DHParser

Formale Grammatiken zur Retro-Digitalisierung bibliographischer Daten

gitlab.lrz.de/badw-it/DHParser - dhparser.readthedocs.io

Semi-strukturierte Daten

- Daten in **schematischer Form**
- aber **von Menschenhand** gepflegt
- daher viele **Ausnahmen**, Sonderfälle
- in Textformaten niedergelegt, **ohne semantische Auszeichnungen**
- meist nur ein unvollständiges oder gar **kein explizites Regelwerk** vorhanden

Beispiele:

(Fach-)Bibliografien in Buchform, Druck-Wörterbücher, Kataloge

Formale Grammatiken

- taugen wie reguläre Ausdrücke zur **Erkennung von Textmustern**
- **leistungsfähiger** (durch Rekursion) und übersichtlicher als reg. Ausdrücke

Herausforderungen:

- **Steile Lernkurve** für Nicht-Informatiker
- ohne explizites Regelwerk (siehe oben) muss die Grammatik in einem **Versuch-und-Irrtumsverfahren** (⟶ Testumgebung) rekonstruiert werden
- **keine Erfolgsgarantie**: Ob die Daten überhaupt genug Struktur enthalten, um ihre Syntax formal zu beschreiben, ist nicht von Anfang an absehbar

DHParser

Parsergenerator dhparser.readthedocs.io/en/latest/manuals/01_ebnf.html
(Baum-)Transformationen [1)] dhparser.readthedocs.io/en/latest/manuals/03_transform.html
Testumgebung dhparser.readthedocs.io/en/latest/manuals/05_testing.html

[1)] für Nicht-Baumstrukturen, z.B Graphen oder, wie in den Beispielen unten LOD, müssen noch weitere Bearbeitungsschritte außerhalb von DHParser angefügt werden.

Beispiel 1: Der günstige Fall:

Albert Weber: Bibliographie deutschsprachiger Periodika aus dem östlichen Europa, IOS Regensburg 2013.

```
Titel: SOZIALDEMOKRATISCHE ARBEITERPARTEI
IN BULGARIEN
Erscheinungsort: Sofia
Erscheinungsbeginn: 1907, 1910[AW4]
Ethnische Gemeinschaft: Bulgariendeutsche
GLP ID: 0016
```

- **Daten vom Autor vorbereitet** für die Digitalisierung (beugt auch Fehlinterpretationen der Techniker vor!)
- Klare und weitgehend **eindeutige Struktur**
- Mäßiger **Variantenreichtum, sachlich bedingt** u.a. durch die teils komplizierte Erscheinungsgeschichte der verzeichneten Werke

→ **Arbeitsaufwand Parserbau: 2-3 Werktage**

Beispiel 2: Der ungünstige Fall:

Ralf Schönberger et al.: Repertorium edierter Texte des Mittelalters aus dem Bereich der Philosophie und angrenzender Gebiete

```
   B1700-150/65
ED: Chantelou, Claudius.
IN: Bibliotheca Patrum ascetica sive selecta veterum
Patrum de christiana ac religiosa perfectione
opuscula, I, Paris 1662.
AT: Panereticon, pars prima, Sermones de tempore et
de sanctis complectens
Bernardus Claraevallensis: Vita s. Malachiae.
⇒ B1700-170/20
```

- Daten nicht präpariert: Sinn **machmal nur zu erraten**.
- Teils **uneindeutige Feldabgrenzungen**
- **Idiosynkratischer Variantenreichtum** (insbesondere bei Literatur- Orts- und Zeitangaben)
- **Fehler**, z.B. Rechtschreibfehler in Schlüsselwörtern!

⟶ machbar, aber **Arbeitsaufwand schwer abschätzbar**

Dr. Eckhart Arnold, Bayerische Akademie der Wissenschaften, München
www.eckhartarnold.de
Dr. Albert Weber, Leibnitz-Institut für Ost- und Südosteuropa Forschung, Regensburg
leibniz-ios.de/personen/details/albert-weber
Ingo Frank M.A., Leibnitz-Institut für Ost- und Südosteuropa Forschung, Regensburg
leibniz-ios.de/personen/details/ingo-frank