
<copy_text>For nearly half a century, the Iron Curtain obscured from Western eyes a vital
group of national and regional writers. Seen as a whole, the literatures of Eastern Europe
during the second half of the twentieth century are extraordinarily rich, and in recent years
many Eastern European novelists, poets, and playwrights have attracted wider attention
and broader publication in the West. And yet no reference work, embracing all the countries
of this region, including the former East Germany, has brought synoptic analysis to bear on
these literatures—until now.</copy_text>
<table_of_contents>Preface Chronology of Major Political Events, 1944-2001 Journals,
Newspapers, and Other Periodical Literature Note on Orthography, Transliteration, and Titles
Introduction: The Literature of Eastern Europe from 1945 to the Present Authors A-Z Select-
ed Bibliography Author Index</table_of_contents>
<cup_css_link>http://www.columbia.edu/cu/cup/cup_dtd/cup_css_1-1.css</cup_css_link>
</cup_header>
```

DTD: Document Type Definition



D T D

Document

SGML and XML originated from the need to mark up *text* in *documents*, which is not typically structured like information in databases, which have schemas. . . .

XML: Extensible Markup Language



D T D

Type

For particular *types* of documents, which share characteristics like

- Similar **structure**
- Related **semantics**
- Common **metadata** . . .

XML: Extensible Markup Language



D T D

Definition

Formal, explicit, rigorous . . .

So complete and precise even
a computer can understand it . . .

XML: Extensible Markup Language



DTDs are created for:

- Archiving
- Interchange
- Print production
- Online publishing or E-books
- Multipurposing (“Slicing & Dicing”)

One DTD rarely does it all; DTDs need to be customized or adapted for what they’re for.

XML: Extensible Markup Language



Why start with a “standard” DTD?

- Saves “reinventing the wheel”
- Benefit from broad base of experience, evolution
- Expedites interchange to use a known model
- Vendors are already familiar with it
- Some tools are optimized for certain standards
- A standard can be mandated in a given industry

XML: Extensible Markup Language



Why customize a “standard” DTD?

- Some are too simplistic or generic
- Many are much more complex than you need
- Needs and capabilities change over time:
 - Requirements of customers, vendors, partners
 - Capabilities of software, tools, and staff

Typically, you start with a subset of a standard and then add features you need for your situation.

XML: Extensible Markup Language



Need to transform to “output” XML

- XHTML for online publication
- OEB PS (soon OPS/OPF) for e-books
- DTBook for accessibility
- Models required by business partners, e.g., ACLS History E-Book XML

Your XML must be able to produce these outputs, but they're rarely sufficient for production/archiving.

XML: Extensible Markup Language



DTDs must accommodate:

- Needs of customers and partners
- Transformation to other needed models
- Capabilities of:
 - Staff
 - Vendors
 - Available tools and technologies

DTDs constantly evolve as these factors change.

DTDs can be strict . . .



ISO 12083

*The Mother Superior
of DTDs . . .*

The **ISO 12083** DTD



- Brilliant, idealistic, based on theory
- Creation of one individual, Eric van Herwijnen
- Created before the Web, before XML

Most big STM journal DTDs are still 12083-based

. . . or permissive . . .

TEI

*The “Let One Thousand
Flowers Bloom” DTD . . .*



TEI: The Text Encoding Initiative



- Rich, expansive, accommodating
- Collaborative creation: TEI Consortium
- Created for scholarship, not publication
- Enormously useful resource, but:
 - Full suite is overwhelming, not “a DTD”
 - TEI Lite is too simplistic

Most humanities scholarship is TEI-based

```
<?xml version="1.0" encoding="us-ascii"?>
<!DOCTYPE TEI SYSTEM "tei_all.dtd">
<!-- TEI p5\tei-p5-exemplars-0.7\xml\dtd\tei_all.dtd 2007-05-26 Release -->
<TEI>
```

```
<teiHeader type="text">
```

```
<fileDesc>
  <titleStmt>
    <title type="main">Chapter Title</title>
    <author>Chapter Author</author>
  </titleStmt>
  <editionStmt>
    <edition>
      <date value="2007">2007</date>
    </edition>
  </editionStmt>
  <publicationStmt>
    <distributor>
      <address>
        <addrLine>
          <name key="Pub" type="organisation">Publisher</name>
        </addrLine>
        <addrLine>Address</addrLine>
        <addrLine>
          <name type="place">Place</name>
        </addrLine>
        <addrLine>Email</addrLine>
      </address>
    </distributor>
    <idno type="demo">DEMO_Ch1</idno>
    <availability status="free">
      <p>Public domain</p>
    </availability>
```

Our text as TEI . . .

**The header
goes on for
two pages...**

**...preserving
all sorts of useful
information.**

```
<publisher>Publisher</publisher>
<pubPlace>Place</pubPlace>
<date value="2007-06-09">2007-06-09</date>
</publicationStmt>
<notesStmt>
  <note>Prototype TEI header</note>
</notesStmt>
<sourceDesc>
  <p>Chapter in search of a book.</p>
  <p>
    <bibl>Example TEI Chapter 1. Chapter Author.</bibl>
  </p>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <editorialDecl>
    <p>Minimal TEI encoding. Chapter in search of a book.</p>
  </editorialDecl>
  <refsDecl>
    <p>No refs; no IDs assigned.</p>
  </refsDecl>
</encodingDesc>
<profileDesc>
  <langUsage>
    <language ident="en" usage="100">English.</language>
  </langUsage>
</profileDesc>
<revisionDesc>
  <change>
    <list>
      <item><date value="2007-06-09">June 9, 2007</date> Created initial
version.</item>
    </list>
  </change>
</revisionDesc>
```

```
</change>
</revisionDesc>
</teiHeader>
```

```
<text>
  <body>
    <div type="Chapter" n="1">
      <head>Chapter Title</head>
      <opener>
        <byline><docAuthor>Chapter Author</docAuthor>
        <name type="affiliation" rend="italics">Author Identification</name></
byline>
      </opener>

      <p>Here&#x2019;s some text at the beginning of this chapter. Let&#x2019;s
make one more line&#x2019;s worth.</p>
      <div type="level-1">
        <head>Level One Subhead</head>
        <p>Here&#x2019;s some more text. This author&#x2019;s
<hi rend="italics">a pretty nice girl</hi>, but she doesn&#x2019;t have a lot to say.
</p>

        <div type="level-2">
          <head>Level Two Subhead</head>
          <p>The end.</p>
        </div> <!-- end of level-2 div -->
      </div><!-- end of level-1 div -->
    </div><!-- end of chapter div -->
  </body>
</text>
</TEI>
```

Note separation of semantics & formatting

Note nested structure

. . . or utilitarian . . .



DocBook

The “Crank It Out” DTD ...

DocBook



- Common general-purpose book model
- Widely used for technical documents, manuals
- Not as often used for scholarly, trade, reference, or textbooks
- Vendors and technical writers are often most familiar with DocBook

DocBook is often used in structured environments

```
<?xml version="1.0" encoding="us-ascii"?>
<!DOCTYPE chapter PUBLIC "-//OASIS//DTD DocBook XML V4.3//EN"
"http://www.oasis-open.org/docbook/xml/4.3/docbookx.dtd">
```

Our text as DocBook

```
<chapter label="1">
```

```
<chapterinfo>
```

```
<authorgroup>
```

```
<author>
```

```
<firstname>Chapter</firstname>
```

```
<surname>Author</surname>
```

```
<affiliation>
```

```
<shortaffil remap="ITAL">Author Identification</shortaffil>
```

```
<jobtitle></jobtitle><orgname></orgname>
```

```
</affiliation>
```

```
</author>
```

```
</authorgroup>
```

```
</chapterinfo>
```

Author's name & affil. generated from metadata

```
<title>Chapter Title</title>
```

```
<para>Here's some text at the beginning of this chapter. Let's make one more line's worth.</para>
```

Context-sensitive formatting

```
<sect1>
```

```
<title>Level One Subhead</title>
```

```
<para>Here's some more text. This author's
```

```
<emphasis remap="ITAL" role="italics">a pretty nice girl</emphasis>, but she doesn't have a lot to say.</para>
```

Preserving a record of previous markup

```
<sect2>
```

```
<title>Level Two Subhead</title>
```

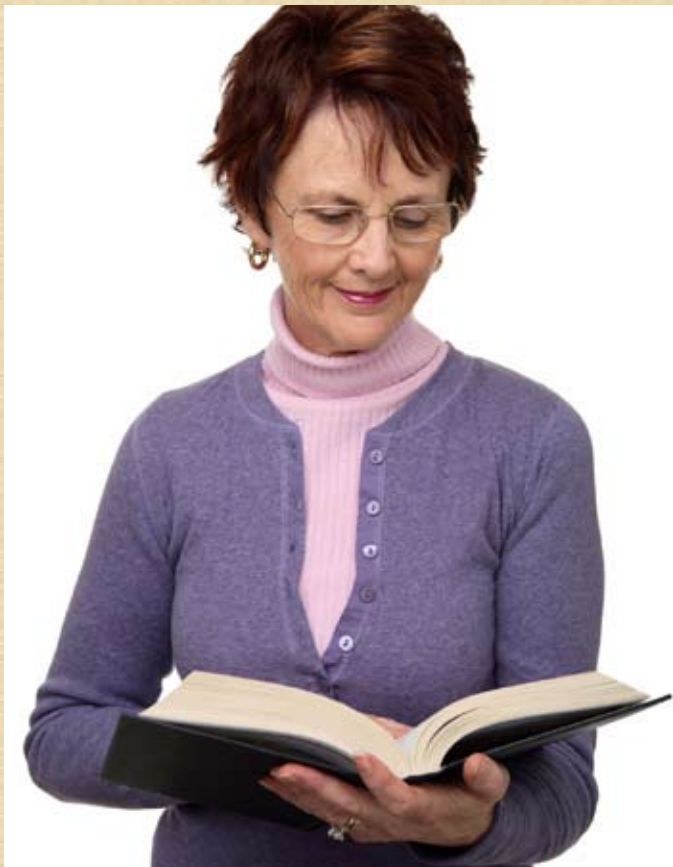
```
<para>The end.</para>
```

```
</sect2>
```

```
</sect1>
```

```
</chapter>
```

. . . or strike a useful balance . . .



NLM

The “Works and Plays Well Together” DTD . . .



The **NLM** Family of DTDs

- Based on a comprehensive study of STM journals
- Designed to be customized, adapted
- Has a permissive “Archival & Interchange” model and stricter “Publishing” and “Authoring” models
- Very widely adopted and well maintained
- NLM Book DTD is just a re-tooled journal DTD

NLM is the “no-brainer” basis for journal DTDs today

```
<?xml version="1.0" encoding="us-ascii"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD Journal Publishing DTD v2.3
20070202//EN"
"journalpublishing.dtd">
```

*Our text as NLM
Publishing XML*

```
<article>
```

```
<front>
```

```
<journal-meta>
```

```
<journal-id journal-id-type="publisher">Publisher</journal-id>
```

```
<issn>0001-0001</issn>
```

```
<publisher>
```

```
<publisher-name>Publisher</publisher-name>
```

```
</publisher>
```

```
</journal-meta>
```

```
<article-meta>
```

```
<article-id pub-id-type="other">1</article-id>
```

```
<title-group>
```

```
<article-title>Chapter Title</article-title>
```

```
</title-group>
```

```
<contrib-group>
```

```
<contrib contrib-type="author">
```

```
<name><surname>Author</surname>
```

```
<given-names>Chapter</given-names></name>
```

```
<aff>Author Identification</aff>
```

```
</contrib>
```

```
</contrib-group>
```

```
<pub-date pub-type="pub">
```

```
<day>9</day><month>June</month><year>2007</year>
```

```
</pub-date>
```

```
<history>
```

```
<date date-type="received">
```

```
<day>9</day>
```

**This model is
for journals!**

```
<month>6</month>
<year>2007</year>
</date>
<date date-type="rev-request">
<day>9</day>
<month>6</month>
<year>2007</year>
</date>
</history>
<abstract></abstract>
</article-meta>
</front>
```

```
<body>
  <p>Here's some text at the beginning of this chapter. Let's
make one more line's worth.</p>
  <sec id="bid_002">
    <title>Level One Subhead</title>
    <p>Here's some more text. This author's a pretty
nice girl, but she doesn't have a lot to say.</p>
    <sec id="bid_003">
      <title>Level Two Subhead</title>
      <p>The end.</p>
    </sec>
  </sec>
</body>
</article>
```

Our text as NLM Book XML

```
<?xml version="1.0" encoding="us-ascii"?>
<!DOCTYPE book-part PUBLIC "-//NLM//DTD Book DTD v2.3 20070202//EN"
"book.dtd">
```

```
<book-part id="bid_001" book-part-type="chapter" book-part-number="1">
```

```
<book-part-meta>
```

```
<title-group>
```

```
<title>Chapter Title</title>
```

```
</title-group>
```

```
<contrib-group>
```

```
<contrib contrib-type="author">
```

```
<name><surname>Author</surname>
```

```
<given-names>Chapter</given-names></name>
```

```
<aff>Author Identification</aff>
```

```
</contrib>
```

```
</contrib-group>
```

```
<history>
```

```
<date date-type="created">
```

```
<day>9</day>
```

```
<month>6</month>
```

```
<year>2007</year>
```

```
</date>
```

```
<date date-type="updated">
```

```
<day>9</day>
```

```
<month>6</month>
```

```
<year>2007</year>
```

```
</date>
```

```
</history>
```

```
<abstract></abstract>
```

```
</book-part-meta>
```

**CN, CT, AU, &
AFF are ONLY in
the metadata**

**Note that
<sec>s are recursive
("nested")**

```
<body>
  <p>Here's some text at the beginning of this chapter. Let's
make one more line's worth.</p>
  <sec id="bid_002">
    <title>Level One Subhead</title>
    <p>Here's some more text. This author's a pretty
nice girl, but she doesn't have a lot to say.</p>
    <sec id="bid_003">
      <title>Level Two Subhead</title>
      <p>The end.</p>
    </sec>
  </sec>
</body>
</book-part>
```

**"bid" attributes
uniquely identify a
chunk of content;
not required, but
usual & useful**

XML: Extensible Markup Language

**Is XML making publishing
complicated?**

XML: Extensible Markup Language

**Is XML making publishing
complicated?**

**No, publishing's inherently
complicated.**

XML: Extensible Markup Language

**Is XML making publishing
complicated?**

**No, publishing's inherently
complicated.**

XML is helping to make it easier!

Thanks!



Bill Kasdorf

Vice President, Apex Publishing

bkasdorf@apexcovantage.com

+1 734 904 6252