Zwischen Forschungsdaten und Buchveröffentlichung

Die Plattform Open Text Collections

Sebastian Nordhoff, Mandana Seyfeddinipur & Christian Döhler

Berlin-Brandenburgische Akademie der Wissenschaften

2023-09-28 Open-Access-Tage-Berlin

Sprachwissenschaft

- Verföffentlichungsformen in der Sprachwissenschaft
  - Artikel
  - Bücher
  - Wörterbücher
  - Textsammlungen
  - (Multimedia)

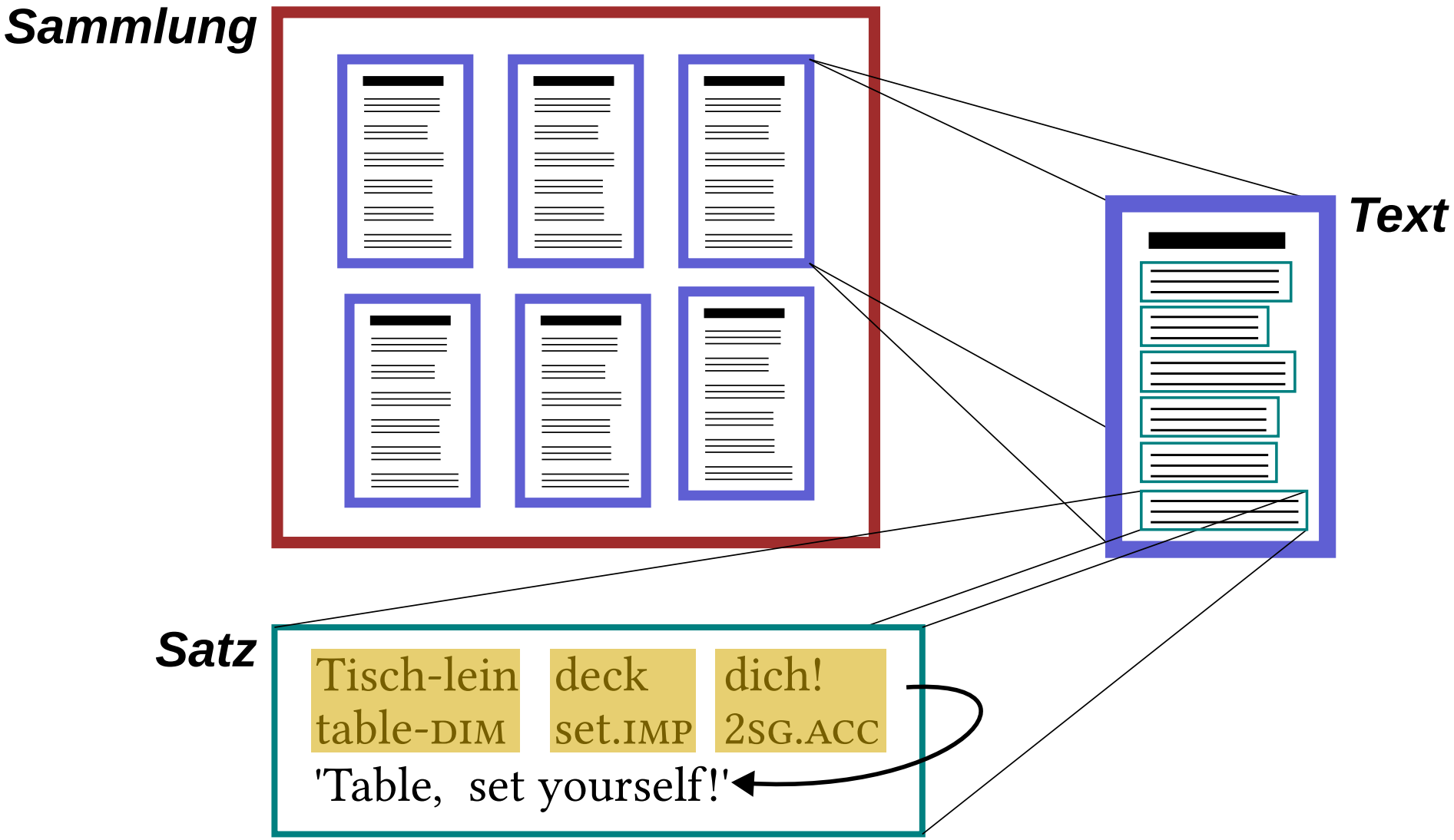
Open Access in den Sprachwissenschaften

- Generell gutes Bewusstsein für OA in der Disziplin
- Viele Sprachen werden in ökonomisch weniger wohlhabenden Weltregionen gesprochen
- Zugang ist für Sprecher und Forscher dort besonders wichtig.
- Artikel
  - `oaling.wordpress.com`
  - Glossa (Diamond OA)
  - Zeitschrift für Sprachwissenschaft (Diamond OA)
- Bücher
  - Language Science Press (Diamond OA)
- Wörterbücher
  - Dictionaria
- Textsammlungen
  - ???

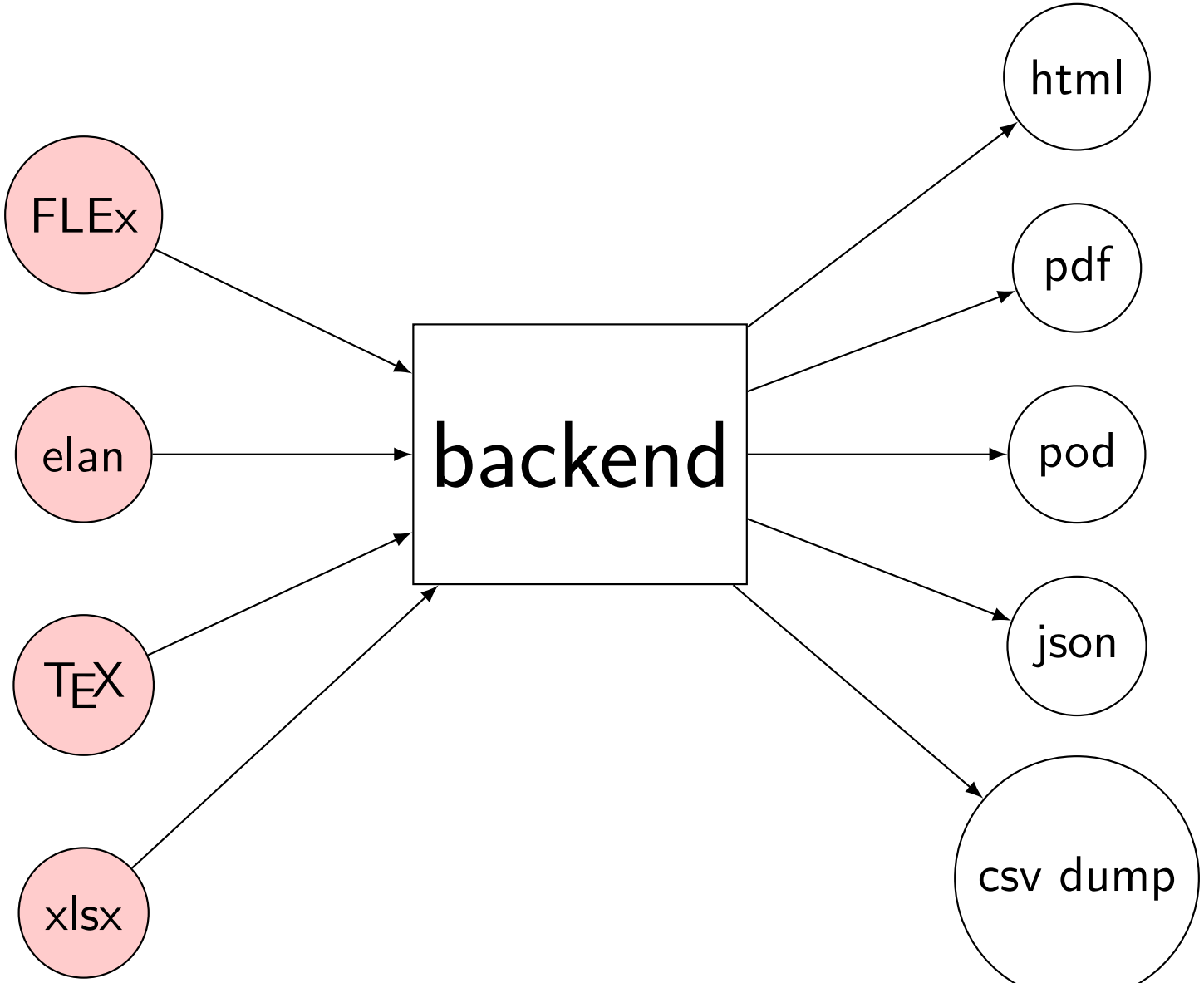
Das Projekt *Open Text Collections*

- Start letzte Woche
- DFG-Föderung 3 Jahre
- Danach konsortiale Finanzierung via Language Science Press
- innovativ, selektiv, international
- 5 Regional Boards:
  - Africa, Eurasia, Caucasus, Papunesia, South America

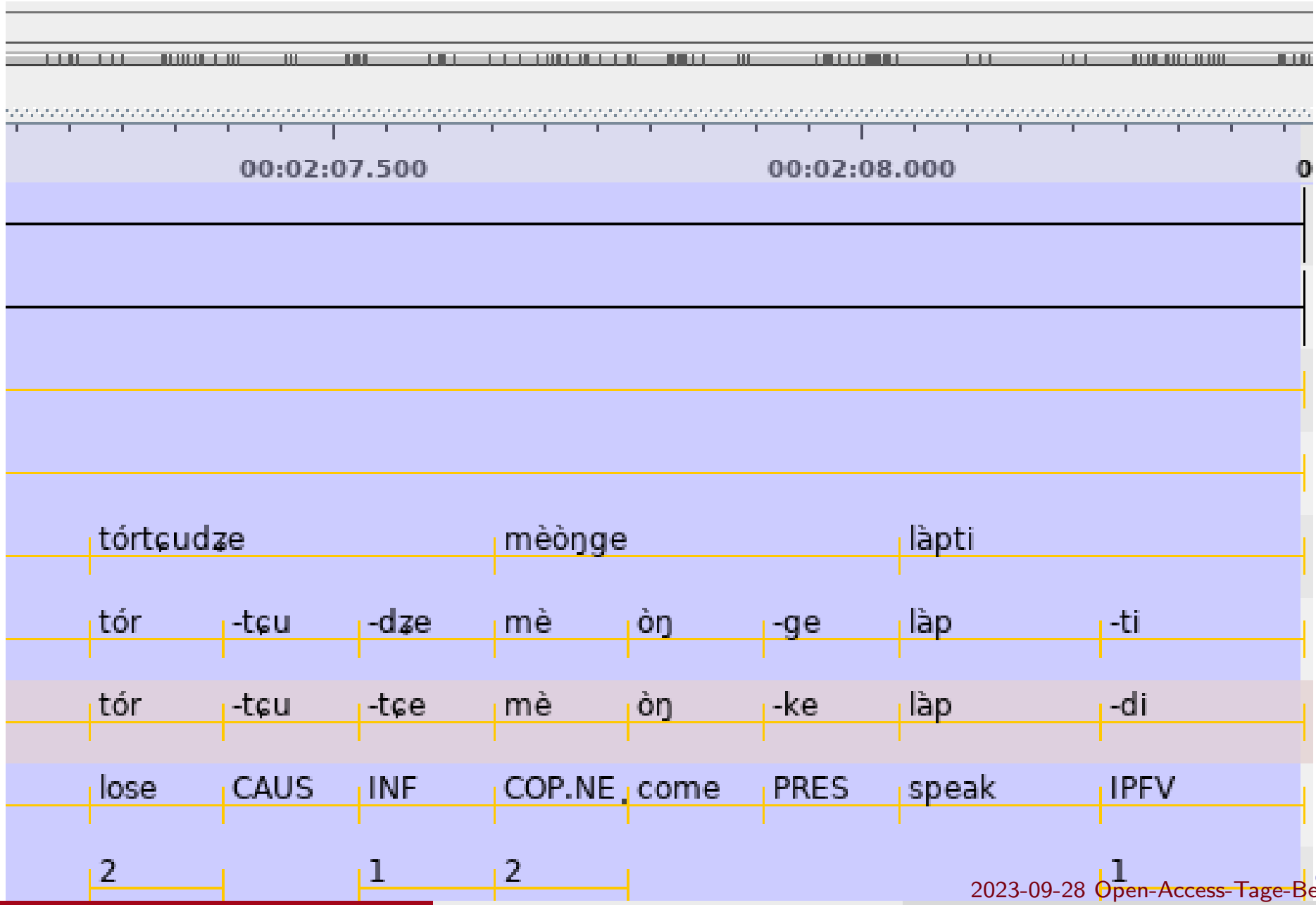
Was ist eine Textsammlung?



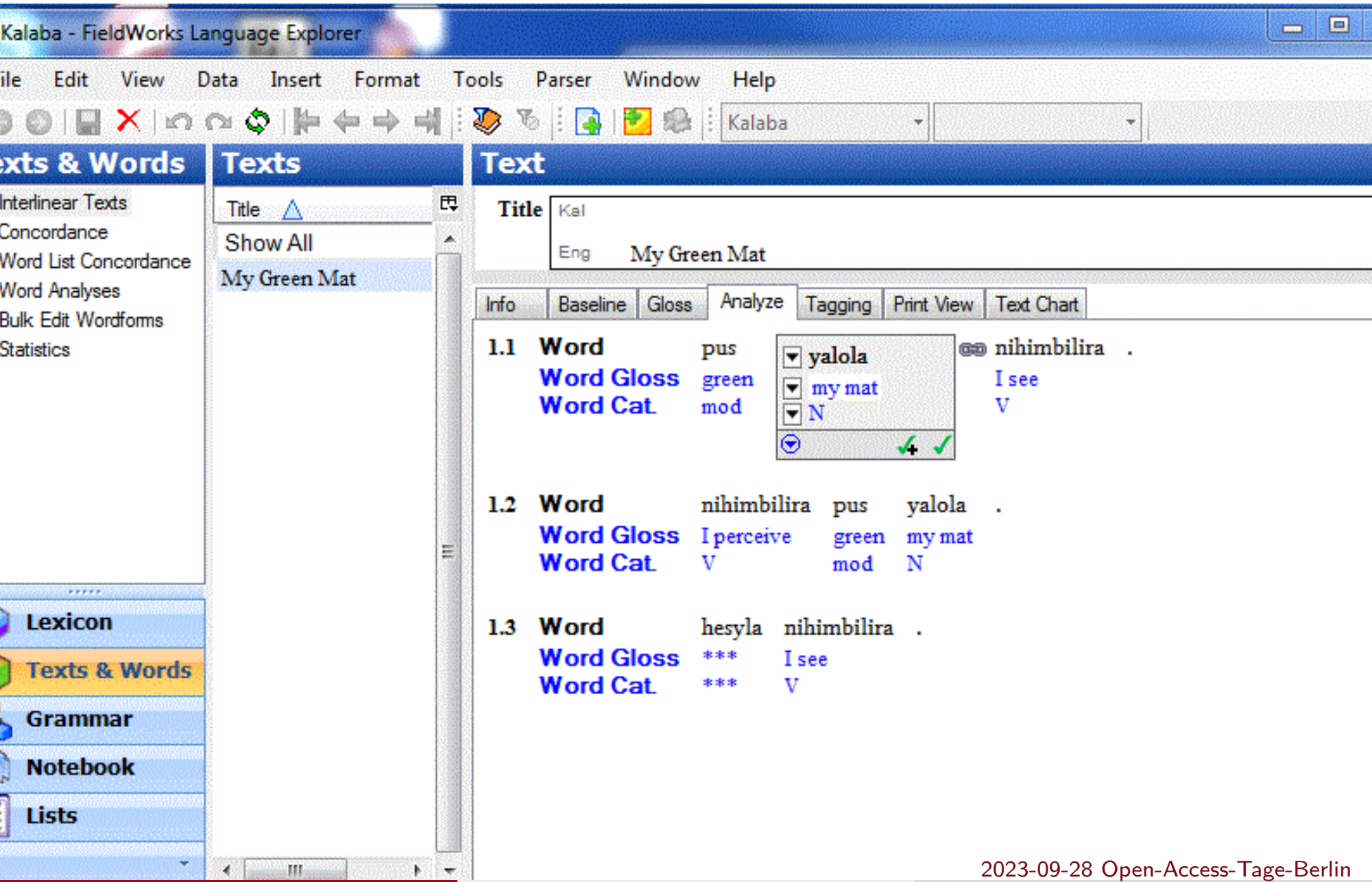
Datenverarbeitung



Inputformate: ELAN



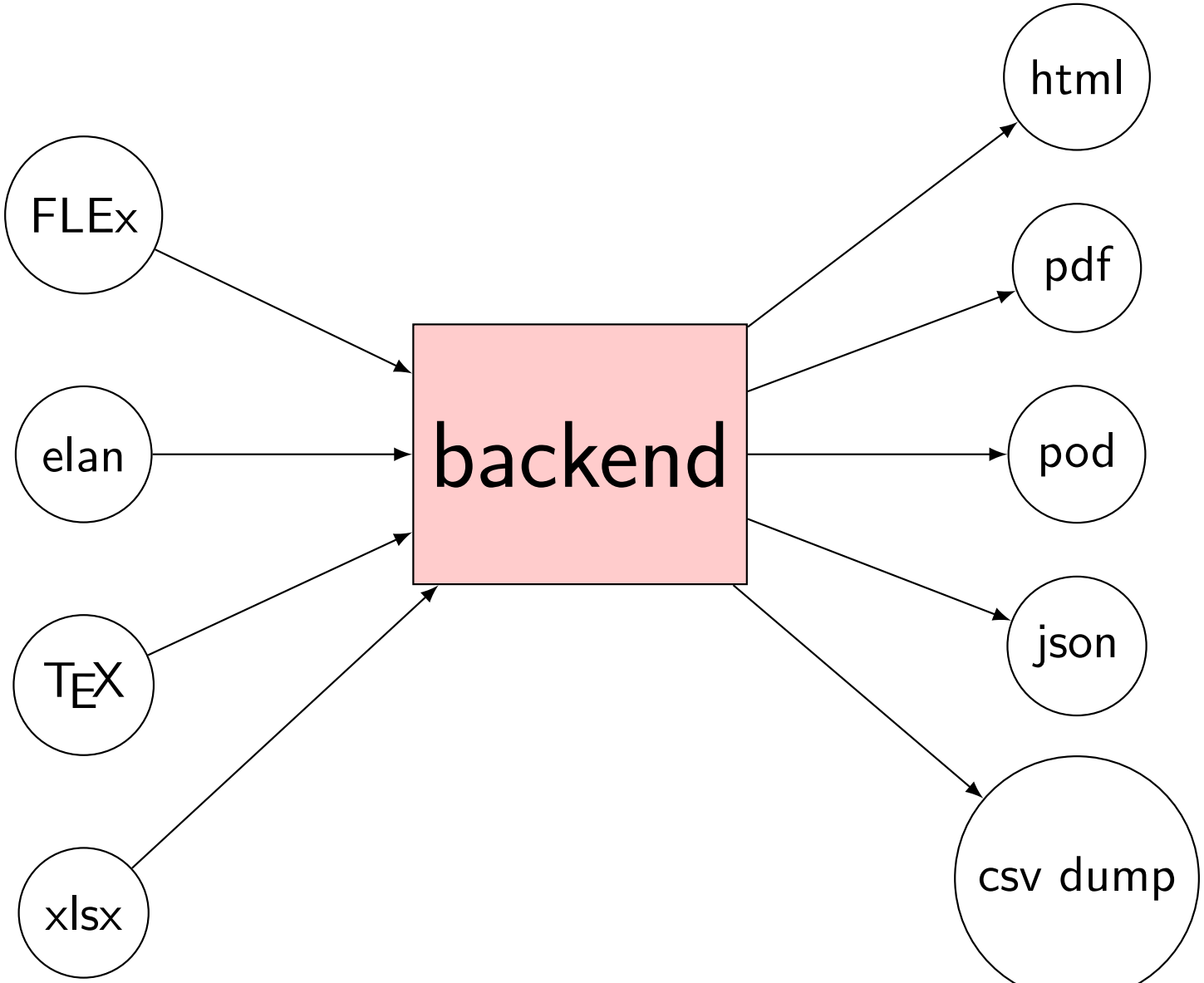
Inputformate: FLEx



Inputformate: tex

```
\ex a bɛth thɔk lɛ ka bɛrɛ\\
\gll a    bɛth  thɔk  lɛ    ka    bɛrɛ\\
\textsc{1sg}  cut  tree  \textsc{def}  with  axe\\
\glt `I cut the tree with an axe.' (P67 K:2)
```

Backendformat

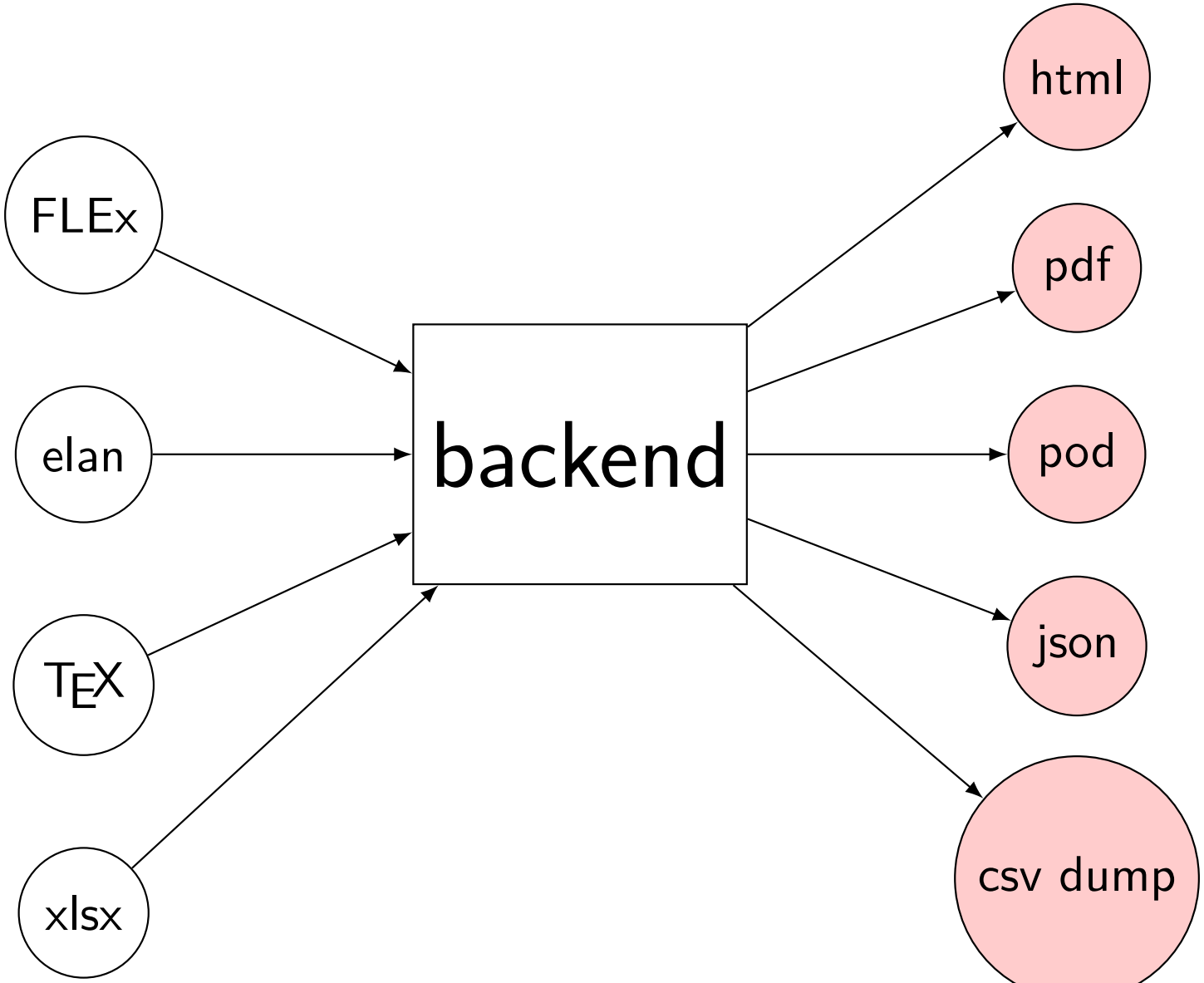


Backendformat

- Cross-Linguistic Data Format
  - `https://cldf.clld.org`
- CSV-basiert
- sich abzeichnender Quasistandard in der Sprachypology

	A	B	C	D	E
1	aɪ↦bɛtʰ↦tʰɔk↦lɛ↦kaɪ↦bɛrɛ	1SG↦cut↦tree↦DEF↦with↦axe	'I cut the tree with an axe.'	P67 K:2	sher1258
2					
3					
4					
5					

Outputformate



Outputformate: HTML

day passed

Clear search

142 results
found in 17ms

Page size

5 | 10 | 25

Sorting

Filters

Language

iso6393

- aaz 3
- ayl 1
- cbw 1

Mijał dzień za dniem.

pass.3SG.PST day.SG.NOM after day.SG.INS

Day after day passed. ()

Citation: docekal:wagiel:ed:21

Outputformate: PDF

(11) *Yup m'é' sój d'öb k'ët yő' mah, tíhàn tih ídíh.*
yup m'éʔ sój d'öb-k'ët-yőʔ=mah, tíh-àn tih ʔíd-íh.
that carajuru brilliant descend-stand-SEQ=REP 3SG-OBJ 3SG speak-DECL
'Standing there looking down, brilliant with *carajuru*, it's said, he spoke to her.'
'Ficando lá olhando para ela, brilhante com carajuru, dizem, ele falou para ela.'

(12) *Yít páh, "Hồp ámàn ãh käk w'öb péét, d'ő'őy ám páh?" nóóy mah.*
yít páh, hồp ʔám-àn ʔãh käk-w'öb-pé-ét, d'őʔ-őy
thus PROX.CNTR fish 2SG-OBL 1SG pull-set-go.upstream-DECL take-DYNM
ʔám páh? nó-óy=mah.
2SG PROX.CNTR say-DYNM=REP
'And then, "Where I went upstream catching fish and setting them out for you; have you taken them?" he said, it's said.'
'Aí, "Lá onde fui rio acima, pescando e deixando peixe, você pegou?" ele falou, dizem.'

Outputformate: JSON-LD


```
https://imtvault.org/content/static/ligt-0.2.ttl#hasWords:
  0:
    @id:                                                    "https://imtvault.org/langsci316-6e388cec49_wt"
    @type:
      0:                                                    "https://purl.org/liodi/ligt#WordTier"
    https://imtvault.org/content/static/ligt-0.2.ttl#item:
      0:
        @id:                                                "_:langsci316-6e388cec49_0"
        https://imtvault.org/content/static/ligt-0.2.ttl#Word:
          0:
            @language:                                      "und"
            @value:                                         "Mijał"
          1:
            @language:                                      "en-x-lgr"
            @value:                                         "pass.3SG.PST"
        https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:
          0:
            @id:                                            "_:langsci316-6e388cec49_1"
      1:
        @id:                                                "_:langsci316-6e388cec49_1"
        https://imtvault.org/content/static/ligt-0.2.ttl#Word:
          0:
            @language:                                      "und"
            @value:                                         "dzień"
          1:
            @language:                                      "en-x-lgr"
            @value:                                         "day.SG.NOM"
        https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:
          0:
            @id:                                            "_:langsci316-6e388cec49_2"
      2:
        @id:                                                "_:langsci316-6e388cec49_2"
        https://imtvault.org/content/static/ligt-0.2.ttl#Word:
          0:
            @language:                                      "und"
```

Textsammlungen als Hybrid zwischen Buch und Forschungsdaten

- Aus der Makroperspektive sind Textsammlungen Bücher
  - Autor*in
  - Titel
  - Auflage
  - ISBN
  - Einband, ...
- Aus der Mikroperspektive handelt es sich um strukturierte Forschungsdaten
  - standardisierte Darstellung
  - Beziehung zwischen Elementen
  - Metadaten
- Ist die Bibliothek zuständig weil OA oder das Rechenzentrum weil Forschungsdatenmanagement?
- Was sind mögliche Finanzierungsmodelle?

Nachnutzung

- automatisch generierte/extrahierte Wörterbücher
- typologische Forschung
- maschinelles Lernen („künstliche Intelligenz“) für kleinere Sprachen

Rechtliche und ethische Aspekte

- OA bedingt den Export angelsächsischer/westlicher Konzepte
  - Eigentum
  - geistiges Eigentum
  - Rechtsstreit
  - verbriefte Rechteabtretung
  - regional große Unterschiede
    - in Australien „gehören“ Sprachen den Sprechern
    - in Nordamerika große Vorbehalte gegen nicht-indigene Forschende
    - in den anderen Regionen weniger problematisch

Publikationsanfragen Stand heute

Danke

https://www.opentextcollections.org
https://fedihum.org/@otc
https://twitter.com/OpenTextColl