Revrit –
Automatisierte Retransliteration von Hebraica-Titeldaten

Ziel

Revrit ermöglicht es, transliterierte Titeldaten von Hebraica mit Originalschrift anzureichern und verbessert Nachweis und Auffindbarkeit der betreffenden Werke.

Zwar wurde die Katalogisierung von Hebraica in Originalschrift flächendeckend seit ca. 2010 eingeführt, Hebraica-Titeldaten liegen aber weiterhin mehrheitlich nur in transliterierter Form vor. Die ursprünglich geltenden Umschriftregeln gehen auf die Preußischen Instruktionen zurück und unterlagen über die Jahre weitreichenden Veränderungen. Sie sind für Nutzer*innen nur bedingt nachvollziehbar und haben wenig Bedeutung für wissenschaftliche Publikationspraktiken. Zudem erwarten Nutzer*innen inzwischen die Möglichkeit zur Suche und Anzeige in Originalschrift. Eine intellektuelle Überarbeitung der betreffenden Titeldaten wäre sehr aufwändig und angesichts fehlenden Personals mit entsprechenden Spezialkenntnissen kaum umsetzbar.

Funktionsweise des Revrit-Verfahrens

Das Revrit-Konversionstool (Deromanize) wurde zunächst auf Basis der Titeldaten der Universitätsbibliothek J.C. Senckenberg entwickelt, die über die größte bibliothekarische Hebraica- & Judaica-Sammlung verfügt. In zwei Testphasen wurden das Konversionsverfahren optimiert und um eine JSON-API sowie ein Auditing Webtool erweitert. Zu den Datengebern gehörten in einer ersten Testphase die Staatsbibliothek zu Berlin, die Universitätsbibliothek Potsdam, in einer zweiten Phase die Bayerische Staatsbibliothek und die Staats- und Landesbibliothek Sachsen-Anhalt.

Das Konversionstool prüft, ob Titeldaten Transliterationen enthalten, identifiziert den jeweiligen Transliterationsstandard und konvertiert ausgehend davon und unter Anwendung heuristischer Methoden die Titelstrings in hebräische Originalschrift. Da Hebräisch eine Konsonantensprache ist und die Transliterationsregeln Katalogisierungsfehler begünstigen, ist die Konversion auf Wortebene nicht eineindeutig. Revrit generiert daher Vorschlagslisten für jedes einzelne Wort, die auf Basis von lexikalischen Daten priorisiert werden (Ranking). Die Genauigkeit der Rekonversion eines Wortes ist von verschiedenen Faktoren abhängig (z. B. der Wortlänge) und beträgt durchschnittlich 96%.

Revrit stellt eine umfassende Diagnostik auf der Ebene der Subfelder bereit, welche die Qualität der Konversionsergebnisse bewertet. Zudem werden die konvertierten Titeldaten mit einer Referenzdatenbank abgeglichen, die auf dem Hebraica-Katalog der National Library of Israel (NLI) beruht. Im Falle einer Übereinstimmung mit der Referenzdatenbank werden die orginalschriftlichen Titeldaten der NLI übernommen. Insbesondere bei der Verantwortlichkeitsangabe werden damit Fehler auf Basis der Transliteration drastisch reduziert.

Revrit teilt die prozessierten Daten in drei Kategorien ein: *verified*, *unverified*, *error*. Für Titeldaten der Kategorie *verified* und *unverified* gibt Revrit Handlungsempfehlungen für die weitere Bearbeitung: *Display* empfiehlt die Einspielung in anzeigbare Datenfelder. Dies trifft auf alle verifizierten Titeldaten zu, kann aber auch nicht-verifizierte Titeldaten einschließen, die auf Basis der Diagnostik als hochqualitativ eingestuft werden. *Search* empfiehlt die Einspielung von Titeldaten in indexierbare Datenfelder. Die Konversionsergebnisse werden als ausreichend gut bewertet, um die Auffindbarkeit von Hebraica zu optimieren.

Kontrolle und manuelle Bearbeitung

Für die weitere Prozessierung speziell von nicht verifizierten Retransliterationen steht das Revrit Auditing Webtool zur Verfügung. Es ermöglicht die Begutachtung der Konversionsergebnisse und Empfehlungen sowie die manuelle Bearbeitung und Korrektur der Konversionen.

Revrit – Spezifikationen und Charakteristika

Datenformat	JSON	basierend auf der Dublin Core Metadata Initiative (DCMI)
Erforderliche Datenfelder	title (mandatory)	Haupttitel, Titelzusätze und Verantwortlichkeitsangabe
	ispartof (optional)	Gesamttitel mehrteiliger Monographien
	creator (recommended)	verantwortliche Personen
	date (recommended)	Erscheinungsjahr
	publisher (recommended)	Verlag und Ort
	identifier (recommended)	spezifischer Identifier des Datengebers
Prozessierung	automatisierte Identifikation der Transliterationsnorm	heuristisches Verfahren
	automatische Retransliteration	heuristisches Verfahren auf Grundlage der identifizierten Transliterationsnorm (Phyton)
	Ranking der Konversionsmöglichkeiten	Ranking mit Hilfe hebräischer Wörterbücher, dem Cache verifizierter Formen und häufiger Abweichungen/Fehler
	Verifikation	automatisierter Abgleich der konvertierten Daten mit originalschriftlichen Metadaten der National Library of Israel
Datenausgabe	verified	Verifikation mit Referenzdatenbank (Hebraica-Katalog der National Library of Israel) und Angabe der Referenzdaten
	unverified	Titeldaten ohne Übereinstimmung mit Referenzdatenbank mit Angabe des bestmöglichen Treffers Weitere Prozessierung abhängig von der Diagnostik
	error	Keine Konversion – Fehlermeldung für Daten in nicht-lateinischer Schrift oder mit Elementen, die kombinatorische Explosionen liefern (und i. A. keine hebräischen Worte sind).
Diagnostik	Angabe detaillierter diagnostischer Informationen	Die Diagnostik wird für jedes Titelsubfeld separat angegeben. Sie umfasst u. a. die Ausgabe des identifizierten Transliterationsstandards und Angaben zum Vorhandensein von Sonderzeichen und zum Abgleich mit internen Lexika. Bei verifizierten Titeln wird zusätzlich die Übereinstimmung mit dem Match quantifiziert.
	Handlungsempfehlungen	Empfehlung für *verifizierte* Titel: Übernahme der Referenzdaten (*display match*): Beispiel Empfehlung für *nicht-verifizierte* Titel: Beispiel • bei sehr guter Qualität – Übernahme der retransliterierten Daten in anzeigbare Datenkategorien (*display*) • bei guter Qualität – Übernahme in indexierte Kategorien (*search*)
Prüfung und Bearbeitung der Daten	Revrit-Auditing-Webtool	Begutachtung von Konversionsergebnissen und Empfehlungen sowie manuelle Bearbeitung und Korrektur

Titelkonversionen

Mit Hilfe von Revrit wurden in den ersten beiden Förderphasen des FID Jüdische Studien Titeldaten aus Hebis und Fremddaten anderer Verbünde automatisch in hebräische Originalschrift rekonvertiert. Ein großer Teil dieser Titelkonversionen wurde bereits in die entsprechenden Verbundsysteme zurückgespielt. Dabei handelte es sich ausschließlich um verifizierte Retransliterationen.

Im Hebis-Verbund wurden nach Abschluss der ersten Entwicklungsphase (2019) 12.798 originalschriftliche Titel eingespielt (Beispiele). Weitere ca. 5.000 Titel folgen demnächst. Die Staatsbibliothek zu Berlin hat als Datengeber der ersten Testphase 5.923 bearbeitete Titeldatensätze eingespielt (Beispiele). Dem Bibliotheksverbund Bayern (BSB) liegen ebenfalls mit Revrit angereicherte Titeldaten vor: 13.789 Datensätze konnten verifiziert werden und verfügen über die Empfehlung *display match*, weitere 5.381 Datensätzen verfügen auf Basis der Diagnostik über die Empfehlung *display*. Der BVB prüft derzeit in Abstimmung mit der Bayrischen Staatsbibliothek München die Einspielung der angereicherten Daten.

Der FID Jüdische Studien steht derzeit im Austausch mit der BSB sowie dem BVB über die weitere Prozessierung. Weiterhin liegen ca. 2.500 verifizierte Titeldaten aus zwei GBV-Bibliotheken vor. Der zuständige Fachreferent der Universitätsbibliothek Potsdam stand diesbezüglich mit Mitarbeiter*innen des GBV in Kontakt.

Dokumentationen

Die Entwicklung steht unter einer CC0-Lizenz und ist über das GitHub der Universitätsbibliothek Frankfurt am Main mit einer ausführlichen Dokumentation zugänglich.

- Beschreibung von Revrit im Portal des FID Jüdische Studien
- Zugang zur Revrit-API
- Dokumentation des Konversionstools Deromanize
- Zugang zum Revrit Auditing Webtool
- Dokumentation des Revrit Auditing Webtools

Die verlinkten Beispiele sollte in Firefox geöffnet werde, da sie hier in einem JSON-Viewer strukturiert angezeigt werden.

Kontakt

FID Jüdische Studien

Goethe-Universität Frankfurt am Main
Bockenheimer Landstraße 134-138
60325 Frankfurt am Main
E-Mail: info@jewishstudies.de
Web: www.jewishstudies.de

Projektleitung:

Dr. Kerstin von der Krone
Tel.: 069/798-39665
E-Mail: K.vonderKrone@ub.uni-frankfurt.de

Entwicklung Revrit

Aaron Christianson
E-Mail: a.christianson@ub.uni-frankfurt.de

Januar 2023