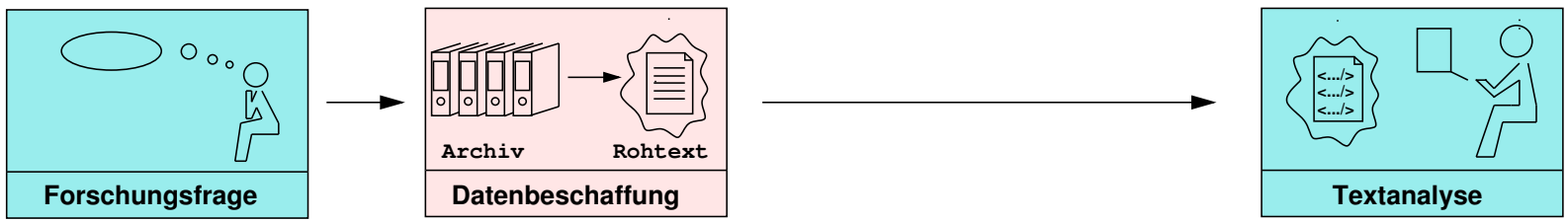
Kurze Einführung zur Demonstration der e-Identity-Werkzeuge

Ulrich Heid

Universität Hildesheim
IwiSt - Institut für Informationswissenschaft und Sprachtechnologie
Universitätsplatz 1, 31141 Hildesheim

Politikwissenschaftliche Textanalyse

Analyse von Pressetexten als Beispiel

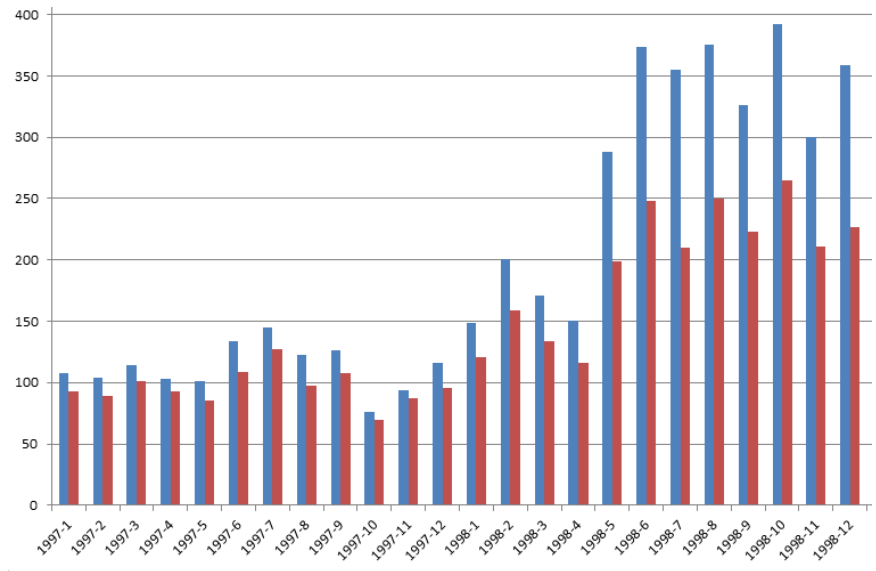


- Ausgangspunkt: politikwissenschaftliche Forschungsfrage
- Datenbeschaffung, z.B. aus Zeitungsarchiven
- Ziele der Textanalyse:
  - Identifikation und Klassifikation relevanter Passagen
  - Beschreibung von Entwicklungen in der Zeit
  - usw. ...

Politikwissenschaftliche Textanalyse

Typen von Fragestellungen: zwei Beispiele

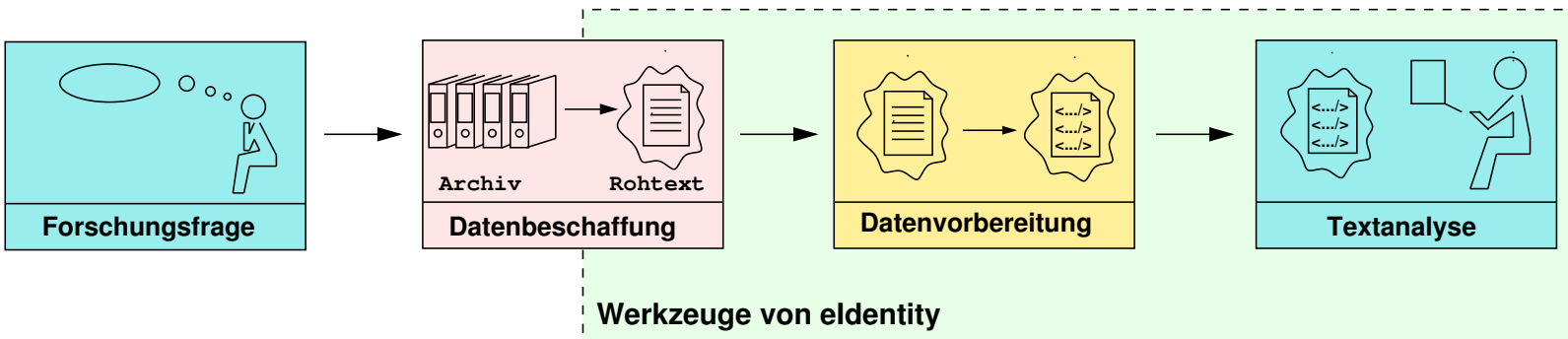
- Relevante Texte zu einer Fragestellung finden, z.B. “Demokratieförderung im Staat X”
- Zeitverlaufsanalysen: Medienaufmerksamkeit
  - Themen der Landtags-debatten
  - “Terror(ismus)” als Begründung für Gesetzesvorlagen
  - “Sprechweisen” bestimmter Akteure, z.B. zum Thema “Nachhaltigkeit”



Computerlinguistische Unterstützung

Die e-Identity-Werkzeuge

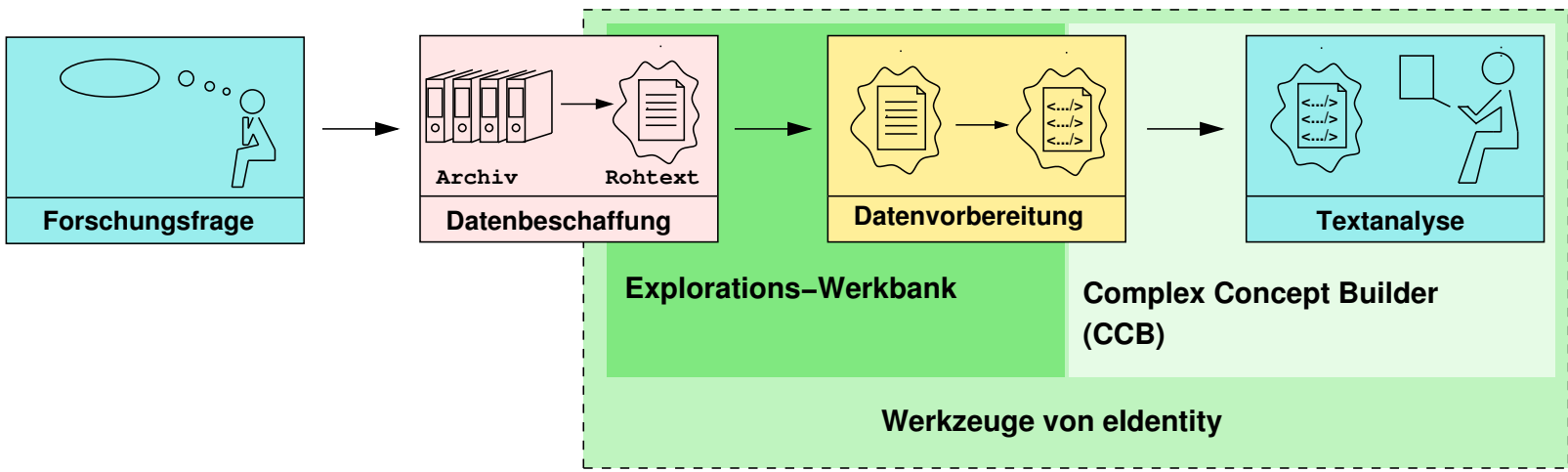
- Vom Rohtext zum Input für die Textanalyse:
  Die e-Identity Explorationswerkbank Kliche et al. 2014
- Unterstützung bei der Textanalyse:
  Der e-Identity Complex Concept Builder, CCB Blessing et al. 2013
  - Klassifikation von Texten: Bericht ↔ Kommentar
  - Sentiment-Analyse: Meinungen in Texten
  - Interaktive Suche in Texten
  - Interaktive Kodierung von Textpassagen



Computerlinguistische Unterstützung

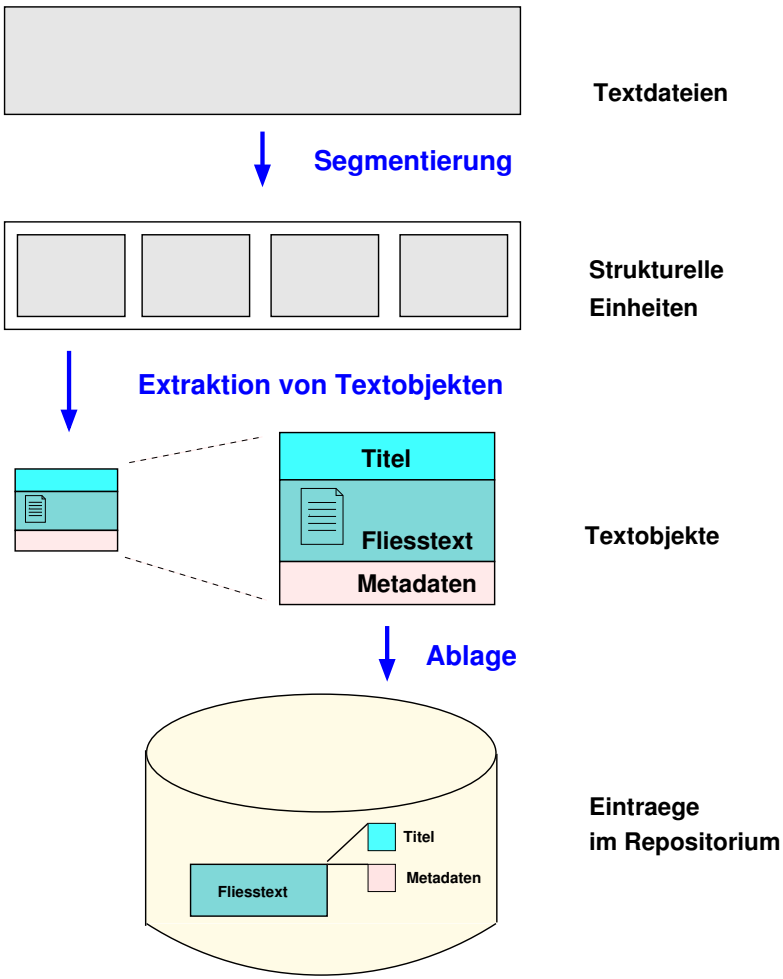
Die e-Identity-Werkzeuge: Demonstrationen

- Status der Werkzeuge: Prototypen
  - Explorationswerkbank: Forschungsprototyp: Funktionen – erste Version der Benutzerschnittstelle
  - CCB: Bausteine mit unterschiedlichem Reifegrad:
    * Prototypen, zum Teil in konkreter Anwendung erprobt
    * Noch keine umfassende Benutzerschnittstelle



Vom Textarchiv zum Korpus

Funktionen der e-Identity-Explorationswerkbank – Beispiele

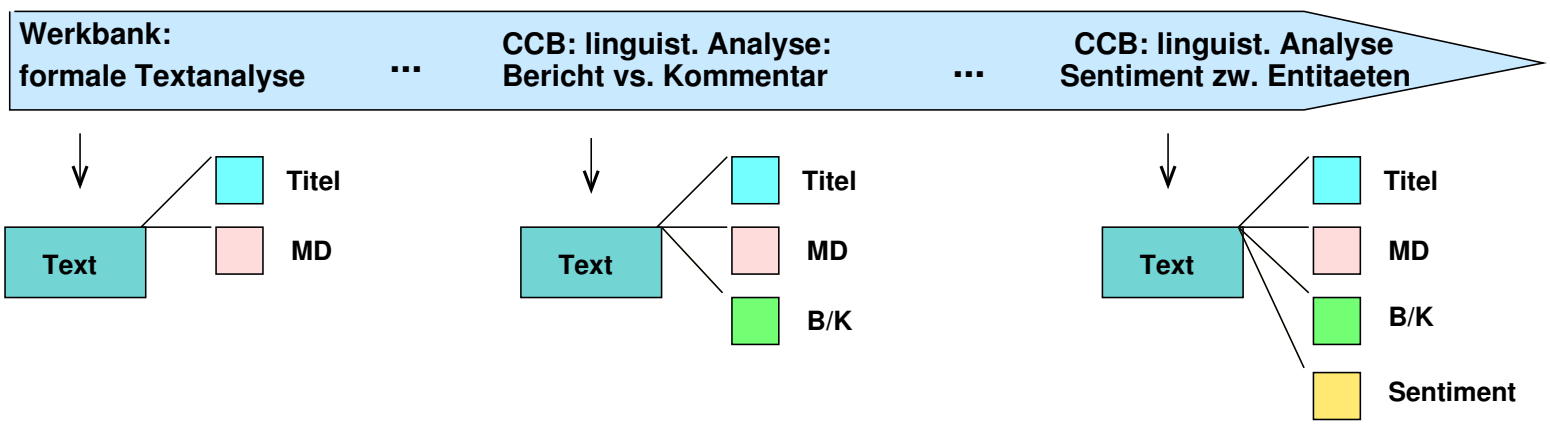


- Daten aus Textdateien
- Relevante Einheiten
- Bausteine der Einheiten: Textobjekte
- Sammlung alles relevanten Materials in Repositorium

Vom Textarchiv zum Korpus

Grundprinzipien der Arbeit mit den Texten (1/2)

- Keine Daten gehen verloren:
  Alle Metadaten und Analysen werden mit den Texten abgelegt
- Computerlinguistische Werkzeuge werden nacheinander benutzt:
  Ablaufketten ('Pipeline')
- Alle Arbeitsschritte werden im Repositorium dokumentiert:
  Nachvollziehbarkeit – Zugriff auf Zwischenergebnisse



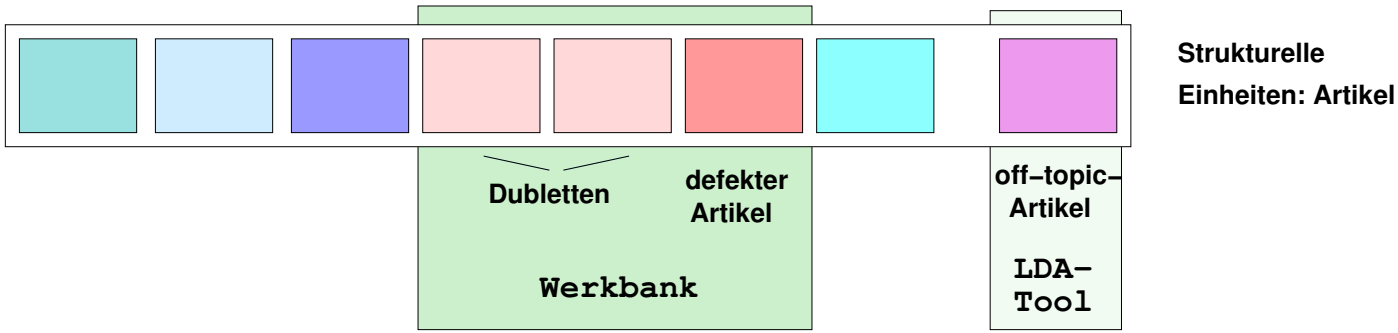
Vom Textarchiv zum Korpus

Grundprinzipien der Arbeit mit den Texten (2/2)

- Arbeitsschritte können sein:
  - automatisch: von Werkzeugen realisiert
  - interaktiv: vom Wissenschaftler manuell realisiert,
    z.B. Kodierung anhand Codebuch
- Ergebnisse manueller Kodierung
  werden mit Ergebnissen automatischer Analysen zusammengeführt
- Langfristig: Ablaufketten werden vom Fachwissenschaftler bestimmt:
  Reihenfolge nicht vorab festgelegt,
  relevante Werkzeuge kommen zum Einsatz, wenn Bedarf ist

Vom Textarchiv zum Korpus

Beispielfall Sample-Bereinigung



- Ziel: Identifikation und Ausschluss irrelevanter Daten
  - Formale Probleme:
    Dubletten – unvollständige/leere Artikel
  - Inhaltliche Probleme:
    Artikel nicht relevant für Forschungsfrage

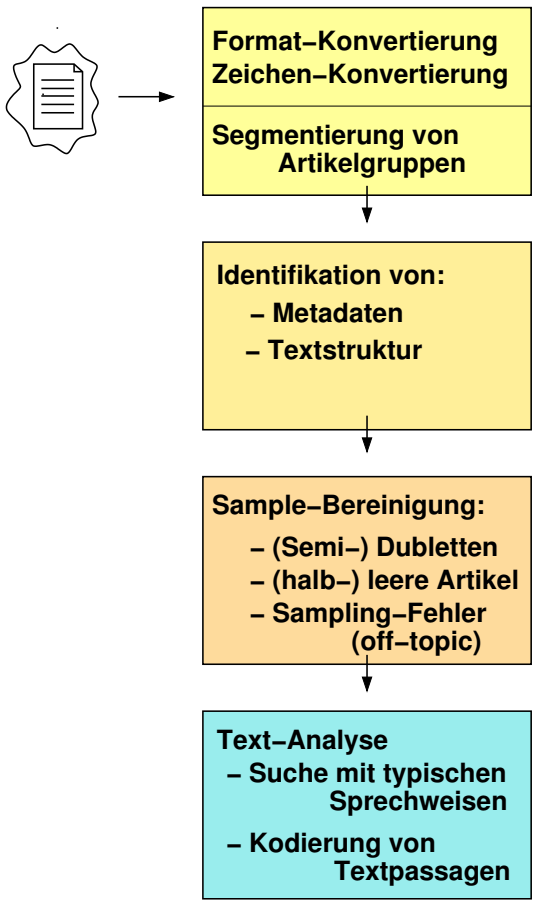
Fußballbericht

Buchkritik

- Ergebnis: statistische Aussagen auf relevantem Material

Funktionen der demonstrierten Werkzeuge

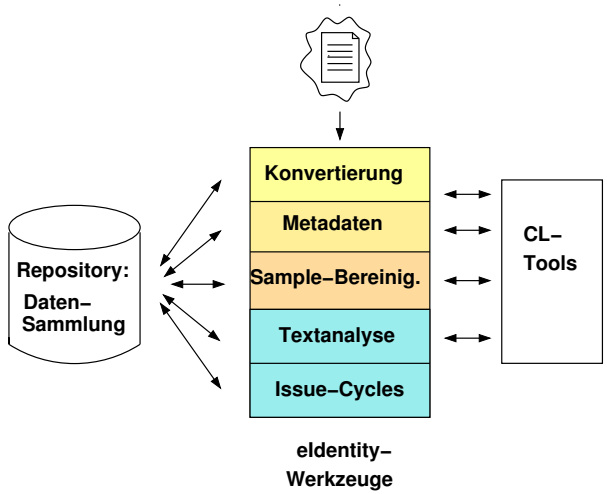
Zusammenfassender Überblick



- Text-Konvertierung
- Segmentierung in Artikel
- Metadaten-Extraktion
- Identifikation der Textstruktur
- Sample-Bereinigung: formal & inhaltlich
- Kodierumgebung: Codebuch-Verwaltung, Ergebnisvisualisierung
- Extraktion/Analyse von Textstellen: Bericht vs. Kommentar, Sentiment

Gesamtarchitektur der Werkzeuge

Funktionen – Werkzeuge – Repositorium



Repositorium:

- Analyseschritte
- Ergebnisdaten
- Prozess-Metadaten

- Übersicht über die Verarbeitungsprozesse:
  Der Nutzer soll immer wissen (können),
  welche Schritte angewandt wurden:
  automatisch/manuell

Mehrwert für die politikwissenschaftliche Textanalyse

Aktueller Stand: Demonstration

- Textanalyse: tief und breit zugleich:
  - Tiefe linguistische Analyse: z.B.
    - Unterscheidung: Bericht $\leftrightarrow$ Kommentar
    - Identifikation von Sentiment
  - Anwendung auf große Datenmengen:
    in e-Identity zur Zeit: total $>$ 800.000 Zeitungsartikel
- Arbeit mit bereinigten Samples:
  - Identifikation und Ausschluss von irrelevanten Daten
  - Samples werden damit statistisch aussagekräftig
- Unterstützung bei der manuellen Kodierung
  - Kodierungshandbücher: interaktiv
  - Ablaufkontrolle und Ergebnis-Visualisierung
  - Geplant: Nutzung von maschinellem Lernen

Mehrwert für die politikwissenschaftliche Textanalyse

Generelle Aspekte

- Werkzeuge sind generisch:
  - Unterstützung für verschiedene Arten von Forschungsdesigns
  - Nutzbar für DE, EN, FR – erweiterbar auf andere Sprachen
- Werkzeuge unterstützen Daten-Nachhaltigkeit
  - Be-/ Verarbeitungsschritte sind dokumentiert und nachvollziehbar
  - Dieselben Texte sind nutzbar für ganz unterschiedliche Fragestellungen
- Status der Werkzeuge:
  - dienstleistungsartige Werkzeuge:
    - Inhaltliche Bereinigung von Textmengen
    - Online-Kodierumgebung:
      Kodier-Handbücher und Management von Kodier-Aktivitäten
    - Visualisierung der Ergebnisse
  - Forschungsprototypen:
    - Formale Bereinigung der Texte: Explorationswerkbank
    - Tiefe linguistische Analyse