Protokoll des Treffens der U-AG „Technische Infrastruktur“ der FID zum Thema Metadatenmanagement am 15.11.2018 an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen

Sitzungsleitung: M. Kittelmann und K. Keßler

Moderation: M. Kittelmann und K. Keßler

Protokoll: K. Keßler

Teilnehmer/innen: 29 Teilnehmer/innen aus 18 Einrichtungen

Vorträge/Berichte des KfL zu den Themen:

- Metadatendienstleistungen des KfL im Rahmen der FID-Lizenzen (M. Kittelmann, KfL) Zusammenfassung
- Erfahrungsbericht Interaktion mit Verlagen (S. Brauns und P. Otto, KfL)

Fragen und Antworten

F: Wie sieht es mit dem Zugriff nach Auslaufen der Lizenz aus?

A: Das ist abhängig von den Lizenzbedingungen. Auf die lizenzierten Zeiträume kann weiterhin Zugriff bestehen. Es kann aber notwendig werden, dass die Lizenzinhalte an anderer Stelle gehostet werden müssen.

F: Wird die FIDELIO-Datenbank in den k10plus übertragen werden?

A: Dies muss noch geklärt werden.

F: Wie hoch ist der Anteil an Aufsatzdaten in FIDELIO?

A: Um die 50%

F: Wie sieht die Qualitätssicherung beim Eingang der Rohdaten aus?

A: Mitarbeiter/in der/die Lieferung entgegennimmt führt Prüfungsskripte aus und gibt Rückmeldung, ob die Daten in Ordnung sind. (Der Artikel unter https://goescholar.uni-goettingen.de/handle/1/15382 beschreibt ausführlich, was wir testen, und wie wir dabei vorgehen.)

F: Wie und wann findet die Vollständigkeitsprüfung der Daten statt?

A: Diese findet im Verlauf der Prozessierung statt. Da sich die Datenlieferungen sehr stark voneinander unterscheiden können, kann besonders bei umfangreichen Produkten ein Abgleich mit Titellisten erst kurz vor oder im Laufe der Konversion stattfinden. Bei Produkten mit geringerem Umfang kann die Vollständigkeitsprüfung in der Regel unmittelbar nach Eingang der Lieferung erfolgen.

F: Wird die FIDELIO-Datenbank in den k10plus übertragen werden?

A: Dies muss noch geklärt werden.

F: Sind vollautomatische Verfahren im Einsatz oder geplant?

A: Noch nicht im Einsatz. Die Konversionen werden für jedes Produkt individuell eingerichtet und angepasst (bei neuen Formaten ggf. neu geschrieben) und kann anschließend für Updates automatisch widerholt werden. (Gelegentlich sind erneute Anpassungen aufgrund von Formatänderungen o.ä. notwendig.)

F: Wie sieht es mit der inhaltlichen Prüfung aus? Es gab z.B. den Fall, dass DOIs zur falschen Ressource auflösten.

A: Bisher wird eine exemplarische Prüfung, insbesondere bei „Sorgenkindern“, durchgeführt. Die exemplarische Prüfung ist wichtig für den Proxy-Zugang um die Funktionalität zu gewährleisten. Die Prüfung ist aber inhaltlich nicht durchgängig machbar. Hinweis von D. Opitz: Eine Prüfung der Titel gegen CrossRef ist für das Bremer System in Planung. Kommentar von Olaf Brandt: Aber auch wie BASE ist CrossRef nicht immer vollständig.

F: Im Workflow gibt es einen Archivierungsschritt. Was passiert in diesem genau?

A: Dieser ist noch in Planung. Eine Lösung sollte Ende des Jahres 2019 vorliegen.

Vortrag zum Thema: Metadatenmanagement und -verarbeitung im FID Politik und in der E-LIB (D. Opitz, SuUB Bremen)

Fragen und Antworten

F: Wie hoch ist der Aufwand in VZE für die Konvertierung?

A: Eine Bibliothekarin

F: Wie funktioniert die Qualitätskontrolle?

A: Bei kleinen Lieferungen werden kleine Skripte mit kleineren Fehlerbehebungen geschrieben. Nach der Konvertierung erfolgt die gängige Normalisierung, die Fehler behebt. Danach wird methodisch nachgeprüft, ob der Fehler in den gelieferten Daten liegt oder in der Verarbeitung liegt.

F: Wie funktioniert die Dublettenerkennung?

A: Diese basiert auf maschinellem Lernen. Durch eine studentische Hilfskraft wurde ein Model trainiert. Das System arbeitet mit Wahrscheinlichkeiten, z.B. ist eine Dublette sehr wahrscheinlich wenn dois übereinstimmen bei nur Gleichheit im Titel ist die Wahrscheinlichkeit viel geringer. Es findet auch eine phonetische Überprüfung statt. Von ca. 3 Millionen Artikel in DB sind nur ca. 1,8 Millionen im Suchindex enthalten. Eine Komplettdeduplizierung dauert ca. 5 bis 6 Stunden.

F: Wie sind die Erfahrungen mit Sperrungen beim Crawling von Daten?

A: Bisher keine Probleme, aber die Projekte waren nur recht klein.

F: Wann kann Nightwatch ausprobiert werden?

A: Es soll bald als Open Source veröffentlicht werden. Ziel ist Frühjahr 2019. Eine erste lauffähige Version sollte im Dezember 2018 vorliegen. Auf Anfrage kann ein Gitlab-Konto erstellt werden.

F: Basiert Nightwatch auf einem Framework oder bestehendem Tool?

A: Es ist eine Eigenentwicklung, die auf eine einfache Darstellung und Transparenz abzielt.

Metadatenmanagement der FIDs der UB Tübingen (O. Brandt, UB Tübingen)

Fragen und Antworten

F: Wie sieht es mit der Datenqualität von Zotero aus?

A: Das ist abhängig vom Verlag. Aber solange Verlag einheitlich liefert, z.B. über COinS, sollte die Qualität adäquat sein. Bei Verlagen bei denen die Datenqualität nicht stimmt, muss man dann wohl zurück zur semi-automatischen Prozessierung.

F: Wie werden im Zotero Workflow Dubletten und hierarchische Beziehungen gehandhabt?

A: Die wesentliche Metadaten des Harvestens werden in einer Datenbank abgelegt, das komplette XML-File wird als XML-BLOB in der DB abgelegt.. Danach findet ein methodisches Vorgehen für den Abgleich statt, allerdings keine komplette Deduplikation mit dem SWB. Im Bereich Theologie, Religionswissenschaften und Kriminologie gibt es allerdings nur wenige Dubletten. Hierarchische Beziehungen werden durch ISSN abgebildet.

F: Gibt es bei Zotero Mengen-/Datengrenzen?

A: Bisher noch an keine Grenzen gestoßen.

F: Wie sieht es mit rechtlichen Problem des Metadatenharvestings aus?

A: Das Harvesting wurde von Juristin als in Ordnung befunden.

Metadatenmanagement im FID Darstellende Kunst der UB Frankfurt (J. Beck, UB Mainz/UB Frankfurt)

Hinweis: Viele Mappings auf Europeana-Datenmodel vorhanden

Fragen und Antworten

F: Von welchem Ausgangspunkt aus werden die Entitäten aufgesplittet?

A: Ausgangspunkt ist des EDM XML.

F: BaseX vs. Luigi?

A: Es wird geprüft werden, ob die Workflow-Engine in BaseX ausreicht oder ein teilweiser Einsatz von LUIGI notwendig ist.

F: Wie sieht es mit der Nutzung von BaseX als Basis für das Frontend aus?

A: Die Performance erscheint zu schlecht zu sein.

F: Wie werden Entitäten verlinkt?

A: Innerhalb des EDM. Eine Graph-DB wäre natürlich toll.

F: Gibt es ein Upload-Tool für die Datenlieferungen?

A: Dies wird noch geprüft. OwnCloud ist eine Option, aber auch eine Nachnutzung des Upload-Tools des Meta-Kataloges.

Diskussionsrunde

Nutzung von Daten innerhalb der FID:

- Die FID der UB Tübingen werden weiterhin selbst die MARC-Daten zur Indexierung benötigen.
- Der FID Move wird die Metadaten auch für die Dokumentlieferung benötigen.

Nutzerverwaltung und Authentifizierung: IP-basiert ist diese natürlich am einfachsten, aber schwierig bei Instituten oder Einzelpersonen als Zielgruppe. Hier erscheint nur die Möglichkeit der verifizierten Nutzer eine Möglichkeit. Das KfL-ERMS bietet z.B. diese Funktionalität. Dieses Thema wurde im Detail auf dem Workshop im Februar 2018 besprochen. Die Unterlagen dazu können bei Bedarf von Kristof Keßler zugesandt werden.

Wie sieht es mit Rückmeldungen aus der Forschung zur Nutzung der Portale hinsichtlich Massendaten vs. tiefere Erschließung:

- Kriminologie: Hier ist Masse und direkter Zugriff wichtig.
- Theologie, Religionswissenschaften: Die Suche auf einer Plattform ist wichtig.
- Auch wichtig ist die Betonung des Mehrwertes zu Google:
    - Die fachspezifischen Inhalte sind dabei ein gutes Argument. Man findet nur Fachliteratur selbst wenn man einen allgemeinen Begriff eingibt, wie z.B. state (Politik).
    - Die Community nimmt Bibliographien gut an bei der Kriminologie. Ca. 1200 Nutzungen pro Woche, bei der Theologie mehr. Google ist keine Konkurrenz, da dies sehr spezielle Zielgruppen und Publikationen sind.
    - Die Daten des FID Darstellende Kunst waren vorher in der Regel nicht im Internet verfügbar. Ein wissenschaftlicher Beirat kann eine Eigendynamik entwickeln. Durch den übergeordneten Dienst entstehen Strukturen und gemeinsame Projekte.
    - Beim FID Move werden die Datensets nach Filterkriterien für bestimmte analytische Zwecke und Training beim maschinellen Lernen zur Verfügung gestellt.

Nachnutzbarkeit von FIDELIO-Daten: Ja, diesen können weiterverwendet werden. Bitte das KfL ansprechen. Wichtig: Einfache und nach Standard selektierbare Exporte auch aus dem K10plus, z.B. nach FID-Sigel, sind vorhanden. Die Planung ist, dass FIDELIO Ende 2018 online geht.

Dem KfL würden die Pläne der FID bei der eigenen Datennutzung weiterhelfen. Wenn ein FID besondere Vorhaben in der Datennutzung hat, dann melden Sie sich bitte beim KfL.

Lightning Talks

- FID Pharmazie: Deduplizierung und PICA+-Generierung auf Basis von XSL-Transformationen (K. Keßler, UB Braunschweig)
- Erleichterte Suche nach freien Verfügbarkeiten in Suchportalen (C. Poley, ZB MED LIVIVO)
    - Wie wird die Verfügbarkeit basierend auf dem Standort bestimmt? Auf Basis der IP, bei dieser Methode kann es aber zu Datenschutzproblematiken kommen.
    - Ist der Dienst Journals Online & Print massenverarbeitungstauglich? Leider im Moment nicht.
    - Wie ist der Turnus des regelmäßigen Abgleichs? Einmal im Monat.

Abschluss

- Wiederholung in ca. einem Jahr
- Nachtrag zur Erinnerung: Weitere Workshops sind in Planung:
    - Softwareentwicklungstools/-methoden
    - Verfügbarkeitsprüfung