

# Curation Technologies for a Cultural Heritage Archive: “Project Tongilbu”

Peter Bourgonje, Julián Moreno-Schneider, Georg Rehm,  
Cosima Wagner, Martin Lee

Corresponding author: [peter.bourgonje@dfki.de](mailto:peter.bourgonje@dfki.de)

DFKI GmbH  
Speech and Language Technology, Berlin

July 8th, 2019

# Overview

- Background: Digital curation technologies
- Original project and data set
- Processing pipeline and components (German & Korean)
- Interactive curation workbench

# Digital Curation Technologies

- **DKT** (2015-2017)
- **Qurator** (2018-2021)



<http://digitale-kuratierung.de>



<http://qurator.ai>

- “use AI methods to improve quality, efficiency and cost-effectiveness of individual *curating activities* and to convert them into practical industry solutions”
- 10 project partners (Qurator) representing different domains (journalism, public archives, museums and exhibitions, health and life science, legal and compliance, etc.)

# Original Project and Data Set

- **Project Tongilbu:** “Sharing German government’s documents on unification and Integration, and Building a data-base on German unification”
  - Institute of Korean Studies, FU Berlin, 2010–2016
  - Funded by the Ministry of Unification, Republic of Korea
- Collected official political documents regarding the German reunification process to make them available for research and planning processes in the context of a potential future reunification process of Korea
- Documents were collected, intellectually curated, analysed, interpreted, partially translated and published.
- <https://www.geschkult.fu-berlin.de/e/tongilbu/index.html>

Freie Universität



Berlin

# Data Set

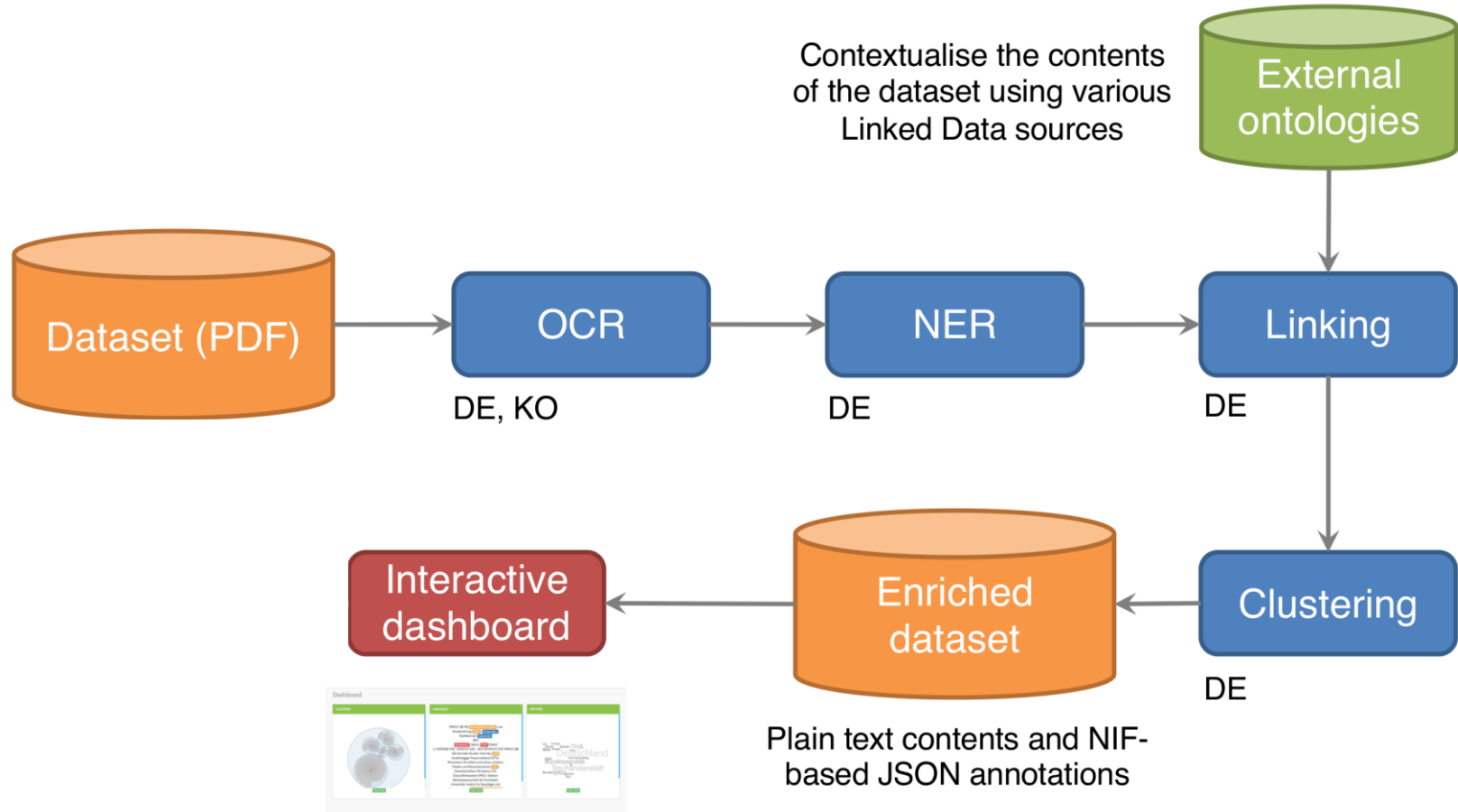
- **51 volumes**, mostly **PDF** files. Largest, most comprehensive manually curated data set on the German unification.
- Collection: transcripts of debates in the German Parliament, minutes of committee meetings, reports, proceedings etc.
- All primary documents written in German. The researchers added summaries and analyses in both German and Korean.
- **>138k pages** in both German and Korean
  - DE: 96k pages, 38m words
  - KO: 41k pages, 15m words
- File types:
  - vast majority in PDF
  - auxiliary documents in tables, Excel sheets, Word documents etc.
- **10 Gigabytes**, approx. 200 files



# Objectives

- High quality, information-rich data set with limited exploration and navigation possibilities
- Can we make this more accessible to scholars by providing the collection of PDFs inside a (curation) platform that semantically analyses, enriches and visualizes the data?
- Focus on entities (persons, organisations, locations), temporal expressions and thematic contexts.
- Integrate this use case in the curation platform also used in the DKT/Quator contexts

# Processing Pipeline



# OCR

## Einleitung

Fünfzehn Jahre nach der deutschen Vereinigung liegt die Wirtschaftskraft der neuen Länder noch immer deutlich niedriger als in Westdeutschland. Im Jahre 2003 belief sich das Bruttoinlandsprodukt je Einwohner in den neuen Ländern auf rund 64% des westdeutschen Vergleichswertes; etwas höher ist – mit knapp 73% – der Angleichungsstand bei der Arbeitsproduktivität. Die Arbeitnehmereinkommen – die sich langfristig an der Produktivität orientieren müssen – liegen demgegenüber schon Niveaus. Dies geht einher mit einer Quote der registrierten Arbeitslosigkeit nicht nur durch einen Mangel an rentablen Investitionen verursacht, sondern in nahezu gleichem Maße auch durch den geringeren Erwerbsteilnahme in den neuen Ländern.

■ 약어색인

AA	외무청 (독일연방공화국 외무부)
ABC-Waffen	핵, 세균 및 화학 무기 (대량 살상 무기) (Atomare, biologische und chemische Waffen)
ABM	일자리 창출 조치 (Arbeitsbeschaffungsmaßnahmen)
ADN	(동독) 공영 독일 통신사 (Allgemeiner Deutscher Nachrichtendienst)
AfD	독일 연합 (Allianz für Deutschland)
AGCK	동독 기독교회 노조 (Arbeitsgemeinschaft der Kirchen in der DDR)
AM	외무장관 (Außenminister)
Anm	주석 (Anmerkung)
Az	파일 번호
BARd	연방자료보관소
BEK	동독 기독교 연맹 (Bund der Evangelischen Kirchen (DDR))

- Input PDFs are high quality (mostly)
- they include typewriter fonts, with the occasional table,
- typically multi-column layout,
- and mixed German/Korean

Vertrag  
über die Beziehungen  
zwischen der Bundesrepublik Deutschland  
und den Drei Mächten

Convention  
on Relations between the Three Powers  
and the Federal Republic of Germany

Convention  
sur les relations entre les Trois Puissances  
et la République Fédérale d'Allemagne

DIE BUNDESREPUBLIK DEUTSCHLAND	THE UNITED STATES OF AMERICA,	LA RÉPUBLIQUE FRANÇAISE,
einserseits und	THE UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND	LES ÉTATS-UNIS D'AMÉRIQUE
DIE VEREINIGTEN STAATEN VON AMERIKA, DAS VEREINIGTE KÖNIGREICH UND NORDIRLAND	and THE FRENCH REPUBLIC,	LE ROYAUME-UNI DE GRANDE-BRETAGNE ET D'IRLANDE DU NORD
andererseits	of the one part, and THE FEDERAL REPUBLIC OF GERMANY,	d'une part, et LA RÉPUBLIQUE FÉDÉRALE D'ALLEMAGNE
DIE FRANZÖSISCHE REPUBLIK	of the other part;	d'autre part.

der Erwägung,  
friedliche und blühende  
Völkergemeinschaft, die  
kennst zu den Grund-  
sätzen der Vereinten Na-  
tionen anderen freien Völ-  
ker verbunden ist, nur  
te Förderung und Ver-  
r gemeinsamen Freiheit  
sinnamen Erbes verwerk-  
lich

WHEREAS a peaceful and pro-  
perous European community of nations  
firmly bound to the other free nations  
of the world through dedication to  
the principles of the Charter of the  
United Nations can be attained only  
through united support and defence  
of the common freedom and the com-  
mon heritage;

CONSIDÉRANT qu'une communauté  
européenne pacifique et prospère,  
étroitement liée aux autres nations  
libres du monde par un attachement  
commun aux principes de la Charte  
des Nations-Unies, ne peut être réa-  
lisée que par l'union des efforts des  
nations d'Europe en vue de défendre  
leur liberté et leur patrimoine com-  
mun;

WHEREAS it is the common aim of  
the Signatory States to integrate the  
Federal Republic on a basis of equali-  
ty within the European Community  
which is included in a developing At-  
lantic Community;

CONSIDÉRANT que les États Si-  
gnataires ont pour objectif commun  
d'intégrer la République Fédérale sur  
une base d'égalité dans la Commu-  
nauté Européenne, elle-même incluse  
dans une communauté atlantique en  
développement;

### 1. Der Weg zu einer demokratischen Verfassung Brandenburgs

Wie lehren uns Gott ist eine Verfas-  
sung für alle Dienstvernehmer der Bürger  
von höchstem Belang. Als staatliches  
Grundgesetz greift sie mittelbar oder un-  
mittelbar in die Lebensverhältnisse jedes  
Menschen ein.

In der Verfassung regelt ein Land seine  
staatliche Grundordnung. Sie bestimmt die  
staatliche Organisation und Struktur und  
die grundsätzlichen Beziehungen zwischen  
Staat und Bürgern mit dem Ziel der Regu-  
lierung innerer Konflikte. Deshalb enthal-  
ten die meisten Verfassungen einen Kata-  
log von Freiheits- und Gestaltungsrechten.  
Weil dies in der Verfassung formulierte  
Recht höherwertig ist als die ihr nachge-  
ordnete Rechtssetzung und sie wegen der  
von ihr getroffenen Regelung grundlegender  
Probleme für die staatliche Gemeinschaft  
von besonderem Gewicht ist, wird eine  
dauerhafte Verfassungsnormierung  
angestrebt. Deshalb ist für das Zustandekommen  
einer Verfassung meist eine besondere  
Form vorgeschrieben, z.B. ein Volksentscheid  
oder eine Zweidrittelmehrheit.

Die erste Verfassung Brandenburgs  
1947

Als nach dem Zusammenbruch des nationa-  
lsozialistischen Regimes die sowjetische  
Siegmacht daranging, in dem von ihr  
beherrschten Teil Deutschlands eine neue  
Staatsmacht zu installieren, spielten Ver-  
fassungsfragen vorerst nur eine unterge-  
ordnete Rolle. Die schweren Aufgaben des  
täglichen Überlebens hatten für die meis-  
ten Menschen Priorität. Und die Politik  
der von der KPD/SED dominierten neuen  
Machorgane in der sowjetischen Besatzungs-  
zone zielt im Sinne der Lehren Le-  
nins darauf ab, vollendete Tatsachen ins-  
besondere in strukturell-personellen Be-  
ziehungen der staatlichen Macht und in den  
Eigentumsverhältnissen zu schaffen. Sie  
sollten eine in der Folgezeit zu beschle-  
dende Verfassung von vornherein weitge-  
hend fertigen und zukünftige - durchaus  
an bestimmten parlamentarischen Tradi-  
tionen formell orientierte - gewählte Ver-  
tretungskörperschaften inhaltlich binden.



# OCR

- **Apache Tesseract** to convert PDF to plain text
  - 4.0 LSTM (,best' model) for DE
  - 3.5 for KO (segmentation issues when using 4.0)
- Evaluation using four most frequent content types
- Ground truth created using Transkribus, evaluated using ocrevalUAtion:
  - DE: **2.870** words
  - KO: **2.483** words

# OCR

Content Type	German		Korean	
	WER	CER	WER	CER
Machine-readable (one column)	12.99	2.05	43.19	19.27
Machine-readable (two columns)	65.73	27.05	–	–
Tables	56.64	20.72	85,39	68,99
Typewriter scans	12.23	2.24	–	–

table from Rehm et al. DATeCH-2019, Brussels, Belgium

- Results vary greatly depending on content type
- Improving OCR results for these formats would be beneficial but for our downstream applications, the impact remains relatively limited:
  - Majority of our data set's content is single column.
  - Pipeline relies on entities, probably minimally affected by the incorrect interpretation of individual rows in a table.
- Future improvements: using XML/hOCR as Tesseract output format (instead of plain text) to preserve structural information.

# NER, Entity Linking, temporal expressions

- Recognising **persons**, **locations** and **organisations** in the plain text OCR output
- For German:
  - **OpenNLP** trained with WikiNER (>1m wikipedia articles) for spotting, **GND** (Gemeinsame Normdatei) for linking
  - **HeidelTime** and custom set of patterns for temporal expressions and normalisation
- For Korean (experiments so far unsuccessful):
  - **OpenNMT** using opensubtitles parallel data for translating KO > DE
    - far out of domain (but not much else available), much left untranslated, German output not very relevant
  - Korean NER (<https://github.com/digitalprk/KoreaNER>)
    - Missing models/training data, sparsely documented
  - Lower level processing (POS-tagging; <http://konlpy.org>)
    - No relevant tags found that reliably relate to entities

# Clustering

- Clustering done based on URIs of recognised entities
- **WEKA** (Expectation Maximization)
- Allows theme-based exploration of the data set (starting off with a certain region, cluster of people, etc.)
- Because of challenges with upstream tasks for Korean, using German entities only

# Curation workbench

The dashboard is titled "Dashboard" and features a navigation sidebar on the left with icons for home, power, download, building, arrow, book, folder, folder, and settings. At the top right, there is a user profile for "Julian Administrator".

The dashboard is divided into three main sections:

- CLUSTERS:** A bubble chart showing a large cluster at the bottom and several smaller clusters above it, each labeled with a number (1, 2, 3, 4, 5, 6). A "View more" button is at the bottom.
- HIGHLIGHT:** A text-based highlight section with a "View more" button at the bottom. The text includes:

1990년 2월 9일 **Bundesminister für Arbeit** und Sozialordnung (**BMA**), **Norbert Blüm**; Bundeskanzler **Helmut Kohl** 출처 **Bundesarchiv**, BArch, **B 136** /21660 2 사회현장을 위한 기본원칙과 입장 - 동독 원탁회의의 토론 1990년 3월 5일 Zentraler Runder Tisch der **DDR**; Unabhängiger Frauenverband (UFV); Ministerium für Arbeit und Löhne; Initiative Frieden und Menschenrechte (**IFM**); Gewerkschaften; Ministerium für Gesundheitswesen (MfG); Sektion Rechtswissenschaft der Humboldt-
- ENTITIES:** A word cloud of entities including "Bundesrepublik", "Deutschland", "Deutsche Bundestag", "Berlin", "Korea", "EG", "DIW", "CDU", "Bundes Sachsen", "Europa", "Bonn", "Bremen", "Staates", "SPD", "THA", "Treuhandans", and "RWI". A "View more" button is at the bottom.

© 2019 DFKI. All Rights Reserved | Design by w3layouts.

# Summary & Conclusions

- Dashboard for the curation of cultural heritage data
  - Goal: intuitive analysis, exploration, visualisation of data
- Tools readily available for German, much less support for Korean
  - OCR (Apache Tesseract) works well for both, challenge lies mostly in proper handling of layout/text formatting
  - NER:
    - Not much available, what is available could be improved upon (issues with documentation, availability of required models/data)
    - Alternative approaches (MT > annotation projection) not feasible due to lack of training data

Thank you for your attention!