Durm XML Markup

1. Overview

The formal DTD used within the Durm Corpus is available for download. Here, we briefly describe the meaning of the various elements.

1.1. Tags Resulting from the TUSTEP Markup

These tags are already present in HdA.txt (Tustep markup).

P

Paragraph, like HTML. Can have an attribute `pagebreak` denoting that it continues on the next page (value "firstpart") or is such a continuation (value "secondpart").

Formel

Formulas have not been transcribed. Thus, in place of a formula or equation there are multiple underscores in the Tustep file, followed by an empty `Formel` tag.

negEZ

Seems to be some kind of list, numbered with letters.

Biaoge

Has something to do with tables, not investigated further yet.

tspb

Has something to do with tables, not investigated further yet.

1.2. Tags Introduced by the durm2xml Tool

H1

Heading level 1, like HTML. Used for section headings ("a) ...").

H2

Heading level 2, like HTML. Used for subsection headings ("1) ...").

i

Italic, like in HTML.

hr

Horizontal line, like in HTML. Can have an attribute `footnotes` denoting that it separates the footnotes of a page from the main text. These are in fact the only `<hr>`s occuring.

center

Centering of text, like in HTML.

book

The top level element, enclosing the whole file.

chapter

A chapter. This information is added by `durm2xml`.

title

The title of a chapter is explicitly marked up by this tag.

section

A chapter. This information is added by `durm2xml`.

Grafik

A container for `<Figure>`s. `durm2xml` quotes its attribute values correctly.

Fig

A single figure. Only occurs inside `<Grafik>`, even when there's only one figure. In the Tustep file, they are already present, but look like this: "Fig1", "Fig2" etc. The number denotes the order of the figures inside the Grafik container, it is extracted and transformed into the value of an attribute `no` by `durm2xml`.

break

Linebreaks are removed by `durm2xml` by replacing the first part of the word by the complete word and removing the second part. This word is then enclosed in `<break>`.

fontsize

A change in font size. What kind of change is given by the attribute `change`, with values of the form "{ + | - }{1, 2, 3}".

footnote

Either a reference to a footnote in the text (attribute `to`) or a footnote itself (attribute `from`), in both cases with the footnote number as value.

highlow

Subscript (attribute `type` = "low") or superscript (attribute `type` = "high").

indent

Indentation, how far is given by the attribute `level` with values "1" or "two".

line

A line in the original print edition. To avoid all sorts of incorrect nesting problems, it is an empty element, occuring at the start of the line. Has the line number and the page the line is on as attributes.

page

A new page in the original print edition. Has the number as value of the attribute `no`.

sidenote

A side note/marginal note.. The attribute `margin` denotes on which margin ("left" or "right") it appears in the original print.

smallcaps

SMALL CAPS.

spacing

Emphasis by wider letter spacing.