Universität Stuttgart
Universitätsbibliothek

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Karlsruher Institut für Technologie

ZAR
Zentrum für Angewandte Rechtswissenschaft

Umgang mit TDM-Korpora nach dem neuen § 60d Urheberrechtsgesetz im bibliothekarischen Alltag

XSample
access, reuse, advance

Sibylle Hermann, Felicitas Kleinkopf, Markus Gärtner

XSample

- Förderprogramm des Ministeriums für Wissenschaft, Forschung und Kunst BW-BigDIWA Wissenschaftliche Bibliotheken gestalten den Digitalen Wandel
- Projektlaufzeit 09/2019 - 08/2021 (verlängert bis 31.03.2022)
- Projektpartner:
  - Universität Stuttgart:
    - Universitätsbibliothek
    - Institut für Maschinelle Sprachverarbeitung (IMS)
  - KIT:
    - Zentrum für angewandte Rechtswissenschaften (ZAR)

Motivation

XSample
access, reuse, advance

- Die Forschung an Texten ist in den digitalen Geisteswissenschaften durch Defizite des deutschen Urheberrechts konfrontiert
- Beschränkung der Weitergabe, Archivierung und Zugriffsmöglichkeiten auf Forschungskorpora des Text und Data Mining (TDM)
- Konsequenz: kaum Forschung auf urheberrechtlich geschützten Texten

- Idee: Juristische Prüfung der technischen Umsetzung eines Auszugkonzeptes anhand von konkreten Use Cases.

Use Case 1: Unzuverlässiges Erzählen

Zeichnung von E. T. A. Hoffmann zu seinem Buch „Der Sandmann“

- Unzuverlässiges Erzählen ist ein in einigen literarischen Erzählungen auftretendes Phänomen, bei dem den Aussagen des Erzählers nicht zu trauen ist
- Faktenbezogene Unzuverlässigkeit, die mithilfe von automatischen und manuellen Annotationen von Primär- und Sekundärtexten untersucht wird
- Breites Spektrum unterschiedlicher literarischer Werke, sowohl urheberrechtlich geschützt als auch gemeinfrei und in diversen typischen Publikationsformaten

Use Case 2: Wissenschaftssprache

XSample
access, reuse, advance

- Wissenschaftssprache der geisteswissenschaftlichen Disziplinen Literaturwissenschaft, Linguistik und Philosophie werden in einem datengeleiteten Verfahren miteinander verglichen
- Datengrundlage: Korpus aus 45 Artikel Zeitschriftenartikeln pro Fach
- Die Daten werden mit linguistischen Annotationen zu Lemmata, Wortarten und syntaktischen Dependenzen versehen
- Auf Grundlage dieser Annotationen (und der einfachen Ebene der Wortformen) ist eine gezielte Auswahl von Textabschnitten möglich
- Zugriff auf die Kontexte, wie er durch das Auszugsprinzip geleistet werden kann, ist hierzu unbedingt notwendig

§ 60c UrhG

XSample
access, reuse, advance

- § 60c UrhG erlaubt, zu Zwecken der nicht-kommerziellen wissenschaftlichen Forschung, Auszüge von bis zu 15% bzw. vollständige Werke geringen Umfangs zu nutzen

- Idee: Anwendung des § 60c auf § 60d UrhG

§ 60d UrhG

- Ab 01.03.2018: TDM - Erlaubnis zugunsten der nicht-kommerziellen wissenschaftlichen Forschung
- Auf Basis dieses Rechtsrahmens war es möglich, zu Forschungszwecken TDM-Analysen vorzunehmen
- Forschungsergebnisse mussten allerdings **gelöscht** oder einer **archivierenden Institution übergeben** werden
- Seit 07.06.2021: **Löschung der Daten** auf Datenträgern der Forschenden selbst ist nun **nicht mehr erforderlich**, allerdings ist auch die **ausdrückliche Erlaubnis der Datenweitergabe an archivierende Institutionen weggefallen**
- Defizit: Rechtslage zum Umgang mit den Korpora nach Abschluss der Forschungsarbeiten

Technischer Ansatz

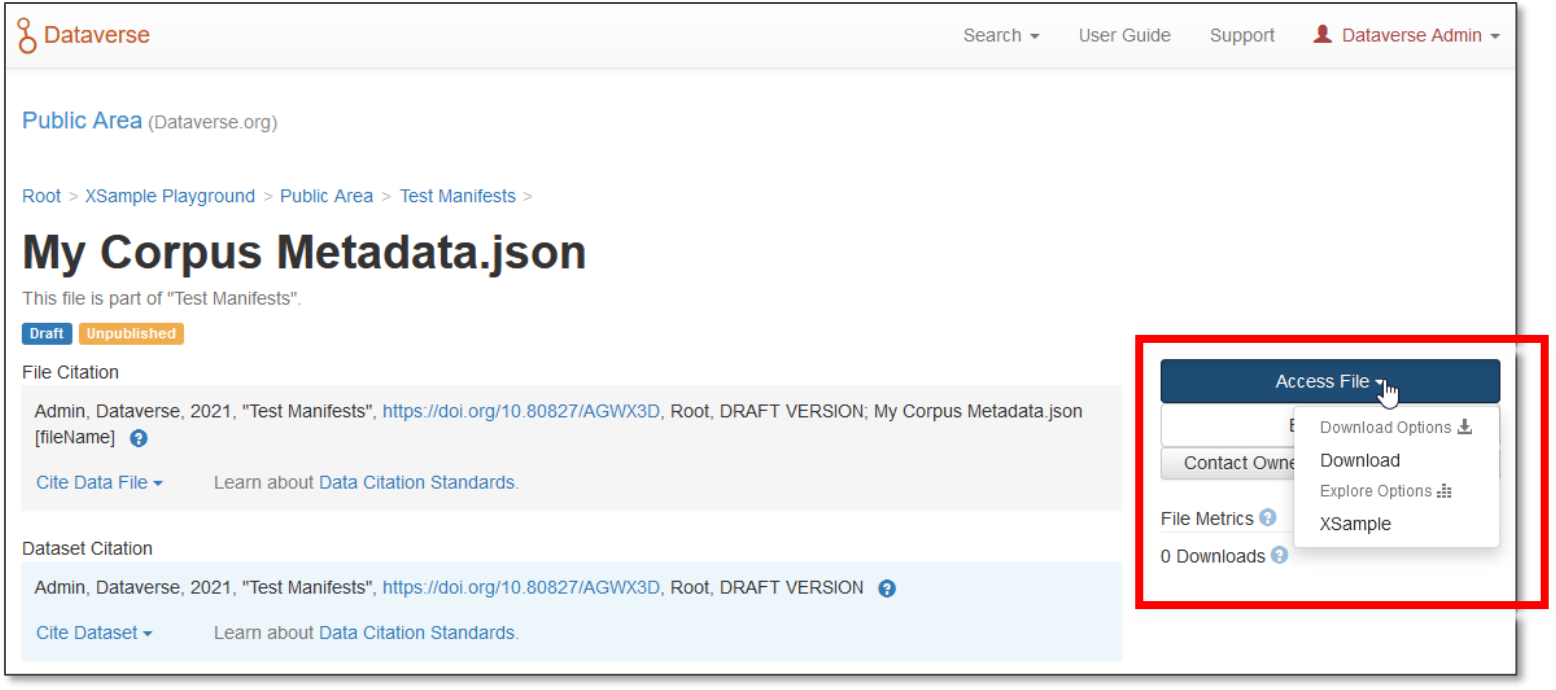
XSample
access, reuse, advance

- Annahme: geschützte Ursprungsdaten und daraus erzeugte Korpora liegen zusammen mit Metadaten in einem Repositorium
- Einstiegspunkt: öffentlich auffindbare Metadaten
- Auszugserstellung: Über eine Plugin-Schnittstelle, direkt über die Weboberfläche des Repositoriums

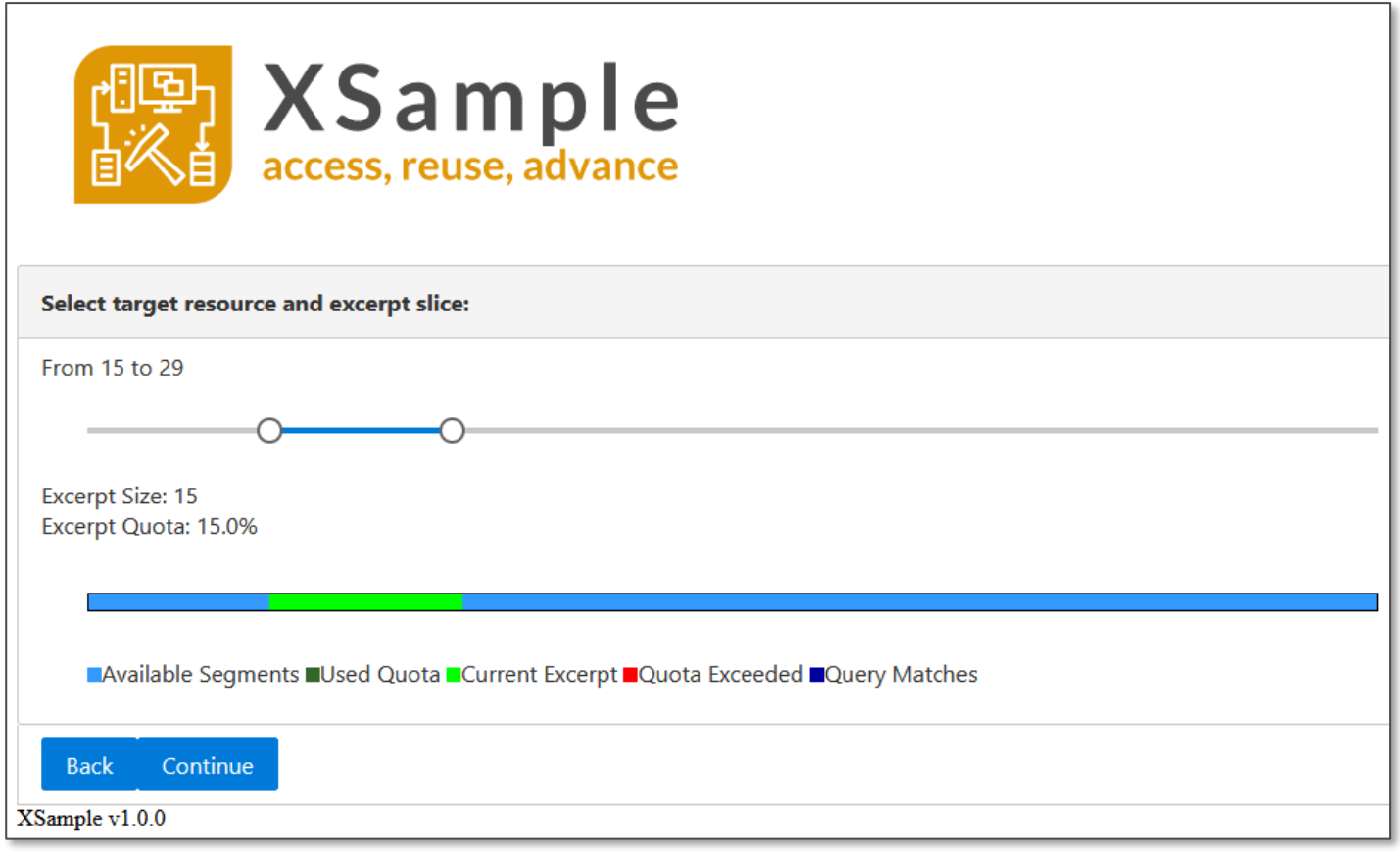
Auszugerstellung

- Statisch: z.B. die ersten 15% oder zusammenhängende Abschnitte (z.B. der Bereich 50-60%) eines Korpus als Auszug
- Dynamisch: feinere Abstufungen, die Suchanfragen auf Basis der im Korpus enthaltenen Annotationen ermöglichen
- Basierend auf diesen Suchergebnissen werden die auszugebenden Segmente der Primärdaten (zumeist Seiten) bestimmt

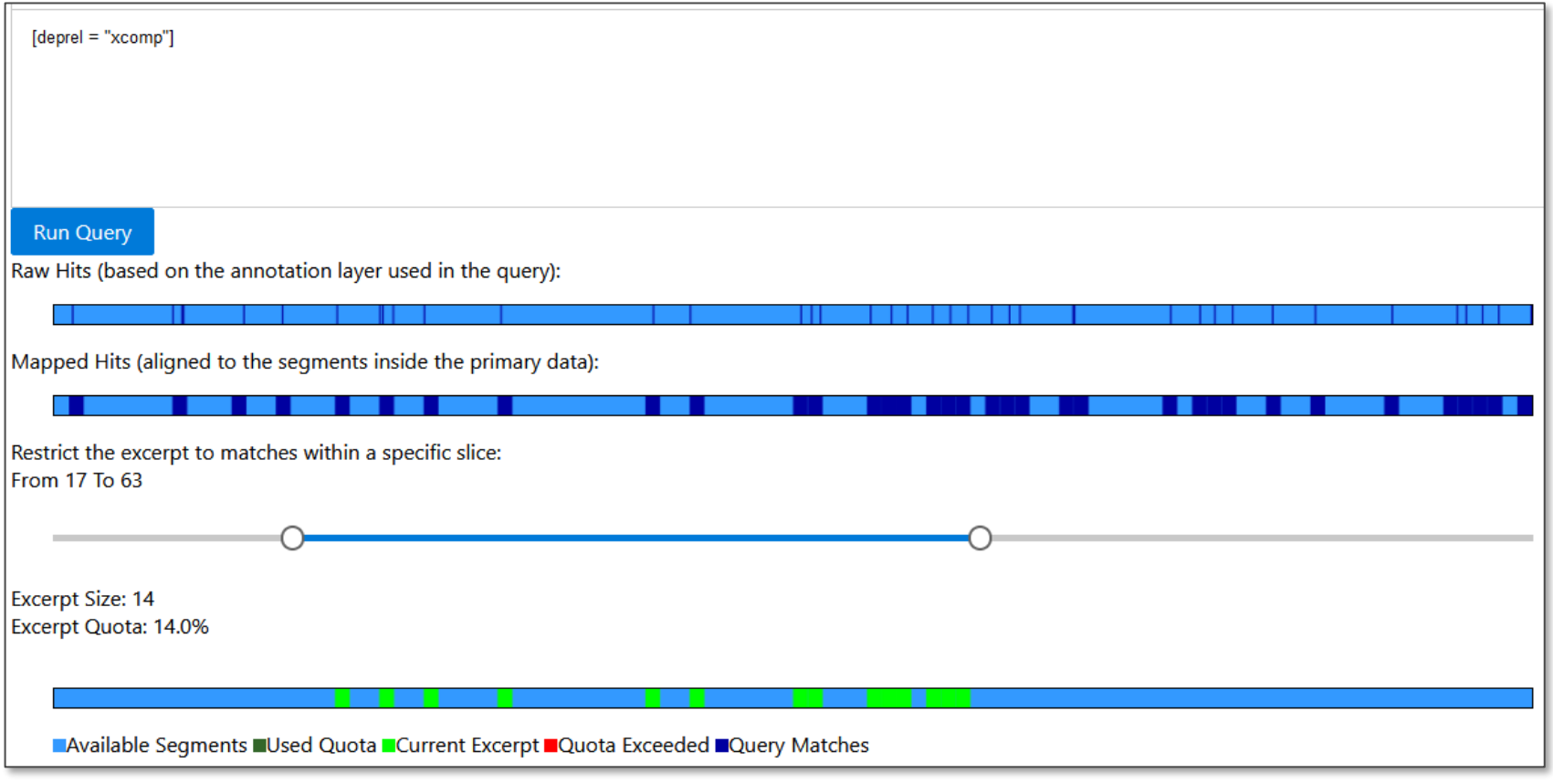
Auszugerstellung - Einstiegspunkt



Auszugerstellung - Abschnitt



Auszugerstellung - Query



Rechtliches Gutachten

TDM-Korpora zur Verfügung stellen

XSample
access, reuse, advance

Workflow

- Die Forschungsdaten sollten bereits während der Forschungsarbeiten auf den Servern der Infrastruktur gespeichert werden.
- Solange Zugriff auf Forschungsgruppen beschränkt.

Archivierung

- Archiviert werden dürfen nur die Vervielfältigungen für das TDM, d. h. die Korpora, z. B. in den Formaten CoNLL und TEI sowie ggf. TXT.
- Ursprungsdaten müssen gelöscht werden, wenn die Forschung abgeschlossen ist. Sie dürfen nur enthalten sein, wenn diese Nutzungen lizenziert sind, dazu zählen auch Open Access publizierte Werke.
- Wenn die Korpora kürzer oder länger als zehn Jahre aufbewahrt werden sollen, muss das plausibel begründet werden.

TDM-Korpora nachnutzen

- Nachnutzende benötigen einen authentifizierten DaRUS- bzw. Dataverse-Zugang. Dabei müssen nicht-kommerzielle, wissenschaftliche Zwecke verifiziert werden.
- Über die Suchmaske kann das betreffende Korpus aufgefunden werden.
- Um zum XSample-Auszugs-Tool zu gelangen, muss eine individuelle Anfrage gestellt werden.
- Wenn der Zugriff gewährt wurde, kann ein Auszug interessengerecht erstellt und heruntergeladen werden.
- Für die Weiternutzung der Korpusauszüge gelten die Vorgaben des Urheberrechts.

Infrastruktur bereitstellen

- Verfügbar gemacht werden dürfen nur die Korpora.
- Ursprungsdaten dürfen nur dann verfügbar gemacht werden, wenn diese Nutzung lizenziert wurde.
- Die Korpora müssen in Dataverse dergestalt eingestellt werden, dass auf sie, auch mittels des XSample-Tools, nur nach individueller Anfrage zugegriffen werden kann.
- Die standardmäßige Aufbewahrungsfrist für TDM-Korpora beträgt zehn Jahre, sie kann mit entsprechender Begründung aber kürzer oder länger bemessen sein. Nach Ablauf sind die Korpora zu löschen.

Nachhaltigkeit – Erste Überlegungen

XSample
access, reuse, advance

- Wie bekommen wir dieses aber auch andere Projekte nachhaltig in unsere Infrastruktur?

  - Stufe 0 - WÄHREND der Projektlaufzeit (= es gibt noch Ansprechpartner für das Tool mit Ressourcen)
  - Stufe 1 - Betrieb OHNE Maintenance (= Bewährungsphase - wird das Tool genutzt?)
  - Stufe 2 - (Kurz-)Betrieb MIT Maintenance (= wir haben vorübergehend Ressourcen für ein bewährtes Tool)
  - Stufe 3 - Dauerbetrieb bei FoKUS (= es gibt offiziell Dauerstellenanteile für den Betrieb des Tools)

Weitere Informationen

- https://github.com/ICARUS-tooling/xsample-server
- F. Kleinkopf, J. Jacke, und M. Gärtner, „Text- und Data-Mining Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora bei computergestützten Verfahren und digitalen Ressourcen“, *Erstpublikation: MMMR Zeitschrift für IT-Recht und Recht der Digitalisierung*, Bd. 3, S. 196 ff., 2021, doi: 10.18419/opus-11445.
- Gutachten wird demnächst veröffentlicht
- Artikel, der das ganze Projekt zusammenfasst ist bei der ZfdG - Zeitschrift für digitale Geisteswissenschaften https://zfdg.de/ eingereicht
- https://www.izus.uni-stuttgart.de/fokus/fdm-projekte/xsample/