XML Models for Books

It's all about whatcha got and whatcha wanna do with it. . . .

Bill Kasdorf

Vice President, Apex Content Solutions

General Editor, *The Columbia Guide to Digital Publishing*

There's a reason why DTDs and schemas are called "models."

Some common book “models”

- Scholarly monograph
- Textbook
- Reference book (but encyclopedia ≠ dictionary)
- Directory
- Catalog
- Technical manual (but programming manual ≠ auto repair manual ≠ Boeing 737 documentation)
- Trade book (but cookbook ≠ coffeetable book)

Some common book “models”

- Scholarly monograph
- Textbook
- Reference book (but encyclo
- Directory
- Catalog
- Technical manual (but progra

auto repair manual ≠ B2 bom

- Trade book (but cookbook ≠

These models have different:

- Structures
- Semantics
- Purposes
- Audiences
- Type/design conventions

DTDs can be strict . . .

APEX

ISO 12083

The Mother Superior of DTDs . . .

The ISO 12083 DTD

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

Most big STM journal DTDs are still 12083-based

or permissive . . .

APEX

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
- Own table model (can invoke CALS or XHTML)
- Can invoke TeX or MathML for math
- Enormous resource; TEI Lite is too simplistic

Most humanities scholarship is TEI-based

or utilitarian . . .

APEX

DocBook

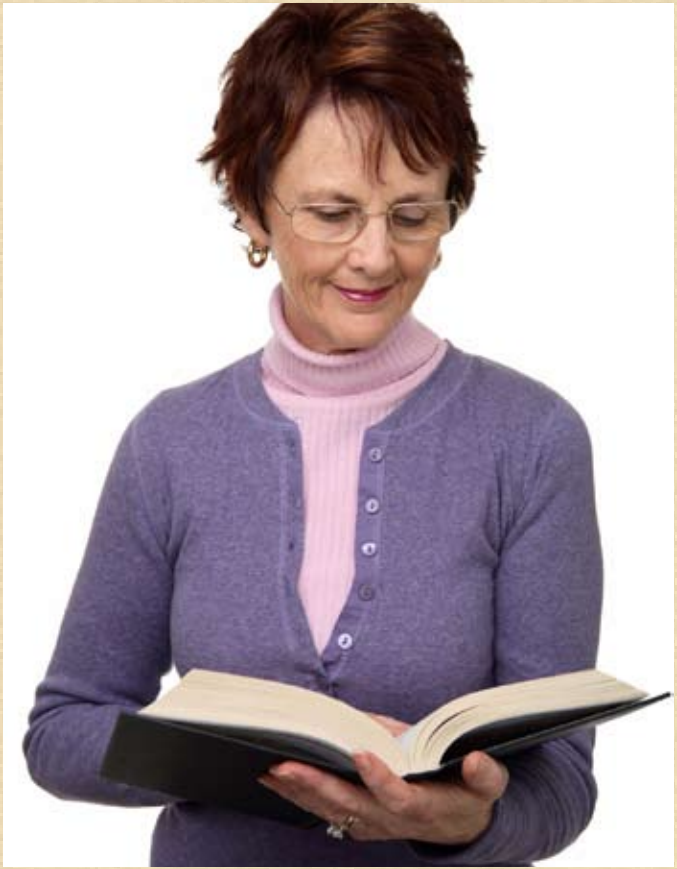
The "Crank It Out" DTD . . .

DocBook

- Common general-purpose book model
- Widely used for technical documents, manuals
- Not often used for scholarly/trade/ref/textbooks
- CALS tables (can invoke XHTML)
- Own math model (can invoke MathML)
- Vendors and tech writers familiar with DocBook

DocBook is often used in structured environments

or strike a
useful balance . . .



NLM

The "Works and Plays Well Together" DTD . . .

The **NLM Book DTD**

- Created for NCBI Bookshelf; now called the "**Book and Book Collection Tag Set**"
- Not based on broad study of books, as the journal models were on journals
- Robust metadata/semantics
- XHTML or CALS tables, MathML for math
- Appealing when mixed with NLM journal XML
- Recently updated: v. 3.0 released 11/21/08

The **NLM Book DTD**

For example . . .

- **<citation-type>** eliminated,

replaced with three attributes:

- **publication-format** (e.g., print vs. online)
- **publication-type** (e.g., journal vs. book)
- **publisher-type** (e.g., stds. body, gov't)

al

- Appealing when mixed with NLM journal XML
- Recently updated: **v. 3.0 released 11/21/08**

or serve a particular purpose . . .

APEX

DTBook

The most important DTD people have never heard of . . .

The **DTBook DTD**

- Part of DAISY/NISO “Digital Talking Book” standard
- Now part of IDPF’s new .epub format for e-books
- First priority: structure—Enables access, navigation, subsetting; accommodates flat or nested structures
- The degree of markup is not mandated; markup needed for print is DAISY’s recommended minimum
- XHTML tables, images and alt attribute for math

The **DTBook DTD**

NIMAS: US National File Format for Education

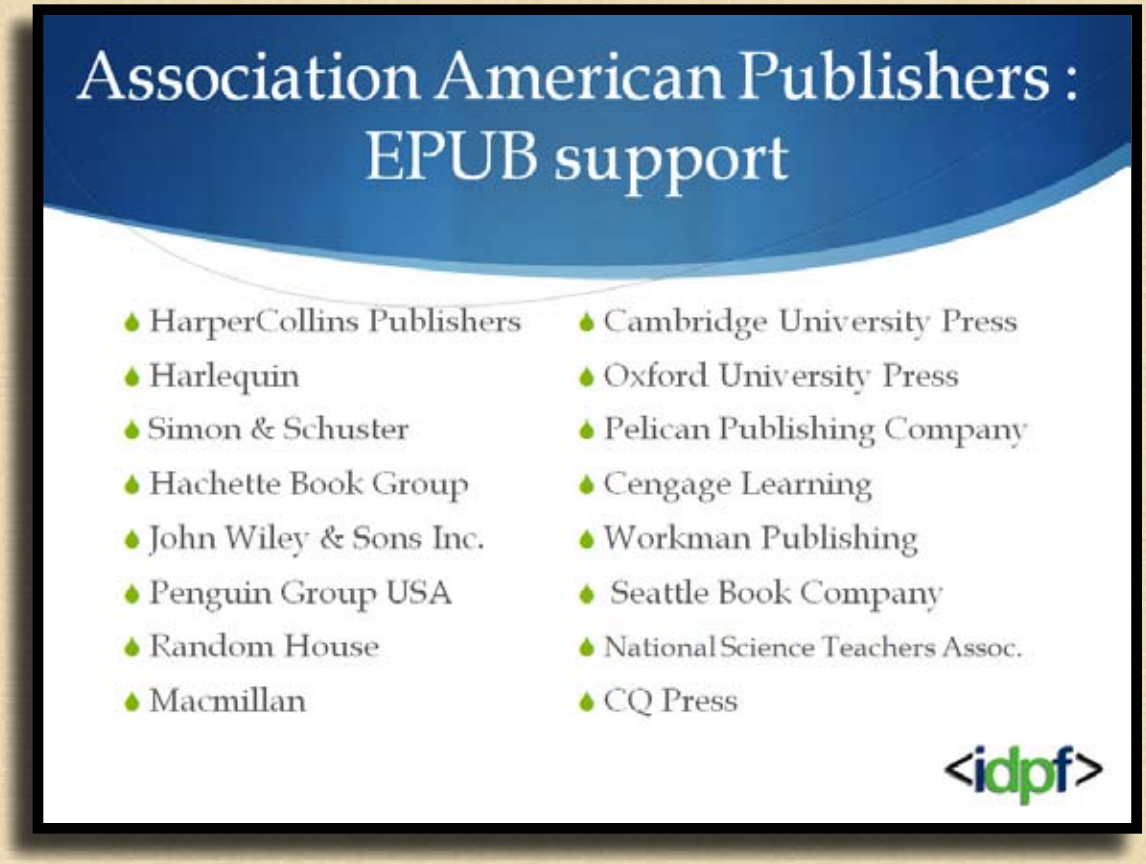
- Implementation of DTBook for US education
- Baseline Element Set (min. requirement, nested): publishers must supply this XML (+ PDF for visual reference, + package file)
- Optional Element Set (rest of DTBook set)
- “Guidelines for Use” follow DAISY, but stricter

The new **.epub** standard from IDPF

- Successor to OEB (Open eBook) standard
- **OPS 2.0** (Open Publication Structure):
  Text markup standard (XHTML + DTBook)
- **OPF 2.0** (Open Packaging Format):
  How the components of a digital book are related
- **OCF 1.0** (Open Container Format):
  How to encapsulate an .epub w/ optional files

APEX

The UK went "straight to EPUB"



APEX

+ Sony Reader, Adobe Digital Editions, and Stanza for iPhone

There are some **.epub issues** . . .

- **Formatting issues:** Should the e-book . . .
  —Look "exactly" like the print? *[Don't go there . . .]*
  —Reflect the print format somewhat? *[Feasible]*
  —Use standard tagging and CSS? *[Good idea!]*
- **Rights issues:** Embedded fonts can be pirated; IDPF is working on "font mangling" spec for .epub
- **Linking** within and between e-books
- **Annotations, notes**—esp. for HE and STM

or, for something completely different . . .

DITA

The "Slice & Dice" DTD . . .

DITA

- DITA = Darwin Information Typing Architecture
- Designed for modular information
- Content is created in “topics,” not documents
- Topics are assembled & reassembled by “maps”
- Becoming the new standard for tech docs

DITA is ideal for granular, modular information—updating a topic updates all docs it’s used in

. . . not to mention (okay, I will) models used in books . . .

Models used as **components** in other models

- MathML for math equations
- CALS/Oasis table model
- SVG—Scalable Vector Graphics
- XHTML (modular XHTML2 is being developed)
- Dublin Core (basic bibliographic metadata)
- ONIX (for marketing/distribution & other info)
- OAI-PMH—Open Archives Initiative Protocol for Metadata Harvesting (no, not just for free content!)

It's very nice not to have to reinvent these wheels!

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 may be **mandated** in a given industry

Why **customize** a standard DTD?

- Too simplistic or **generic** for your needs
- Or, more **complex** than you need or can handle
- Needs and capabilities **change** over time:
  —Requirements of customers, vendors, partners
  —Capabilities of software, tools, and staff
- **Semantics** to enable, enhance, and expedite discovery, navigation, and use = **VALUE**

Example: **Cookbook** content

Could you tag this with a standard model? Sure.

Disaster

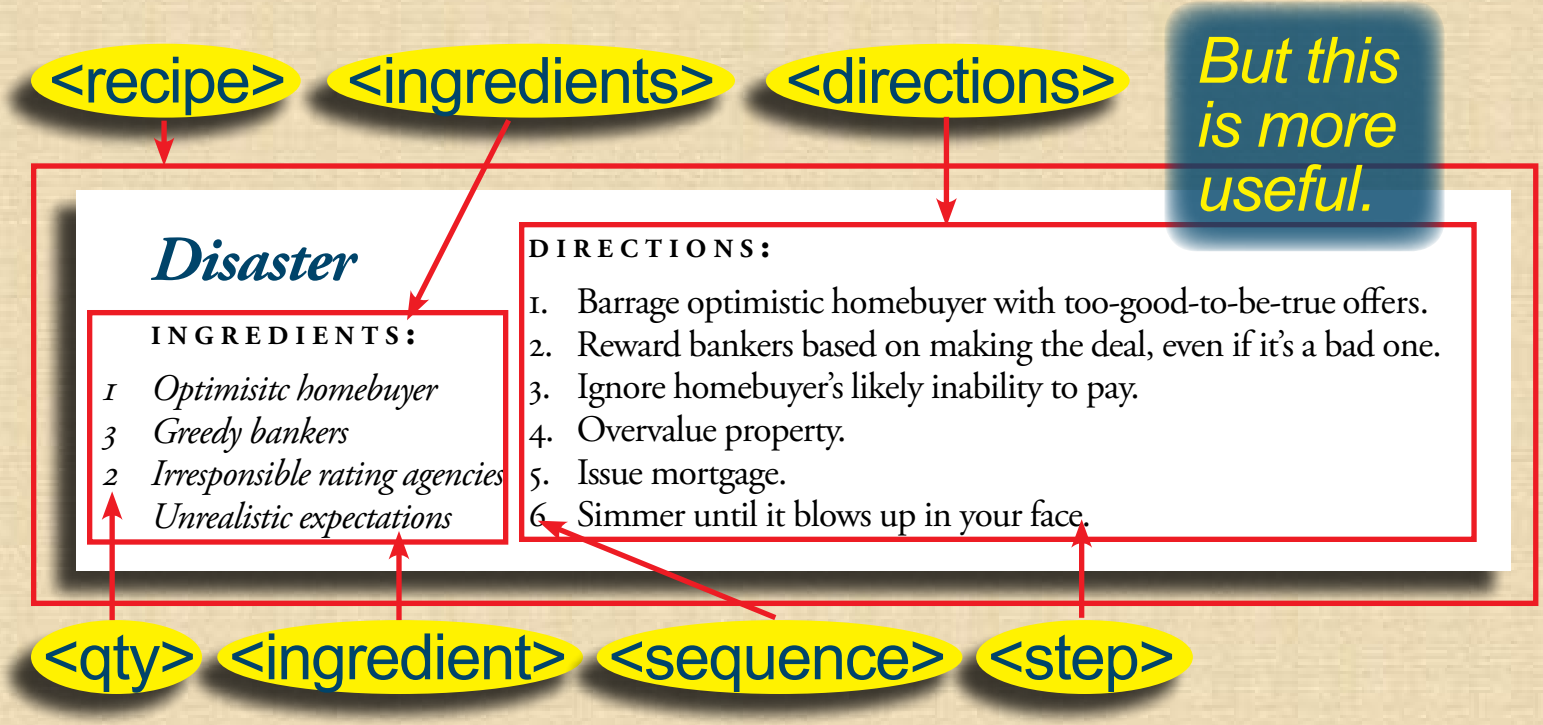
INGREDIENTS:

1 *Optimisitc homebuyer*
3 *Greedy bankers*
2 *Irresponsible rating agencies*
Unrealistic expectations

DIRECTIONS:

1. Barrage optimistic homebuyer with too-good-to-be-true offers.
2. Reward bankers based on making the deal, even if it's a bad one.
3. Ignore homebuyer's likely inability to pay.
4. Overvalue property.
5. Issue mortgage.
6. Simmer until it blows up in your face.

Example: **Cookbook** content



XML Models for Books

[Optimist says:]

What a wealth of options!

[Pessimist says:]

Clear as mud!

XML Models for Books

It's not XML's fault
this is complicated.
Books are messy.

Thanks!

Bill Kasdorf

Vice President, Apex Content Solutions

bkasdorf@apexcovantage.com

+1 734 904 6252