



MODELL ROMANTIK  
Variation • Reichweite • Aktualität

DFG Deutsche  
Forschungsgemeinschaft

# Compiling Digital Fragments of *ALZ*, a Major Textual Resource for German Romanticism Research, into a Single Comprehensive Text Corpus

Tinghui Duan<sup>1,2</sup> and Udo Hahn<sup>2</sup>

<sup>1</sup> Graduate School "Romanticism as a Model. Variation - Scope - Relevance"

<sup>2</sup> Jena University Language & Information Engineering (JULIE) Lab  
Friedrich-Schiller-Universität Jena, Jena, Germany

Jena  
University  
Language &  
Information  
Engineering  
Lab

FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA

## What is *ALZ*?

- *Allgemeine Literatur-Zeitung* (General Literature Gazette, *ALZ*)
- Published between 1785 and 1849 in Jena/Halle, Germany
- Consists of anonymous reviews of contemporary publications
- Most comprehensive review organ around 1800 in Germany
- Invaluable historical resource for literary research on German romanticism

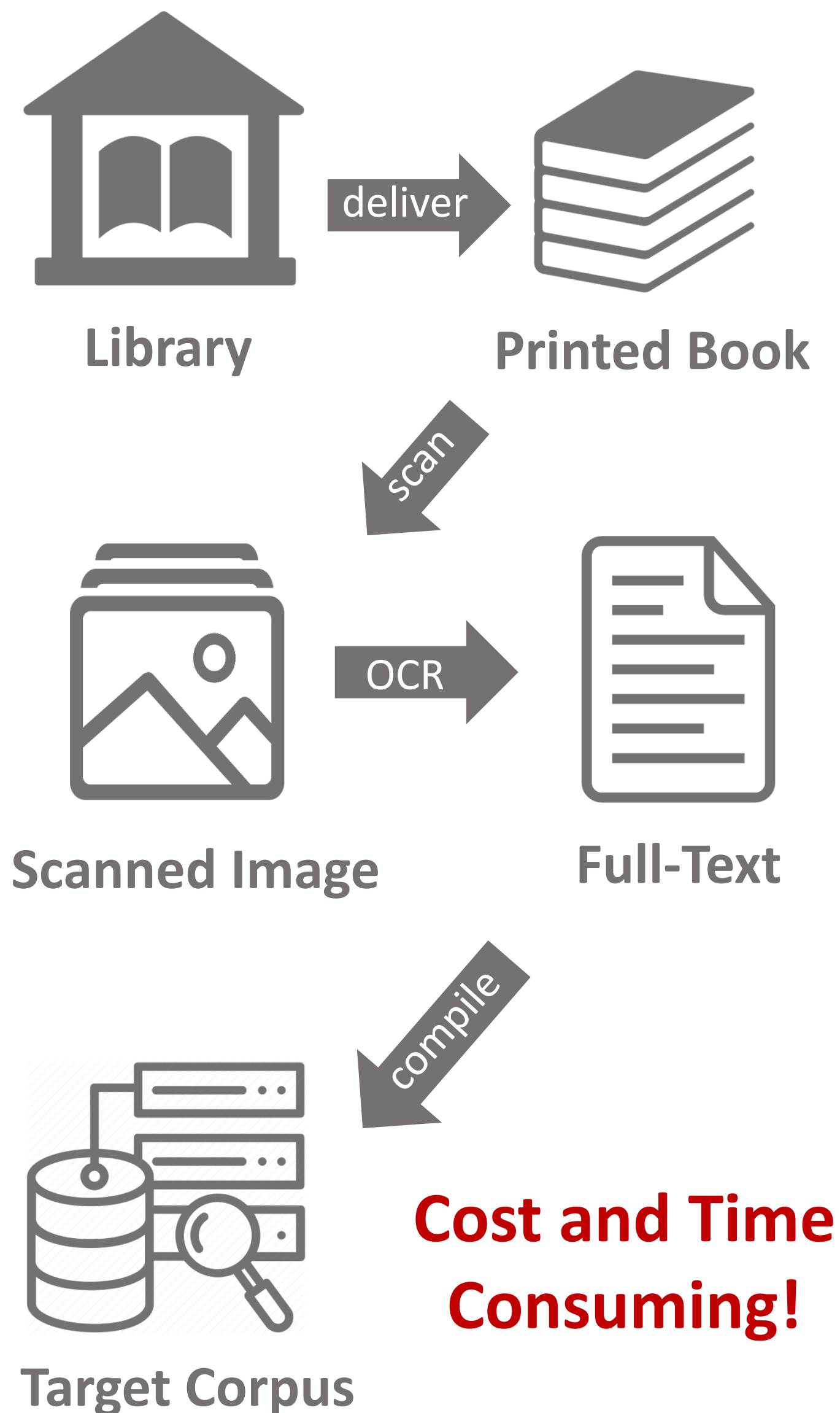
## Task

- Create a comprehensive full-text corpus of *ALZ* for research on German romanticism

## Scattered Digital Fragments of *ALZ* Hosted by Various Digital Libraries



## Traditional Workflow



## Future Work

- Adding Missing Volumes ( $\approx 18\%$ )
- OCR Error Correction
- Named Entity Recognition
- Social Network Analysis
- Topic Modeling

## Overview of All Collected Full-Text Resources

Library	Google Books			Internet Archive			Bavarian State Library		
	Volumes	Pages	Tokens	Volumes	Pages	Tokens	Volumes	Pages	Tokens
Bavarian State Library	1	285	289,269	-	-	-	152	78,675	76,568,508
University of Lausanne	4	2,008	1,969,395	-	-	-	-	-	-
Harvard University	4	3,056	2,645,386	4	3,054	2,520,832	-	-	-
Indiana University	67	33,721	32,255,992	49	24,923	25,146,053	-	-	-
New York Public Library	103	48,949	47,909,730	94	42,636	42,970,913	-	-	-
Princeton University	38	36,854	35,245,420	29	29,096	30,036,319	-	-	-
Stanford University	2	887	869,272	3	1,102	1,216,273	-	-	-
University of California	45	28,69	28,032,026	54	33,248	33,179,736	-	-	-
University of Illinois	-	-	-	4	3,192	2,820,205	-	-	-
University of Michigan	144	71,681	68,554,109	308	147,559	146,025,563	-	-	-
University of Oxford	46	20,995	20,506,977	56	26,181	25,185,788	-	-	-
<b>Total</b>	<b>454</b>	<b>247,126</b>	<b>238,277,576</b>	<b>601</b>	<b>310,991</b>	<b>309,101,682</b>	<b>152</b>	<b>78,675</b>	<b>76,568,508</b>

## Deduplicating

## Compiled Corpus (XML-Format)

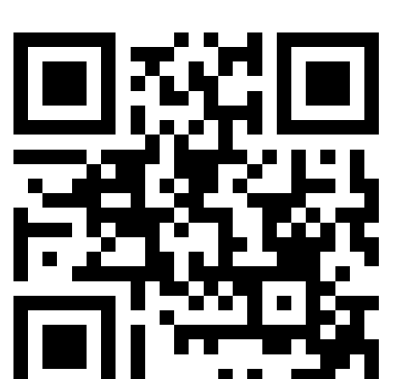
Library	Volumes	Pages	Tokens
Bavarian State Library	152	78,675	76,568,508
University of Lausanne	-	-	-
Harvard University	2	1,528	1,322,693
Indiana University	3	2,641	2,547,872
New York Public Library	4	2,08	2,043,980
Princeton University	15	8,038	7,267,803
Stanford University	-	-	-
University of California	9	3,354	3,162,025
University of Illinois	2	1,512	1,299,100
University of Michigan	64	25,538	23,276,002
University of Oxford	8	3,246	2,881,022
<b>Total ( <math>\approx 82\%</math> of all )</b>	<b>261</b>	<b>126,612</b>	<b>120,369,005</b>

Encoding

Compiling

## Survival of the Best Duplicate\*

$L \sim \text{Library}$   
 $E \sim \text{OCR Engine}$   
 $S \sim \text{Sources} \subseteq L \times E$   
 $F = \text{Fulltexts}$   
 $F \sim \{F_1, F_2, \dots, F_n\}$   
 $F_i \sim \text{Duplicates of text } i$   
 $R \subseteq F \times S \sim \text{Text exists in source}$   
 $f_{best}^i = \arg \max_{f \in F_i} \{ocrQuality(s): fRs\}$   
Minimum Edit Distance  
Manually Generated Ground Truth



Reference: Udo Hahn, Tinghui Duan. Corpus Assembly as Text Data Integration from Digital Libraries and the Web. In: *JCDL '19. Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries*, Urbana-Champaign, IL, USA, June 02–06, 2019. Pages 25–28.

Corpus Available under: <https://github.com/julielab/alz>

\* The authors would like to thank Erik Fässler for his idea to formalize the step of deduplication.



CL 2019  
International Corpus Linguistics Conference  
Cardiff, Wales, UK  
July 22–26, 2019