

# **Guidelines for the conversion of print originals to DTBook XML**

**Version 1.3  
2<sup>nd</sup> October 2007**

## Table of contents

About the guidelines.....	4
Guideline contents.....	4
Example collection.....	4
A short description of TPB's production process .....	4
1 About the standard .....	5
1.1 The DAISY Consortium and the DAISY standard .....	5
1.1.1 DTBook.....	5
1.1.2 Which version of the standard is to be used?.....	5
2 Guidelines for scanning text and images.....	6
2.1 Scanning .....	6
2.2 OCR and proofreading .....	6
2.3 Images from scanned print originals .....	6
2.3.1 General .....	6
2.3.2 Handling of specific image types .....	6
2.3.3 Colour images .....	7
2.3.4 Greyscale images.....	7
2.3.5 Images containing text .....	7
2.3.6 Editing images.....	7
2.3.7 Delivery.....	8
3 General guidelines for mark up.....	8
3.1 TPB applies a subset of the DTBook standard.....	8
3.2 A complement to the DTD .....	8
3.3 Elements allowing both text and elements as children (%flow) .....	9
3.3.1 Nested lists .....	9
3.4 Placement of "floating" elements.....	10
3.5 The XML file in general.....	10
3.5.1 File name .....	10
3.5.2 Processing instructions.....	10
3.5.3 XML declaration .....	10
3.5.4 Document type declaration.....	10
3.5.5 Character encoding.....	10
3.5.6 xml:lang.....	11
3.5.7 Representation of punctuation, typographic characters and special characters. 11	
3.5.8 Hyphenation .....	13
3.5.9 Empty elements .....	13
3.6 Explanation of wording .....	14
4 Detailed guidelines for markup .....	15
4.1 Guidelines for mark up of structural elements .....	15
4.1.1 dtbook.....	15
4.1.2 head .....	15
4.1.3 book.....	15
4.1.4 frontmatter.....	15
4.1.5 bodymatter.....	17
4.1.6 rearmatter .....	17
4.1.7 level1-6.....	17
4.2 Guidelines for the mark up of block elements - listed alphabetically.....	19
4.2.1 annoref & annotation.....	19

4.2.2	blockquote .....	19
4.2.3	doctitle .....	19
4.2.4	docauthor .....	19
4.2.5	div and the class="pgroup" attribute .....	19
4.2.6	h1 - h6.....	21
4.2.7	imggroup .....	21
4.2.8	meta .....	23
4.2.9	note & noteref.....	24
4.2.10	p.....	25
4.2.11	pagenum .....	26
4.2.12	poem.....	27
4.2.13	sidebar .....	27
4.3	Guidelines for mark up of inline elements – listed alphabetically .....	27
4.3.1	br.....	27
4.3.2	code .....	27
4.3.3	em.....	27
4.3.4	strong.....	28
4.3.5	sub .....	29
4.3.6	sup .....	29
4.4	Guidelines for the mark up of lists .....	29
4.4.1	list.....	29
4.4.2	dl.....	31
4.5	Guidelines for the mark up of tables .....	32
4.5.1	Table.....	32
4.5.2	Caption .....	33
4.5.3	tr .....	33
4.5.4	th.....	33
4.5.5	td.....	33
5	Guidelines for the mark up of unified content .....	33
6	Guidelines for the notation and mark up of mathematics, physics and simpler chemistry	
	33	
6.1	Mark up of formula based content .....	34
6.1.1	Block .....	34
6.1.2	Inline.....	34
6.2	Notation.....	35
7	Changes .....	36
	Version 1.3 - 2007-10-02 .....	36
	Authors:.....	37

## **About the guidelines**

These guidelines describe those quality requirements established by TPB to be applied by suppliers producing text in DTBook XML format from print originals.

### ***Guideline contents***

The guidelines are divided into five sections:

1. About the standard
2. Guidelines for scanning text and images
3. General guidelines for mark up
4. Detailed guidelines for mark up
5. Guidelines for the mark up of unified content
6. Guidelines for the notation and mark up of mathematics, physics and simpler chemistry

### ***Example collection***

A collection of examples has been established illustrating principles described within these guidelines for a range of different material. The collection will be expanded continually during the contract period as solutions to problems are established.

The collection is available in both .PDF and .xml formats and is distributed by TPB.

### ***A short description of TPB's production process***

Since 2005, TPB has had a unit producing files in DTBook XML format. Said production has required the development of a process similar in detail to those services to be purchased by TPB.

The following is a short description of TPB's production process:

1. Removal of the print original's binding.
2. a) Printed pages converted, via scanning, to PDF format or  
b) Printed pages converted, via scanning, directly to RTF format. This method involves an OCR process that is suitable for only the simplest material.
3. PDF file zoned, zones then OCR processed using Abbyy FineReader 8.0. Resulting files saved in Microsoft Word 2003 .doc format.
4. The saved file is examined for a subset of DTBook elements that are marked up using formatting templates for Microsoft Office Word 2003 developed at TPB.
5. The edited file is converted to a basic DTBook XML.
6. The basic DTBook XML is refined in Altova XMLSpy, utilising a number of XSL transformations and operator oversight, until the required level of DTBook conformity and quality is achieved.
7. Late-stage correction is performed using the Authentic view in XMLSpy. Authentic can also be used as a free-standing program

# 1 About the standard

## 1.1 *The DAISY Consortium and the DAISY standard*

The DAISY Consortium assumed the role of maintenance agency for the DAISY 3 standard in 2005, the formal name for the standard is *DAISY/NISO z.39.68-2005*. The standard has been published in its entirety and can be found at <http://www.daisy.org/z3986/2005/z3986-2005.html>

### 1.1.1 DTBook

DTBook is that part of the DAISY 3 standard describing the text content of a DAISY book.

The use of the DTBook standard means a radical change for TPB. For the first time documents that are richly marked up using a homogenous format can act as a source document for the production of all text based media at TPB. While the DTBook file was developed for the production of DAISY talking books, it can also be used, though not exclusively, for the following media:

- Braille volumes
- TPB's proprietary e-book format Textview
- Automatic conversion to HTML, PDF and other e-book formats.

The DTBook standard has two main influences:

- *The Chicago Manual of Style*, the de facto-standard for the design of printed materials and the naming of elements of content
- *XHTML 1.0*, the accepted standard for mark up of web pages, developed by the World Wide Web Consortium

The DTBook standard is not meant to be a blanket solution for all possible content. It is, instead, a collection of the most common and applicable elements present in texts. In particular, teaching materials tend to include content that does not lend itself to accurate mark up using the elements available in the DTBook standard. Examples of such content are included in these guidelines.

### 1.1.2 Which version of the standard is to be used?

Production on behalf of TPB shall use the current DAISY standard, i.e. DAISY/NISO z.39.68-2005. See also [www.daisy.org](http://www.daisy.org)

## 2 Guidelines for scanning text and images

### 2.1 Scanning

The properties of print originals that affect scanning will vary from title to title and settings require adapting to suit each case. Two passes, a text specific and an image specific, might be necessary to achieve necessary quality, depending on the scanner being used (See Section 2.3 *Scanning and editing images*). Text content, regardless of background colour, must be scanned in as a black and white image. Images must be reproduced in accordance with the print original. The majority of print originals can have their binding removed to facilitate scanning. The exception to this regards titles that have been loaned to TPB as opposed to being purchased. These need to be scanned using a flatbed scanner. A closer description of the handling of titles loaned to TPB will be included with relevant orders.

### 2.2 OCR and proofreading

OCR (Optical Character Recognition) involves the conversion of images of text to machine-editable text.

Text produced using OCR must be proofread. The aim of proofreading is to reduce the amount of errors introduced into the text by the scanning and OCR processes. Proofreading does not include the correction of spelling errors present in the print original. TPB places no demands for the correction of obvious errors in the print original, leaving this decision to each producer. The resulting digital text is expected to mirror the print original (with reservation for the preceding sentence) to 100%. This requirement results in the right for TPB to return material found to have proofreading errors.

### 2.3 Images from scanned print originals

To facilitate a simplified handling of images in print originals the following guidelines are to be applied:

#### 2.3.1 General

The practice of including images as a supplement to text content is prevalent. The relevance of included imagery to the text in question cannot be judged generally though three elements can be used as a guide:

- What is the nature of the text?
  - Social sciences, Humanities, Biography and so on
- Is the image directly referenced in the text?
- Is the image necessary for the understanding or appreciation of the text?

#### 2.3.2 Handling of specific image types

Four types of image can occur in print originals:

- Informational – i.e. images pertaining to the subject of the text that represent facts in a visual manner, e.g. diagrams and charts

## Guidelines for conversion of print originals to DTBook xml

- Biographical-Ornamental – i.e. images that while they are not informational in the manner of diagrams or charts are connected to and may even be referenced by the text, e.g. portrait and still life photography
- Logotypes – a design or symbol used by a company to advertise its products
- Vignettes – graphics used to separate sections or chapters, decorate borders or flyleaves, jackets or panels
- Iconic – i.e. an image that stands for its object by virtue of a resemblance or analogy to it
- Formatting – i.e. images that have no connection to the subject matter and are purely an artefact of layout and design

Handling of the preceding types of images should take place as follows:

- Informational – should be scanned and linked to via DTBook markup
- Biographical-Ornamental - should be scanned and linked to via DTBook markup
- Logotypes – images of this type are not to be included in the DTBook markup and consequently not delivered as images
- Vignettes – images of this type are not to be included in the DTBook markup and consequently not delivered as images
- Iconic
  - Inline – where iconic images occur in an inline context they are considered to be crucial to the understanding of the text. In these cases producers of DTBook files are required to contact the ordering entity for advice
  - Other – iconic images not occurring in an inline context should be judged for their relevance to the text. In these cases producers of DTBook files are recommended to contact the ordering entity for advice
- Formatting – images of this type are not to be included in the DTBook markup and consequently not delivered as images

### 2.3.3 Colour images

Colour images present in the print original are to be reproduced with no observable degradation in colour rendering.

### 2.3.4 Greyscale images

Greyscale images present in the print original are to be reproduced without introducing visible compression artefacts, e.g. banding.

### 2.3.5 Images containing text

Images in the print original containing a preponderance of text should be reproduced using the same properties applied for scanning of text in general, optimally 400dpi though no less than 300dpi. Results must be subsequently proofread to ensure reproduction quality.

### 2.3.6 Editing images

Scanned images often require some degree of editing.

- Images are to be cropped, leaving them free of all text external to the image itself, e.g. captions, headers and footers etc.
- Images skewed as a result of the scanning process must be rectified.

### 2.3.7 Delivery

1. Images size
  - All reasonable attempts should be made by producers to maintain consistency of size between resulting images and their source in the print original
  - Resulting images must not exceed 600 pixels in width regardless of orientation or the preceding point
2. The ratio of height to width in resulting images must agree with the original.
3. Images must be delivered in JPEG format.

## 3 General guidelines for mark up

Marking up text using an XML grammar involves the placement of so-called tags around defined elements within a document. Text elements might have particular attributes that need to be represented, this can be achieved by including them in the relevant element tag. Rules for element tags and their allowed attributes are collected in a 'Document Type Definition' or 'DTD'. Elements present in a text include, though not exclusively:

- headings
- page numbers
- paragraphs
- notes
- tables
- lists
- sidebars

### 3.1 *TPB applies a subset of the DTBook standard*

TPB intends to apply a subset of those elements and attributes found in the DTBook standard that are described in the DAISY Structure Guidelines. All elements selected in this subset are described in Section 4 Detailed guidelines for markup.

An exception to this concerns production requiring adaptation of a pedagogic nature, primarily production on behalf of SIT – The Swedish Institute for Special Needs Education. Such production may include instructions for the supplier directing the use of particular attribute values for certain types of content.

In those cases where the DAISY Structure Guidelines recommend markup that has not been described by TPB a choice of element/attribute providing the most applicable markup should be made.

### 3.2 *A complement to the DTD*

The dtbook-2005-2.dtd, used to validate markup produced for TPB, allows a great deal of leeway in its application; this is a condition of its format. In the majority of cases, a document requires further editing after passing validation. The following example is typical, TPB requires that roman numerals be present in the text node (i.e. the text contained within element tags) of `<pagenum page="front">` markup. The DTD is unable to do so and can

therefore not be used to validate such a requirement. See Section 4 *Detailed guidelines for markup*.

The DTD also exists as a so-called LiveDTD, a hypertext document that provides a more user-friendly version of the standard. A LiveDTD för dtbook-2005-1.dtd can be found at <http://se.daisy.org/riktlinjer/>. Conversion software to create a LiveDTD can be downloaded from <http://www.sagehill.net/livedtd/download.html>

### 3.3 *Elements allowing both text and elements as children (%flow)*

Another example of leeway within the structure of the DTD concerns %flow in elements, e.g.:

- **<sidebar>**
- **<list>**
- **<table>**

Block and inline elements are required not to be combined in an inappropriate manner, even when allowed by the DTD.

Two examples of acceptable allowed markup using **<sidebar>**:

Example 1:

```
<sidebar>  
Oh, see. Oh, see Jane. Funny, funny Jane.  
</sidebar>
```

Example 2:

```
<sidebar>  
<p> Oh, see. Oh, see Jane. Funny, funny Jane.</p>  
</sidebar>
```

An example of valid to the DTD though unacceptable markup using sidebar:

```
<sidebar>  
Oh, see. Oh, see Jane. <p>Funny, funny Jane.</p>  
</sidebar>
```

The example above is also valid according to the DTD, though mixing text and block elements. Yet it is semantically unclear and such mark up is to be avoided.

#### 3.3.1 **Nested lists**

A list item may contain text and a new list. However, as above, this is undesirable and instead the **<p>** element should be employed as follows:

```
<list>
  <li>Text</li>
  <li><p>Text</p>
    <list>
      <li>Text</li>
      <li>Text</li>
    </list>
  </li>
</list>
```

### ***3.4 Placement of "floating" elements***

The correct placement of **<table>**, **<sidebar>**, **<imggroup>** and **<annotation>** in DTBook documents is not a given. It might be the case that an image or table 'floats', i.e. it is external to the flow of the text. In such cases, relevant elements are to be placed at their most logical contextual anchor point within the text flow. If such an anchor point cannot be identified the relevant elements are to be positioned as close as possible to the closure tag of the current containing mark up. Such relocation must not break DTD rules for placement and must attempt to maintain optimal conditions for reading comprehension.

### ***3.5 The XML file in general***

#### **3.5.1 File name**

DTBook files produced on behalf of TPB must be given the production number provided by TPB when ordered and have the .xml extension. File names and paths must exclusively contain the following characters: 0-9, a-z, underscore, and hyphen.

#### **3.5.2 Processing instructions**

Processing instructions, e.g. style sheet paths, must not be included in the delivered file.

#### **3.5.3 XML declaration**

The following XML declaration must be used:

```
<?xml version="1.0" encoding="utf-8"?>
```

#### **3.5.4 Document type declaration**

The following document type declaration must be included:

```
<!DOCTYPE dtbook PUBLIC "-//NISO//DTD dtbook 2005-2//EN"
"http://www.daisy.org/z3986/2005/dtbook-2005-2.dtd">
```

NOTE: The abovementioned declaration may change as the standard is updated.

#### **3.5.5 Character encoding**

DTBook documents produced on behalf of TPB must be saved using the UTF-8 character encoding and *not* include a byte order mark (BOM).

### 3.5.6 xml:lang

Longer text extracts (at least one whole sentence) whose language is not that stated in the root element are required to be marked up with the **xml:lang** attribute. Single words or shorter inline quotes are not required to be marked up using the **xml:lang** attribute.

Examples of elements that may require the **xml:lang** attribute:

**<blockquote>**  
**<p>**  
**<list>**  
**<table>**  
**<sidebar>**  
**<note>**  
**<annotation>**

Segments of the print original such as references, indices and appendices, may include terms, tiles, names and the like that also deviate from the documents main language. These do not require marking up with the **xml:lang** attribute.

### 3.5.7 Representation of punctuation, typographic characters and special characters.

This section describes those requirements placed upon producers with regard to the representation of particular types of character sensitive to the scanning process and OCR.

TPB requires producers to contact ordering entities for a clarification of requirements when punctuation, typography and special characters beyond the scope of these guidelines occur.

This section of the guidelines may be expanded and/or updated to facilitate the effective production of DTBook files.

#### 3.5.7.1 Introduction

Due to the nature of production based on print originals errors may be introduced into DTBook files. TPB considers the accurate conversion of all characters present in print originals of the highest importance. A particular group of characters has shown itself to be vulnerable to error: punctuation, typographic and special characters.

#### 3.5.7.2 Punctuation

TPB requires that particular care be paid to the accurate representation of the following punctuation characters and any whitespace surrounding them:

1. **Minus character (U+2212)**
2. **Hyphen character (U+2010)**
3. **Figure dash character (U+2012)**
4. **En dash character (U+2013)**
5. **Em dash character (U+2014)**
6. **Quotation dash character (U+2015)**

##### 3.5.7.2.1 Punctuation character representation for TPB

TPB allows producers the following choice:

- that certain of the abovementioned group of characters be substituted or

- that the abovementioned group of characters be correctly represented in the resulting DTBook file in their entirety.

Note that TPB does not require Producers to interpret the usage of the listed characters in those circumstances where typographer choice differs from usage descriptions provided herein.

### 3.5.7.2.2 *Substitution*

TPB requires the following representation:

- The Minus character (**U+2212**), Hyphen character (**U+2010**) and Figure dash character (**U+2012**) can acceptably be represented with Hyphen-Minus (**U+002D**). However, if producers choose to make such a substitution TPB requires that all instances of each of the abovementioned characters be substituted in the file; producers are required to ensure that mixing of methods does not occur.
- En dash characters are required to be represented using En dash (**U+2013**).
- Em dash characters are to be represented using Em dash (**U+2014**).
- Quotation dash characters (**U+2015**) can acceptably be represented using En dash (**U+2013**).

However, if producers choose to make such a substitution TPB requires that all instances of each of the abovementioned characters be substituted in the file; producers are required to ensure that mixing of methods does not occur.

### 3.5.7.2.3 *Non-substitution*

TPB requires that the following characters be accurately represented:

7. Minus character (**U+2212**)
8. Hyphen character (**U+2010**)
9. Figure dash character (**U+2012**)
10. En dash character (**U+2013**)
11. Em dash character (**U+2014**)
12. Quotation dash character (**U+2015**)

### 3.5.7.3 *Typographic characters*

TPB requires producers to contact ordering entities for a clarification of requirements when punctuation, typography and special characters beyond the scope of these guidelines occur.

### 3.5.7.4 *Special characters*

#### 3.5.7.4.1 *Arrows*

Print originals containing arrows in the text are required to have such characters represented accurately. Producers are allowed to represent such characters using the appropriate Unicode code point from the span U+2190 to U+2199.

### 3.5.7.5 *Other*

#### 3.5.7.5.1 *Phonetics*

TPB requires producers to contact ordering entities for a clarification of requirements when phonetic symbols are used in print originals.

#### 3.5.7.5.2 *Pictograms, Ideograms and Logograms*

TPB requires producers to contact ordering entities for a clarification of requirements when characters belonging to the headed groups occur; commonly Chinese, Japanese or Korean, though hieroglyphics are also an example of such.

### **3.5.8 Hyphenation**

Hyphenation, occurring due to line breaks or page breaks at margins in the print original, is expected to be corrected in the resulting XML document.

### **3.5.9 Empty elements**

Empty elements may not occur in the XML document, exceptions to this are:

**<img>**

**<br>**

**<meta>**

**<link>**

**<col>**

**<th>**

**<td>**

**<dd>**

### 3.6 Explanation of wording

Wording	Explanation
<p>	The element name is coloured red and bracketed by “less than” and “greater than” characters.
Attribute	Attribute names are coloured red.
"attribute value"	Attribute values are written with red text and quotation marks.
h[x]	Elements and possible variables – brackets contain possible values
"should"	Implies compliance by suppliers, though invites judgement calls based on best practice. Non-compliance implies no recourse to return of material.
"is to be", "shall", "require"	Implies compliance without the need for judgement by suppliers. Non-compliance invites return of delivered materials as incomplete.
<div style="border: 1px solid black; padding: 5px; display: inline-block;"> <p>&lt;p&gt;Text&lt;/p&gt;</p> </div>	Code examples are enclosed in a box.

## 4 Detailed guidelines for markup

### 4.1 Guidelines for mark up of structural elements

#### 4.1.1 dtbook

The **<dtbook>** element is the DTBook documents root element and consequently contains all other element tags.

*Attributes:*

**version, xml:lang, xmlns**

The following values are to be used for the **version**, **xml:lang** and **xmlns** attributes

- **version="2005-2"**. This value may change if TPB adopts an updated version of the standard for production purposes. TPB will inform suppliers of any such change during the contract period.
- **xml:lang ("sv", "en-GB" etc.)**. TPB requires the use of the IETF RFC 3066 standard for language identification.

If the language is included in the ISO 639-1 standard TPB requires the use of the two-letter code. If the language is not included TPB require the use of the three-letter code found in the ISO 639-2 standard.

Print originals containing a mix of languages should be given an xml:lang attribute reflecting the majority language; this may often be reflected in that language represented by the title of the print original.

English requires a more detailed two part description, i.e. the requisite language code must be coupled with the appropriate country code to differentiate between American and British English, i.e.: **xml:lang="en-US"/ xml:lang="en-GB"**

Print originals containing both British and American forms of English and not having a discrete balance of usage are allowed to be marked up with the language code only, i..e. **xml:lang="en"**

- **xmlns**. The following namespace is required:  
**<http://www.daisy.org/z3986/2005/dtbook/>**

#### 4.1.2 head

The **<head>** element comes immediately after the **<dtbook>** and contains the documents metadata, **<meta>**.

#### 4.1.3 book

The **<book>** element comes after (though at the same level of hierarchy as) **<head>** inside **<dtbook>**. **<book>** can be compared to **<body>** in HTML and contains the contents of the print original.

#### 4.1.4 frontmatter

**<frontmatter>** contains that information presented before core content in the print original, e.g. copyright information, foreword, preface, table of contents and the like.

To simplify production of adapted materials from the DTBook document TPB requires that content contained in preliminary pages is structured in a uniform manner, see Sections 4.1.4.1 to 4.1.4.4 below.

### **4.1.4.1 Title and author**

The first elements contained by **<frontmatter>** must always be **<doctitle>** followed by **<docauthor>**. This applies regardless of how the print originals title page or cover is formatted. See Section 4.2 *Guidelines for mark up of block elements - listed alphabetically*.

### **4.1.4.2 Title page**

Details of the title, author, and publisher can often be found repeated at the beginning of the print original. Such repeated content is superfluous and is to be excluded from the XML document. **<doctitle>** and **<docauthor>** contain the necessary information.

### **4.1.4.3 Colophon and subsequent content**

The print original's colophon, dedication and the like are to be marked up using the **<level1>** element and one of the class attribute values described in Section 4.1.7.1 class attributes for **<level1>**.

The colophon is required to have details specific to the printing of the work removed. Typical texts to be removed are:

- Publisher adverts
- Publisher contact details
- Details of paper type and quality
- Printer's details including location
- Number combinations detailing printing year

Details that are required to be retained in the colophon are:

- Publisher details
- Copyright texts
- Copyright year
- Print year
- Edition
- ISBN
- Previously published works
- Translator
  - In those cases where details concerning the translator of a work are found elsewhere, e.g. on the title page, the text is to be moved to the colophon
- Editor
- Illustrator/photographer

### **4.1.4.4 Introduction**

If an "Introduction" or material of a similar nature *is numbered in accordance with the print originals other headings* then such material should be included in the **<bodymatter>** element. Otherwise, this type of material is required to be included in the **<frontmatter>** element.

#### 4.1.5 bodymatter

The **<bodymatter>** element contains the printed materials core content. This can be defined as the parts, chapters sub-headings found in the print original. Note that content such as epilogue, conclusions and the like are to be contained within the **<bodymatter>** element.

#### 4.1.6 rearmatter

**<rearmatter>** shall contain all content appended to the core text. Examples of material that can be found in **<rearmatter>** include, though not exclusively: appendices, glossaries, references, notes and indices.

##### 4.1.6.1 Back cover and dust jacket copy

Both the back cover copy and copy placed on the dust jacket flaps of the print original are to be included in **<rearmatter>** and be marked up using **<level1>** and a **class** attribute according to the following:

```
<level1 class="backCoverText">  
<p>[rear cover copy]</p>  
</level1>  
<level1 class="other">  
<p>[dust jacket flap copy]</p>  
</level1>
```

No heading is required to be introduced.

Dust jacket copy is required to be included in the same level and included after any back cover text. An exception to this is copy consisting of advertising and/or review. Text about or by the author is required to be included, as is an image of the author.

#### 4.1.7 level1-6

The **<level[x]>** element represents the heading *levels* of the print original. **<level[x]>** elements are to be associated with their respective headings in the print original, if such exist. [x] mirrors the structure of the headings and must have a value from the span 1 to 6.

The **<level>** element, i.e. the level element not assigned a structure number, is *not* to be used.

A relatively common layout practice is the inconsequent use of fonts when setting headings. This practice leads to discrepancies between a print original's actual structure as opposed to the visual representation of it's structure. For example, an introduction may have a font that is applied to sub-headings in the remaining text. Application of **<level[x]>** must not represent this habit – attempts will produce DTD invalid markup. Note also that use of otherwise empty **<level[x]>** elements as a container for inappropriate **<level[x]>** mark-up in an attempt to achieve validity is also disallowed.

##### 4.1.7.1 class attributes for <level1>

The **class-** attribute is required to be applied to the **<level1>** element and is used to describe the type of content contained by the element. The exception to the above concerns print

original's utilising a primary structure of Part [x]. In such cases the **class** attribute is also to be applied to **<level2>** elements in **<bodymatter>**.

#### 4.1.7.1.1 *frontmatter*

The following values are to be applied to the **class** attribute in **<level1>** contained in **<frontmatter>**:

- **"briefToc"** – used to identify any table of contents providing an outline in those cases where two table of contents occur
- **"colophon"** – used to identify content in the beginning of the print original pertaining to the publisher, printer and the like
- **"dedication"** – used to identify author/editor dedications in the print original
- **"glossary"** – used to identify glossaries, lists of abbreviations and the like
- **"introduction"** – used to identify an introduction not numbered in accordance with the print original's other chapter numbering (see below)
- **"other"** – used to identify content within **<level1>** contained in **<frontmatter>** not corresponding to the above, e.g. "List of figures" and so on
- **"preface"** – used to identify an author/editor preface
- **"toc"** – used to identify the print original's table of contents

#### 4.1.7.1.2 *bodymatter*

The following values are to be applied to the **class** attribute in **<level1>** and **<level2>** contained in **<bodymatter>**:

- **"part"** – used to identify content when a primary structure of Part [x] is used
- **"chapter"** – used to identify chapter structure and any introduction included in the chapter numbering scheme
- **"introduction"** – used to identify an introduction not numbered according to the chapter numbering scheme. See Section 4.1.4.4 *Introduction*
- **"other"** – used to identify content within **<level1>** contained in **<bodymatter>**, e.g. prologue, epilogue and the like
  - Note that in shorter works that do not contain a great deal of hierarchical structure, in particular children's literature, it may be the case that only one (1) **<level1>** can be defined in the **<bodymatter>**. Such should be marked using the **"other"** attribute in the following manner: **<level1 class="other">**

#### 4.1.7.1.3 *rearmatter*

- The following values are to be applied to the **class** attribute in **<level1>** contained in **<rearmatter>**:
- **"appendix"** – used to identify appendices
- **"backCoverText"** – is to be used to identify back cover and dust jacket copy. Dust jacket copy is required to be included in the same level and included after any back cover text. An exception to this is copy consisting of advertising and/or review. Text about or by the author is required to be included, as is an image of the author.
- **"bibliography"** – used to identify a bibliography
- **"colophon"** – used to identify content the print original, pertaining to the publisher, printer and the like, that is on occasion included in back matter
- **"footnotes"** and **"rearnotes"** – used to identify a section of notes
- **"glossary"** – used to identify glossaries
- **"index"** – used to identify indices
- **"other"** – to be used to identify **<level1>** elements in **<rearmatter>**
- **"toc"** – used to identify the print original's table of contents

## 4.2 Guidelines for the mark up of block elements - listed alphabetically

### 4.2.1 annoref & annotation

The **<annoref>** and **<annotation>** elements resemble the **<noteref>** and **<note>** elements. However, annotation (considered a floating element) is to be used when marking up shorter marginalia in the print original. A reference number or character (**<annoref>**) is associated, principally found in text parallel to the marginalia. See also **<noteref>** and **<note>**.

#### 4.2.1.1 annoref

The **<annoref>** element is required to have the following attribute:

- **idref**, this refers to the **id** attribute applied to the **<annotation>** element. Note that the value must be prefixed with the # character to fulfil URI requirements.

#### 4.2.1.2 annotation

Each **<annotation>** is required to have at least one associated **<annoref>**.

##### 4.2.1.2.1 Placement of annotation

See Section 3.4. Placement of "floating" elements

### 4.2.2 blockquote

The **<blockquote>** element is used to mark up quotes broken out of the text flow, i.e. constituting their own paragraph. Note that the **xml:lang** attribute is required to be applied to the **<blockquote>** element if the quote is in a different language than that identified in the root element. Inline quotes are not required to be marked up with **<q>** though are required to be reproduced in accordance with the print original, e.g. with **<em>** or quotation marks.

### 4.2.3 doctitle

The first element within **<frontmatter>** must be **<doctitle>** and use the following content format: [title] – [sub-title].

Any text pertaining to the edition of the work is *not* to be included.

### 4.2.4 docauthor

The second element within **<frontmatter>**, after **<doctitle>**, must be **<docauthor>** and is required to use the following content format: [author/editor first name + other names + surname]. If the print original has more than one author/editor, each name is to be marked up using individual **<docauthor>** elements. Comma separation or other text before or between author names (e.g. "by", "and") is not to be included, regardless of usage in the print original. The following markup is required in those instances where an editor fulfils the role of author:

**<docauthor>**Gregory Stones (ed.)**</docauthor>**

### 4.2.5 div and the class="pgroup" attribute

It is common practice to utilise indentation or empty lines to indicate a shift between paragraphs. It is the case that authors may apply both practices in an attempt to associate groups of paragraphs. TPB require this formatting to be marked up.

The **<div>** element and the applied **class="pgroup"** attribute are to be used to identify all such formatting contained within any **<level[x]>** element. The following example uses **<p>** elements associated using indentation in the print original. Note that content association can be made using other methods.

```
<level[x]>
<h[x]>[Heading]</h[x]>
<div class="pgroup">
<p>[Paragraph]</p>
</div>
(empty line)
<div class="pgroup">
<p>[Possible indented paragraph]</p>
<p>[Indented paragraph]</p>
<p>[Indented paragraph]</p>
</div>
(empty line)
<div class="pgroup">
<p>[Possible indented paragraph]</p>
<p>[Indented paragraph]</p>
<p>[Indented paragraph]</p>
</div>
</level[x]>
```

#### 4.2.5.1 Exceptions to class="pgroup" mark up

The **<div class="pgroup">** element is not to be used if a **<level[x]>** element *does not* contain associated **<p>** elements.

It is common for empty lines to separate paragraphs from other types of content, e.g. sidebars, quotes and the like. **<div class="pgroup">** is not to be used in these cases. For example:

```
<level[x]>
<h[x]>Heading</h[x]>
<p>[Possible indented paragraph]</p>
<p>[Indented paragraph]</p>
<p>[Indented paragraph]</p>
(empty line)
<blockquote/>
(empty line)
<p>[Possible indented paragraph]</p>
<p>[Indented paragraph]</p>
<p>[Indented paragraph]</p>
</level[x]>
```

#### 4.2.6 h1 - h6

The **<h1>-<h6>** elements are used to identify headings in the print original. Note that **<h[x]>** must be contained within their respective **<level[x]>** element.

Typographic formatting of a heading is not required to be mirrored in markup. An exception to this rule concerns single words in headings that are italicised or bold, this are to be marked up using **<em>** and **<strong>** respectively.

This also applies to sidebar, list and table headings (**<hd>**, **<h[x]>** and **<caption>**).

For example:

```
<h2>A word is <em>italicised</em></h2>
```

Standard rules apply for note or annotation references that appear in headings.

For example:

```
<h2>A word has a <noteref>note reference</noteref> in the heading</h2>
```

In cases where the author of a text is included directly under the heading TPB requires that these are marked up using **<p>**. Such text is not to be included in the preceding **<h[x]>** element.

Chapters not possessing a heading in the print original are required to be identified with the appropriate **<level[x]>** element and do not require markup with **<h[x]>**.

#### 4.2.7 imggroup

An image, its caption, and any associated image description must always be contained within the **<imggroup>** element. Exceptions to this rule are, for example, icons or images without captions or image descriptions that are contained within table cells.

An example of **<imggroup>** mark-up:

```
<imggroup>  
  
<caption>Caption text</caption>  
<prodnote render="optional">Image description</prodnote>  
</imggroup>
```

##### 4.2.7.1 Placement of *imggroup*

TPB requires that the **<imggroup>** element not be placed in an inline context, e.g. contained within the **<p>** element. It follows then that an image placed inline, i.e. in the text itself, of the print original should be relocated when the requisite **<imggroup>** mark-up is made. The most appropriate placement in the majority of cases is directly after end of paragraph. See also Section 3.4 Placement of "floating" elements

Images stretching over two (2) pages are to be scanned as an individual image per page, i.e. images are required to be divided in two at the point of change from recto to verso. The following mark-up is to be employed:

```
<imggroup>

<caption>Caption text</caption>
<prodnote render="optional">Image description</prodnote>
</imggroup>
<pagenum id="page-[x]" page="normal">[x]</pagenum>
<imggroup>

<caption>Caption text</caption>
<prodnote render="optional">Image description</prodnote>
</imggroup>
```

NOTE: Remain aware of those requirements contained in Section 3.4 Placement of "floating" elements

#### 4.2.7.2 *img*

The **<img>** element requires the following attributes:

**alt** which must include the value="image"

**src** which must include a reference to an image file.

Images must be located in the same catalogue as the DTBook file.

Filenames for images must only contain the following characters: 0-9, a-z, underscore and hyphen.

#### 4.2.7.3 *caption*

The **<caption>** element contains that text associated with an image. **<caption>** must always be placed directly after the appropriate **<img>** element. If the **<caption>** text describes a series of images it must be placed after the last image in the series.

#### 4.2.7.4 *prodnote*

A **<prodnote>** element contained within an image group is used to mark up an image description. However, standard practice is to provide image descriptions after production of the DTBook file. TPB therefore requires that a "dummy" text be placed within **<prodnote>** elements for image groups. The **<prodnote>** element is to have the **render="optional"** attribute applied and is considered mandatory where image group mark-up occurs. The **<prodnote>** is to be marked up as follows:

```
<prodnote render="optional">Image description</prodnote>
```

It is important to remember not to mix both block and inline mark up in elements that can contain %flow. See Section 3.3 Elements allowing both text and elements as children (%flow).

Image description prodnotes are always placed last in image groups, after any **<caption>** that may occur.

#### 4.2.8 meta

The **<meta>** element contains information about the DTBook document. Meta information is used for identification purposes and to aid index compilation.

##### 4.2.8.1 Table of attributes, contents and schemes

Attribute		
name=""	content=""	scheme=""
Content		
"dtb:uid"	[identification]	
"dc:Title"	[title] : [subtitle]	
"dc:Creator"	[surname], [first name] [other name]	
"dc:Date"	[Date of file completion]	[YYYY]-[MM]-[DD]
"dc:Publisher"	[ordering entity]	
"dc:Language"	[xx(x)](-[XX])	
"tpb:Supplier"	[Company Name]	
"tpb:SuppliedDate"	[Date of file completion]	[YYYY]-[MM]-[DD]
"dc:Identifier"	[identification]	[ordering entity]

##### 4.2.8.2 Meta element format:

**<meta name="dtb:uid" content="[identification]"/>** Contains the document's identification and is supplied by TPB when production is ordered.

**<meta name="dc:Title" content="[title] : [subtitle]"/>** Contains the title of the print original and is to be taken from order details provided by TPB. In cases where more than one author is present TPB require that a **<meta>** element be inserted *for each author*.

**<meta name="dc:Creator" content="[surname], [first name] [other name]"/>** Contains the authors name and is to be formatted "surname, first name". Individual **<meta>** mark up is to be provided for each author if more than one is indicated.

**<meta name="dc:Date" content="[YYYY]-[MM]-[DD]"/>** Indicates the date production of the DTBook file is completed, to be formatted as follows: Year (four figures) – Month (two figures) – Day (two figures)

**<meta name="dc:Publisher" content="[TPB]"/>** Used to indicate the ordering entity.

**<meta name="dc:Language" content="[xx(x)](-[XX])"/>** Used to indicate the main language of the document. TPB requires the use of the IETF RFC 3066 standard for language identification. If the language is included in the ISO 639-1 standard TPB requires the use of

the two-letter code. If the language is not included TPB require the use of the three-letter code found in the ISO 639-2 standard. English requires a more detailed two part description, i.e. the requisite language code must be coupled with the appropriate country code to differentiate between American and British English, i.e.: `xml:lang="en-US"/ xml:lang="en-GB"`.

**`<meta name="tpb:Supplier" content="[Company Name]" />`** Included only once and must contain the name of the contracted supplier of the DTBook file.

**`<meta name="tpb:SuppliedDate" content="[Date of file completion]" scheme="YYYY-MM-DD" />`** Included once on initial delivery of the DTBook file. In the event of returns of faulty DTBook files to the Supplier for correction TPB require that an additional instance be included with the new delivery date as it's content. Deletion of the initial instance is not allowed. An additional instance of this element is required for each return of a faulty DTBook file for correction.

Certain DTBook file productions, details of which are supplied when ordered, require the following `<meta>` elements:

**`<meta name="dc:Identifier" content="[identification]" scheme="[ordering entity]" />`**

The value of [identification] is supplied when ordered and can differ from the [identification] value for `dtb:uid`.

#### 4.2.9 note & noteref

Notes are comprised of two connected elements, `<noteref>` and `<note>`. The actual text of the note is marked up using the `<note>` element and the note reference is marked up using the `<noteref>` element.

Three types of note can be identified in the print original:

- notes placed in the footer of the page
- collected at the end of the respective chapter
- in a note section appended to the main text

Note references to the above occur within the text.

##### 4.2.9.1 noteref

The following attributes are to be applied to the `<noteref>` element:

- **idref** refers to the notes **id** attribute. Note that the value must be prefixed with the # character to fulfil URI requirements.

- **class** this must contain one of the following values, **"endnote"** for notes occurring at the end of a chapter or **"rearnote"** for notes appended to the main text.

Example:

```
<noteref idref="#fn_2_1" class="rearnote">1</noteref>
```

##### 4.2.9.1.1 Note references stating an interval

Note references stating an interval are to be handled as follows; each number in the interval is to be marked up with an individual `<noteref>`, though separators (hyphens, commas and the like) are not to be included.

For example:

<p>Print original: "In the midway<sup>1-3</sup> of this our mortal life..."</p> <p>DTBook: In the midway&lt;noteref&gt;1&lt;/noteref&gt;&lt;noteref&gt;2&lt;/noteref&gt; &lt;noteref&gt;3&lt;/noteref&gt; of this our mortal life...</p>
--

#### 4.2.9.2 Note

Each **<note>** element is required to have at least one associated **<noteref>** element.

The following attributes are required to be applied to **<note>** elements:

- **id** refers to an associated **idref** attribute applied to a **<noteref>** element.
- **class** is required to have the same value as that applied to the **<noteref>** element.

#### 4.2.9.3 Placement of notes

All notes present in the DTBook document, excepting chapter notes (endnotes), are to be placed last in **<rearmatter>** in an appended section of notes. If rear cover copy is present, the note section should be placed before this.

##### 4.2.9.3.1 Footnotes and table footer notes

Notes placed in, for example, page footers or table footers are to be converted to rearmatter notes in the DTBook document, **<note class="rearnote">**. These notes are to be placed in a section appended to the main text in the rearmatter section. If rear cover copy is present, the note section should be placed before this. This note section is to be marked up separately (**<level1 class="rearnotes">**). Notes are required to be sorted in the same order as their respective initial **<noteref>**.

Example of a note section:

<pre>&lt;level1 class="rearnotes"&gt; &lt;note class="rearnote" id="fn_1"&gt; &lt;p&gt;Note text&lt;/p&gt; ... &lt;/level1&gt;</pre>
--

##### 4.2.9.3.2 Endnotes

As stated, notes already placed at the end of a chapter are not required to be moved.

##### 4.2.9.3.3 Notes originally placed in back matter (rearnotes)

Notes already present in the back matter of the print original are not required to be moved though they are to be marked up using **<note class="rearnote">**. The note section itself is to be marked up using **<level1 class="rearnotes">**, existing headings are to be preserved and no new headings inserted.

#### 4.2.10 p

The **<p>** element identifies a paragraph.

#### 4.2.11 pagenum

The **<pagenum>** element identifies the change from recto to verso and consequently the change in pagination, the text node contains, barring exceptions, the page identification; almost exclusively a number.

All pages present in the print original included in the pagination, even those not provided with a printed number, are required to be marked up with the **<pagenum>** element.

TPB's requirements are illustrated by the following; if page 148 is followed by an un-numbered page which is in turn followed by page 150 TPB requires that the un-numbered page be numbered 149.

##### 4.2.11.1 An exception with regard to flyleaves and redundant content

An exception to Section 4.2.11 concerns print originals containing flyleaves and/or superfluous duplicate information about author and title (see Section 4.1.4.2 Title page). Such pages occurring are *not* to be identified with **<pagenum>**. Generally speaking mark-up with **<pagenum>** should begin at **<level1 class="colophon">**.

##### 4.2.11.2 Attributes

**page** with the values "front", "normal", "special"  
**id** with the value format "page-[page number]"

Three types of page are described in the DTBook standard: front, normal and special.

The attribute value "front" is to be used where the print original utilises roman numerals in its front matter. Roman numerals are to retain that letter case used in the print original. **<pagenum page="front"/>** is not allowed outside of the **<frontmatter/>** element; though **<frontmatter/>** may contain **<pagenum page="normal"/>**.

The attribute value "normal" is to be used for pages occurring in the body and back matter of the print original.

The "special" attribute value is rarely applicable. An example is an appendix not numbered in a standard manner, e.g. a1, a2, a3 and so on. The "special" value is also to be used in those cases where un-numbered inserts, for example suites of photography, occur; these are to be marked up as follows:

```
<pagenum page="special" id="unnum-[page no.]">Un-numbered page</page>
```

In those cases where pagination of a text cannot be effectively represented using the above rules the *Supplier* is required to contact TPB.

##### 4.2.11.3 Placement of pagenum

The **<pagenum>** element can be used as either block or inline and can therefore be placed precisely where the print original changes page regardless of content mark-up, e.g. within a paragraph, list or table.

The **<pagenum>** element is also to be placed within words breaking over pages. It is allowed though, to place the **<pagenum>** element directly after words hyphenated due to page change. If such a relocation of the **<pagenum>** element is made TPB requires that

eventual hyphens, with reservation for those found in compound words, be removed. Relocated `<pagenum>` elements should always be *preceded* by a space character.

When a page begins with a `<level[x]>` element the `<pagenum>` element is to be placed between the `<level[x]>` element and the `<h[x]>` element.

I de fall en boks sidnumrering är väldigt speciell och inte inkluderas i ovanstående resonemang, t.ex. om sidnumren repeteras (t.ex. s. 1-18 för Del 1, s. 1-33 för Del 2 och s. 1-28 för Del 3), ska producenten kontakta TPB för vägledning om uppmärkningen.

### ***4.2.11.4 Print originals free of pagination***

Un-paginated print originals, particularly shorter children's literature and picture books, are required to be provided with pagination. Note that standard verso-recto order is to be maintained.

### **4.2.12 poem**

The `<poem>` element is to be used to mark up poems, song texts and the like.

The following elements are to be used where appropriate within the `<poem>` element: `<hd>`, `<linegroup>` and `<line>`.

### **4.2.13 sidebar**

The `<sidebar>` element is used to mark up sidebar content, i.e. content that while contextually coherent may graphically or positionally differentiated from the main flow of the text. The `<sidebar>` element can contain the majority of other block elements, e.g. p, list, table and so on. Headings occurring within the `<sidebar>` element are to be marked up using the `<hd>` element.

#### ***4.2.13.1 Attribute***

TPB requires that sidebar has the `render` attribute applied. The `render` attribute is to have the value `"optional"`.

#### ***4.2.13.2 Placement of sidebar***

See Section 3.4 Placement of "floating" elements.

## ***4.3 Guidelines for mark up of inline elements – listed alphabetically***

### **4.3.1 br**

The `<br/>` element identifies a line break. It is to be used as an exception and only where appropriate structural mark-up is unavailable.

### **4.3.2 code**

The `<code>` element is used to identify program code.

### **4.3.3 em**

The `<em>` element identifies italicised text.

The **<em>** element is not to be used to emphasize headings in those instances where the use of italics is entirely typographic, though individual words in headings that are italicised are required to be marked up.

#### 4.3.3.1 *Handling spaces and punctuation in conjunction with <em>*

Two usages of italicisation occur that can generate mark-up errors when using **<em>**, partial italicisation of text within a sentence and italicisation of an entire sentence.

In those cases where only parts of a sentence are italicised TPB requires the following rules to be applied when using **<em>**:

- The start **<em>** tag must not be followed by any whitespace character, new line, carriage return or punctuation
- The closing **</em>** tag must not be preceded by any whitespace character, new line, carriage return or punctuation

The following example illustrates correct usage:

```
<p>This is a sentence <em>where part of the text</em> is italicised. This sentence has no italicisation.</p>
```

In those cases where an entire sentence is italicised TPB requires the following rules to be applied when using **<em>**:

- The start **<em>** tag must not be followed by any whitespace character, new line, carriage return or punctuation
- The closing **</em>** tag must not be preceded by any whitespace character, new line or carriage return
- The closing **</em>** tag must be preceded by associated punctuation

```
<p>This sentence is not italicised. <em>This is sentence is completely italicised.</em>  
This sentence is not italicised.</p>
```

#### 4.3.4 strong

The **<strong>** element identifies bold text.

The **<strong>** element is not to be used to emphasize headings in those instances where the use of italics is entirely typographic, though individual words in headings that are italicised are required to be marked up.

##### 4.3.4.1 *Handling spaces and punctuation in conjunction with <strong>*

Two usages of italicisation occur that can generate mark-up errors when using **<strong>**, partial italicisation of text within a sentence and italicisation of an entire sentence.

In those cases where only parts of a sentence are italicised TPB requires the following rules to be applied when using **<strong>**:

- The start **<strong>** tag must not be followed by any whitespace character, new line, carriage return or punctuation
- The closing **</strong>** tag must not be preceded by any whitespace character, new line, carriage return or punctuation

The following example illustrates correct usage:

`<p>This is a sentence <strong>where part of the text</strong> is in boldface. This sentence does not contain bold text.</p>`

In those cases where an entire sentence is italicised TPB requires the following rules to be applied when using `<strong>`:

- The start `<strong>` tag must not be followed by any whitespace character, new line, carriage return or punctuation
- The closing `<strong>` tag must not be preceded by any whitespace character, new line or carriage return
- The closing `<strong>` tag must be preceded by associated punctuation

`<p>This sentence does not contain bold text. <strong>This is sentence is completely in boldface.</strong> This sentence does not contain bold text.</p>`

#### 4.3.5 sub

The `<sub>` element identifies subscripted text, for example H<sub>2</sub>O is required to be marked up as H`<sub>`2`</sub>`O.

Mark-up with `<sub>` must not be contain any whitespace character, new line, carriage return or punctuation.

#### 4.3.6 sup

The `<sup>` element identifies superscripted text, for example X<sup>2</sup> is required to be marked up as X`<sup>`2`</sup>`.

Mark-up with `<sup>` must not be contain any whitespace character, new line, carriage return or punctuation.

Note that use of `<sup>` to identify note references is incorrect, see Section 4.2.9 note & noteref

## 4.4 Guidelines for the mark up of lists

### 4.4.1 list

The `<list>` element is used to mark up lists and organised registers. A list item may or may not be identified with a number or bullet of some kind.

A list item can consist of more than one paragraph (`<p>`).

The `<list>` element must not occur within the `<p>` element. The `<p>` element is to be closed before beginning `<list>` mark-up and a new `<p>` element opened, if necessary, on completion of the `<list>` mark-up.

Borderline cases occur, for example, where an apparent list stretches over a number of pages, where `<list>` mark-up is inappropriate. Such cases must be judged as and when they occur, with the readers understanding of the text guiding any decision made.

Lists of terms and definitions are not to be marked up using **<list>**. In these cases, i.e. definition lists, the **<dl>** element is required to be used. See Section 4.4.2 *dl*.

Standard lists consist mainly of two or more **<li>** elements. Where headings occur they are to be marked up using the **<hd>** element.

Lists can contain the **<pagenum>** element.

#### 4.4.1.1 *Attributes*

The **<list>** requires the following attributes to be applied:

The **type** attribute identifies three varieties of list:

- **"ol"** identifies ordered lists; each item is preceded by an alphanumeric character in series. If **"ol"** is applied, TPB requires that the alphanumeric character be removed from the text node of the list item.
  - The **enum** attribute is required by TPB for lists applying **type="ol"** and has the values:
    - **"1"** for numbers
    - **"a"** for lower case characters
    - **"A"** for upper case characters
    - **"i"** for lower case roman numerals and
    - **"I"** for upper case roman numerals.
  - Note that **"ol"** lists without a specified enum value default to **enum="1"**
- **"ul"** identifies un-ordered lists; each item is preceded by a symbol. If **"ul"** is applied, TPB requires that the symbol and any associated whitespace be removed from the text node of the list item.
- **"pl"** identifies lists not possessing any type of leading character, e.g. bullets or numerals. Typical examples of this kind of list are a table of contents and indexes. In certain circumstances **"pl"** is to be used to identify lists that do possess leading characters where it is important to preserve the specific character in use. For example, if the text refers to particular symbol types used **"pl"** is required to be used in such cases instead of **"ul"** or **"ol"** with the character and associated whitespace preserved.

#### 4.4.1.2 *Content to be marked up with list*

The following content is required to be marked up with **<list>**:

- Table of contents, see Section 4.4.14 *lic*
- Indices – TPB requires that these be marked up as one list with nesting as appropriate. No nesting is expected for those lists divided alphabetically
- References – TPB requires that these be marked up as one list, no nesting is required to identify reference type

#### 4.4.1.3 *li*

The **<li>** element is to be used to mark up individual items in a list.

#### 4.4.1.4 *lic*

For tables of contents, brief tables of contents, lists of illustrations, lists of tables and the like TPB requires the use of the **<lic>** element for mark up of constituent parts of a list item.

Indices and other registers do not require the use of the **<lic>** element. The purpose of such mark-up is to differentiate between a heading and its page number. This is achieved with **<lic class="entry">** and **<lic class="pagenum">** respectively.

See the following example:

```
<level1 class="toc">
<h1>Content</h1>
<list type="pl">
<li><lic class="entry">Chapter 1</lic><lic class="pagenum">1</lic></li>
<li><lic class="entry">Chapter 2</lic><lic class="pagenum">15</lic></li>
<li><lic class="entry">Chapter 3</lic><lic class="pagenum">30</lic></li>
</list>
</level>
```

#### 4.4.1.5 *hd*

The **<hd>** element is used to mark up headings not associated with a level element. Occurrence of this type of heading is generally limited to the **<list>** and **<sidebar>** elements.

#### 4.4.2 *dl*

The **<dl>** element is used to mark up lists of terms and their definitions, examples range from glossaries to lists of acronyms and the like.

Definition lists are generally made up of coupled **<dt>** and **<dd>** elements.

```
<dl>
<dt>Term</dt>
<dd>Definition</dd>
</dl>
```

The **<dl>** element is not allowed in the **<p>** element. Preceding **<p>** elements are required to be closed before marking up definition lists. A new **<p>** should be opened, if necessary, on completion of the definition list mark-up.

##### 4.4.2.1 *dt*

The **<dt>** element is used to mark up those terms occurring in definition lists.

##### 4.4.2.2 *dd*

The **<dd>** element is used to mark up definitions corresponding to relevant terms in a definition list. In those cases where empty definitions occur, essentially in lists that illustrate that a term has no definition, the following mark-up should be used, **<dd/>** or **<dd></dd>**. Such mark up may also occur where definition lists apply empty tags for formatting purposes.

## 4.5 Guidelines for the mark up of tables

### 4.5.1 Table

The **<table>** element is required when content in the print original is presented in tabular format or table-like format. Suppliers are required to judge whether table content can be defined as informative and illustrative or simply pictorial, i.e. whether the table in question provides the reader with information necessary for understanding an argument or can be seen as incidental. In those cases where table content is considered incidental TPB require that the content be scanned as an image and provided with a **<prodnote>**. An illustration of the above is a train timetable where, unless specific cells are referenced by the text, the content can acceptably be included as an image with relevant image description mark-up.

Table captions require mark up with the **<caption>** element, this element is placed directly after the **<table>** element regardless of placement in the print original.

#### 4.5.1.1 Empty table cells

The following practice is to be applied when marking up empty table cells occurring in the print original, either **<td/>** or **<td></td>**.

#### 4.5.1.2 Tables covering more than one page

In those instances where tables stretch over more than one page the following mark up is to be applied: the **<pagenum>** tag is placed within a **<p>** tag that is itself contained within the first **<td>** or **<th>** element on the new page. Note that the text node of the **<td>** or **<th>** must also be included within the **<p>** tag.

For example:

```
<td>
<p><pagenum id="page-[x]" page="[x]">[x]</pagenum>
Text node</p>
</td>
```

#### 4.5.1.3 Placement of tables

See Section 3.4. *Placement of "floating" elements.*

#### 4.5.1.4 Table notes

Notes found in or associated with tables are required to be moved to a note section according to that practice outlined in section 4.2.9 *note & noteref.*

#### 4.5.1.5 Tables within tables

Tables occurring within table cells are not required to be marked up using the **<table>** element. TPB requires that such content be marked up using elements appropriate to the nested table's content, e.g. **<p>**, **<list>** and so on.

#### 4.5.2 Caption

The following mark up is to be applied where two captions occur in a table: a **<caption>** element is opened, each caption is then contained with a **<p>** element, and the **<caption>** element is then closed.

#### 4.5.3 tr

The **<tr>** element is used to mark up table rows.

#### 4.5.4 th

The **<th>** element is used to mark up column or row headings. In those cases where a **<th>** element stretches over more than one column or row, TPB requires that the **colspan** and **rowspan** attributes be applied.

#### 4.5.5 td

The **<td>** element is used to mark up table cell content.

In those cases where a **<td>** element stretches over more than one column or row TPB requires that the **colspan** and **rowspan** attributes be applied.

## 5 Guidelines for the mark up of unified content

Certain print originals contain text and image content that are integrated to such a degree that separation of the two is considered overly problematic. In such cases of unified content TPB requires that the entire image, including text content, be captured as a JPEG file. TPB also recommends that the entire page be captured if this is considered optimal.

The **<caption>** element is considered unlikely to occur in such material, though TPB requires inclusion where necessary.

The **<prodnote>** element used for image descriptions continues to be required as in standard image mark-up. Text content is required to be included using applicable mark-up.

Note that TPB requires that marked up text content be placed before the **<imggroup>** element. Examples of such mark-up are included in the *Example Collection*.

## 6 Guidelines for the notation and mark up of mathematics, physics and simpler chemistry

Structurally complex formulae found in mathematics, physics, chemistry and other disciplines using such, are to be handled using the notation currently applied on behalf of TPB and SIT for production of Textview formatted titles.

This facilitates production and delivery of different e-text formats, e.g. Textview and Full text DAISY DTB.

The DTBook file can also be utilised Braille production, though a great deal of editing is required to realise this.

The mathematical notation currently in use has been developed for both the sighted and visually-impaired who use tactile presentation, and while requirements placed by speech synthesis have been taken into account current notation is a compromise.

The notation described in these guidelines cannot be used to express formulae in all circumstances. Structurally complex formulae unsuitable to notation are required to be scanned in and marked up as images, see Section 4.2.7 `imggroup`. In certain circumstances, in particular for the blind, tactile images may also be produced to complement the finished adapted title.

Future production requiring adaptation of mathematics, physics, chemistry, and the like will be based entirely on available XML grammars, specifically MathML. A project aimed at delivering production methodologies for the above has been started by TPB, successful development also been made by the DAISY Consortium. A proposed goal for the TPB project is delivery of MathML solutions for production in time for the next purchasing round.

Production of DTBook documents containing formula based content will require manual editing in the post-scanning process.

### ***6.1 Mark up of formula based content***

All occurrences of formula based content to be adapted using notation is to be marked up according to the following:

#### **6.1.1 Block**

Formulae formatted as paragraphs, i.e. separated with blank lines, are to be marked up as follows:

**<p class="tpb-specialnotation">**

Example:

$$\sum_{k=0}^3 (-2)^k = (-2)^0 + (-2)^1 + (-2)^2 + (-2)^3 = -5$$

**<p class="tpb-specialnotation">**  
sum{k=0;3;(-2)^k}= (-2)^0 + (-2)^1 + (-2)^2 + (-2)^3 = -5  
**</p>**

#### **6.1.2 Inline**

Formulae found in an inline context are to be marked up as follows:

**<span class="tpb-specialnotation">**

Example:

And that being able to simplify oneself – simplify. Let us say that we are developing an algorithm or implementing someone else's. Let us also say that we have also arrived at a point when a certain task is to be performed regarding whether  $y \neq z$  or whether  $x = 0$  or whether both  $x \neq 0$  and  $y = z$ . At this point it is a benefit if one is conversant with the rules of logic so that one can assure oneself that this task should *always* be performed.

**<p>**And that being able to simplify oneself – simplify. Let us say that we are developing an algorithm or implementing someone else's. Let us also say that we have also arrived at a point when a certain task is to be performed regarding whether **<span class="tpb-specialnotation">** $y \neq z$ **</span>** or whether **<span class="tpb-specialnotation">** $x = 0$ **</span>** or whether both **<span class="tpb-specialnotation">** $x \neq 0$ **</span>** and **<span class="tpb-specialnotation">** $y = z$ **</span>**. At this point it is a benefit if one is conversant with the rules of logic so that one can assure oneself that this task should **<em>**always**</em>** be performed.**</p>**

## 6.2 Notation

The notation required by TPB/SIT is defined in Appendix 5 Mathematical Notation.

TPB require that a dialogue be established between the supplier and TPB in those instances where the notation defined in Appendix 2 does not cover formulae found in the print original but suppliers judge that conversion is necessary and possible.

Note that the notation described in Appendix 5 Mathematical Notation and Section 6 Guidelines for the notation and mark up of mathematics, physics and simpler chemistry, are not required unless specifically ordered. In those cases where no order of notation is made but simpler mathematics and the like appear in the print original it is required to be marked up using appropriate elements, e.g. **<sup>** for powers.

## 7 Changes

### *Version 1.3 - 2007-10-02*

<b>2.3.2</b> Handling of specific image types	Clarification regarding image types.
<b>2.3.7</b> Delivery	Clarification regarding image size.
<b>3.5.7</b> Representation of punctuation, typographic characters and special characters.	Re-definition of punctuation handling.
<b>4.1.4.3</b> Colophon and subsequent content	Clarification of content that should be retained or removed.
<b>4.2.11</b> pagenum	New text requiring Suppliers to contact TPB with regard to special pagination.
<b>4.2.13</b> sidebar	Clarification.

## **Authors:**

Linus Ericson  
Markus Gylling  
Joel Håkansson  
Tomas Johansson  
Jesper Klein  
Cecilia Mattsson  
Gita Nabavi  
Björn Nyqvist  
Anne Stigell  
Richard Stones  
Björn Westling