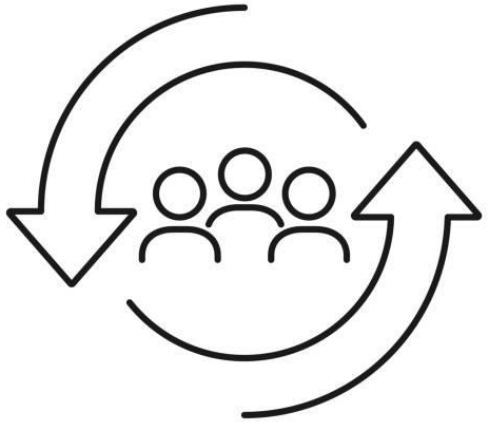
Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK
FID SLAWISTIK

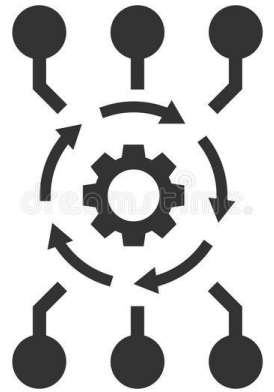
Metadaten
(direkter) Austausch, Export, Nachnutzung

AG FID Netzwerk Philologien
Vladimir Neumann
11. Mai 2023

- Austausch von Metadaten
  - auf dem zentralen Wege über das Kompetenzzentrum für Lizensierung (KfL)
- direkter Austausch von Metadaten
  - Notwendigkeit und die Spezifika
- Export-Import: Austauschformate, Austauschplattform
- Datenkonversion (Werkzeuge) und Datenfilterung (Mechanismen), Nachnutzungsszenarien in den lokalen Discovery Systemen bzw. Datenbanken
- Wie sieht es konkret aus?
- Diskussion / Fragen



- Metadaten aus den Lizenzverträgen des KfL (Nationallizenzen, FID-Lizenzen)
- Geschäftsgang:
  - Vertrag > Metadatenlieferung > KfL-ERMS (grobe Produktbeschreibung, Zugangslinks)
  - Validierung der Metadaten (SBB-PK) > Weitergabe an VZG (zum Formatieren nach MARC)
  - Einspielung der Daten in GBV-Datenbanken (in erster Linie FIDELIO), die dann über diverse Schnittstellen und Filtermechanismen in eigenen FID-Suchsystemen angefragt werden können

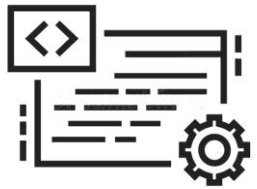


- Ist ein direkter Metadatentausch notwendig?
- Gründe:
  - Metadaten in diversen Sprachen können von **Sprachexperten** konvertiert, angepasst, veredelt werden
  - hochspezialisierte Daten erfordern **hochspezialisierte** Behandlung
    - Unicode-Handling, Diakritika in den slawischen Sprachen, Transliteration- und Transkriptionsformate
  - Mappings für **fremdsprachliche** Sacherschließungssysteme können bedient werden
  - Filtermechanismen können aufgrund spezifischer **Sprach- bzw. Fachkenntnis** implementiert werden
  - Konversionsdaten (aus lokalen Projekten) haben eine komplexere Struktur als „gewöhnliche“ bibliographische Daten (z.B. Retransliterations, **Language Detection**)
- => Zeitfaktor / Profilschärfe

- Werkzeuge: Arbeiten mit nativen **P-Sprachen** und Entwicklung von transparenten Funktionen, im Sinne einer „WhiteBox“
- SP-Werkzeuge: benutzerdefinierten **Funktionen** erlauben das sequenzielle Laden/Manipulieren/Speichern von Daten sowie die Verarbeitung von beliebig codierten Zeichen
- Mechanismen: **Filterung** und Anreicherung der Daten
- Ziel und **Ausgabeformat** ist sehr oft das XML in diversen Ausführungen, meistens das SOLR-XML für direkte Indizierung
- Integration in lokale **Discovery Systeme** oder lokale Datenbanken
  - für SOLR-basierte DS: MARC bzw. MARCXML
  - für alle anderen SOLR-Server: SOLR-XML (inkl. Datenschema)
  - Herausforderung: Integration von Volltexten in DS (SP-Prototyp)

Metadatenaustausch
Slavistik-Portal: Arbeiten mit Daten in der Praxis: Übersicht

- Datenquellen: bibliographisch und in Volltext (knapp 6. Mio. DS)
- Alle Quellen werden unter Beibehaltung der Originalschrift nach Unicode konvertiert
- Konvertierung von gedruckten Bibliographien in Datenbankformat [BibDatSlav at al.]
- Harvesting von Daten aus standardisierten Schnittstellen (SRU, XML, OAI, JSON, Z39.50) [FBC, Hrčak, RNB]
- Harvesting von Daten aus nichtstandardisierten Schnittstellen (Webharvesting, Datenbank- und Katalogharvesting) [Runivers.ru, COBIB.SR, „Baza Artykuły z czasopism polskich“]
- Filterung und Herauslösung von Strukturen aus dem geharvesteten Material [Uknol.ua, Chtyvo.ua]
- Konvertierung von Daten aus „alten“ Datenbanken in überholten Formaten [SorBib, RussGus, KempgenDB]
- Data Mining (aus reinen Textdateien werden mit Hilfe von RegEx die Datenstrukturen herausgelöst) plus Anreicherung mit ZDB-Daten – [BibMatSlaw]
- Korrektur von fehlerhaften Zeichen und Metadaten-Strukturen [UDB-EDU]
- Neuzusammensetzung von „alten“ Daten [OLC Slawistik], [GBV Slawistik-ToCs]
- Maschinelle multidirektionale Übersetzung via Translation-APIs [BibSlavArb]
- Anreicherung von Daten mit Sprachangaben aus „Language Detection“ und mit Kategorien mit Hilfe von Soundex-Algorithmus [BibSlavArb]
- Gewinnung von Volltextdaten durch Extrahieren der Textschicht aus PDF-Content (Apache Tika) [SovSlav, KievSta, WdSl]
- Gewinnung von Volltexten aus Wikisource-Quellen über die XML-API
- Anreicherung von Textdaten mit Annotationen via UDPipe-API [Kirchenslavica-Corpus]



- Wie könnte ein Metadatentausch **in der Praxis** aussehen?

  1. Notwendig ist eine **Übersicht** darüber, welche Metadaten die einzelnen FIDs im Einsatz haben
  2. ein FID benennt den **Content**, der für ihn interessant ist
  3. das gewünschten **Ziel-Format** wird festlegt
  4. die Filterung- / Anreicherung-**Kriterien** werden festgelegt

- => Die Metadaten **werden** dann nach Wunsch aufbereitet und auf einem FTP-Server **bereitgestellt**

Vielen Dank für die Aufmerksamkeit!

Ich freue mich auf Ihre Fragen!