Marc Kupietz und Nils Diewald

SCHNITTSTELLEN ZUR NUTZUNG DER KORPUSANALYSEPLATTFORM KORAP

KobRA-Abschlusstagung, 30.10.2015

Bundesministerium
für Bildung
und Forschung

ÜBERBLICK

1. IDS im KobRA-Projekt
2. KorAP
3. Schnittstellen
4. Protokoll
5. Resümee

1. IDS IM KOBRA-PROJEKT

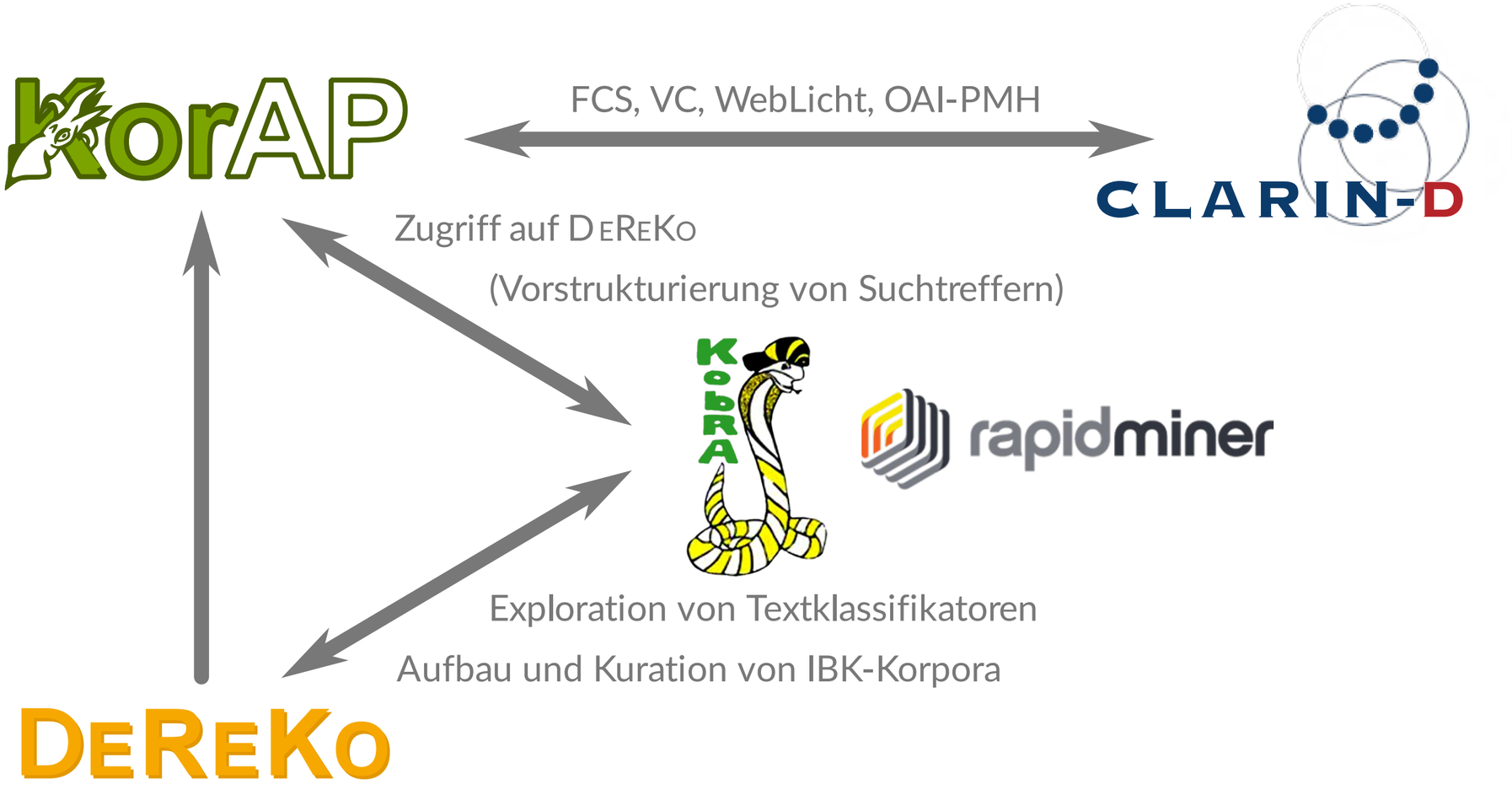
ROLLEN DES IDS IM KOBRA-PROJEKT

1. Bereitstellung von Korpora:
   - Deutsches Referenzkorpus DeReKo
2. Integration von KobRA-Entwicklungen in dauerhafte Infrastrukturen:
   - KorAP
   - CLARIN

DEREKO

- *Ur-Stichprobe* des gegenwärtigen Schriftsprachegebrauchs
- wird seit 1964 aufgebaut
- dient als empirische Grundlage für die germanistische Sprachwissenschaft
- umfasst z. Zt. 28 Milliarden Wörter
- wird permanent erweitert
  - Wachstumsrate ca. 1,7 Milliarden Wörter / Jahr
- Nicht zuletzt auch dank/aufgrund der KobRA-Kooperation

ZUSAMMENSPIEL



2. KORAP

KORAP: HINTERGRUND

Korpusanalyseplattform der nächsten Generation

- Nachfolgesystem zu COSMAS
  - COSMAS seit 1992 im Einsatz
  - > 38.000 registrierte Nutzer
- KorAP seit 2011 am IDS in Entwicklung

KORAP: HERAUSFORDERUNGEN

- Skalierbarkeit
  - beliebig große Datenmengen
  - beliebige Annotationsschichten
  - mehrere Anfragesprachen
- genaue Abbildbarkeit von Lizenzen
  - optimales Ausschöpfen von Nutzungsrechten
- Tragfähigkeit für die nächsten 15-20 Jahre
  - Wartbarkeit
  - Erweiterbarkeit
    - auch durch externe Entwicklungen

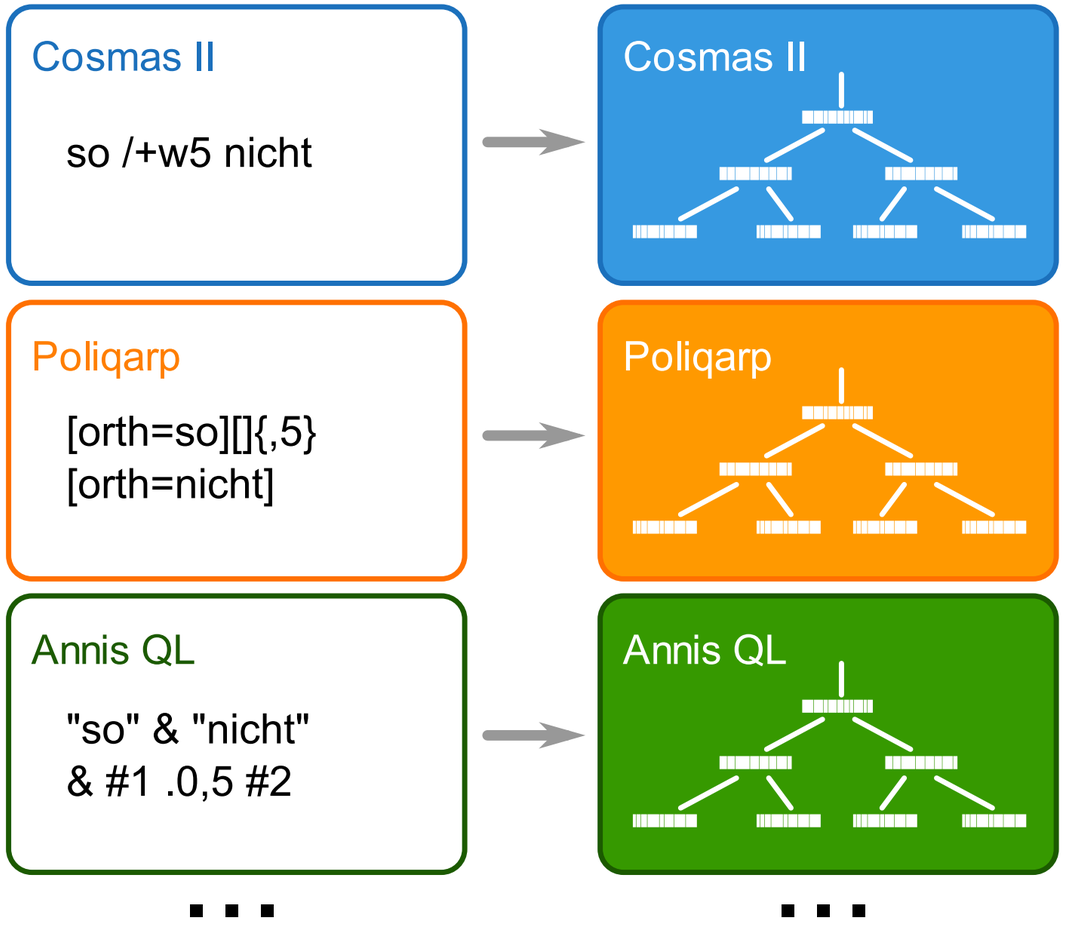
LÖSUNGSANSÄTZE

- horizontale Skalierbarkeit
  - wenn das System zu langsam ist, wird ein weiterer Rechner hinzugefügt
- Komponenten
  - möglichst einfach
  - austauschbar
  - erweiterbar
  - durch neue ergänzbar
- Schnittstellen
  - auch durch externe Komponenten nutzbar

KORAP: ARCHITEKTUR

https://github.com/KorAP/

KORAL: QUERY-SERIALIZER



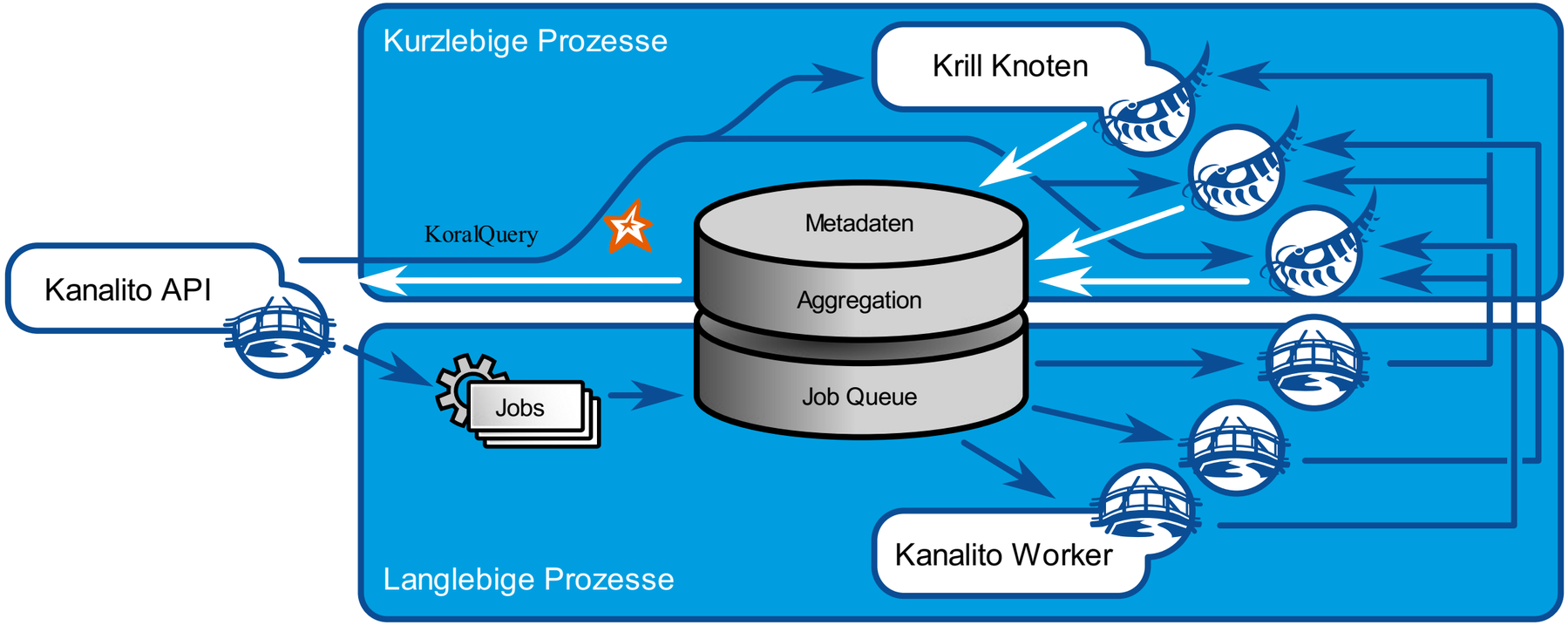

```
{
  "context" : "http://ids-mannheim.de/ns/
               KorAP/json-ld/v0.1/context.jsonld",
  "query" : {
    "@type" : "korap:group",
    "operation" : "operation:sequence",
    "inOrder" : true,
    "distances" : [{
      "@type" : "korap:distance",
      "key" : "w",
      "min" : 0,
      "max" : 5
    }],
    "operands" : [{
      "@type" : "korap:token",
      "wrap" : {
        "@type" : "korap:term",
        "key" : "so",
        "layer" : "orth",
        "match" : "match:eq"
      }
    },{
      "@type" : "korap:token",
      "wrap" : {
        "@type" : "korap:term",
        "key" : "nicht",
        "layer" : "orth",
        "match" : "match:eq"
      }
    }]
  }
}
```

Koral

KRILL: SUCH- UND ANALYSEBACKEND

- Anfragen nur in KoralQuery
- Unterstützt ...
  - Volltextsuche, Reguläre Ausdrücke, Annotationssuche, Positionssuche, Distanzsuche …
  - Kombination und Verschachtelung von Suchoperationen
  - mehrere Annotationsebenen

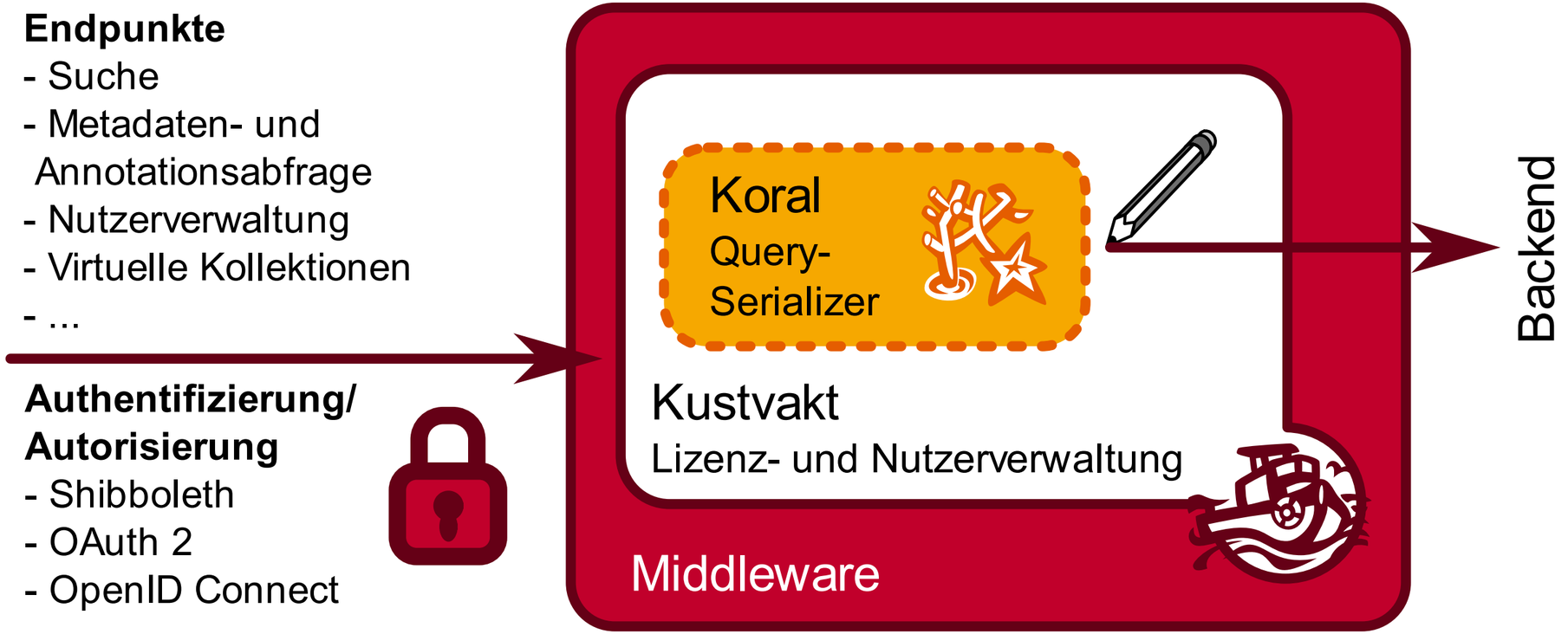
KRILL: VERTEILTES SYSTEM



KUSTVAKT: LIZENZ- UND NUTZERVERWALTUNG

Endpunkte
- Suche
- Metadaten- und Annotationsabfrage
- Nutzerverwaltung
- Virtuelle Kollektionen
- ...

Authentifizierung/ Autorisierung
- Shibboleth
- OAuth 2
- OpenID Connect



KALAMAR: VISUELLE NUTZERSCHNITTSTELLE

3. SCHNITTSTELLEN

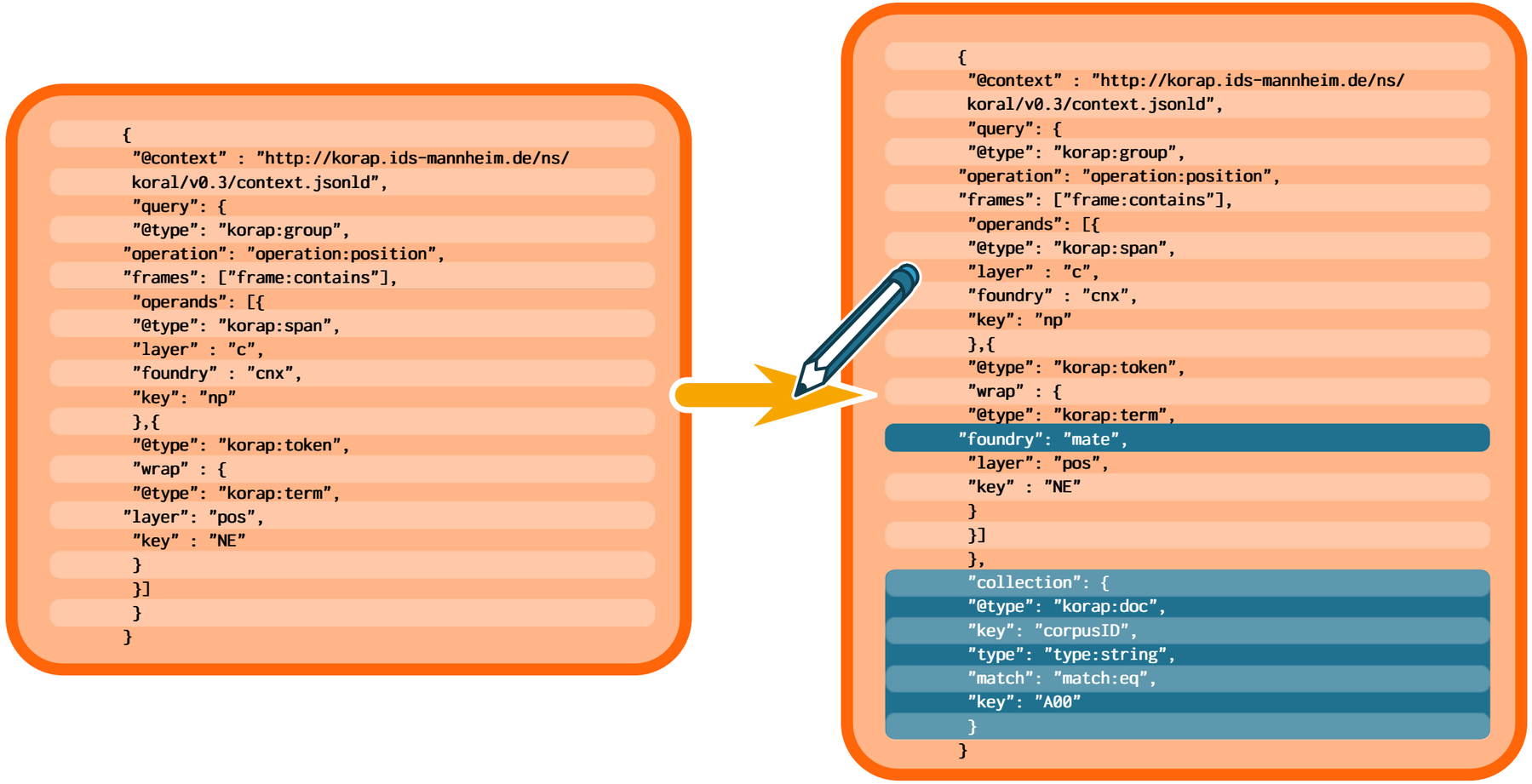
DEMO

4. PROTOKOLL

KORALQUERY: GENERAL CORPUS QUERY PROTOCOL

- Zentrales Kommunikationsprotokoll aller Komponenten
- JSON-LD-Basis (leicht erweiterbar)
- Unabhängig von …
  - Forschungsfragen
  - Daten
- Dokumentanfragen zur Erstellung virtueller Kollektionen
- Vorkommensanfragen zur Suche in Texten
- Dokumentiert: https://korap.github.io/Koral/

KORALQUERY: REWRITES



```
{
 "@context" : "http://korap.ids-mannheim.de/ns/
 koral/v0.3/context.jsonld",
 "query": {
 "@type": "korap:group",
"operation": "operation:position",
"frames": ["frame:contains"],
 "operands": [{
 "@type": "korap:span",
 "layer" : "c",
 "foundry" : "cnx",
 "key": "np"
 },{
 "@type": "korap:token",
 "wrap" : {
 "@type": "korap:term",
"layer": "pos",
 "key" : "NE"
 }
 }]
 }
}
```

```
{
 "@context" : "http://korap.ids-mannheim.de/ns/
 koral/v0.3/context.jsonld",
 "query": {
 "@type": "korap:group",
"operation": "operation:position",
"frames": ["frame:contains"],
 "operands": [{
 "@type": "korap:span",
 "layer" : "c",
 "foundry" : "cnx",
 "key": "np"
 },{
 "@type": "korap:token",
 "wrap" : {
 "@type": "korap:term",
"foundry": "mate",
 "layer": "pos",
 "key" : "NE"
 }
 }]
 },
 "collection": {
 "@type": "korap:doc",
 "key": "corpusID",
 "type": "type:string",
 "match": "match:eq",
 "key": "A00"
 }
}
```

5. RESÜMEE

AKTUELLER STAND

- seit Februar 2014 im IDS-internen Testbetrieb
- öffentlicher Parallelbetrieb zu COSMAS II
  - geplant ab März 2016
- noch nicht veröffentlicht:
  - Kustvakt (Lizenz- und Nutzerverwaltung)
  - Kanalito (“zoo-keeper”)
- http://korap.ids-mannheim.de

SCHNITTSTELLEN

- Anbindung an KobRA bzw. RapidMiner
  - Schnittstelle implementiert
  - Autoriserung zum Zugriff auf geschützte Korpora noch in Entwicklung
- Einbindung in CLARIN-Federated-Content-Search (FCS)
  - Schnittstelle und Anfragesprache (CQL) implementiert
  - Referenzimplementation zu Advanced-FCS mit CQP/Poliqarp-Anfragesprache geplant
- Anbindung an CLARIN-Virtual-Collection-Registry
  - in Vorbereitung
- WebLicht-Schnittstelle noch über COSMAS II realisiert

ZUSAMMENFASSUNG

- KorAP-Ziele
  - Skalierbarkeit
  - Nachhaltiges
  - Erweiterbarkeit
- Lösungsansätze
  - horizontale Skalierbarkeit
  - einfache, austauschbare Komponenten
  - Schnittstellen zur Anbindung externer Entwicklungen
  - Einladung am Open-Source-Projekt KorAP mitzuarbeiten

6. REFERENZEN

REFERENZEN I

Bański, Piotr/Bingel, Joachim/Diewald, Nils/Frick, Elena/Hanl, Michael/Kupietz, Marc/Pęzik, Piotr/Schnober, Carsten/Witt, Andreas (2013):

KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt/Uszkoreit, Hans (Hrsg.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. S. 586-587 - Poznań: Fundacja Uniwersytetu im. A., 2013. →IDS-Publikationsserver

Bingel, Joachim/Diewald, Nils (2015):

KoralQuery – a General Corpus Query Protocol. In: Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, Vilnius, Lithuania, May 11-13, 2015, pp. 1-5.

REFERENZEN II

Bodmer, Franck (1996):

Aspekte der Abfragekomponente von COSMAS-II. In: LDV-INFO 8. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung. Hrsg. vom Institut für deutsche Sprache. Redaktion: Irmtraud Jüttner, Robert Neumann. S. 112-122 - Mannheim: 1996. (LDV-INFO 8)

Kupietz, Marc/Lüngen, Harald (2014):

Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (Hrsg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). S. 2378-2385 - European Language Resources Association (ELRA), 2014. →Text