

DHPs

gitlab

Semi-strukturi

arser

b.lrz.de/badw-it/DHPPars

rierte Daten

Formale Gr zur Retro-D bibliograph

ser - [dhparser.readth](https://github.com/dhparser/readth)

DH

Grammatiken Digitalisierung nischer Daten

hedocs.io

Parser

g

1

- Daten in **schematisierten** Formaten
- aber **von Menschen** erstellt
- daher viele **Ausnahmen**
- in Textformaten nicht **strukturiert**
ohne semantische Informationen
- meist nur ein **unvollständiges** Schema
kein explizites RDBMS

Beispiele:

(Fach-)Bibliografien

Druck-Wörterbücher

stischer Form
menhand gepflegt
ahmen, Sonderfälle
niedergelegt,
he Auszeichnungen
vollständiges oder gar
Regelwerk vorhanden

n in Buchform,
er, Kataloge

Parser

dhparsen.readthedocs.io

(Baum-)Tra

dhparsen.readthedocs.io/

Testu

dhparsen.readthedocs.io/

¹⁾ für Nicht-Baumstrukturen, z.B Graphen oder, wie
Bearbeitungsschritte außerhalb von DHParser ang

Beispiel 1: I

Albert Weber: Bibliograph
aus dem östlichen Europa

Titel: SOZIALDEMOKRA
IN BULGARIEN

Erscheinungsort: So

Erscheinungsbeginn:

Ethnische Gemeinschaft

ergenerator

cs.io/en/latest/manuals/01_ebnf.html

ansformationen ¹⁾

cs.io/en/latest/manuals/03_transform.html

umgebung

cs.io/en/latest/manuals/05_testing.html

wie in den Beispielen unten LOD, müssen noch weitere
angefügt werden.

Der günstige Fall:

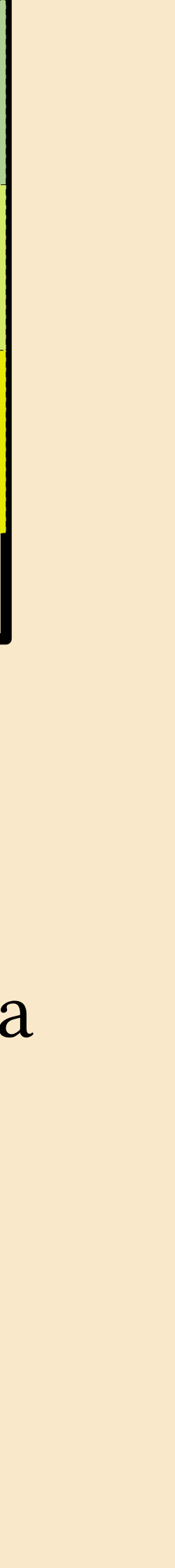
ographie deutschsprachiger Periodika
a, IOS Regensburg 2013.

RATISCHE ARBEITERPARTEI

ofia

: 1907 , 1910 [AW4]

haft: Bulgariendeutsche



a

Formale Gram

- taugen wie reguläre
Erkennung von
- **leistungsfähiger**
und übersichtliche

Herausforderungen

- **Steile Lernkurve**
- ohne explizites Re

Grammatiken

Äre Ausdrücke zur

Textmustern

r (durch Rekursion)

er als reg. Ausdrücke

en:

e für Nicht-Informatiker

egelwerk (siehe

- **Daten vom Autor vor**
(beugt auch Fehlinterpretation vor)
- Klare und weitgehend einheitliche
- Mäßiger **Variantenreichtum**
durch die teils komplizierten
verzeichneten Werke
→ **Arbeitsaufwand** Par

Beispiel 2: D

Ralf Schönberger et al.: R
Mittelalters aus dem Bereich
angrenzender Gebiete

haft: Bulgariendeutsche

**rbereitet für die Digitalisierung
retationen der Techniker vor!)**

eindeutige Struktur

ichtum, sachlich bedingt u.a.

erte Erscheinungsgeschichte der

urserbau: 2-3 Werktage

Der ungünstige Fall:

**Repertorium edierter Texte des
reich der Philosophie und**

g
b

r

oben) muss die Gr
Versuch-und-Irr
(→ Testumgebung

- **keine Erfolgsgar**
überhaupt genug S
um ihre Syntax fo
nicht von Anfang

Dr. A

Ingo

Grammatik in einem
Erkennungsverfahren

(g) rekonstruiert werden
Garantie: Ob die Daten
Struktur enthalten,
normal zu beschreiben, ist
an absehbar

Dr. Eckhart Arnold, Bayerische Akademie der Wissenschaften

www.ea.de

Albert Weber, Leibniz-Institut für Computerlinguistik

leibniz-ios.de/pers

Dr. Frank M.A., Leibniz-Institut für Computerlinguistik

B1700-1707/03

ED: Chantelou, Claudius.

IN: Bibliotheca Patrum a

Patrum de christiana ac
opuscula, I, Paris 1662.

AT: Panereticon, pars pr
de sanctis complectens

Bernardus Claraevallensi

⇒ B1700-1707/20

- Daten nicht präpariert:
- Teils **uneindeutige Felder**
- **Idiosynkratischer Variablen**
Literatur- Orts- und Zeita
- **Fehler**, z.B. Rechtschreibf
→ machbar, aber **Arbeitsa**

Akademie der Wissenschaften, Münch
[ueckhartarnold.de](http://www.zentrum-berlin.de/personen/ueckhartarnold.de)

Ost- und Südosteuropa Forschung, R
[ersonen/details/albert-weber](http://www.zentrum-berlin.de/personen/details/albert-weber)

Ost- und Südosteuropa Forschung, R

3.

ascetica sive selecta veterum
e religiosa perfectione

2.

prima, Sermones de tempore et

is: Vita s. Malachiae.

t: Sinn machmal nur zu erraten
dabgrenzungen

iantenreichtum (insbesondere bei
angaben)

bfehler in Schlüsselwörtern!

saufwand schwer abschätzbar

nchen

Regensburg

Regensburg

Lizenz:

CC-BY 4.0

en.

Regensburg

