DE Specs Working Group Meeting

Klaus Thoden

12 September 2008

1 Introduction

In this meeting, Wolfgang and Klaus presented their list of things that should be considered writing the DE Specs.[1] It was pointed out that the specifications should be fairly general to cover a large set of books.

2 Things to be marked up

Based on examples from ECHO, the following points were discussed. Structural markup means how text is organized on the page. Positional markup means how the text is formatted.

2.1 Structural markup

2.1.1 Markup done by the digitizers

Not many things will be marked up by the digitizers. This applies mainly to headings, paragraphs, columns and marginal notes. All of these will be marked by beginning and end tags.

Marginal notes should be written where they occur on the page so that they already roughly anchored to a certain place.

When page numbers are found on the page, they will be put as an argument into the header of the page break. Page breaks will be coded as milestones.

2.1.2 Things to be ignored

Catchwords and signatures at the bottom of the page will be ignored, because they do not carry any useful information.

Sentences or other semantic units will not be marked up by the digitizers, because it is too difficult.

[1] This wiki-page shows the major issues:
`https://itgroup.mpiwg-berlin.mpg.de:8080/tracs/mpdl-project-content/wiki/SampleTexts`

2.2 Positional markup

2.2.1 Ligatures

A list of ligatures will be handed to the digitizers which shows them how they should be resolved.

2.2.2 Markup of special characters

In order not to have the digitizers type too many tags, special characters could be marked up more easily. Thus, text in italics or small caps could be surrounded by an underscore (_). They have to be used with care, as texts might actually contain these characters (especially books from the 20th century).

2.2.3 Punctuation and spatia and hyphens

The spatia in the books are not consistent, be it between words, letters or letters and punctuation. As a rule, the digitizers are told not to write a spatium before a punctuation, even if it is in the text.

As for spatia inside words, nothing can be done to get the digitizers recognize words. Such errors will have to be emendated by NLP-tools. This applies also to missing hyphens.

2.2.4 Physical damage

Text might be rendered unreadable by folds, creases or even holes. In these case, the digitizers are supposed to mark these locations by a special tag.

3 Things to keep in mind

- The specifications have to be clear and simple
- You cannot code everything!

4 Next steps

A first draft version will be delivered on Friday, 19th September. Version 1.0 is due September 29.

The authors themselves, as well as willing students, are going to type some text using the DESpecs for evaluation purposes.