PB138 – Markup Languages

Tomáš Pitner

May 6, 2013

Obsah

DocBook as an example of a more complex markup

- big project, one complex markup for all programmmers documentation
- now many other purposes - writing papers (article), books (book), chapters (chapter), sections (section, sectX)
- authored by Norman Walsh (formerly Sun Microsystems Inc.)
- details, DTD, help, software, styles, see docbook.org (`http://docbook.org`)
- probably the biggest markup for technical documentation ever
- there is the TDG (DocBook: The Definitive Guide) - also as Windows Help (`/~tomp/xml/tdg-en-2.0.7.chm`)

What is Docbook?

- Docbook is a XML (and SGML) markup for writing documents, namely of technical nature (computer/software manuals, technical documentation).
- Originally as a tool to cope with large UNIX-systems documentation.
- In principle, DB is a logical (semantic) markup (i.e. visual representation is not of importance when writing the source. Text is created using semantic elements for:
    - big text blocks (book, paper, chapter, section, paragraph, screen...)
    - smaller in-line parts (emphasized, link, product name, command,...)
    - multimedia elements (images, videos, sounds...)
    - helper elements and metadata (title, authoring, date of creation, copyright, index items, ToC...)

Advantages of Docbook

- Easy processing:
  - visualization (using CSS, using XSLT for transf. to HTML, via LaTeX or XSL:FO to PDF, but also PostScript, PDF, RTF, DVI and plain-ASCII...), or documentation/help formats (HTML Help, Microsoft CHM, man-pages)
  - selected parts or elements can be extracted separately (take the intro chapter, generate the book ToC...) or connect more texts into one

Origin

- Docbook since beginning of 90s (1991), as a SGML markup that time.
- After introduction of XML as de-facto standard for semistructured data (W3C spec. XML in 1998) is Docbook predominantly encoded in XML – mainly because of plethora of tools available.
- Further development under OASIS (`http://www.oasis-open.org`) (The Organization for the Advancement of Structured Information Standards).
- Jirka Kosek (`http://www.kosek.cz`) is involved in the development, the editor of specs. is Norm Walsh (`http://norman.walsh.name`).

Storing files

Usuale extension for files containing Docbook documents is `.dbk`, or simply `.xml`

MIME type for Docbook is `application/docbook+xml`

Document categories

The nature (purpose, size) of the document is mainly determines by using certain *structural elements*. The categories include:

set collection of (`book`) or other collections – may be nested.

book book containing (`chapters`), papers (`article`) or parts (`part`), may contain indices (index), appendices etc.

part part containing one or more chapters, may be nested, may contain intro texts.

article paper, may contain a sequence of block element (like chapters, paragraphs).

chapter named and usually numbered section of a bigger document (book, paper).

appendix příloha

dedication decication of a certain element

Block elements

- paragraphs
- tables
- lists
- examples
- figures, etc.

these block elements are visualized in the order they will be read, ie. – top-down in Western languages, but left-right in Chinesse.

In-line elements

contained in block elements:

- emphasized text (`emphasis`...)
- links (eg. `link`, `ulink`, `olink`...)
- meaning (keyword, command, file name...)

Example of Docbook 5 document

Docbook 5 is the latest but still developed standard. It uses *XML Namespaces* and no DOCTYPE declaration.

```
<?xml version="1.0" encoding="UTF-8"?>
<book id="simple_book" xmlns="http://docbook.org/ns/docbook
   <title>Very simple book</title>
   <chapter id="chapter_1">
     <title>Chapter 1</title>
     <para>Hello world!</para>
     <para>I hope that your day is proceeding <emphasis>spl
   </chapter>
   <chapter id="chapter_2">
     <title>Chapter 2</title>
     <para>Hello again, world!</para>
   </chapter>
</book>
```

Still Docbook 4.x is predominatnly used mainly for legacy docs

The same in Docbooku 4.4

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE book PUBLIC "-//OASIS//DTD DocBook XML V4.4//EN"
                    "http://www.oasis-open.org/docbook/xml/
<book id="simple_book">
   <title>Very simple book</title>
   <chapter id="chapter_1">
     <title>Chapter 1</title>
     <para>Hello world!</para>
     <para>I hope that your day is proceeding <emphasis>spl
   </chapter>
   <chapter id="chapter_2">
     <title>Chapter 2</title>
     <para>Hello again, world!</para>
   </chapter>
</book>
```

Version 5.x or 4.y?

1. Either, or... You won't do a big mistake still using 4.y, since there is plethora of tools and docs.
2. Conversion to DB 5 any time later...

DocBook: layers and customization

- DocBook can be used as basic (Full)
- or simplified (Simplified) or to make a
- customization.

Which means:

- modify schema
- evt. modify (XSL) styles
- XSL styles by importing the original style and overriding selected templates

Docbook Layers - Simplified

derived languages/markups can be created by reduction or extension of allowed elements:

Simplified Docbook from a family of elements just one is preserved/left: `programlisting`, but not `screen`
No "big things" like books, just articles
Any doc. in Simplified Docbook is also a (full) Docbook doc.
Docs for Simplified Docbook online (`http://www.docbook.org/schemas/simplified`)

Docbook Slides

- Extension :-) of Simplified Docbook
- For writing (PowerPoint-like) presentations – "foils".
- XSLT styles allow to make static- or JavaScript-enabled web/HTML pages.
- Modern browsers can even navigate through the structure (go to next slide, toc, etc.).

Editors

- In the worst case, any *plain-text editor* can be used if supporting the required charset and encoding (eg. Unicode/UTF-8).
- Better to use any editor with auto-closing (or even auto-completion) of elements.
- If an on-the-fly validation is supported - the best!
- Ideally an WYSIWYG producing a valid Docbook text – eg. XML Mind (XXE) or oXygen.

Available editors

xmlmind http://xmlmind.com (http://www.xmlmind.com/) of Pixware powerfull WYSIWYG editor for Docbook, DITA, XHTML and other formats including ebooks, can be further customized, suitable for enterprise environment and integration. Professional- and Evaluation- license.

oXygen Synchro Soft SRL's (http://www.oxygenxml.com/) oXygen Editor/Developer/Author.

GNU Emacs with (http://www.thaiopensource.com/nxml-mode/)nxml-mode

Validation Tools

- Docbook 4.x was DTD-constraint/defined
- Docbook 5.x uses namespaces and is RelaxNG/Schematron-constraint
- for transition, see http://docbook.org/docs/howto/ (http://docbook.org/docs/howto/)
- and complete reference (http://docbook.sourceforge.net/release/xsl/current/doc/) to use Docbook XSL

Transformation Tools

Mainly for conversion into other document formats (”Office-like” as Office Open XML, Open Document Format, RTF, Wordprocessing XML) or visualization via PDF, PS, XSL:FO, or web formats (XHTML 1.x, XHTML 5)

- Fundamental tools are Docbook XSL (`http://en.wikipedia.org/wiki/DocBook_XSL`) styles
- well parametrized, rich, modifiable
- a book on Docbook XSL by Sagehill (`http://www.sagehill.net/docbookxsl/index.html`) publishers
- complete reference (`http://docbook.sourceforge.net/release/xsl/current/doc/`) to use Docbook XSL

Co je TEI

Iniciativa směřující k vytvoření a aplikacím podpory zachycování textů různé povahy ve standardizované formě

- dnes v XML syntaxi (P5), dříve SGML (po P3) nebo obojí (P4)
- rozsáhlé značkování (ještě větší počet elementů než např. Docbook)
- lépe podporuje metadata dokumentů a jejich životní cyklus (vznik, revize)
- používá se pro různorodé dokumenty (texty pořizované na počítači, skenované texty, historické dokumenty, dokumenty v neevropských jazycích)
- značkování je modulární - lze sestavit na míru potřebám

Aplikace TEI značkování

- příklady textů v TEI (http://wiki.tei-c.org/index.php/Samples) (především XML)
- Manuál (Guidelines (http://www.tei-c.org/Guidelines/P5/)) pro TEI P5

Darwin Information Typing Architecture (DITA)

IBM and the Consortium OASIS have introduced DITA (`http://docs.oasis-open.org/dita/v1.0/archspec/ditaspec.toc.html`) architecture as:

- Nástroj pro tvorbu tematicky orientovaného značkovaného obsahu s možností specializace pro zvláštní účely.
- Není to, na rozdíl např. od Docbooku, jedno pevné značkování.
- Využívá se principů podobných jako v objektových jazycích.
- Specializace znamená podědit vlastnosti (např. formátování) a konkretizovat je.
- Používá se tam, kde se tvoří rozsáhlý, vysoce strukturovaný, znovupoužitelný obsah s přesně vymezenou sémantikou.

Historie a současnost

- od roku 2001 DITA vyvíjena společností IBM (motivace: pevná značkování nestačí...)
- 2004 – IBM daruje standard do správy OASIS
- O vývoj se stará OASIS DITA Technical Committee (http://www.oasis-open.org/committees/dita/).
- Duben 2005 – Version 1.0 of the DITA specification:
  - OASIS Darwin Information Typing Architecture (DITA) Language Specification (http://xml.coverpages.org/DITAv10-OS-LangSpec20050509.pdf)
  - OASIS Darwin Information Typing Architecture (DITA) Architectural Specification (http://xml.coverpages.org/DITAv10-OS-ArchSpec20050509.pdf)

Základní pojmy

topic téma – jednotka informace daná názvem a obsahem; dostatečně malá, aby byla dále nedělitelná z hlediska obsahu a pořízení (menší už by nedávala ucelený smysl) – např. odpověď na jednu otázku

map dokument organizující témata do větších jednotek se zachycením vztahu mezi tématy, vč. např. obsahu

specialization specializace – je technika umožňující definovat nové strukturální typy nebo nové informační domény) s maximálním znovupoužitím existujícího návrhu a kódu, důraz je kladen na snižování nákladů přechodu na nové typy (výměna dat, migrace, správa)

structural vs. domain specialization *strukturální specializace* – umožňuje tvořit nové typy témat (topic types) nebo map (map types)

doménová specializace – dovoluje vznik nového značkování použitelného pro více strukturálních typů

Příklad

CambridgeDocs nabízí řešení pro pořizování a správu dokumentů navržených podle DITA – xDoc Pro (`http://www.cambridgedocs.com/solutions/dita.htm`).