OTTo - Online Transcription Tool

A Tool for Diplomatic Transcription of Historical Texts

Stefanie Dipper - Lara Kresse
Martin Schnurrenberger - Seong-Eun Cho

Institute of Linguistics
Bochum University

Introduction

OTTo[1] is being developed in the context of the DFG-funded project "Reference Corpus Middle High German (1050—1350)".[2] This project aims at creating a reference corpus of Middle High German which is annotated with morpho-syntactic information. The corpus will be made available to the research community via the web-based corpus search tool ANNIS.[3]

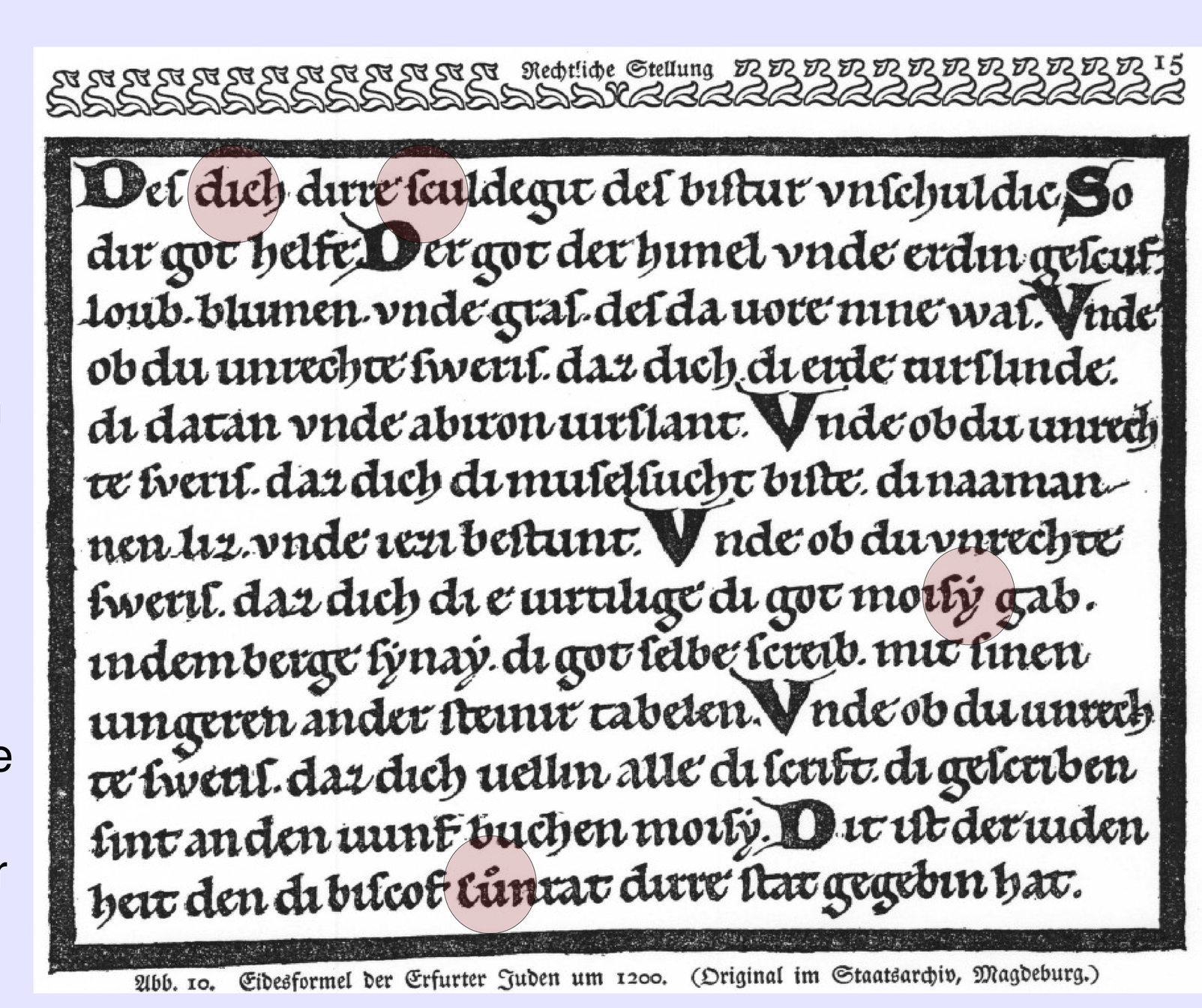
The project is part of a large initiative whose goal is to bring a diachronic German corpus into being that allows for searches through texts from different regions and centuries.

OTTo is designed for editing, viewing and storing information of historical language data. It is used through a standard web browser and supports distributed, collaborative working of multiple parties.

Historical Manuscripts

In the following excerpt, dating back at the beginning of the 13th century, a manuscript of the oldest preserved German Jewish oath is presented. This oath had to be taken by Jews in court. It says that Jews who do not honestly take this oath will be punished with diseases and death.

"Dessen, wofür dieser Dir Schuld gibt, bist Du unschuldig, so Dir Gott helfe, der Gott, der Himmel und Erde erschuf, Laub, Blumen und Gras, das zuvor nicht war. Und wenn Du unrecht schwörst, dass Dich die Erde verschlinge, die Datan und Abiran verschlang. Und wenn Du unrecht schwörst, dass Dich der Aussatz befalle, der Maeman verließ und Gehasi befiel. Und wenn Du unrecht schwörst, dass Dich die Gesetze vertilgen, die Gott Moses gab auf dem Berge Sinai, die Gott selbst schrieb mit seinen Fingern auf die steinerene Tafel. Und wenn Du unrecht schwörst, dass Dich zu Fall bringen alle Schriften, die geschrieben sind in den fünf Büchern Moses. Das ist der Juden Eid, den Bischof Konrad dieser Stadt gegeben hat."



(Gaby Herchert: Recht und Geltung, 2003, pp. 57-58)

This text will be used as an example to demonstrate what kinds of meta-information are stored in the OTTo database and how it is transcribed.

Challenges

Early manuscripts exhibit a large amount of peculiarities. Many characters are not easily encoded by, e.g., the ASCII encoding standard. Hence, an important issue with historical corpora is the transcription and encoding of special characters, e.g.:

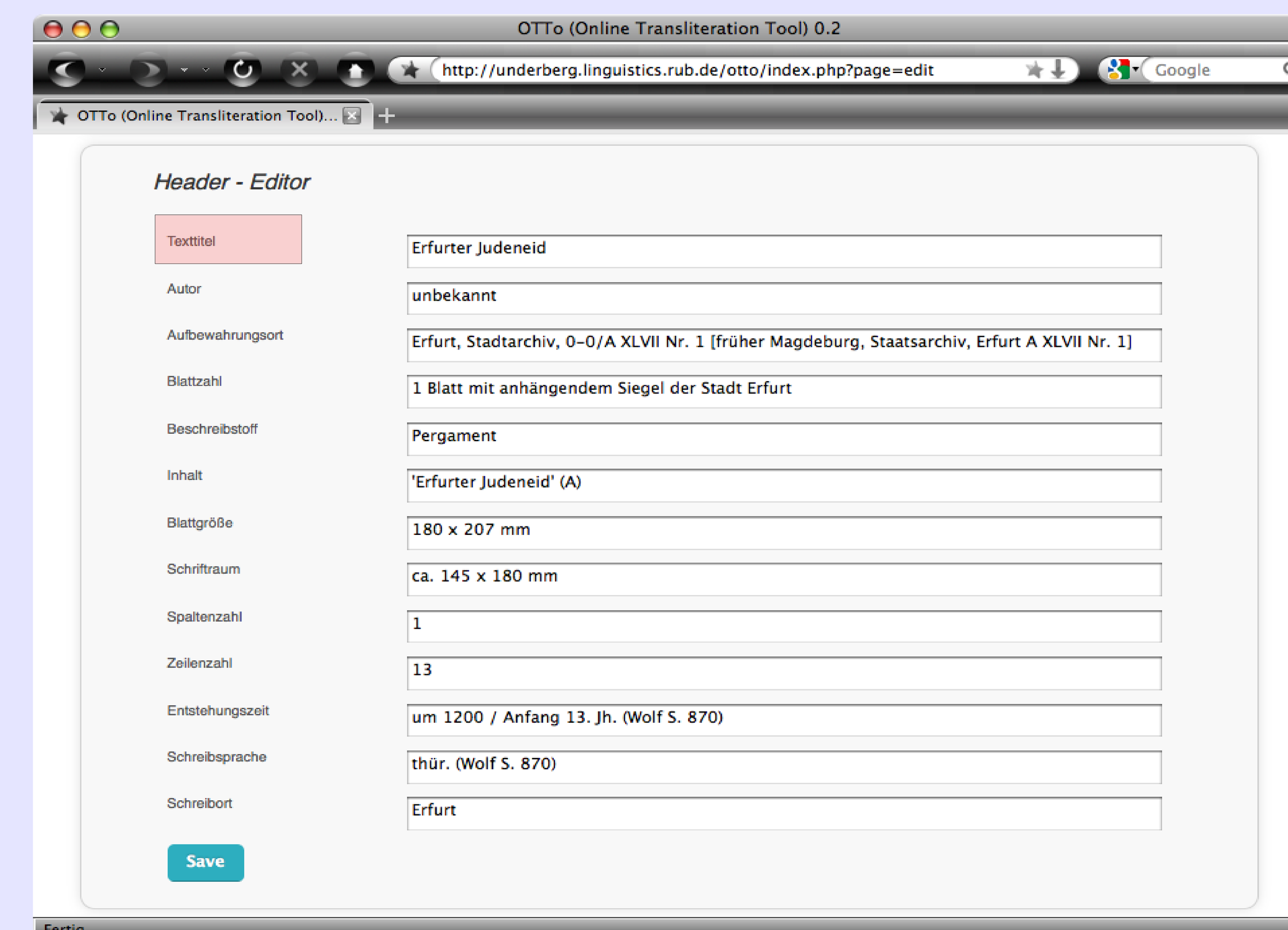
- Special letters, e.g. "long s" (ſ)
- Combination of characters, e.g. ligatures (æ), diacritics (â á à ā ă ą ā ä ẫ ẳ ầ ậ), other combinations (aͫ aͧ aͪ)
- Punctuation marks
- Abbreviations
- Different sizes of initials (A **A**)

Further attributes have to be taken into consideration:

- Glosses
- Later additions (manuscripts that were modified by other people)
- Layouts (e.g. use of columns)
- Bad conditions of the manuscripts

User-defined Header / Meta-information

Title
Author
Depository
Number of pages
Material
Content
Size (page)
Size (textfield)
Number of columns
Number of lines
Time of origin
Language
Location



Text Editor

M117 , *15* , 11	**Text M117** , *Page 15* , Line 11
Green window	The transcriber types characters and substitutes
Red window	Shows a live transformation of the entered line into its actual diplomatic transcription form, using the set of transcription rules.

User-defined Transcription Rules

Transcription rules replace some characters (e.g. $) by diplomatic Unicode characters (e.g. ſ).

- Global Rules: Project-wide set of rules, which are applied to all transcriptions in the OTTo database.
- Local Rules: Text-specific rules that encode scriber-specific idiosyncracies

	Encoding	Character	Unicode Code Point	Unicode name (or MUFI name)
1	$	ſ	U+017F	LATIN SMALL LETTER LONG S
2	v\-	v̄	U+0076 U+0304	LATIN SMALL LETTER V + COMBINING MACRON
3	a_e	æ	U+00E6	LATIN SMALL LETTER AE
4	a_e\-	ǣ	U+01E3	LATIN SMALL LETTER AE WITH MACRON
5	y\:	ÿ	U+00FF	LATIN SMALL LETTER Y WITH DIAERESIS
6	w\^	ŵ	U+0077 U+0302	LATIN SMALL LETTER W + COMBINING CIRCUMFLEX ACCENT
7	u\o	uͦ	U+0075 U+0366	LATIN SMALL LETTER O + COMBINING LATIN SMALL LETTER O
8	%.	·	U+00B7	MIDDLE DOT
9	'	'	U+F152	MUFI descriptive name: COMBINING ABBREVIATION MARK SUPERSCRIPT ER

These rules can also be used to remove abbreviations and to replace them by their expanded forms.

Future Work

- Transparent overlay of the manuscript scan on the transcription field to facilitate collating
- Improved facilities to enter special characters
  - Addition of a virtual keyboard
  - Glyph editor for designing new fonts
- TEI Export

Reference

Stefanie Dipper and Martin Schnurrenberger (2009) *OTTO: A Tool for Diplomatic Transcription of Historical Texts.* In Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 516-520. Poznan, Poland.

URL

[1] http://www.linguistics.rub.de/otto/
[2] http://www.linguistics.rub.de/mhd/
[3] http://www.sfb632.uni-potsdam.de/annis/