DEUTSCHE NATIONAL BIBLIOTHEK

Christa Schöning-Walter, Sandro Uhlmann

Herausforderungen der Textextraktion im Produktivbetrieb der DNB

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

1. **Hintergrund**
2. **Beispiele typischer Fehlerfälle**
3. **Unsere Fragen an die Teilnehmenden**

DEUTSCHE NATIONAL BIBLIOTHEK

Hintergrund

- Die Inhaltstreue der Textgrundlage – in unserem Fall die fehlerfreie Gewinnung von Plain Text aus den Dateiformaten PDF und EPUB – ist Grundvoraussetzung für gute Ergebnisse der maschinellen Erschließung.
- Im Massenbetrieb der DNB gelingt dies noch nicht gut genug: Der Text, der aus den digitalen Publikationen erzeugt wird, entspricht nicht immer dem Inhalt der Publikationen.

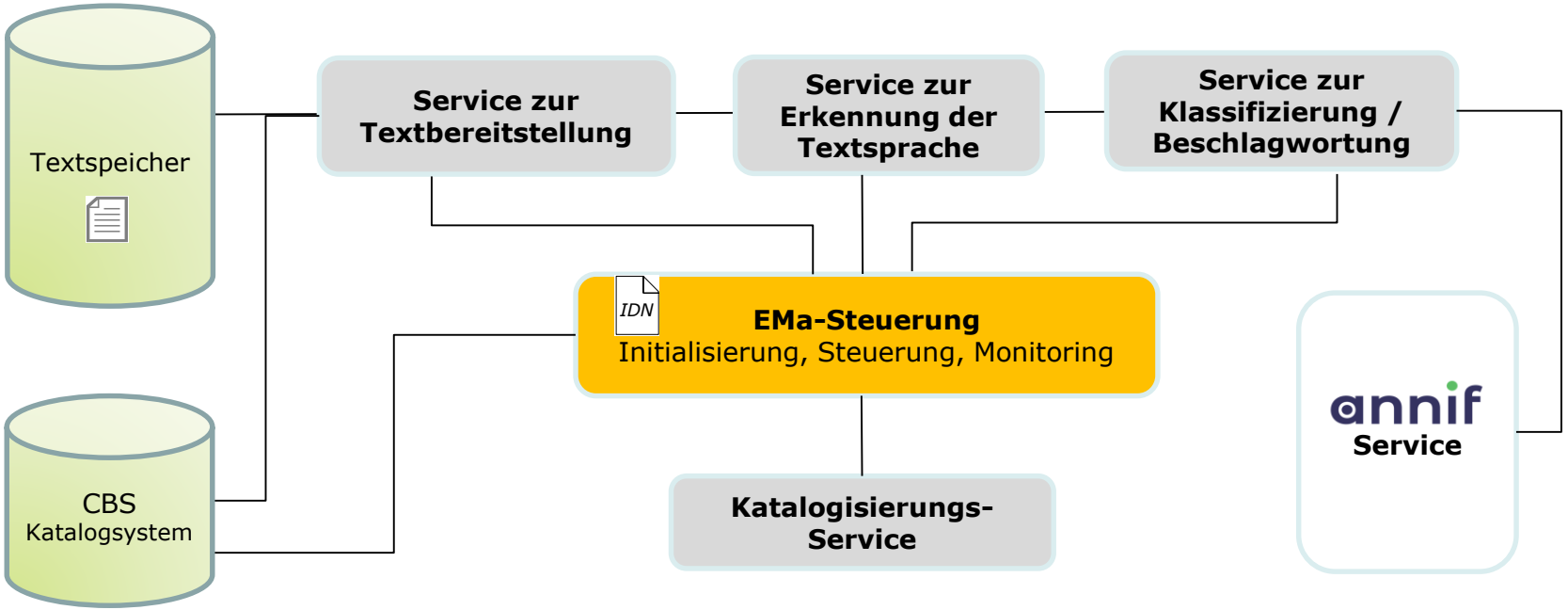
DEUTSCHE NATIONAL BIBLIOTHEK

Hintergrund

- Derzeit setzt die DNB für die Textextraktion eine ältere Version einer Open Source-Software für PDF (Stand 2013) und eine Eigenentwicklung für EPUB ein.
- Verarbeitet werden (aktuell veröffentlichte oder digitalisierte) Monografien und Zeitschriftenartikel:
  - sehr große Mengen,
  - sehr große Vielfalt der Gestaltungsmerkmale (Layout, Schrift etc.).

DEUTSCHE NATIONAL BIBLIOTHEK

Modulare Erschließungsmaschine EMa



DEUTSCHE NATIONAL BIBLIOTHEK

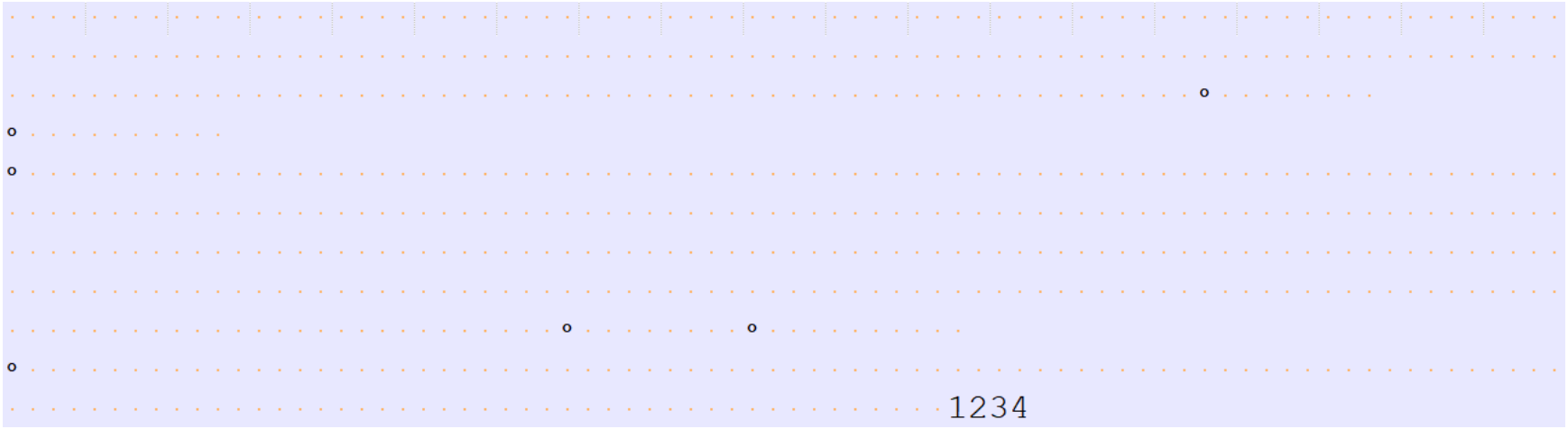
PDF-Datei ohne Unicode-Codierung von Umlauten und 'ß‘

Textbereitstellung liefert Leerzeichen statt Umlaute oder 'ß'

```
Wolfgang Dohle Funktionalisierte Heterocyclen durch eine
Halogen-Magnesium-Austauschreaktion M nchen 2002
Dissertation zur Erlangung des Doktorgrades der Fakult t f r
Chemie und Pharmazie der Ludwig-Maximilians-Universit t M nchen
Funktionalisierte Heterocyclen durch eine
Halogen-Magnesium-Austauschreaktion von Wolfgang Dohle aus
Winterberg M nchen 2002 Erkl rung Diese Dissertation wurde im
Sinne von § 13 Abs.3 bzw. 4 der Promotionsordnung vom 29. Januar
1998 von Professor Dr. Paul Knochel betreut. Ehrenw rtliche
Versicherung Diese Dissertation wurde selbst ndig und ohne
unerlaubte Hilfe erarbeitet. M nchen, am 18. September 2002
```

Hinterlegter Text enthält nur Steuerzeichen bzw. nicht darstellbare Zeichen

Textbereitstellung liefert nur nicht darstellbare Zeichen



DEUTSCHE NATIONAL BIBLIOTHEK

PDF-Datei enthält Text mit falsch gewähltem Zeichensatz

Textbereitstellung liefert nur Ascii-Zeichen

```
!q h C n n !O q t c n k v { < ! v j g ! G z c o r n g ! q h !E c v q !v j
g !G n f g t / / // / // / // / // // / // / // / // // / / / // / // // /
// / // / // // / // / // / / / 2 1 8
;/!Jgkmg!PKSWGV<!Vjg!Kfgcn!qh!vjg!Ugpcvqtkcn!Citkewnvwtkuv!cpf!Tgcnkv{
Fwtkpi!vjg!Tqocp!Tgrwdnke!cpf!Gorktg////////////////////////////////////
////////////////////// 232 Vcdng!qh!Eqpvgpvu 7
21/!Jcpu!Łłti!JGPPGEMG<!Nc{gtu!qh!Oqtcnkv{!cpf!Vjgkt!Eqphnkev
kp!vjg!Jkuvqtkqitcrjke!Yqtm!qh!Itgiqt{!qh!Vqwtu
////////////////////////////////////////////246 22/!Uqplc!OCT\KP\KM
< !I t c x g . I qq f u ! k p ! E q p x g t u k q p ! R g t k q f !c p f !
N c v g t D w t k c n u ! ! c ! E c u g ! q h ! G c t n { ! O g f k g x c
n ! T g n k i k q w u !F q w d n g ! U v c p f c t fu A / / // / // // /
// / // / / // / // / // 2 5;
```

DEUTSCHE NATIONAL BIBLIOTHEK

Typische Herausforderungen

- keinerlei Text hinterlegt
- extrahierte Zeichen bilden keine sprachlichen Einheiten
- Sonderzeichen sind in der falschen Unicode-Normalform (NFD, NFC, NFKD oder NFKC) codiert
- Silbentrennung beim Zeilen- oder Seitenwechsel
- Gesperrtschreibung

DEUTSCHE NATIONAL BIBLIOTHEK

Unsere Fragen

- Welche Methoden und Werkzeuge setzen Sie für die Textextraktion aus PDF- und EPUB-Dateien ein?
- Welche OCR-Verfahren empfehlen Sie für die Texterkennung und welche Qualität ist erreichbar?
- Sind Methoden und Metriken bekannt, mit denen es möglich ist,
  - fehlerhaften Text zu identifizieren?
  - die Qualität der extrahierten Texte zu bewerten oder sie zu verbessern?
- Können Sie uns Tools oder Softwarebibliotheken empfehlen?