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

DFKI GmbH
Speech and Language Technology, Berlin

July 8th, 2019

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

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

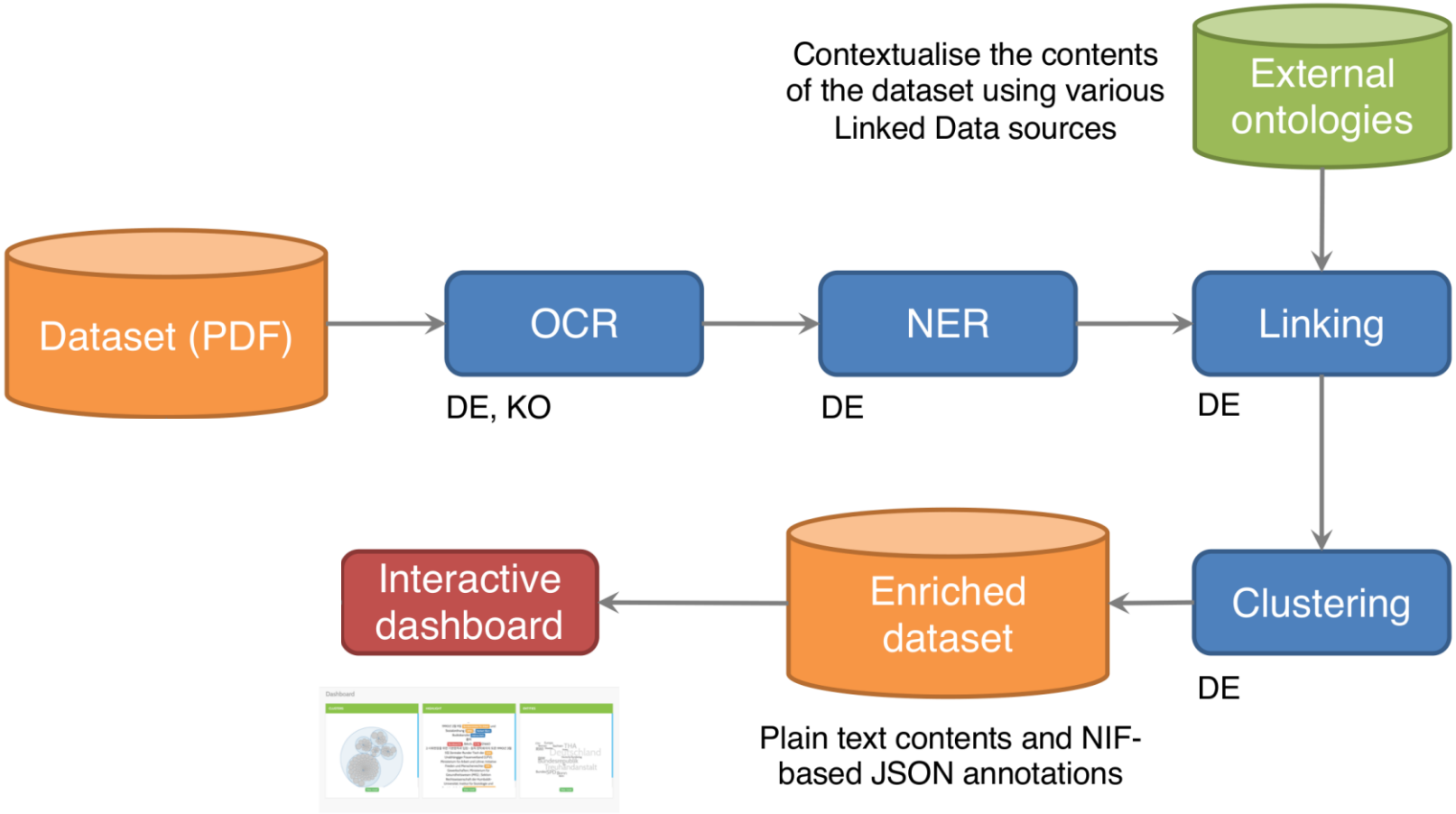
Data Set

- **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

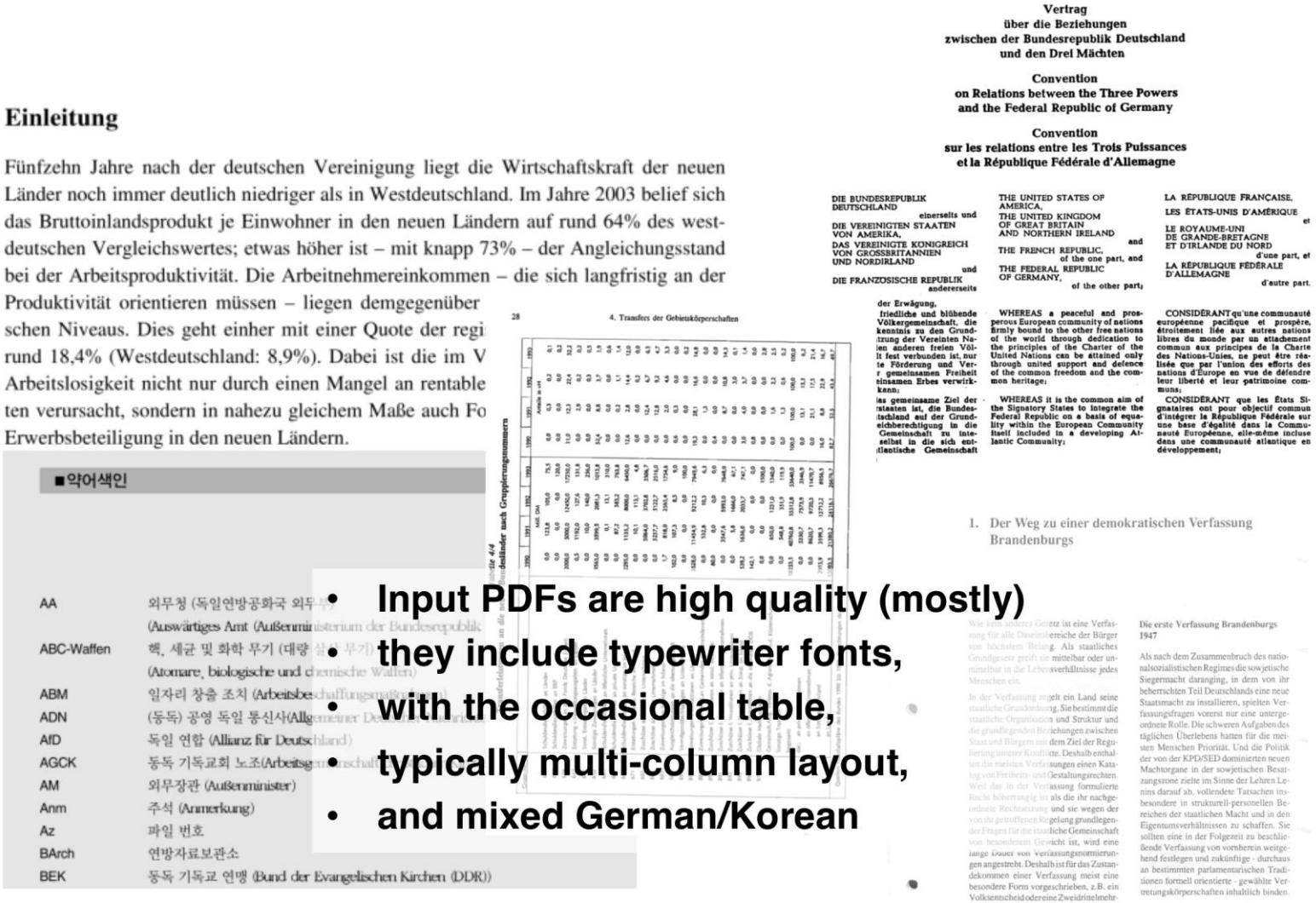
Objectives

- 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/Qurator 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
schen Niveaus. Dies geht einher mit einer Quote der regi
rund 18,4% (Westdeutschland: 8,9%). Dabei ist die im V
Arbeitslosigkeit nicht nur durch einen Mangel an rentable
ten verursacht, sondern in nahezu gleichem Maße auch Fo
Erwerbsbeteiligung in den neuen Ländern.

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
einerseits und
DIE VEREINIGTEN STAATEN VON AMERIKA,
DAS VEREINIGTE KONIGREICH VON GROSSBRITANNIEN UND NORDIRLAND
und
DIE FRANZOSISCHE REPUBLIK
andererseits

THE UNITED STATES OF AMERICA,
THE UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND
and
THE FRENCH REPUBLIC,
of the one part, and
THE FEDERAL REPUBLIC OF GERMANY,
of the other part;

LA RÉPUBLIQUE FRANÇAISE,
LES ÉTATS-UNIS D'AMÉRIQUE
et
LE ROYAUME-UNI DE GRANDE-BRETAGNE ET D'IRLANDE DU NORD
d'une part, et
LA RÉPUBLIQUE FÉDÉRALE D'ALLEMAGNE
d'autre part.

der Erwägung,
friedliche und blühende Völkergemeinschaft, die
kenntnis zu den Grund-
tzung der Vereinten Na-
len anderen freien Völ-
lt fest verbunden ist, nur
te Förderung und Ver-
r gemeinsamen Freiheit
einsamen Erbes verwirk-
kann;

las gemeinsame Ziel der
staaten ist, die Bundes-
tschland auf der Grund-
eichberechtigung in die
Gemeinschaft zu inte-
selbst in die sich ent-
tlantische Gemeinschaft

WHEREAS a peaceful and prosperous 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 common heritage;

WHEREAS it is the common aim of the Signatory States to integrate the Federal Republic on a basis of equality within the European Community itself included in a developing Atlantic Community;

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 réalisée que par l'union des efforts des nations d'Europe en vue de défendre leur liberté et leur patrimoine communs;

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

28 4. Transfers der Gebietskörperschaften

Gebietskörperschaften nach Gruppierungsnummern

■약어색인

AA	외무청 (독일연방공화국 외무
	(Auswärtiges Amt (Außenmi
ABC-Waffen	핵, 세균 및 화학 무기 (대량
	(Atomare, biologische und c
ABM	일자리 창출 조치 (Arbeitsbe
ADN	(동독) 공영 독일 통신사(Allg
AfD	독일 연합 (Allianz für Deutsc
AGCK	동독 기독교회 노조(Arbeitsg
AM	외무장관 (Außenminister)
Anm	주석 (Anmerkung)
Az	파일 번호
BArch	연방자료보관소
BEK	동독 기독교 연맹 (Bund der Evangelischen Kirchen (DDR))

1. Der Weg zu einer demokratischen Verfassung Brandenburgs

Die erste Verfassung Brandenburgs 1947

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

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.

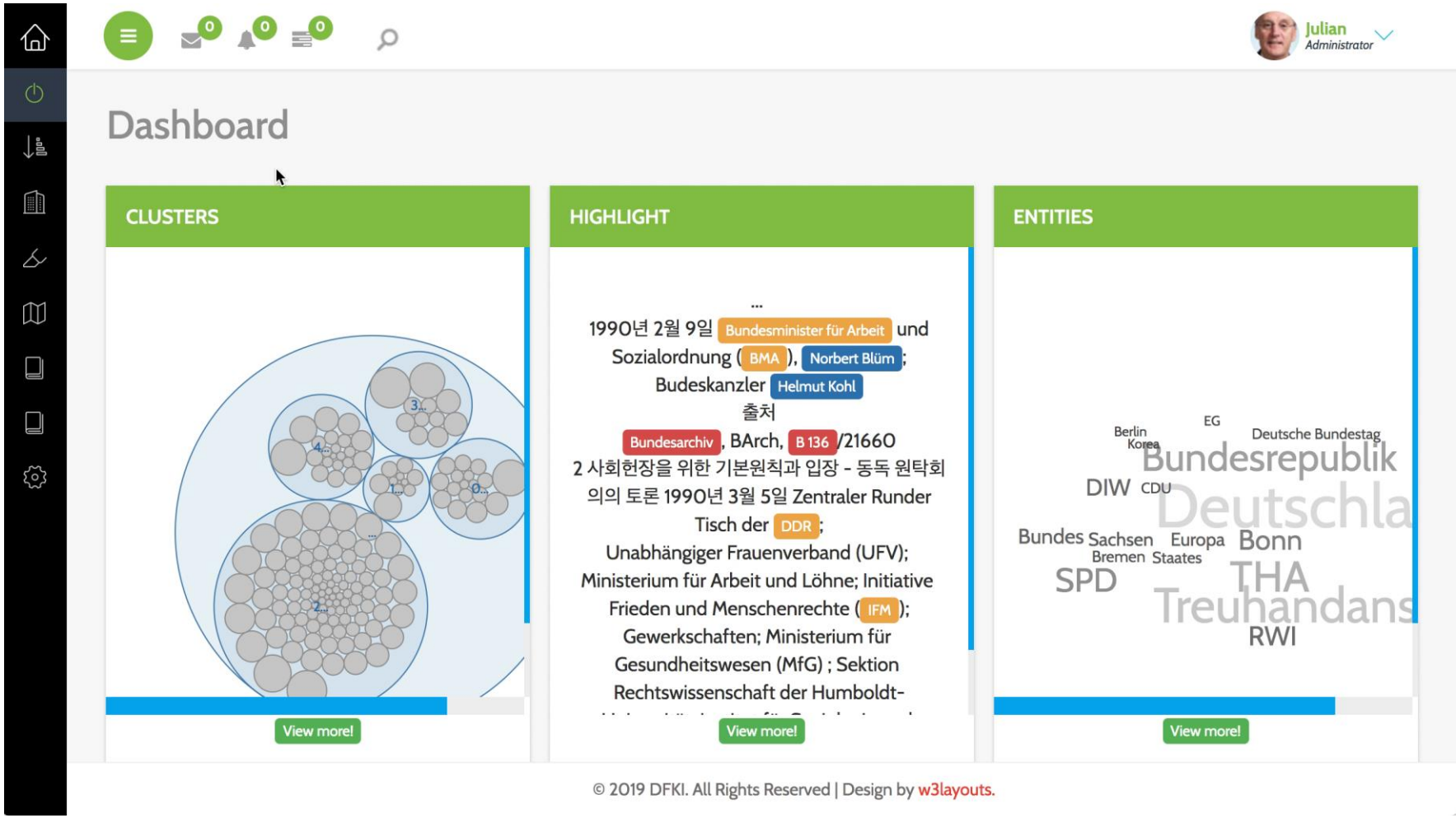
NER, Entity Linking, temporal expressions

- 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



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!