The Durm Corpus

toc_collapse=0; Table of Contents

1. Overview

Siamesischer Elefantenstall [404]).
1/100 w. Gr.

und, wegen des vorhandenen Holzreichtumes, als naturgemäße Bauweise sich erhalten hat, zur Herstellung kleiner Nebengebäude und von Scheide- und Abteilungswänden in Stallungen.

c) Hohle Fachwerkwände.

Die hohle Fachwerkwand unterscheidet sich grundsätzlich von der gewöhnlichen, in Kap. 6 besprochenen nur dadurch, daß die konstruierte Ausfüllung der Gefache wegfällt und daher zum Abschluß stets eine Verkleidung des Gerippes, wenigstens auf einer Seite, notwendig ist. Die Wände mit einer Ausfüllung der Gefache mit losen Stoffen, um die umschlossenen Räume gegen die Einwirkungen von Wärmeänderungen unempfindlicher zu machen, erfordern stets eine beiderseitige Verkleidung. Wände mit einseitiger Verkleidung werden nur bei ganz untergeordneten Baulichkeiten benutzt, so bei Schuppen, für Speicher, bei denen ein Luftdurchgang erwünscht ist, für Scheidungen in Kellern

As part of the Durm project, we digitized a single volume from the historical German *Handbuch der Architektur* (Handbook on Architecture), namely:

> **E. Marx:** *Wände und Wandöffnungen* (Walls and Wall Openings). In "Handbuch der Architektur", Part III, Volume 2, Number I, Second edition, Stuttgart, Germany, 1900.
> Contains 506 pages with 956 figures.

The corpus developed in this project is made available under a free document license in several formats: scanned page images, Tustep format, and XML format. Additionally, an online version and tools for transforming the various formats are available as well.

2. Background Information

For details on the background of this project and the developed corpus, please have a look at the Durm project page and our LREC 2008 paper, *A Semantic Wiki Approach to Cultural Heritage Data Management*.

The following resources are currently available:

Scanned Page Images
: The complete book was scanned at the Library of the University of Karlsruhe. For each page, a grayscale image in TIFF format with 600dpi resolution is available.

TUSTEP Format
: The page images were then transformed into a machine-readable format (see the LREC 2008 paper for details), namely

TUSTEP. The complete book is available as a single file.

XML Format

As the TUSTEP format is rather cumbersome to use with contemporary NLP tools (again, read our LREC 2008 paper for details), we developed tools for transforming it into XML format. The tools allows different conversions; one of them (complete book in single file) is offered here for download as well, together with a DTD.

PDF Format

For printing or reading the book offline, we combined the scanned page images into a single (very large!) PDF file.

3. License

All of the different versions of the Durm corpus are distributed under the terms of the GNU Free Documentation License, version 1.2 or any later version, as published by the Free Software Foundation.

4. Download

The following data files are currently available for download.

4.1. Scanned Page Images

The tar archive (gzipped) contains a single TIFF file for each physical book page. Note that the file numbers are sequentially ordered, but do not correspond to the physical page numbers:

- durm_tiff.tgz **(Note: 185 MB)**

4.2. TUSTEP Data

The complete book text in a single file, in TUSTEP markup:

- HdA.txt (v1.64, 1.7 MB)

Some further information on the TUSTEP markup is also available.

4.3. XML Data

For automated processing, we developed a conversion tool to transform the TUSTEP markup into XML (see the Durm Corpus Tools page for more information). The following file contains the complete text (with additional markup) in a single file:

- HdA.xml (2.3 MB)
- The corresponding DTD (v1.4)

Some further information on this XML format is available as well.

4.4. PDF Format

All scanned pages combined into a single PDF file:

- durm.pdf **(Note: 170 MB)**

5. Feedback

For questions, comments, etc., please use the Durm Forum.

6. Acknowledgments

Development of the Durm Corpus was funded by the German research foundation (DFG) under the title "Entstehungswissen" (LO296/18-1).